To God, whose love and blessing powered me in each step of my progress towards the successful completion of my research work.

To my beloved grandfather Hassan, who has firmly supported my advancement in both my academic, professional, and personal life.

To my selfless late uncle Mohamed, who has been an unshatterable backbone even in his absence.

To the rest of my family, whose encouragement and prayers led me to where I am today.
ABSTRACT

The preliminary engineer’s estimate for public highway projects has long been a deciding factor on whether State Transportation Agencies (STAs) can proceed with projects that are essential for the public’s well-being. Most transportation and infrastructure projects are funded from a limited reservoir provided by federal, state, and local government programs, and the preliminary engineer’s estimate acts as a benchmark for the spending of funds in said reservoir. Therefore, it is of paramount importance that the drafting of the preliminary engineer’s estimate considers the market conditions and is reflective of the contractor bids, for the proper allocation of funds to projects governed by STAs. Cost estimating in transportation and infrastructure projects is a dynamic process that transforms along the major phases of highway construction projects. The phases are broken out between planning, project development, final design, right-of-way acquisition, construction, and operation and maintenance. This research primarily explored the engineer’s estimate prepared during the final design phase in highway construction projects, which is referenced in this paper as the “engineer’s estimate”. The Federal Highway Administration (FHWA) measures the effectiveness and accuracy of the engineer’s estimate in terms of the percentage deviation of the low bid from the engineer’s estimate and recommends an accuracy defined by at least 50 percent of low bids falling between ±10 percent of the that estimate. Despite commendable efforts from STAs, high deviations of estimates from low bids remain a persistent problem that public agencies face. The Wisconsin Department of Transportation (WisDOT) requested the support of the Construction and Materials Support Center (CMSC) at the University of Wisconsin - Madison in running an estimating peer exchange with fellow STAs to determine underlying causes behind the high deviation of the final design engineer’s estimate from low bids. One important influencing factor on the effectiveness and accuracy of the estimate identified, is the method of cost estimation, which includes historical bid-based estimating, cost-based estimating, and combination estimating. The highway and infrastructure industry has no precise analysis or conclusion on the impact of the different methods of cost estimation on the estimates developed during the initial stages of a project and lack a universally accepted methodology for the choice of the method of cost estimation. Thus, there is a need for STAs to evaluate the effect of using the different methods of cost estimation.
on the estimate accuracy, as defined by the FHWA, to identify the most suitable approach for all project types.

This research utilizes expert opinion from the estimating peer exchange and data-driven based algorithms for STAs to predict the better suited method of cost-estimation for the estimates created during the early stages of the project stages to better allocated funds from public agencies. Data was collected from eleven participating STAs, during the estimating peer exchange, as well as five other STAs using a survey. The data collected is related to the best scoping, cost estimation, and risk assessment practices during early stages of the project, as well as performance of the states’ engineer’s estimate accuracy from the year 2018 to 2020. Additionally, both qualitative and quantitative analysis was performed to evaluate the variation in estimate precision and accuracy using the method of cost estimation using the average yearly data from the STAs. Among the number of bidders, geographic location, shortage of estimators, and economic volatility, the method of cost estimation was identified as a majorly impactful factor on the engineer’s estimate accuracy, as defined by the FHWA. Historical bid-based estimating was found to be the most common method, followed by combination estimating, and finally cost-based estimating. The methods averaged at 47%, 48%, and 53% respectively, of the low bids falling within ±10 of the engineer’s estimate. While cost-based estimating results in the highest accuracy, it requires extensive training of the estimating personnel.

The yearly average dataset was insufficient in concluding which method of cost estimation is better suited for the highway and infrastructure sector. Consequently, prediction machine learning algorithms were employed to predict the optimum method of cost estimation depending on project related variables and economic variables. Raw data was collected from six STAs, Montana Department of Transportation (MDT), Nebraska Department of Transportation (NDOT), North Dakota Department of Transportation (NDDOT), Tennessee Department of Transportation (TDOT), Washington State Department of Transportation (WSDOT), and Wisconsin Department of Transportation (WisDOT). The data obtained only included observations for projects estimated using historical bid-based estimating, and combination estimating with 5-10% line items estimated using cost-based estimating. The dataset spanned 11 unified project types, and was trained using the following machine learning algorithms, multiple linear regression (ML), logistic regression (LOGIT), classification and regression trees (CART),
and random forests (RF) to predict the most suitable method of cost estimation. The gathered data were separated into two groups: one for training the model and the other for testing purposes. Using the same dataset, the models were developed, and then their performances were evaluated based on the area under the receiver operating curve (AUC).

ML was used as the standard statistical analysis to evaluate the need for more complex machine learning models. It was unable to capture non-linear relationships, which proved to be a governing factor behind its low model performance. Economic variables were found to be the most influential on the optimal method of cost estimation, primarily the prime loan rate with a feature coefficient of -11.8611. The project types loosely followed behind with feature coefficients ranging between 0.1787 and 0.6571.

LOGIT was found to be substantially better than the ML method in many respects, including the flexibility around linear relationships, and a obtained a significantly higher performance. Three models were developed using LOGIT, a base model, l1-regularization model, and l2-regularization model. All three models obtained a classification accuracy of 89%, but the l2-regularization model reduced the feature correlation and bias in the model, so it was deemed more fitting for predicting the method of cost estimation with a low risk of overfitting. The prime loan rate was again found to be of highest importance with a coefficient on -84.9338, followed by the project types ranging between 1.102 to 4.907.

One CART model was then developed due to its flexible and non-parametric modeling properties, meaning that there are no strict assumptions. It was able to better capture nonlinearities between the features and the target variables than both the ML and LOGIT models. Using hyperparameter tuning, a maximum model accuracy of 0.99 was obtained using a maximum depth of 9, minimum samples per leaf of 4, and minimum samples per split of 8. CART models are notoriously susceptible to overfitting, and even with the hyperparameter tuning the CART model was deemed not optimal for the prediction of the method of estimation. The projects under the maintenance or minor upgrades type were ranked at the top of the list with a coefficient of 2.0997 followed by the crude oil prices at a coefficient of 0.7586.

Similar to CART, RF was not sensitive to linear relationships between the features and the target variable. Multiple CART trees are combined, with an additional incorporated hyperparameter related to the number of CART trees to tackle the overfitting of the singular
CART trees. The hyperparameter tuning resulted in a maximum depth of 2, minimum samples per leaf of 8, minimum samples per split of 9, and an optimum number of trees of 219. The model obtained a classification accuracy of 90%, which was the highest accuracy across all ML algorithms.

The RF model was deemed the most suitable for the purpose of predicting the most optimal method of cost estimation. The data-driven model can be used by STAs to allocate teams of estimating professionals with varying degrees of experience in estimating. Estimators with a higher understanding of project cost, can be assigned to projects that require the use of cost-based estimating, such that the burden of training estimating personnel on cost-based estimating as a method of estimation is lightened. Hence, the funds available to STAs can be more optimally allocated for the benefit of the public and the economy.

Moreover, economic related factors in all the models consistently exceeded the influence of project related factors. Primarily in the form of the prime loan rate, and the crude oil prices. The project type was the leading influence among the project related features with safety and traffic control, maintenance and minor upgrades, environmental mitigation, roadway redesign, road or culvert replacement, earthwork, and resurfacing project types favoring the use of combination estimating while bridge construction and bridge replacement projects consistently elected historical bid-based estimating as the preferred method of estimation.
ACKNOWLEDGEMENT

I would like to express the deepest appreciation to my advisor and the chair of the Construction Engineering and Management program, Dr. Awad S. Hanna, who has entrusted me with the research related to this thesis and steered me in the right direction when needed. He also enlightened me with the excitement of studying and teaching at the University of Wisconsin – Madison. I would also like to extend to the following people my immeasurable appreciation and deepest gratitude for their support, advice, and encouragement in my fulfillment of this master’s thesis.

I had the pleasure of working with Gary C. Whited, the manager of the Construction and Materials Support Center. His contribution to my knowledge in the transportation industry was of paramount importance for the success of this thesis.

I would like to recognize Fred C. Schunke, Estimating Engineer at the Wisconsin Department of Transportation (WisDOT) for introducing me to the gap in the industry that became the inspiration behind this research project. I am also grateful to the rest of the WisDOT team, the Montana Department of Transportation (MDT), the North Dakota Department of Transportation (NDDOT), the Nebraska Department of Transportation (NDOT), the Tennessee Department of Transportation (TDOT), and the Washington State Department of Transportation (WSDOT) for granting me the data needed for this research.

I am deeply indebted to Michael Mejia, my mentor and manager in the professional construction field. He taught me the craft of estimating and was always willing to extend his wealth of knowledge in the industry in support of the success of my project.

I am also thankful for Dr. Zhenhua Zhu, Joseph G. Walsh and Justin Boutilier, who have all expanded my academic horizon through their courses during my graduate studies.

I could not have undertaken this journey without the support of my best friends, Enaba, Hoda and Roubana, who have believed in my capabilities and offered all the support they could muster. Thank you for constantly inspiring me to pursue all my goals and continue to grow.

I would also like to extend thanks to Cassidy, my one companion in Madison, Wisconsin. My research journey would have been infinitely less bright without our food adventures.

vii
Many thanks to my colleagues in the Construction Engineering and Management program. Anson, Calob, Eyad and Joe, it was a pleasure learning and sharing my work with you.

Lastly, I would be remiss in not mentioning my family, my parents Abeer and Ahmed, my sister Nada, and brother Mohamed. A special thank you for showering me with support and keeping my spirits and motivation high during this process.
# TABLE OF CONTENTS

ABSTRACT ......................................................................................................................... iii

ACKNOWLEDGEMENT ........................................................................................................ vii

LIST OF FIGURES ............................................................................................................... xii

LIST OF TABLES ................................................................................................................ xvi

LIST OF ABBREVIATIONS ................................................................................................. xvii

CHAPTER 1: INTRODUCTION ........................................................................................... 1
  1.1 Organization of the Research .................................................................................. 1
  1.2 Research Background ............................................................................................. 2
  1.3 Problem Statement ................................................................................................. 4
  1.4 Research Objectives ............................................................................................... 6
  1.5 Research Methodology ........................................................................................... 7

CHAPTER 2: LITERATURE REVIEW ............................................................................... 10
  2.1 Introduction ............................................................................................................. 10
  2.2 Major Phases of Highway Construction Projects ................................................... 10
    2.2.1 Planning Phase .................................................................................................... 11
    2.2.2 Project Development Phase ............................................................................ 12
    2.2.3 Final Design ....................................................................................................... 13
    2.2.4 Right-of-Way-Acquisition ............................................................................... 14
    2.2.5 Construction ...................................................................................................... 14
  2.3 Cost Estimation ........................................................................................................ 14
    2.3.1 Definition ........................................................................................................... 14
    2.3.2 Methods of Cost Estimation ............................................................................ 15
    2.3.3 Advantages of Cost-Based Estimating vs. Historical Bid-Based Estimating ..... 18
  2.4 Prior Research on the Preconstruction Cost Estimating Services (PCS) for
    Transportation Projects ............................................................................................... 19
5.3.3 Model Training and Results .......................................................... 87
5.4 Classification and Regression Trees Model ............................................. 98
  5.4.1 Background .................................................................................. 98
  5.4.2 Verification of Assumptions ............................................................ 100
  5.4.3 Model Training and Results .......................................................... 100
5.5 Random Forests Model ........................................................................ 105
  5.5.1 Background .................................................................................. 105
  5.5.2 Verification of Assumptions ............................................................ 108
  5.5.3 Model Training and Results .......................................................... 108
5.6 Model Comparison ............................................................................... 113
  5.6.1 Algorithm Performance Accuracy .................................................. 113
  5.6.2 Feature Importance ....................................................................... 114

CHAPTER 6: DISCUSSION AND CONCLUSIONS ............................................. 116
  6.1 Conclusions ..................................................................................... 116
  6.2 Limitations & Recommendations ...................................................... 121

APPENDICES ......................................................................................... 123
  A: References ....................................................................................... 123
  B: Estimating Peer Exchange Survey .................................................... 132
  C: Multiple Linear Regression Python Code ........................................... 136
  D: Logistic Regression Python Code ...................................................... 136
  E: CART Python Code ............................................................................ 141
  F: Random Forest Python Code ............................................................ 146
  G: Model Comparison Python Code ...................................................... 152
LIST OF FIGURES

Figure 1: Research Methodology ........................................................................................................... 9
Figure 2: Project Development Phases in the Highway Transportation Sector .................................. 11
Figure 3: Geographic Distribution of Survey Respondents ................................................................. 34
Figure 4: Do you track the engineer’s estimate accuracy? ................................................................. 34
Figure 5: Estimate Preparation Process ................................................................................................ 35
Figure 6: Goal of the Engineer’s Estimate ............................................................................................. 36
Figure 7: Primary Estimating Approach Distribution ........................................................................... 37
Figure 8: Use of Indexing to Adjust Historical Bid-Based Items ......................................................... 37
Figure 9: Average no. of Bidders from 2018 to 2020 .......................................................................... 38
Figure 10: Number of State DOTs Using Different Percentages of Cost-Based Estimating .............. 39
Figure 11: Average Percentage of Low Bids Within ±10% of the Engineer’s Estimate ..................... 40
Figure 12: Average Percent of Bids within ±10% of the Engineer’s Estimate by Method of Estimating ........................................................................................................................................................................... 41
Figure 13: Demographic Distribution of State DOTs Involved in Stage 2 ........................................... 44
Figure 14: EDA Process ......................................................................................................................... 51
Figure 15: Boxplot of Advertising Time ............................................................................................... 53
Figure 16: Stage 1 Pair-wire Feature Correlation .................................................................................. 54
Figure 17: Stage 2 Pair-wire Feature Correlation .................................................................................. 55
Figure 18: Project Distribution by Method of Estimating ...................................................................... 57
Figure 19: Distribution of Engineer’s Estimate Variation from Low Bid Based on Project Type 58
Figure 20: Project Distribution by Location Type .................................................................................. 58
Figure 21: Accuracy of Engineer’s Estimate by Location Type ............................................................... 59
Figure 22: Cost Distribution in Dataset .................................................................................................. 60
Figure 23: Cost Distribution by Location Type ....................................................................................... 61
Figure 24: Data Distribution by Forecasted Duration Range ................................................................. 62
Figure 25: Cost Distribution by Duration Range .................................................................................... 63
Figure 26: Distribution of Number of Bidders in Dataset ...................................................................... 64
Figure 27: Data Distribution by Project Type .......................................................................................... 65
Figure 28: Engineer’s Estimate Distribution per Project Type ............................................................... 66
Figure 29: Distribution of Number of Bidders by project Type ............................................................. 68
Figure 30: Feature Engineering Components ................................................................. 71
Figure 31: Receiver Operating Curve for Binary Problems .................................................. 76
Figure 32: QQ Plot of Linear Regression Model ................................................................... 78
Figure 33: Shapiro Wilk Test Result .................................................................................. 79
Figure 34: ML Regression Model Summary ........................................................................ 80
Figure 35: Sigmoid Function with Threshold ...................................................................... 83
Figure 36: % Diff between the Engineer’s Estimate and Low Bids Log Odds Linear Plot ...... 84
Figure 37: Number of Bidders Log Odds Linear Plot ............................................................ 85
Figure 38: Construction Spending Log Odds Linear Plot ..................................................... 85
Figure 39: Prime Loan Rate Log Odds Linear Plot ............................................................... 86
Figure 40: Crude Oil Prices Log Odds Linear Plot ............................................................... 86
Figure 41: Logistic Regression Results and Weights with no Regularization ....................... 89
Figure 42: AUC Score for Testing and Training Sets for Base Logistic Regression Model...... 90
Figure 43: Confusion Matrix of Testing Set in Base Logistic Regression Model ................. 91
Figure 44: AUC Score for Testing and Training Sets for L1-Regularization Logistic Regression Model ...................................................................................................................... 93
Figure 45: Confusion Matrix of Testing Set in Base Logistic Regression Model ................. 94
Figure 46: AUC Score for Testing and Training Sets for L2-Regularization Logistic Regression Model ...................................................................................................................... 96
Figure 47: Confusion Matrix of Testing Set in Ridge Logistic Regression Model ................. 97
Figure 48: Decision Tree Visualization ............................................................................... 98
Figure 49: Maximum Depth Hyperparameter Effect on the CART AUC Score .................. 101
Figure 50: Minimum Number of Samples per Leaf Hyperparameter Effect on the CART AUC Score ....................................................................................................................... 101
Figure 51: Minimum Number of Samples per Split Hyperparameter Effect on the CART AUC Score ....................................................................................................................... 102
Figure 52: Split Level 3 Tree in CART Model ..................................................................... 103
Figure 53: AUC Score for Testing and Training Sets for CART Model .............................. 104
Figure 54: Confusion Matrix of Testing Set in CART Model ............................................... 105
Figure 55: Maximum Depth Hyperparameter Effect on the RF AUC Score ....................... 109
Figure 56: Minimum Number of Samples per Leaf Hyperparameter Effect on the RF AUC .... 109
Figure 57: Minimum Number of Samples per Split Hyperparameter Effect on the CART AUC Score ................................................................. 110
Figure 58: Impact of Number of Trees in Decision Tree Model on AUC Score.......................... 110
Figure 59: AUC Score for Testing and Training Sets for RF Model................................. 112
Figure 60: Confusion Matrix of Testing Set in RF Model...................................................... 113
Figure 61: Comparison of AUC Algorithm Scores for Estimate Predictions...................... 114
LIST OF TABLES

Table 1: NCHRP Report 574 – Factors Causing Cost Escalation .............................................. 21
Table 2: Summary of Independent Variables Identified in Literature ........................................ 25
Table 3: Summary of Research on the Engineer’s Estimate in the Transportation Industry ...... 32
Table 4: Combined Project and Economic Variables to be Used in the Research .................... 50
Table 5: Exploratory Data Analysis and Corrective Action ......................................................... 53
Table 6: Features Selected for Prediction Model ......................................................................... 56
Table 7: Summary Statistics for Project Distribution ................................................................. 60
Table 8: Summary Statistics for Cost Distribution of Projects by Location Type ..................... 62
Table 9: Summary Statistics for Cost Distribution by Forecasted Duration Range ................... 63
Table 10: Summary Statistics of Distribution of Number of Bidders ........................................ 64
Table 11: Summary Statistics for Project Distribution by Project Type ................................. 68
Table 12: Summary Statistics for Project Distribution by Project Type ................................... 70
Table 13: Feature Engineering Process ...................................................................................... 72
Table 14: Feature Coefficients and Statistical Significance from 0 ......................................... 81
Table 15: Feature Coefficients and Statistical Significance from 0 ......................................... 90
Table 16: Feature Coefficients and Statistical Significance from 0 ......................................... 93
Table 17: Feature Coefficients and Statistical Significance from 0 ......................................... 96
Table 18: Feature Importance Ranking ..................................................................................... 103
Table 19: Feature Importance Ranking ..................................................................................... 111
Table 20: Feature Importance Ranking by Model ................................................................. 115
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AACE</td>
<td>Association for Advancement of Cost Engineering</td>
</tr>
<tr>
<td>AASHTO</td>
<td>American Association of State Highway and Transportation Officials</td>
</tr>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>AIC</td>
<td>Corrected Akaike Information Criterion</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under Receiver Operating Curve</td>
</tr>
<tr>
<td>BF</td>
<td>Basis Function</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>BP</td>
<td>Breusch-Pagan</td>
</tr>
<tr>
<td>BPNN</td>
<td>Back Propagation Neural Network</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and Regression Trees</td>
</tr>
<tr>
<td>CBR</td>
<td>Case-Based Reasoning</td>
</tr>
<tr>
<td>CPI</td>
<td>Consumer Price Index</td>
</tr>
<tr>
<td>CPPR</td>
<td>Contractor's Past Performance Rating</td>
</tr>
<tr>
<td>CS</td>
<td>Construction Spending</td>
</tr>
<tr>
<td>DF</td>
<td>Degree of Freedom</td>
</tr>
<tr>
<td>DJIA</td>
<td>Dow Jones Industrial Average</td>
</tr>
<tr>
<td>DOT</td>
<td>Department of Transportation</td>
</tr>
<tr>
<td>DOR</td>
<td>FHWA Federal-aid division offices</td>
</tr>
<tr>
<td>DRISI</td>
<td>Caltrans Division of Research, Innovation and System Information</td>
</tr>
<tr>
<td>DW</td>
<td>Durbin-Watson</td>
</tr>
<tr>
<td>EIA</td>
<td>U.S. Energy Information Administration</td>
</tr>
<tr>
<td>FDOT</td>
<td>Florida Department of Transportation</td>
</tr>
<tr>
<td>FHWA</td>
<td>Federal Highway Administration</td>
</tr>
<tr>
<td>FRED</td>
<td>Federal Reserve Economic Data</td>
</tr>
<tr>
<td>GBM</td>
<td>Geometric Brownian Motion</td>
</tr>
<tr>
<td>GCV</td>
<td>Generalized Cross-Validation</td>
</tr>
<tr>
<td>GRNN</td>
<td>General Regression Neural Network</td>
</tr>
<tr>
<td>IDOT</td>
<td>Illinois Department of Transportation</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>IDOT</td>
<td>Iowa Department of Transportation</td>
</tr>
<tr>
<td>KAPI</td>
<td>Kentucky Asphalt Price Index</td>
</tr>
<tr>
<td>KDPI</td>
<td>Kentucky Diesel Price Index</td>
</tr>
<tr>
<td>KTC</td>
<td>Kentucky Transportation Cabinet</td>
</tr>
<tr>
<td>LAR</td>
<td>Least Angle Regression</td>
</tr>
<tr>
<td>LASSO</td>
<td>Least Absolute Shrinkage and Selection Operator</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MAPE</td>
<td>Mean Absolute Percentage Error</td>
</tr>
<tr>
<td>MARS</td>
<td>Multivariate Adaptive Regression Splines</td>
</tr>
<tr>
<td>MDOT</td>
<td>Massachusetts Department of Transportation</td>
</tr>
<tr>
<td>MDOT</td>
<td>Michigan Department of Transportation</td>
</tr>
<tr>
<td>MDT</td>
<td>Montana Department of Transportation</td>
</tr>
<tr>
<td>MLR</td>
<td>Multiple Linear Regression</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>MnDOT</td>
<td>Minnesota Department of Transportation</td>
</tr>
<tr>
<td>MPE</td>
<td>Mean Percentage Error</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>NDDOT</td>
<td>North Dakota Department of Transportation</td>
</tr>
<tr>
<td>NDOT</td>
<td>Nebraska Department of Transportation</td>
</tr>
<tr>
<td>NCDOT</td>
<td>North Carolina Department of Transportation</td>
</tr>
<tr>
<td>NCHRP</td>
<td>National Cooperative Highway Research Program</td>
</tr>
<tr>
<td>NHCCI</td>
<td>National Highway Construction Cost Index</td>
</tr>
<tr>
<td>NJDOT</td>
<td>New Jersey Department of Transportation</td>
</tr>
<tr>
<td>NMAD</td>
<td>Normalized Median Absolute Deviation</td>
</tr>
<tr>
<td>ODOT</td>
<td>Oklahoma Department of Transportation</td>
</tr>
<tr>
<td>ODOT</td>
<td>Ohio Department of Transportation</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary Least Square</td>
</tr>
<tr>
<td>PCS</td>
<td>Preconstruction Service</td>
</tr>
<tr>
<td>PE</td>
<td>Preconstruction Service</td>
</tr>
<tr>
<td>PERT</td>
<td>Program Evaluation and Review Technique</td>
</tr>
<tr>
<td>PLR</td>
<td>Prime Loan Rate</td>
</tr>
<tr>
<td>PMBOK</td>
<td>Project Management Body of Knowledge</td>
</tr>
<tr>
<td>PMI</td>
<td>Project Management Institute</td>
</tr>
<tr>
<td>PPI</td>
<td>Producer Price Index</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>QA</td>
<td>Quality Assurance</td>
</tr>
<tr>
<td>QC</td>
<td>Quality Control</td>
</tr>
<tr>
<td>RA</td>
<td>Regression Analysis</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forests</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Curve</td>
</tr>
<tr>
<td>ROW</td>
<td>Right-of-Way</td>
</tr>
<tr>
<td>RSS</td>
<td>Residual Sum of Squares</td>
</tr>
<tr>
<td>SCC</td>
<td>Schwarz Bayesian Information Criterion</td>
</tr>
<tr>
<td>SCDOT</td>
<td>South Carolina Department of Transportation</td>
</tr>
<tr>
<td>SDDOT</td>
<td>South Dakota Department of Transportation</td>
</tr>
<tr>
<td>SSR</td>
<td>Sum of Square Residuals</td>
</tr>
<tr>
<td>SST</td>
<td>Total Sum of Squares</td>
</tr>
<tr>
<td>STA</td>
<td>State Transportation Agencies</td>
</tr>
<tr>
<td>STIP</td>
<td>State Transportation Improvement Plan</td>
</tr>
<tr>
<td>STP</td>
<td>State Transportation Plan</td>
</tr>
<tr>
<td>TDOT</td>
<td>Tennessee Department of Transportation</td>
</tr>
<tr>
<td>US</td>
<td>United States of America</td>
</tr>
<tr>
<td>VBA</td>
<td>Visual Basic Application</td>
</tr>
<tr>
<td>VIF</td>
<td>Variance Inflation Factor</td>
</tr>
<tr>
<td>WisDOT</td>
<td>Wisconsin Department of Transportation</td>
</tr>
<tr>
<td>WSDOT</td>
<td>Washington State Department of Transportation</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION

1.1 Organization of the Research

This section outlines the architectural framework of the thesis over 6 chapters as shown below.

- Chapter 1: Introduction
  Designed to outline the research background related to the engineer’s estimate in the transportation industry and mentions the need and motivation for this current thesis study. Identifies the research objectives of the thesis.

- Chapter 2: Literature Review
  Provides the framework behind the highway construction project phases and the cost estimates developed at each phase. Discusses the importance of estimating, specifically in the highway sector, and the prior research on the engineer’s estimate.

- Chapter 3: Estimating Peer Exchange
  Demonstrates the findings from the estimating peer exchange organized by WisDOT and the CMCS at UW-Madison.

- Chapter 4: Data
  Represents a data description of the combined dataset obtained from STAs and introduces the various variables and their importance. Outlines the data collection process, data preparation, and the EDA.

- Chapter 5: Development of Cost Estimating Prediction Models
  Presents the different prediction models created for each machine learning algorithm of the following: OLS, logistic regression, CART, and RF. The models help STAs select the use of historic bid-based estimating on its own or integrate cost-based estimating for the preparation of their engineer’s estimate. Compares the model accuracy of each of the machine learning models created.

- Chapter 6: Discussion and Conclusion
Wraps up the research and outlines the major results and findings of this extensive study. Provides recommendations for future research that addresses the limitations of this study.

1.2 Research Background

In 2021, it was estimated that there were over 4 million miles of public roadways in the United States of America, serving the transportation of the general public and freight. The number of vehicle miles traveled on the public roadways is ever increasing. In fact, there was an 18% increase from the year 2000 to 2019 in vehicle miles traveled, and the number is expected to grow at a faster rate as the need for transportation increases. Consequently, the state of the public roadways is left less than ideal with 43% of roadways’ condition rated as poor and mediocre (ASCE 2021). The burden of repairs falls upon federal, state, and local transportation agencies, which are funded by United States taxpayers. In 2019, the federal government spent $100.4 billion and state and local governments spent $191.1 billion on infrastructure projects (USDT 2022).

The ASCE performed a study in 2017 that deemed available funds for public highway maintenance and improvements as insufficient. There currently exists a $786 billion backlog of road and bridge capital maintenance needs. An increase of 29% is required to address the current and anticipated backlog in highway infrastructure spending. Given the limited funds, State Transportation Agencies (STAs) are responsible for allocating the funds received from the federal and state governments to highway improvement projects based on their urgency (GAO 2022).

A highway project’s lifecycle normally contains six phases: (1) planning, (2) project development (preliminary design), (3) final design (4) right-of-way acquisition, (5) construction, and (6) operation and maintenance (FHWA 2017, FHWA 2007). The process of cost estimating follows that of the project lifecycle, with the first conceptual estimate occurring during the planning phase and the preliminary estimate occurring during the preliminary design and environmental review phase. The preliminary cost estimate is done during the scoping period with minimal project scope definition and is usually the deciding factor on whether highway projects are done (Anderson et al. 2007). It is a state fiscal funding requirement that STAs submit their conceptual engineer’s cost estimate for
approval of funds four years in advance (AASHTO 2013, FHWA 2021). Unreliable preliminary cost estimates can negatively impact the financial operations of STAs, create inconvenience to the public, negatively affect the STA’s reputation, and add cost and time for the contractor. Public highway projects are consistently coming in at costs higher than what the engineer’s estimate budgets for. Cost overruns are an international phenomenon, with approximately 90% of international transportation projects and 55% of US-based transportation projects experiencing high completion costs (Ammar et al. 2022, Anderson et al. 2007, Herrera et al. 2020, Flyvbjerg et al. 2002). Cost underruns are seldom found in transportation projects; however, they pose the same threat as cost overruns since projects might be erroneously rejected due to failing the cost-benefit analysis based on an incorrect value (AASHTO 2013). There are multiple variables that result in an unrepresentative engineer’s estimate in transportation projects. Literature identifies the following as the most common variables: price variation of materials, legal and political issues, inadequate bidding method, difference between winning bid and second bid, project location, and type (Ammar et al. 2022, Herrera et al. 2020, Zhang 2017).

The bidding methods in the transportation industry heavily relies on competitive bidding. Most STAs only consider the lowest responsible bidder when studying the difference between the engineer’s estimate and the awarded bid (Anderson et al. 2007, Schexnayder et al. 2003). A clear movement towards a best value bidding strategy is starting to be implemented in the transportation industry due to concerns that the low-bid strategy gives an incentive to contractors to cut bid prices by using cheaper material or putting unrealistic values to win projects (Gransberg 2003).

Furthermore, the final design engineer’s estimate, which is the subject of study of this research, is quintessential for measuring the performance of STAs estimating practices over the years as well as reviewing and analyzing the low bids submitted by contractors for the different project types by the STAs (AASHTO 2013). Guidelines posted by the FHWA prior to the year 2021 indicated a minimum threshold of the engineer’s estimate should be within ±10% of the low bid for at least 50% of the projects within that year. The recommendation was that the low bid is to be used as the base for this measure of performance (FHWA 2015). Some STAs have used this guideline as a minimum threshold
and challenged their capabilities using more constrained guidelines. WisDOT, for example, uses a threshold of 60% of the estimates to be within 10% of the low bid (60/10) and, 75% of the estimates to be within 15% of the low bid (75/15) (Nassereddin 2016).

In October 2021, the FHWA updated the guidelines on preparing engineer's estimate, bid reviews, and evaluation. They no longer dictate the guideline that 50% of projects in a year should fall within ±10% of the low bid. It is now merely a recommendation, and the base in the comparison was changed to the engineer's estimate rather than the low bid (FHWA 2021).

There has been an admirable effort on the part of STAs to improve the engineer’s estimate by closely monitoring their historical data, exchanging data with their STA peers as well as looking at the reliability of their methods of cost estimating. AASHTO breaks out the most common estimating methods in the transportation sector from order of most common to least common:

- Historical bid-based estimation
- Conceptual estimation
- Risk-based estimation
- Cost-based estimation
- Combination estimation: Mix of historical bid-based and cost-based estimation

Historical bid based estimating uses data from previously bid projects by the STAs to estimate the cost of upcoming projects (Anderson et al. 2009, FHWA 2021, FHWA 2022). Conceptual estimation depends on a feasibility study by the STA and is not within the scope of this research. Risk-based estimation is primarily used in projects of a large size and is recently gaining traction in the industry to limit uncertainty in projects. Cost-based estimation is a detailed breakdown of all factors affecting the estimate. It is therefore time- and cost-consuming (AASHTO 2013, Anderson et al. 2009, FHWA 2021, FHWA 2022).

1.3 Problem Statement

Even though the FHWA worked to create guidelines to assist STAs in the development of the engineer’s estimate, the guidelines are mostly recommendations with
only a few trainings. STAs are prompted to develop their own cost estimating methodology that they implement at the state level. This means that a lot of their resources are spent on researching and investigating the performance of their estimating methodology (Kermanshachi et al 2018).

WisDOT has allocated a lot of their resources to improving their engineer’s estimate accuracy. Like many other STAs, WisDOT has developed a standardized method of estimate documentation in the form of an Estimate Documentation Report (EDR). The documentation includes information about how the estimate was prepared as well as data related to the proposal bids, and acts as a resource pool for WisDOT to analyze their engineering estimate accuracy and monitor their performance (Nassereddin 2016).

One study by WisDOT in August 2021 aimed to cross-check their estimating methodology and resources against other STA peers. They organized a peer estimating exchange with the help of the Construction and Materials Support Center (CMSC) at the University of Wisconsin – Madison. Goals of the peer exchange were to improve the accuracy of engineer’s estimates, identify options and tools for ensuring a strong competitive bidding environment, reduce unbalanced bidding and improve construction estimates for local program objectives. Representatives from 11 state DOTs and the FHWA participated in the peer exchange.

During the peer exchange, it became obvious that the areas of concern that affect the engineer’s estimate were common among the participating states. There was no definitive answer from states on whether historic bid-based estimating or cost-based estimating results in estimates that are closer to the low bids or awarded bids. The following concerns were identified in the peer exchange:

- Cost estimating intent
- Level of competition (number of bidders)
- Project size
- Construction price fluctuation
- Geographic location
- Lack of experience of the estimator
• Lack of construction knowledge

A great deal of research has been committed to cost prediction models and construction cost indices (CCI) for highway projects. The previous research includes both data driven approaches as well as statistical analysis methods. There is also a plethora of qualitative research identifying reasons behind the difficulty of obtaining accurate engineer’s estimates, particularly when using the definition promoted by the FHWA in the past. The reasons identified in the literature are in line with those identified in the estimating peer exchange. Nevertheless, STAs are still facing difficulty obtaining an engineer’s estimate that comes close to market conditions and the bids submitted by contractors. Additionally, there is close to no research that compares the cost estimating methods primarily used during the engineer’s estimate and whether they impact the accuracy of the engineer’s estimate. Nor is there research that compares data from multiple STAs. It is also noteworthy that STAs involved in the estimating peer exchange were unable to identify a cost estimating methodology that can be objectively deemed superior to the other. As such, this thesis investigates the impact of each cost estimating methodology, historic bid-based estimation vs. combination estimation, and the accuracy of the engineer’s estimate depending on the project’s characteristics, as well as the economic characteristics that directly impact the bidding trends and the cost estimate.

1.4 Research Objectives

The study presented by this thesis on the engineer’s estimate can be summarized by the following objective and sub-objectives:

The primary objective of this thesis is to establish a rigorous machine learning prediction model using data-driven techniques to help identify the most suitable method of cost estimating in terms of the engineer’s estimate, depending on set project variables and economic variables. The goal is to develop models that can help STAs be more comfortable in their estimating protocol for each project, screen unreasonable bids, and quickly set budgets with high reliability. To accomplish this primary objective, and maximize the contribution of this study, the following sub-objectives have been established:
• Analyze the average data from the year 2018 to 2020 submitted by the STAs involved in the estimating peer exchange to produce a list of qualitative and quantitative analysis of the factor(s) that produce(s) more accurate cost estimates.

• Identify influencing variables in the literature that impact the cost of transportation projects as well as factors that affect the engineer’s estimate accuracy. Evaluate the relationships between the identified variables and eliminate repetitive variables.

• Obtain raw data from STAs and develop multiple machine learning models to select the ideal cost estimating methodology based on the identified variables. The machine learning algorithms to be investigated are multiple linear regression (ML), logistic regression, classification and regression trees (CART), and random forests (RF).

• Compare the accuracy of the models and verify the effectiveness of the use of machine learning algorithms to identify the most accurate method of cost estimation.

1.5 Research Methodology

This section is devoted to the development stages of this research paper. The research started with the development of the initial survey for the estimating peer exchange conference organized by WisDOT and the CMSC. The survey served as a documentation of qualitative and quantitative factors that impact transportation project costs and the engineer’s estimate. During the estimating peer exchange conference, more questions were posed, which were later added to the survey and sent to the STAs that were not a part of the estimating peer exchange.

The factors identified through discussions with industry professionals were cross-checked with the available literature. A comprehensive review of existing research was executed to fully understand the governing factors behind transportation projects and their costs. The research team also explored whether research related to the most common cost estimating strategies already exists in the literature. The research team was able to identify an abundance of research related to the creation of a construction cost indices to be used to
better the engineer’s estimate, as well as research related to the qualitative factors that impact transportation project costs. However, there was no quantitative research that compared historic bid-based estimating to cost-based estimating.

The research team then set out to obtain data from the year 2012 to 2020 from all STAs in the US. All 50 states were sent a formal request for data. Unfortunately, due to time constraint and concerns about legality, only 6 states were able to provide the data needed by the research team. A total of 8,635 raw data points between the years 2012 and 2021 were obtained. The research team then combined the project data with economic data related to their location and year of implementation that were obtained from the Federal Reserve Economic Data (FRED) and the U.S. Energy Information Administration (EIA).

After obtaining the dataset, the research team was able to verify the data variables previously identified in the literature and select the optimum features for the analysis. This was followed by an Exploratory Data Analysis (EDA) that provided insight into the dataset and helped the research team detect anomalies and create newly summarized parameters to be used in the dataset. To avoid errors resulting from inconsistent feature scales, the research team applied feature normalization to the variables using a min-max scaler. EDA was also used to discover preliminary relationships between project types and the deviation of the engineer’s estimate from low bid. To further analyze the statistical significance of the variables used in the research, a simple ML model was used to find a relationship between the project bid price and other variables including the engineer’s estimate prepared by the participating STAs. To verify the validity of using ML, the research team underwent tests to verify the assumptions of a linear model. ML assumes that:

1. Features are known exactly and not subject to error
2. Features all have a linear relationship to the target variable
3. Constant variance or homoscedasticity is fulfilled
4. The error terms can be, but are not required to be, normally distributed.

Since the data used was directly obtained from STAs or from verified published economic databases, assumption (1) is satisfied. The research team used partial regression plots to verify assumption (2) and (3) and QQ Plots to check assumption (4).
The research team then moved on to the formulation of the machine learning models with a binary prediction target of whether historic bid-based estimating or the integration of cost-based estimating would be the ideal choice. One model was developed for ML regression, three models for logistic regression, and one model each for classification and regression trees (CART) and random forests (RF). The model performance accuracy was then compared using the area under the receiver operating curve (AUC) to identify the most suitable model to be used by the STAs to decide on whether historic bid-based estimating is sufficient on its own, or some cost-based estimating is required to improve the accuracy of the preliminary engineer’s estimate. The summary of stages is shown in Figure 1 below.

<table>
<thead>
<tr>
<th>Stage 1</th>
<th>Develop Peer Estimate Exchange Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Obtain data averages from participating states</td>
</tr>
<tr>
<td></td>
<td>• Identify factors affecting transportation project costs</td>
</tr>
<tr>
<td></td>
<td>• Draw conclusions on the state of practice of state DOTs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 2</th>
<th>Literature review</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Highway construction phases</td>
</tr>
<tr>
<td></td>
<td>• Cost Estimation</td>
</tr>
<tr>
<td></td>
<td>• Research on preconstruction cost estimating in the highway sector</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 3</th>
<th>Raw data from departments of transportation (DOTs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Summarize project variables</td>
</tr>
<tr>
<td></td>
<td>• Feature selection and explanatory data analysis (EDA)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 4</th>
<th>Machine Learning Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Ordinary Least Square Regression</td>
</tr>
<tr>
<td></td>
<td>• Logistic Regression</td>
</tr>
<tr>
<td></td>
<td>• Classification and Regression Trees</td>
</tr>
<tr>
<td></td>
<td>• Random Forests</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 5</th>
<th>Assessment of developed models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Model performance</td>
</tr>
</tbody>
</table>

*Figure 1: Research Methodology*
CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Cost estimating is a dynamic process in construction projects that is constantly being updated throughout the project’s lifecycle. At early project stages, the levels of expected accuracy are lower than is required at later stages (AASHTO 2013). However, the burden placed on the cost estimates at the early project stages can sometimes be higher than the estimates developed towards the end since it is sometimes the basis that determines whether project owners can afford to execute a project (Anderson et al. 2007, PMBOK 2018). The highway transportation sector follows standardized steps for project development. This section reviews the major phases of highway construction, the principles of cost estimation, and the timeline of cost estimation as it relates to the highway transportation sector. This section also explores the research performed using the state-of-the-art estimating methods published by academics and by STAs in highway construction.

The literature review places an emphasis on the following for the success of this thesis study:

- Variables in previous research that impact transportation project costs
- Proposed methods to improve the engineer’s estimate accuracy
- Research gap concerning the estimating methods detailed by AASHTO and the FHWA
- Limitations of using traditional cost indexing systems and why a shift of focus towards the estimating methods is required

2.2 Major Phases of Highway Construction Projects

The project phases in the transportation sector are divided over 5 major stages that takes the project from planning all the way to project completion. The stages are outlined in the figure below according to the AASHTO and FHWA guidelines. (AASHTO 2013, FHWA 2017, FHWA 2007).
The FHWA notes that there is considerable overlap between the phases of highway planning and development. As such, STAs may slightly deviate from the phases outlined by the FHWA if they fulfill the subdivisions in each of the phases. The sub-sections below identify the subdivisions according to the FHWA guidelines.

### 2.2.1 Planning Phase

The planning phase is specific for the feasibility study of a project and its needs. The FHWA defines this stage as the initial definition of the improvement needs for a highway or bridge project. The improvement can be broken out into:

- Structural repair
- Highway expansion to accommodate future capacity needs
- Improvements due to safety concerns
- Improvements due to bottlenecks and pressure points.

Identification of the problem that prompted the need for improvement is decided by the funding agency. When all three levels of government - federal, state, and local - are involved, all three levels are required to reach a consensus on the type of improvement before the project can move forward from the planning stage. This encompasses the state planning, which is specific for the state’s department of transportation (DOTs), regional
planning, which includes the Metropolitan Planning Organizations (MPOs), and local planning, which can be broken out by cities or geographic state regions (Anderson et al. 2007, FHWA 2017, Zhang 2017).

The planning stage witnesses the first round of cost estimating. The cost estimate at this stage is a rough order of magnitude conceptual estimate that depends on the feasibility study performed by the three levels of government. This estimate serves as the first guideline to the STAs on whether projects are worth pursuing after considering the funds made available to the STAs (Anderson et al. 2009, Anderson et al. 2007, FHWA 2017, USDT 2022). The U.S. Government Accountability Office emphasizes the criticality of the initial cost estimates because they provide valuable information related to the feasibility of a program, how it should be designed, and the resources needed to actualize it (GAO 2022). This estimate is not the primary focal point of the study, but it is still a critical element that impacts the feasibility of projects in the highway transportation sector.

### 2.2.2 Project Development Phase

This stage works on refining the feasibility study in the planning phase with a specific focus on the environmental analysis. STAs are required to explore alternative designs and their impact on cost, and the natural environment. Consequently, industry professionals are brought in to evaluate the alternatives and refine the major features of the projects. Depending on the project size, such professionals may include engineers, architects, archaeologists, historians, and environmental specialists. In some cases, the general public’s opinion on the improvements is evaluated (FHWA 2017). Ideally, the estimating team should include individuals with experience estimating the cost of projects of a similar nature to guarantee that the need for funds is properly expressed (GOA 2022).

The first form of the engineer’s estimate in the highway transportation sector is created in the project development stage. With the major features of the project and alternatives identified, estimating professionals can establish a baseline for the cost. This estimate involves many forms of cost estimating, most commonly, historical bid-based estimating and cost-based estimating (Anderson et al. 2007, FHWA 2017). STAs establish a baseline cost which is used as the project budget. The method of estimating used has historically depended on the complexity of the project, but most STAs rely on historic bid-

2.2.3 Final Design

Cost estimates are of a dynamic nature. The engineer’s estimate does not stop at the project development phase, but instead continues through to the final design. During the project development phase, the project definition completion typically starts at 30% completion up to 90% completion. The final design bridges the final 10% definition to establish a set of project plans and specifications at a significant level of detail for project letting (AASHTO 2013).

The considerations of final design represent an alloy of the requirements and interests posed by the project stakeholders. Common considerations at this stage are:

- Developing a concept
- Considering the project’s scale
- Detailing the design

The multidisciplinary team involved in final design gives importance to designs addressing community values and develops a set of plans and specifications that reflect the details behind the design (FHWA 2017).

Cost estimating does not have a guideline of occurrence by either AASHTO or the FHWA, however STAs usually create milestones at which the engineer’s estimate is updated (Zhang 2017).

Another major difference between the engineer’s estimate in the project development phase and final design is the consideration of the market conditions and an economic analysis. STAs also include probabilistic risk-based estimating during this stage of the development of the engineer’s estimate if a project’s estimated cost is $100 million or more (Anderson et al. 2007, WsDOT 2018). The end of the final design period marks the engineer’s estimate that is used to evaluate incoming contractor bids (Nassereddine 2016, Pakalapati 2018, Zhang 2017).
2.2.4 Right-of-Way-Acquisition

Right-of-way refers to the need of extra land for project completion. STAs or regional authorities purchase lands needed and minorly adjust the engineer’s estimate to reflect the adjustments (FHWA 2017).

As soon as land is acquired, STAs can establish a letting and bidding program leading up to construction. The program includes the completion of the bid package and project advertisement to the public (Anderson et al. 2006). Letting periods are highly dependent on the urgency and size of the project. A manageable letting period is around six to eight weeks, however, funding status, administrative considerations and environmental clearances can majorly impact it (Anderson and Blaschke 2004). Bids are analyzed using the engineer’s estimate developed in the final design phase. STAs sometimes resort to variance reports for differences between the engineer’s estimate and the bid price (FHWA 2014).

2.2.5 Construction

The construction phase covers the physical execution of the project. It includes the selection of the lowest responsive bidder, initiation of contract, general, and site management, and site construction (Anderson et al. 2007). Changes to cost during construction usually come from design changes, improper planning, or unforeseen conditions (WSDOT 2018). The frequency of cost estimates increases during construction. They are conducted to budget spending and monitor project progress, and to prevent cost escalation (GAO 2020, Kermanshachi et al. 2016).

2.3 Cost Estimation

2.3.1 Definition

Cost estimation is a vital component of public project acquisition and execution. The cost estimates serve as the number one guideline behind the need for funding and is sometimes the determining factor behind pursuing one project over another (GAO 2020). Public agencies in the U.S. require cost estimates to be reflective of the actual cost of projects. The U. S. Government Accountability Office (GAO) defines cost estimation as “the summation of individual cost elements, using established methods and valid data, to estimate the future costs of a program, based on what is known today”. Public
organizations are required to develop cost-estimating and budgeting-related policies, procedures, and guidelines to track and analyze cost estimates (AASHTO 2013, FHWA 2021, GAO 2020, PMBK 2018). STAs have strived to optimally allocate limited federal and state funding to highway projects that are essential for the advancement of the economy as well as the lives of the public (ASCE 2021, Liu et al. 2013). Nevertheless, STAs and FHWA Federal-Aid Division Offices (DOs) have faced major challenges when creating the engineer’s estimate (FHWA 2022). These challenges include:

- Shortage of time to develop cost estimates
- Unfamiliar bid items to the estimating team
- Staff shortages
- Lump sum items
- Rework required for shelved projects due to limited funding
- Lack of details in bid packages advertised for bids

STAs started the movement towards a more accurate engineer’s estimate in the 1980s at the request of the Federal Construction Council (FCC) (Morris 1990). To this day, STAs continue to review and analyze the cost estimating strategies to ensure an accurate estimate (Herbsman et al. 1983, FHWA 2022).

The type of cost estimates varies in international and national institutions. The breakout of types primarily depends on the viewpoints of the users in each institution. Consequently, the Association for Advancement of Cost Engineering (AACE), the Project Management Institute (PMI), and the FCC all have separate breakouts of the types of cost estimates (Morris 1990).

2.3.2 Methods of Cost Estimation

The engineer’s estimate serves as a benchmark for contractor bid analysis and approval (FHWA 2021). The estimate can be generated by private spreadsheets created by individual STAs or estimation software such as AASHTO’s Trans.port© software (Anderson et al. 2006). There are four main methods of cost estimate development: conceptual estimating, historic bid-based estimating, cost-based estimating, and risk-based estimating (AASHTO 2013). Estimates generated during the final design stage of the
highway project development life cycle use historic bid-based estimating, cost-based estimating, risk-based estimating, or a combination of each (FHWA 2021). A study by WSDOT in 2018, at the request of the FHWA, found that risk assessment and risk-based estimating is essential to producing engineer’s estimates that are close to the low bid for projects that are more than $100 million dollars in value. The research suggests using quantitative risk analysis for projects that are $10 million dollars in value or more but suggest that the effort should be proportional to the size of the project (WSDOT 2018). In spite of that, there are only 7 state DOTs that are currently applying quantitative risk analysis and risk-based estimating (DRISI 2015). For this research study, only historic bid-based estimating, cost-based estimating, and a combination of both are considered.

The general trend with STAs on cost estimation approaches proves that historical bid-based estimating is the most used approach, followed by a combination of both, and, finally, cost-based estimating (FHWA 2022). The trend of the popularity of the cost estimation methodologies among STAs has been similar through time. In 2007, Niedwecki et al. used a survey to identify the primary cost estimation approaches used by state DOTs. The responses showed that 60% of the respondents used historical bid-based estimating, 30% used a combination, and only 10% relied primarily on cost-based estimating (Niedwecki et al. 2006). The estimating peer exchange organized by WisDOT in 2021 also found the trend to be the same with only 2 of the 11 states using a majority of cost-based estimating and 2 using a combination. The rest of the state DOTs only relied on historical bid-based estimating.

2.3.2.1 Historical Bid-Based Estimating

This method of estimation requires a pre-existing database of old project line items and unit costs. Competitive bids from previous projects are used as the basis for estimated unit price development for future projects. The unit prices are adjusted for project conditions such as project location, size, quantities, and design, as well as economic conditions. The engineer’s estimate is the summation of all the single line items total prices depending on the unit prices generated from the database (FHWA 2021). The typical period range for the bid data used in generating historical bid-based estimating is 3 to 5
years. If a line item is less commonly used, a larger lookback period is recommended (Anderson et al. 2007).

Public agencies dedicate a large amount of energy and resources to sift through competitive bid data obtained from past projects. The data collected by STAs is overwhelming in size, and given the scarcity of resources at the agencies, it’s not utilized to its utmost potential (Woldesenbet 2015). The collection efforts are not backed up by a comparable amount of effort spent on data digestion and knowledge generation. If STAs chose to balance out the extensive efforts spent on data collection with effort spent on knowledge generation, they could benefit from the smaller set of data (Hawkins and Smadi).

Contrastingly, historical bid-based estimating requires the least amount of time and experience to produce an adequate estimate. The trainings related to historical bid-based estimating are also straightforward and do not exhaust the STAs resources (AASHTO 2013, FHWA 2021).

Even though the data collection requires a great deal of effort from STAs, this method is deemed the least resource- and cost-consuming cost estimation method for projects that are based on competitive bidding (Anderson et al. 2009).

2.3.2.2 Cost-Based Estimating

This method of cost estimation is tailored to the project type and represents an actual estimate of each line item in the cost estimate. It considers elements of the cost such as time, equipment, labor, production rates, subcontractors, materials, overhead, and profit to develop a detailed breakdown for each line item (AASHTO 2013, FHWA 2021).

Unlike historical bid-based estimating, the estimator working on the cost-based estimate is required to have a good working knowledge of construction methods and equipment to use this approach. Data from old projects related to production rates are utilized in this method of estimation to achieve a proper estimate of the time and labor component, but the reliance on data collection is not critical (Woldesenbet 2015). It is primarily used in projects that are bid through non-competitive bidding, or projects of a unique nature and insufficient historical bid-data (AASHTO 2013).
While cost-based estimating is generally taken to yield estimates of a higher accuracy than bid-based estimating, it exhausts all the resources available at STAs due to its time consuming nature and the need for highly qualified estimators (AASHTO 2013, FHWA 2021, FHWA 2014).

2.3.2.3 Combination of Historic Bid-Based and Cost-Based Techniques

This method combines the advantages of both historic bid-based estimating and cost-based estimating. Line items that comprise 75% of the total cost are estimated using cost-based estimating, while the rest of the project line items use the pre-existing database with line item costs adjusted according to the specific project characteristics (AASHTO 2013). In combination estimates, the percentage of cost-based estimating can range from 5% up to 60% of all estimate line items (FHWA 2014).

2.3.3 Advantages of Cost-Based Estimating vs. Historical Bid-Based Estimating

Choosing a more appropriate method of estimating projects in the transportation sector has been a difficult task for STAs (AASHTO 2013, FHWA 2021, GAO 2020, PMBK 2018). As discussed earlier, agencies are more likely to resort to historical bid-based estimating due to the time-consuming nature of cost-based estimating (FHWA 2022). However, research has proved that cost-based estimating produces more accurate estimates. This section explores the existing body of research comparing both methods of estimating.

At the request of the South Carolina DOT (SCDOT), Niedzwiecki et. al (2006) performed research on their engineer’s estimates and produced a list of pros and cons for both methodologies. The research team contacted multiple STAs for information regarding both estimating methods. Key findings of this research concluded that more manhours expended could be expected when developing an estimate using the cost-based estimating approach. It was also found that implementing a cost-based estimating approach could require more monetary risk and increased labor. There are many perceived advantages to cost-based estimating, but most of the states that are implementing this method have been using it for over ten years and can be considered experts with this type of estimating. The main advantage of cost-based estimating is that it yields more accurate estimates.
However, there was no quantitative measure on the increase in accuracy when using cost-based estimating to identify its significance.

2.4 Prior Research on the Preconstruction Cost Estimating Services (PCS) for Transportation Projects

2.4.1 Factors Affecting Highway Construction and Preliminary Engineering Costs

The factors impacting the cost of the highway transportation sector have been the subject of study ever since the late 1980s. The list of factors has been consistently updated by STAs over the years to keep up with the changes in the industry. Previous studies offered insights regarding the factors with the highest impact on cost in the highway construction industry. The first reference for the factors was developed by Herbsman (1983-1986). The factors were categorized as technological and organizational factors. The technological factors included (1) design factors, (2) location factors, and (3) material factors. On the other hand, the organizational factors were dependent on (1) construction techniques, (2) transportation methods, (3) human aspects such as management behavior, and (4) employee motivation.

In 1988, the agencies that sponsor the Federal Construction Council asked the Building Research Board (BRB) to review their current practices. The study took 2 years to complete, and the findings were reported in the year (1990). The project findings were that 35% of federal projects were highly exceeding the budget due to multiple factors. The most influential factors from order of most influential to least influential were: (1) poor definition of user needs, (2) long duration between estimating and construction, (3) poor management, (4) poor estimators, and (5) inadequate estimating procedures. Closely after, Moselhi and Hegazy, (1990) concluded that (1) degree of hazard, (2) degree of difficulty, (3) type of job, (4) uncertainty in estimate, (5) historical profit, (6) current workload, (7) risk of investment, (8) rate of return, (9) project owner, and (10) project location, where the governing factors in their cost prediction model depended on artificial neural network (ANN).

Hegazy and Ayed (1998) ascertain 10 most influential factors that in their development of an ANN parametric cost estimating. The primary factors were (1) project
type, (2) project size, (3) year of construction, (4) project location, (5) capacity, (6) project scope, (7) construction season, (8) duration, (9) water body, and (10) soil condition. The research was followed by an extensive study by Elhag and Boissebaine, (1999), where they managed to identify 67 factors that impact the construction cost. The factors were grouped into 6 categories based on their relationship to (1) client characteristics, (2) consultant and design parameters, (3) contractor attributes, (4) project characteristics, (5) contract procedures and procurement methods, and (6) external factors and market conditions.

Turochy et al. (2001) established 10 variables in their creation of a predictive cost estimating model for the Virginia Transportation Research Council. The variables were broken out into objective and subjective variables and included (1) length, (2) number of lanes, (3) earthwork volume, (4) number of intersections, (5) number of grade-separated interchanges, (6) construction cost index, (7) functional classification, (8) location, (9) work type, and (10) incorporation of technology.

AASHTO’S involvement in the cost estimating practices of STAs was first documented in (2003). Schexnayder et al. deduced issues in DOT’s practices that impact the engineer’s estimate. The issues were (1) premature optimism on cost, (2) lack of experience related to risk factors and contingency, and (3) inadequate estimate review protocols.

Sodikov (2005) explored the issues related to preliminary cost estimation from a developing countries perspective. Using data from 65 developing countries, the study investigated the relationship between transportation project cost and other variables in statistical cost prediction models. The common variables with the highest impact identified were (1) work duration, (2) pavement width, (3) shoulder width, (4) ground rise fall, (5) average site clear/grub, (6) earthwork volume, (7) surface class, and (8) base material.
The first practical guidebook for cost estimation in the highway sector (NCHRP 574) was founded in (2007) by the National Cooperative Highway Research Program (NCHRP). This guidebook provided STAs with suitable practices to generate realistic cost estimates and manage the developed cost estimates during the project development process. The factors were separated into internal vs. external categories. Table 1 is adopted from the NCHRP report 574 and signifies the factors causing cost escalation of highway projects.

<table>
<thead>
<tr>
<th>Source</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>1. Bias</td>
</tr>
<tr>
<td></td>
<td>2. Delivery/Procurement Approach</td>
</tr>
<tr>
<td></td>
<td>3. Project Schedule Changes</td>
</tr>
<tr>
<td></td>
<td>4. Engineering and Construction Complexities</td>
</tr>
<tr>
<td></td>
<td>5. Scope Changes</td>
</tr>
<tr>
<td></td>
<td>6. Scope Creep</td>
</tr>
<tr>
<td></td>
<td>7. Poor Estimation</td>
</tr>
<tr>
<td></td>
<td>8. Inconsistent Application of Contingencies</td>
</tr>
<tr>
<td></td>
<td>9. Faulty Execution</td>
</tr>
<tr>
<td></td>
<td>11. Contract Document Conflicts</td>
</tr>
<tr>
<td>External</td>
<td>1. Local Concerns and Requirements</td>
</tr>
<tr>
<td></td>
<td>2. Effects of Inflation</td>
</tr>
<tr>
<td></td>
<td>3. Scope Changes</td>
</tr>
<tr>
<td></td>
<td>4. Scope Creep</td>
</tr>
<tr>
<td></td>
<td>5. Market Conditions</td>
</tr>
<tr>
<td></td>
<td>6. Unforeseen Events</td>
</tr>
<tr>
<td></td>
<td>7. Unforeseen Conditions</td>
</tr>
</tbody>
</table>

Table 1: NCHRP Report 574 – Factors Causing Cost Escalation

Follow-up research by the NCHRP was conducted in (2008) to investigate the performance of STAs that have employed the guidelines in NCHRP Report 574. More internal and external factors that impact the engineer’s estimate were identified, namely, (1) shortage of experienced estimators, (2) political changes in the middle of projects, (3) inadequate tools for cost estimating, (4) lack of consistent policy and guidance, (5) limited skills related to risk management, (6) inadequate engagement of project stakeholders, and (7) shortage of contractors. NCHRP Report 625 (2009) closely followed with a focus on right-of-way (ROW) cost estimating and cost management. The report is used by practitioners to track and manage ROW cost from planning to final design phase.
Molenaar et al. (2010) performed an in-depth risk-analysis on the management practices of STAs in the next attempt by the NCHRP to tackle the trend of underestimating in the highway transportation sector. NCHRP Report 658 identified the factors from the NCHRP Report 574 that were directly related to uncertainty and risk. The subset of factors was comprised of (1) inconsistent application of contingencies, (2) inadequate risk tracking, (3) ineffective risk management for projects of different magnitudes, (4) unrefined organizational structure, (5) political issues, and (6) limitations by the regulatory environment.

Mahamid and Bruland (2010) examined the statistical relationship between estimated and actual cost of highway construction using a regression model. The variables explored in the research are the (1) road length, (2) pavement width, (3) base coarse width, (4) terrain condition, (5) soil drill ability, and (6) soil suitability.

The FHWA sponsored statistical research in (2011) in which 461 North Carolina DOT bridge projects and 188 roadway projects, let between 2001 through 2009, were analyzed. From 42 variables, there were 11 that were responsible for 80% of the data variability. These variables were (1) geographical area of state, (2) planning document responsible party, (3) NEPA document classification, (4) project construction scope, (5) federal funding utilized, (6) right of way cost to STIP estimated construction cost, (7) roadway percentage of construction cost, (8) project length, (9) number of lanes, (10) metropolitan area designation, and (11) length of structures within project.

Jeong and Woldesenbet (2012) developed procedures and models for estimating of the preconstruction costs using 26 factors that fall under the categories of (1) project scope, (2) geographical attributes, (3) design attributes, (4) environmental attributes, and (5) external factors.

The NCHRP Reports 574, 625, and 638 served as the first series of attempts from a federal transportation agency to tackle the strenuous responsibility on STAs to produce accurate preliminary estimates. Each report addressed separate concerns posed by state DOTs. In (2013), AASHTO consolidated the three reports and developed a further detailed guideline. The guideline does not directly address the critical variables that affect the cost of projects in the transportation sector, but it instead provides recommendations to combat
their effect. The recommendations included in the guideline are related to the consideration of inflation, letting strategies for cost control, assessment of contractor bids, and evaluation of the accuracy of cost estimates. The NCHRP followed AASHTO’s guidebook with Report 826, (2016) on the estimating highway preconstruction services costs. The report echoed the findings of the NCHRP reports 574, 625, and 638 with a total list of 8 most influential factors on the costs of highway projects: (1) project type, (2) complexity, (3) National Environmental Policy Act (NEPA) classification, (4) early construction cost estimate, (5) length of project, (6) number of bridges involved in the project, (7) number of lanes, and (8) project location.

Nassereddine (2016) at the University of Wisconsin – Madison assisted WisDOT in developing a construction cost index. The index used 9 different variables: (1) fiscal years, (2) calendar years, (3) months, (4) seasons, (5) funding programs, (6) proposal type, (7) region, (8) WisDOT range, and (9) FHWA range.

Gardner et. al (2016) investigated the use of stochastic approaches to pre-construction estimates and created an empirical distribution using ANN to express the cost range. They used a total of 12 variables including, (1) Urban/Rural Indicator, (2) reservations, (3) design AADT, (4) site topography, (5) start and end stations, length, and width, (6) number of bridges in scope, (7) intersection signalization and signage, (8) contract time, (9) typical section (depths of surfacing and aggregate), (10) curb, gutter and sidewalk, (11) bridge deck areas, traffic control, and (12) extent of utility relocations and costs.

The following year (2017), Zhang used 6 project related variables: (1) number of bidders, (2) original contract days, (3) original contract amount, (4) contractor past performance rating, (5) length of project, and (6) number of lanes and weather days. As well as 4 economic indicators: (1) consumer price index, (2) construction spending, (3) producer price index, and (4) prime loan rate to develop models that estimates the contract amounts of transportation projects.

In (2018), Pakalapati used a total of 5 variables to identify the optimum look-back periods when using historical bid-based estimating. The factors were (1) project scale, (2) project year, (3) geographic conditions, (4) level of competition, and (5) complexity. In the
following year, Baek (2019) used logistic regression (LOGIT) to identify factors affecting the low-bid deviation from the agency’s engineer’s estimate using 959 highway projects between 2011 to 2015 in the State of Louisiana. The 4 variables with the most significance on the deviation were (1) number of bidders, (2) number of activities in the contract, (3) crude oil price, and (4) the value of construction put in place of pavement projects.

In surveying STAs, the FHWA found that the most impactful oversight when preparing the engineer’s estimate comes as a result of unsuccessful identification of factors that influence projects cost (FHWA 2022). The literature review conveys that research related to influential factors on the engineer’s estimate has been extensive. The summary of independent variables is presented in Table 2 on the next page. Some variables of similar nature were combined for consistency.
<table>
<thead>
<tr>
<th>Attributes</th>
<th>Herbsman</th>
<th>Morris</th>
<th>Moselhi and Hegazy</th>
<th>Hegazy and Ayed</th>
<th>Elhtag &amp; Boissebaine</th>
<th>Turecy</th>
<th>Schexnayder</th>
<th>Sokolov</th>
<th>Andersen (NCHRP 57/4)</th>
<th>Andersen (NCHRP 62/5)</th>
<th>Molenara and Bruland</th>
<th>Liu (FHWA)</th>
<th>Jeong</th>
<th>Nassereddine</th>
<th>Gardner</th>
<th>NCHRP (Report 826)</th>
<th>Zhang</th>
<th>Pakalapati</th>
<th>Baek</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project size</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project year</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original contract duration</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final contract duration</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project location</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project complexity</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project type</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time between estimate and construction</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management behavior</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change orders</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No of lanes</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of bidders</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Utility adjustment</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Economic state</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 2: Summary of Independent Variables Identified in Literature*
2.4.2 Data-Driven Based Cost Estimating Research in the Highway Sector

The use of quantitative forecasting in the transportation sector to predict future estimates is the primary subject of research on the engineer’s estimate. Quantitative models have also helped researchers identify the most significant factors that affect the cost of projects in the transportation sector as well as provided a less time-consuming method to predicting the engineer’s estimate. The quantitative methods employed in prior research can be categorized into statistical and causal methods. Statistical methods utilize time-series analysis and model curve fitting with a goal to predict the target value as accurately as possible for future observations, while causal methods require the variables and target to have a direct explanatory relationship. In this case, care is needed with respect to model assumptions, otherwise the information extracted from the model may be misleading (at best) or incorrect (at worst) (Hanna and Blair 1993, Touran and Lopez 2006).

The first causal model was developed by Herbsman (1983) to forecast the long-range cost projections after providing weights to the cost elements by order of importance. The model developed was based on data gathered from FDOT on projects from the year 1968 to 1984. At that time, research on relevant variables for transportation cost estimating was scarce, and they relied on (1) cost components such as indirect and direct costs for labor, equipment, materials, etc., (2) indicators from the U.S. Bureau of Labor Statistics (BLS), (3) influencing factors, and (4) prediction processes. In (1986), Herbsman developed another model pursuant to the 1983 version. They utilized a multiple regression algorithm to estimate the cost using influencing factors from two categories: technical and organizational.

Bell and Bozai (1987) followed in suit and used multiple regression to develop bid-based cost estimating models for the Alabama Department of Transportation (ALDOT). Using 174 datapoints, Bell and Bozai’s multiple regression equations calculated project costs per mile with an estimating accuracy ranging from ±17% to ±35%.

The first documented use of ANN in preliminary cost estimation models related to the transportation industry was by Moselhi et al (1990). Their research focused on developing multiple algorithms for different components of the preliminary estimate,
including the evaluation of risks, construction markup, planning, and scheduling. They also used a combination of the components to optimally predict bid prices.

Hegazi and Ayed (1998) developed a parametric cost-estimating model for highway projects using 18 datapoints from highway projects constructed in Newfoundland, Canada. They created two separate models, one for the estimation of the total project cost and the other for the five major cost components of highway projects: cost of site work, excavation, materials, paving, and others.

Williams (2003) used a regression model based on the natural logarithmic scale of both the low-bid and actual project cost for each transportation agency involved in the study. The total project cost could then be predicted using the low bid estimate.

Minchin et al. (2004) provided an overview on the accuracy of roadway projects in comparison to bridge projects. Using a regression model, they were able to forecast the ratio of low bid to the engineer’s estimate. They also discovered that economic factors had a significant impact on the engineer’s estimate.

Mahamid and Bruland (2010) managed to build 11 regression models to predict highway costs with model accuracy of 69% to 87%. The area under the receiver operating curve (AUC) was between 0.92 and 0.98. In (2011), the FHWA analyzed 461 North Carolina DOT (NCDOT) bridge projects and 188 roadway projects, let between 2001 and 2009, and developed statistical models linking variation in preliminary engineering costs and preliminary engineering duration with distinctive project parameters.

El Asmar et al., at the University of Wisconsin-Madison (2011), used 77 projects in 14 highway corridors to describe a statistical approach to producing a reliable conceptual cost estimate. They used an analysis similar to the program evaluation and review technique (PERT) used in scheduling to predict the construction costs in the conceptual stage.

Furthermore, Jeong and the Oklahoma Transportation Center (2012) developed a procedure and model for estimating preconstruction costs of highway projects. The developed system allows engineers to easily manipulate the requirements for a specific plan development task and helps them configure contingencies as to whether any of the
entities are either under/over-estimated or if there exists misallocation of resources. 353 projects from the Oklahoma Department of Transportation (ODOT) were evaluated using 26 factors that fall under the categories of project scope, geographical attributes, design attributes, environmental attributes, and external factors. Jeong used decision tree models, regression models, and neural network models, and found that ANNs had the least error.

On the other hand, Liu et al. (2013) developed an assessment of the cost-estimation strategy for roadway projects using 188 projects between 1999 and 2009 in North Carolina. A multiple linear regression prediction model was developed to estimate the preliminary cost estimate. The research found a significant correlation between the preliminary engineering cost estimate and the preliminary engineering duration. Also, Hollar et al. (2013) created a cost estimating model to predict the cost ratio of bridge projects using 461 projects between 2001 and 2009 in North Carolina. An MLR model for the preliminary engineering estimate was created.

Choi et al. (2014) created a conceptual cost prediction model by combining rough set theory, case-based reasoning, and genetic algorithms to better predict costs in the conceptual planning phase in roadway construction. A total of 191 projects in South Korea were used to develop the model with all variables attributed to project specific details, notwithstanding economic factors.

In addition, Ilbeigi et al. (2014) researched the use of a stochastic process to model the fluctuations of asphalt cement prices, which is one of the leading causes of a low accuracy of engineer’s estimate when compared to low bids. The research was done using the Geometric Brownian Motion (GBM) as the stochastic process to model random variations of asphalt cement price over time using data between 2005 and 2012 from 15 asphalt suppliers in Georgia. A probabilistic approach based on the Monte-Carlo simulation was applied on the GBM model to simulate future random paths for asphalt cement price index. The research’s aim was to help contractors and transportation agencies systematically analyze variations in the price of asphalt cement and develop more accurate estimations for their transportation projects.

At the same time, Ćirilović et al. (2014) developed a cost estimation model for road rehabilitation and reconstruction of projects using 94 projects in 14 countries in Europe.
and Central Asia. The variables used were either related to oil price, country specific, or project specific.

Also, in (2015), Chou et al. investigated the use of optimized artificial intelligence models for predicting the project award price using ANN and multiple regression analysis to be used by contractors to make bidding decisions. During the study, 275 bridge construction projects were evaluated using general linear regression (GLR), genetic-algorithm-based nonlinear regression (GA-NLR), genetic algorithm-artificial neural networks (GA-ANN), and case-based reasoning (CBR). The results showed that the GA-ANN model exhibited optimal prediction performance, indicating that ANNs are superior mathematical models for realistic simulations.

Nassereddine (2016) created a construction cost index for WisDOT using a time series model. The research contained 2,229 observations that represented contracts awarded by the WisDOT from 2009 to 2015. The historical dataset tracks different project characteristics. Relevant variables include the contract ID, engineer’s estimate, low bid price, number of bidders, fiscal year, region, type of work, and funding program.

On the other hand, Zhang (2017) used 741 projects from Florida DOT’s database that occurred between 2003 and 2015 to determine the best data driven model to estimate the contract amounts of resurfacing projects. 5 different types of machine learning algorithms were used, Ordinary Least Square Linear Regression (OLS), Back Propagation Neural Network (BPNN), Least Absolute Shrinkage and Selection Operator Linear Regression (LASSO), Multivariate adaptive regression spline (MARS), and General Regression Neural Network (GRNN). Zhang used 7 project related factors (Variables, No. of Bidders, Original Contract Days, Original Contract Amount, CPPR, Length of Project, Number of Lanes, and Weather Days) and 4 economic indicators (Consumer Price Index, Construction Spending, Producer Price Index, and Prime Loan Rate) to develop the models that estimate the contract amounts. Based on the evaluation, the MARS model outperformed the other 4 machine learning algorithms, and it was proven that economic indicators do in fact have an impact on contract prices of transportation projects.

In (2018), Pakalapati et al. used 2,122 projects from the Alabama DOT’s database between 2011 and 2016 to identify the optimum look-back periods for historical bid-based
highway projects. AASHTO’s guideline is a minimum of 1-2 years look-back period for historical bid-based estimating, but Pakalapati evaluated a range of 1-5 years to identify the optimal look-back period for data retrieval and the most suitable indexing method from 12 indices developed in the research and 8 previously existing indices. The study recommended the use of a 2-year look-back period and the use of quarterly cost indices for historical bid-based estimating.

Baek et al. (2019) used logistic regression (LOGIT) to identify factors affecting the low bid deviation from the agency’s engineer’s estimate using 959 highway projects between 2011 to 2015 in the State of Louisiana. The four variables with the most significance on the deviation found were number of bidders, number of activities in the contract, crude oil price WTI, and the value of construction put in place of pavement projects.

In January (2021), Farshidpour et al. developed a statistical approach to analyzing engineering estimates and bids within the transportation sector with the purpose of identifying methods to prevent errors in the estimates that result from uncertainties in construction methodology, cost, and time. Using 50 projects, the research found that choosing the lowest bid overall was the most reliable approach.

Gaikwad (2021) investigated the disparity between the engineer’s estimate and the bids in 305 DB (Design-Build) highway projects. Contrary to the guidelines set that suggest that the engineer’s estimate should be within ±10% of the lowest bid, the research found that a more accurate percentage for DB highway projects should be 25%. The research alternatively suggested the development of new policies for DB projects to benefit from the project delivery system’s advantages.

There is no shortage of models when it comes to creating predictive models for STAs to estimate the low bid price of highway construction. The literature review conveys that research directly impacted the knowledge STAs have regarding the most significant factors affecting highway construction cost. Summary of previous studies on the engineer’s estimate is presented in Table 3 on the next page.
2.5 Literature Gap

The extant body of literature contains numerous studies concerning the factors that influence the cost of transportation projects and data-driven construction cost indices in support of historical bid-based estimating at multiple STAs. However, there is scarce research on cost-based estimating and the advantages of employing this estimating technique on predicting the cost of construction projects. In fact, all research exploring the methods of cost-based estimating is inconclusive, and merely suggests that cost-based estimating requires training before becoming the wide scale approach to cost estimating in the transportation industry.

The Engineering New Records (ENR) (2022) published a study on the benefits of cost-based estimation in the construction industry. It provides a better understanding of how projects are built, a better chance at estimation reliability, and a better understanding of specific contractor pricing.

In conclusion, both methods of estimation have advantages that prompt STAs to employ them. A combination of both methods at STAs might lead to a higher overall engineer’s estimate accuracy per year. This research aims to address this research gap by analyzing the findings of the estimating peer exchange organized by WisDOT with the help of the CMSC at UW-Madison and conducting a holistic quantitative predictive model to select the optimum method of estimation based on project characteristics and economic characteristics.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Models for the Cost</td>
<td>Analysis/Recommendation</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Cost Estimating Guidelines</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimating type comparison</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factors Impacting Highway Project Cost</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordinal Least Square Linear Regression (OLS)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Regression Analysis</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artificial Neural Network (ANN)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression (LOGIT) or other non-linear regression</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision Trees</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical/ Stochastic Analysis</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Datapoints</td>
<td>3,000</td>
<td>3,000</td>
<td>174</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>461</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>131</td>
<td>461</td>
<td>77</td>
<td>353</td>
<td>-</td>
<td>191</td>
<td>188</td>
<td>461</td>
<td>94</td>
</tr>
</tbody>
</table>

Table 3: Summary of Research on the Engineer’s Estimate in the Transportation Industry
CHAPTER 3: ESTIMATING PEER EXCHANGE

3.1 Introduction

WisDOT hosted a virtual peer exchange meeting on August 3rd and 4th, 2021, to discuss topics related to improving procedures for developing the engineer’s estimate for highway improvement projects. Prior to the exchange, each state was asked to fill out an on-line survey to obtain preliminary information regarding their estimating procedures and practices, organizational approach, measurements of estimating accuracy and bidding history, among other things.

The focus of the peer exchange was to improve accuracy of engineer’s estimates and identify options and tools for ensuring a strong competitive bidding environment and reducing unbalanced bidding. It provided a forum for the exchange of ideas, lessons learned, and opportunities for the development of state-to-state relationships for continued improvements. The main topics of interest included:

- Improved accuracy of engineer’s estimates
- Creating a strong competitive bidding environment
- Reduction in unbalanced bidding
- Improved construction estimates for local program projects

This chapter presents the combined findings of the peer exchange and the survey sent out to STAs regarding the engineer’s estimate accuracy.

3.2 Survey Summary

The survey was developed by the research team with the help of WisDOT. A list of information related to the engineer’s estimate, number of bidders, and estimate accuracy was incorporated in the survey. The survey questions captured quantitative data related to (1) average percentage of cost-based estimating used, (2) percent of bids within ±10% of estimate in the years 2018, 2019, and 2020, and (3) average number of bidders. It also captured qualitative data including (1) estimate documentation, (2) goal of engineer’s estimate, (3) primary estimating approach, and (4) estimate preparation.
3.2.1 Demographic Distribution of Respondents

States from the Midwest, Western, and Southern Regions of AASHTO, (MAASHTO, WASHTO and SASHTO) were invited to participate. Representatives from 16 state DOTs provided data in the survey as shown in Figure 3 below.

![Geographic Distribution of Survey Respondents](image)

Figure 3: Geographic Distribution of Survey Respondents

3.2.2 State Practices for Developing Preliminary Estimates

This section presents the findings related to the estimating practices of the aforementioned states.

- **Tracking Engineer’s Estimate Accuracy:**

![Pie Chart](image)

Figure 4: Do you track the engineer's estimate accuracy?
Tracking the engineer’s estimate accuracy is recommended for all stateDOTs, however, it is not a requirement (FHWA 2022). Figure 4 represents the response of the states involved in the peer exchange on whether they track the engineer’s estimate accuracy. Of the 16 states, 14 states reported tracking the engineer’s estimate across over the years.

- **Estimate Preparation Protocol:**

  States who prepare their own engineer’s estimate follow one of three different approaches to estimating.

  1. Centralized estimate - the estimate preparation group is concentrated within a particular geographical location group.
  2. Decentralized by project staff - the estimate preparation group is spread out across all offices. Each estimate is assigned to a project staff.
  3. Decentralized by designated staff - the estimate preparation group is spread out across all offices. Each office has a designated staff of estimators.

  A centralized approach ensures that a group of highly skilled professionals are in charge of all the estimates. This might be a disadvantage since it places a large load on a small group of individuals. A decentralized approach prioritizes the involvement of all offices; however, it introduces a variability in the experience of estimators creating the estimates (Miller and Evans 2006).

![Figure 5: Estimate Preparation Process](image)

Figure 5 presents the summary of the estimate preparation protocol of the states in the peer exchange. The majority of states use a decentralized estimate
preparation approach. This means that they do not rely on one central location to prepare the estimates. In most cases, the project staff prepare the estimates themselves.

- **Goal of the Engineer’s Estimate:**
  
  The engineer’s estimate is normally compared to the lowest bid price from a list of contractor bids for the project. Some states prefer to target a “fair and reasonable” estimate. The goal for this estimate is to come close to the second or third lowest bid price from a list of contractor bids.

![Figure 6: Goal of the Engineer’s Estimate](image)

About 70% of the states conform to the standard that targets a low bid price, but 30% of the states prefer to obtain a fair and reasonable estimate, as shown in Figure 6.

- **Primary Estimating Approach:**
  
  The primary approach for the engineer’s estimate is divided across historical bid-based estimating, cost-based estimating, and a combination of both estimating methods.
Figure 7 shows that the majority of states resort to using the historical bid-based estimating method when preparing the engineer’s estimate. States that use a combination approach to estimating came second with only 2 of the 16 states using a cost-based estimating approach.

- **Indexing for Historical Bid-Based Approach:**
  
  Under historical bid-based estimating, historical bid data is usually adjusted by the STAs for project conditions and the general market conditions ([AASHTO 2013](#)). The figure below presents the distribution among the state DOTs that use an index to adjust cost and the states that do not.

While the FHWA ([2021](#)) recommends that historical bid-based estimating is accompanied by indexing and the plethora of construction cost indices (CCI) created in the body of literature, the distribution among the states proved that
more than half of the states do not use an index to adjust prices as portrayed in Figure 8.

3.2.3 Quantitative Findings from Survey

- **Average Number of Bidders:**

  The number of bidders has been identified in the literature as one of the main factors that affect the highway construction costs (Anderson et. al 2009, Baek et. al 2019, Elhag and Boissebaine 1999, Pakalapati 2018, Turochy 2001, Zhang 2017).

![Figure 9: Average no. of Bidders from 2018 to 2020](image)

Figure 9 shows the average number of bidders for all respondents in the peer exchange. The 3-year average was 3.2 bidders per project. However, there was a fair amount of variation from a low 2.4 bidders per project to a high of 3.7 bidders per project. The value across the three years for the average number of bidders remains between 3 to 4 bidders.
• **Average Percentage of Cost-Based Estimating Used:**

This metric relates to the percentage of bid line items by cost that is estimated using the cost-based estimating approach. It primarily relates to the STAs that utilize the combination estimate approach (AASHTO 2013, FHWA 2014).

![Bar Chart]

*Figure 10: Number of State DOTs Using Different Percentages of Cost-Based Estimating*

Similar to the trends identified in the body of literature, the majority of state DOTs resort to only using historical bid-based estimating, which is portrayed in the 0% cost-based estimating used bar in Figure 10. There is a comparable number of states using 5-10% cost-based estimating for their line items, while only 2 of the states are primarily using cost-based estimating, with 60-70% of the line items estimated using the cost-based estimating approach.
• **Percent of Bids within ±10% of the Engineer’s Estimate in the Years 2018, 2019, and 2020:**

The FHWA (2021) recommends that 50% of low bid projects in a year should fall within ±10% of the engineer’s estimate. This is the primary method the state DOTs use to measure their performance.

![Figure 11: Average Percentage of Low Bids Within ±10% of the Engineer’s Estimate](image)

Figure 11 shows that the average percentage per state from the year 2018 to 2020 resonates around the 50% mark. It ranges from a low of 30% to a high of 70% depending on the state DOTs. The median percentage across the years is 50%, 54% and 52% respectively. This indicates that the state DOTs are just fulfilling the recommendations by the FHWA.
• **Percent of Bids within ±10% of the Engineer’s Estimate by Method of Estimating:**

The percentage of low bids within ±10% of the engineer’s estimate was also used to compare the accuracy of the three methods of cost estimation in the highway sector. State DOTs were separated by the primary method of estimating.

![Comparison of Bids within ±10% of Engineer's Estimate by Method of Estimating](image)

*Figure 12: Average Percent of Bids within ±10% of the Engineer’s Estimate by Method of Estimating*

There were 8 datapoints for historical bid-based estimating, 4 data points for combination estimate, and 2 datapoints for cost-based estimating in Figure 12. The results show that cost-based estimating yielded the highest estimate accuracy at a low of 50%, high of 56%, and an average of 53%. Historical bid-based estimating and using a combination estimate had a similar mean value for the average percent of bids within ±10% of the engineer’s estimate, however, the combination method of estimating has a lower variability with a low average of 41% vs. 33% for the historical bid-based estimating method.

### 3.3 Factors Affecting Highway Construction Costs

State DOTs used the opportunity to share their findings on the factors that cause a significant variation between the engineer’s estimate and contractor bids.

- **Bid Analysis:**

  Inciting bid competition was the one main conclusive factor identified in the peer exchange. Single bids are problematic for most states and only a few reported penny bids as being an issue. Reviews for award tend to focus on
quantities to ensure the bids are not materially unbalanced and other issues that may not have been accounted for to justify the bid price.

Some initiatives posed by the states to increase the bid competition include:

(1) Small business initiative: This initiative serves small and medium business enterprises. It invites contractors who made under a certain amount the year prior to bid on the projects. This instigates those smaller businesses, who may have trouble finding bids, to stay in business and, in turn, serves the level of competition.

(2) Provide a flexible contract time: For projects with no impact on the safety of the public, state DOTs are providing a flexible contract timeline so that contractors have the flexibility to assign crews to projects who would have otherwise been busy.

(3) Combine small projects into one project: State DOTs are also combining projects of a similar nature that are less than $10M into one contract as an incentive for bidders.

- Geographic Location:
  All states reported that rural geographic locations are more likely to witness a reduction in the number of bidders. Rural locations usually have a limited number of local contractors and aggregate quarries.

- Historical Bid-Based Estimating Limitations:
  State DOTs identified one of the factors that impacts the deviation between low bid and the engineer’s estimate as the use of historical bid-based estimating. The concern relates to the fact that the data relies on past bids and doesn’t consider future market conditions.

- Skilled Estimators and Time Shortage:
  States have also identified that limitations related to the availability of time to perform the engineer’s estimate and the level of experience that the estimators have highly impacted the quality of the engineer’s estimate that is delivered.

- Market Conditions (Volatility):
The state of the economic market is another main factor that the state DOTs identified. Issues related to the supply chain, labor shortages, inflation, etc. can have a direct impact on the project cost.

3.4 Limitations of the Estimating Peer Exchange

The estimating peer exchange provided a deeper insight on the performance of multiple STAs. It also allowed for some quantitative analysis of the performance of the different cost estimating methods. However, the data collected using the survey only allowed for analysis related to the average percentage of low bids that fall within ±10% of the engineer’s estimate. It also outlined the advantages of using cost-based estimating, historical bid-based estimating, and a combination estimating approach, but didn’t draw any conclusive evidence detailing which of the methods is the most appropriate method of estimating in the transportation industry.

In the extensive literature review performed by the research team, an abundance of factors that impact the cost of transportation projects was identified. The estimating peer exchange provided a firsthand insight from industry professionals on the factors they consider of significant impact on cost.

Using this information, the research team embarked on the second stage of this research, which entails collecting another set of data that is focused on the project-based performance of transportation projects between the period 2012 and 2020.
CHAPTER 4: DATA FOR PREDICTION MODEL

4.1 Data Variables

Based on the variables identified in the literature and verification by the professionals through the estimating peer exchange, the research team consolidated a list of variables that impact the cost of transportation projects for the purpose of the second stage of the research. The variables were divided into project related variables and economic indicator variables. Project related variables refer to the variables obtained from the STAs related to project characteristics such as project type and project location. Economic indicator variables relate to the market trends during a given year such as the funding provided to STAs and employment in construction. 6 out of the 50 state DOTs contacted were able to grant the research team project related data with the information needed for the successful completion of this thesis. The states are shown below in Figure 13.

![Demographic Distribution of State DOTs Involved in Stage 2](Figure13.png)

Figure 13: Demographic Distribution of State DOTs Involved in Stage 2

A total of 8,635 raw data points between the years 2012 and 2021 were obtained. The economic data was retrieved from the Federal Reserve Economic Data (FRED) and the U.S. Energy Information Administration (EIA).
Project related variables:

The project related variables were obtained from the data provided by the STAs. This section defines each of the variables and their corresponding values.

- **Contract Type:**
  Contract type relates to the contract pricing method used for the project. The contract type is a governing factor when distributing risk among the project parties (Godwin 2013).
  The contract pricing methods can be:
  - **Lump-Sum:** The contractor in this contract pricing method agrees to perform the work for a predetermined price that includes profit.
  - **Unit Price:** The unit price is predetermined, but the quantities are estimated.
  - **Cost-plus (Fee or Percentage Fee):** The contractor is paid based on the cost of material and labor plus a fixed fee or fixed percentage fee.
  - **Guaranteed Maximum Price:** The cost is determined based on a maximum number and profit.
  - **Target Cost:** All project stakeholders agree on a predetermined target cost for each construction system.

  The contract type is a nominal categorical variable.

- **Type of Work:**
  The type of work identifies the type of construction work to be undertaken. The variable originally contained 1,027 levels when the research team combined the data from all six state DOTs. The research team then summarized the variable into the following 11 levels:
  - **Bicycle/pedestrian:** Construction of new bicycle/pedestrian bridges.
  - **Bridge replacement:** Demolition and reconstruction of bridge structures.
  - **Earthwork:** Construction of earthwork, including drainage and other aggregate work.
• Environmental mitigation: Work related to wetland protection or rehabilitation.
• Maintenance or minor upgrades: Minor repair or upgrade work such as steel painting, roadway painting, signage, or fencing.
• New bridge construction: Construction of new bridge structures.
• Resurfacing: Work related to pavement resurfacing in the form of seal coats, overlays, or hot mix asphalt.
• Road or culvert replacement: Demolition and reconstruction of roadway or culvert structures.
• Roadway redesign: Improvements related to the roadway expansion.
• Safety and traffic control: Traffic signal installation or improvements for pedestrian and traffic safety.
• Utilities: Site lighting, plumbing or mechanical upgrades, or new construction.

• Location Type:
  Location type is an ordinal categorical variable made of two levels. It can be converted into a binary metric indicating urban vs. rural locations.

• Forecasted Duration Range:
  Project size can be measured by the construction duration range a project is expected to take. The duration ranges are divided across 4 levels:
  • Level 1: < 30 days
  • Level 2: 30 ~ 99 days
  • Level 3: 100 ~ 299 days
  • Level 4: > 299 days

• Advertising Time:
  The advertising time is defined from the day of the first bid advertising to the bid opening. It measures the time available for bidders to collect data and create a well-rounded cost estimate that is reflective of market conditions. It is a numerical variable that’s measured in number of days.
• **Project Size:**

Another metric related to the complexity of a project is the project size. This variable is measured by the forecasted estimate range based on the conceptual cost estimate. The ranges are spread over 8 levels as defined by the literature:

- Level 1: < $10,000
- Level 2: $10,000 ~ $49,999
- Level 3: $50,000 ~ $99,999
- Level 4: $100,000 ~ $499,999
- Level 5: $500,000 ~ $999,999
- Level 6: $100,000,000 ~ $999,999,999
- Level 7: $100,000,000 ~ $249,999,999
- Level 8: > $249,999,999

• **% of Cost-Based Estimating Used:**

This measures the utilization per project of the line items estimated using the cost-based estimating technique. This metric is not tracked by DOTs, so the research team instead used the average percentage that each DOT uses for projects as the categories. The data obtained by the research only encompassed DOTs that utilize 0%, 5%, and 10% cost-based estimating. As a result, this metric was converted to a binary ordinal categorical metric that measures whether only historical bid-based estimating was used or a combination of historical bid-based estimating and cost-based estimating.

• **% Difference Between the Engineer's Estimate and Low Bid:**

As defined by the FHWA 2022 recommendation, this variable is a numerical variable, and it is based on the percentage variation of low bid from the engineer’s estimate. It is calculated using the following equation, Equation 1:

\[
\% \text{ Difference} = \frac{\text{Low Bid} - \text{Engineer's Estimate}}{\text{Engineer's Estimate}} \times 100
\]

Equation 1

• **Change Orders:** Measures the change order by dollar value that increased the project scope. It is a numerical variable.
• Weather days: Number of approved contingency days for weather days. It is a numerical variable.

• Bituminous Adjustments: Percentage adjustment due to bituminous cost changes. It is a numerical variable.

• Fuel adjustments: Percentage adjustment due to fuel cost changes. It is a numerical variable.

• Number of lanes: Number of highway lanes that fall within the scope of construction in the project. It is a numerical variable.

• Number of bidders: Number of contractors who bid on the project. It is a numerical variable.

• Contractor Past Performance Indicator (CPPI): The previous score assigned to the awarded contractor. It is a numerical variable.

**Economic Indicator Variables:**

The economic indicator variables reflect the state of the transportation construction market. Using a combination of concerns, from lack of funding to lack of construction laborers, these factors ensure that the state of the economy is well reflected in the research.

(1) **Consumer Price Index:**

As defined by the U.S. Bureau of Labor Statistics (BLS), consumer price index (CPI) is the instrument to measure inflation. It is used to estimate the average variation between two given periods in the prices of products consumed by industry type and location. The CPI used in this research relates to the transportation industry and is a numerical variable.

(2) **Construction Spending (CS):**

Construction spending is an economic indicator that measures the amount of spending toward new construction by million dollars by month. The research team only considered data related to the transportation industry. This variable is a numerical variable.

(3) **Producer Price Index (PPI):**

As defined by the U.S. Bureau of Labor Statistics (BLS), the consumer price index (CPI) is the instrument to measure the average change over time in the selling
prices received by domestic producers for their output. It is determined by industry type and location. The PPI used in this research relates to the transportation industry and is a numerical variable.

(4) Prime Loan Rate (PLR):

The prime loan rate signifies the current interest rate that financial institutions in the U.S. charge public entities or entities with high credit scores. It is a numerical variable that changes by year.

(5) Employment Level in Construction:

This measures the average weekly hours of production by construction laborers and non-managerial positions in the transportation sector. This numerical measure provides an insight on labor shortage in the industry depending on the year of construction.

(6) Average Hourly Earnings:

The hourly earnings are another measure of construction activity in the transportation sector. The focus is on the average dollar amount in earnings in the transportation industry. It is a numerical measure.

(7) Crude Oil Prices:

Transportation of goods requires the utilization of crude oil. This numerical metric measures the spot variation in various barrels of crude oil during a given period. The periods for this research were determined by year.

The combined list of variables identified are shown below in Table 4.
<table>
<thead>
<tr>
<th>Variable Category</th>
<th>Variable</th>
<th>Description</th>
<th>Variable Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Related Variables</td>
<td>Contract Type</td>
<td>Contract pricing method for the project</td>
<td>Nominal Categorical</td>
</tr>
<tr>
<td></td>
<td>Type of Work</td>
<td>Nature of the project. E.g.: Maintenance or minor upgrades</td>
<td>Nominal Categorical</td>
</tr>
<tr>
<td></td>
<td>Location Type</td>
<td>Urban or rural</td>
<td>Nominal Categorical</td>
</tr>
<tr>
<td></td>
<td>Forecasted Duration Range</td>
<td>Forecasted project duration</td>
<td>Ordinal Categorical</td>
</tr>
<tr>
<td></td>
<td>Advertising Time</td>
<td>Time between letting date and award date</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Project Size</td>
<td>Metric based on the cost bracket of the project</td>
<td>Ordinal Categorical</td>
</tr>
<tr>
<td></td>
<td>% of Cost Based Estimating Used</td>
<td>Percentage by line item cost estimated using cost-based estimating</td>
<td>Ordinal Categorical</td>
</tr>
<tr>
<td></td>
<td>% Difference between the Engineer's Estimate and Low Bid</td>
<td>% deviation of the low bid from the engineer's estimate</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Change Orders</td>
<td>Approved change orders in dollars</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Weather days</td>
<td>Number of approved contingency days for weather days</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Bituminous Adjustments</td>
<td>% adjustment due to bituminous cost changes</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Fuel adjustments</td>
<td>% adjustment due to fuel cost changes</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Number of lanes</td>
<td>Number of lanes in project</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Number of bidders</td>
<td>Number of bidders per project</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Contractor Past Performance Indicator (CPPI)</td>
<td>Past rating assigned by STA to contractor</td>
<td>Numerical</td>
</tr>
<tr>
<td>Economic Indicator Variables</td>
<td>Consumer Price Index (CPI)</td>
<td>measures the change in prices paid by consumers for goods and services</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Construction Spending (CS)</td>
<td>measures monthly expenditures toward new construction</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Producer Price Index (PPI)</td>
<td>measures the average change over time in the selling prices received</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Prime Loan Rate (PLR)</td>
<td>the interest rate at which banks lend to customers with good credit</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Employment Level in Construction</td>
<td>Average weekly hours of production</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Average Hourly Earnings</td>
<td>Average hourly $ earnings in the transportation industry</td>
<td>Numerical</td>
</tr>
<tr>
<td></td>
<td>Crude Oil Prices</td>
<td>Measures the spot variation in various barrels of crude oil</td>
<td>Numerical</td>
</tr>
</tbody>
</table>

*Table 4: Combined Project and Economic Variables to be Used in the Research*
4.2 Exploratory Data Analysis

This section details the data preparation process that the research team carried out through Exploratory Data Analysis (EDA). EDA is a framework used to explore and understand data to identify missing data, outliers, and highly correlated features that would impact the bias in the model. Figure 14 details the components that make up the data preparation stage using EDA.

Figure 14: EDA Process

The detailed EDA process is outlined in Table 5 below.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Reflection</th>
<th>Corrective Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Features</td>
<td>There was a severe shortage in the Contract Type feature. A total of 8,635 data points were missing due since the STAs do not track this metric</td>
<td>Contract Type feature was removed from the analysis</td>
</tr>
<tr>
<td>Missing Features</td>
<td>Weather Days was another feature with multiple missing features. A total of 7,734 data points of the 8,635 data points were missing the Weather Days since it's also not tracked</td>
<td>Weather Days feature was removed from the analysis</td>
</tr>
</tbody>
</table>
Missing Features

The Change Orders feature was also severely lacking. 7,063 data points of the 8,635 data points did not contain data related to the Change Orders feature.

Change Orders feature was removed from the analysis

Missing Features

The Bituminous Adjustment and Fuel Adjustment features were also lacking with only 658 and 1,282 data points respectively, of the 8,635 data points, containing data for the features.

Bituminous Adjustments feature and Fuel Adjustment feature were removed from the analysis

Missing Features

Only 865 data points of the 8,635 data points contained data for the CPPR feature.

The CPPR feature was removed from the analysis

Missing Observations

The Location Type feature contained 696 missing data points. This feature is instrumental in the research.

The observations were dropped from the analysis

Missing Observations

The Type of Work feature contained 1,489 missing data points. This feature is instrumental for the success of the machine learning models.

The observations were dropped from the analysis

Missing Observations

866 more data points were missing from 5 features including the Project Size and Number of Bidders features.

The observations were dropped from the analysis

Erroneous Values

Some observations were entered as "0" rather than left empty. A total of 212 data points were erroneous and unrepresentative of actual values.

The observations were dropped from the analysis

Erroneous Values

Some observations in the Location Type feature were entered as "X" rather than "R" for rural or "U" for urban. A total of 302 data points were erroneous and unrepresentative of actual values.

The observations were dropped from the analysis

Erroneous Values

Advertising Time showed unrealistic data with multiple negative values as shown in Figure 15.

The survey responses were revisited and the data related to the award date from NDDOT was unavailable. The Advertising Time feature is calculated based on the award date. NDDOT was unable to provide the research team with the missing data, so the
Irrelevant Data: The data relevant for this research contains a maximum of 10% difference between the engineer’s estimate and low bid. All data with a higher variation was dropped. The final data contains 1,761 data points with the percentage variation less than or equal to 10%. The observations were dropped from the analysis.

Table 5: Exploratory Data Analysis and Corrective Action

**Feature Correlation:**

Highly correlated features can undermine the statistical significance of an independent variable. This stage identifies the multicollinearity between the features such that feature correlation is no more than 0.8, and no less than -0.8. The research team studied the feature correlation in two stages to avoid missed multicollinearity.
Stage 1:
The first pair-wise correlation between features is outlined in Figure 16 below.

The Construction Spending ($) and Producer Price Index features were highly correlated, with a correlation coefficient of 0.88139. The Employment Level in Construction and Prime Loan Rate features also exhibited a high correlation coefficient at 0.873759. Consequently, The Producer Price Index and Employment Level in Construction features were dropped.
• Stage 2:

The second pair-wise correlation between features to check for any missed correlation is outlined in Figure 17 below.

The maximum correlation coefficient is between the Consumer Price Index and Construction Spending features. While the correlation coefficient was less than 0.8, the research team opted to drop the Consumer Price Index feature since the Construction Spending feature already accounted for the monthly expenditures towards new construction in the transportation field, which would take into consideration the change in price paid for the goods.

At the end of the data preparation stage, the research team had identified a total of 10 mutually exclusive features with a total of 1,761 observations. For the success of a prediction model, the number of observations in classification problem should fulfill the minimum given by the Equation 2 below, where C is the number of classes and F is the number of features (Boutilier 2020, Yu 2010.)
The value obtained by Equation 2 is 40 observations. The number of observations available highly surpass the minimum required. Table 6 summarizes the finalized list of features used by the research team.

<table>
<thead>
<tr>
<th>Variable Category</th>
<th>Variable</th>
<th>Variable Type</th>
<th>Used?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Related Variables</td>
<td>Contract Type</td>
<td>Nominal Categorical</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Type of Work</td>
<td>Nominal Categorical</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Location Type</td>
<td>Nominal Categorical</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Forecasted Duration Range</td>
<td>Ordinal Categorical</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Advertising Time</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Project Size</td>
<td>Ordinal Categorical</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>% of Cost Based Estimating Used</td>
<td>Ordinal Categorical</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>% Difference between the Engineer's Estimate and Low Bid</td>
<td>Numerical</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Change Orders</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Weather days</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Bituminous Adjustments</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Fuel adjustments</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Number of lanes</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Number of bidders</td>
<td>Numerical</td>
<td>Yes</td>
</tr>
<tr>
<td>Economic Indicator Variables</td>
<td>Contractor Past Performance Indicator (CPPI)</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td>Economic Indicator Variables</td>
<td>Consumer Price Index (CPI)</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td>Economic Indicator Variables</td>
<td>Construction Spending (CS)</td>
<td>Numerical</td>
<td>Yes</td>
</tr>
<tr>
<td>Economic Indicator Variables</td>
<td>Producer Price Index (PPI)</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td>Economic Indicator Variables</td>
<td>Prime Loan Rate (PLR)</td>
<td>Numerical</td>
<td>Yes</td>
</tr>
<tr>
<td>Economic Indicator Variables</td>
<td>Employment Level in Construction</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td>Economic Indicator Variables</td>
<td>Average Hourly Earnings</td>
<td>Numerical</td>
<td>No</td>
</tr>
<tr>
<td>Economic Indicator Variables</td>
<td>Crude Oil Prices</td>
<td>Numerical</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 6: Features Selected for Prediction Model

4.3 Data Characteristics

This thesis is dedicated to the accuracy of the engineer’s estimate. The scope of the research is to allocate the method of estimating to a project such that the percentage deviation of the low bid from the engineer’s estimate is kept at a minimum. This section explores the characteristics of the dataset and their impact on the project performance in terms of the estimate accuracy.
Method of Estimation:

Figure 18 below provides the first insight on the distribution of the projects. Imbalanced data is usually avoided in machine learning problems due to their tendency to overfit. Imbalanced data can be identified if one class has more than 10 times the observations of another class. As seen in Figure 18, the data is not imbalanced (Boutilier 2020.)

![Figure 18: Project Distribution by Method of Estimating](image)

Based on Figure 18 above, we can conclude that a Combination estimating approach is becoming more integrated by STAs. Of the total 1,761 projects, 1,204 (68%) were obtained from states employing combination estimates, while 557 (32%) projects were estimated using the historical bid-based approach.
Figure 19: Distribution of Engineer’s Estimate Variation from Low Bid Based on Project Type

Figure 19 provides insight on the distribution of data related to the percentage variation of low bid from the engineer’s estimate. The average accuracy using both the historical bid-based estimating approach and the combination estimating approach was found to be similar. A slightly lower variation between the two cost estimates is seen using the combination estimating approach at a mean percentage of 4.62%.

Location Type:

Figure 20: Project Distribution by Location Type
Figure 20 relates to the project distribution by Location Type. In the dataset, the number of observations in rural locations greatly surpass the observations located in urban locations, with 1,477 observations (84%) in rural locations and 284 observations (16%) in urban locations.

The accuracy of engineer’s estimate is illustrated in Figure 21. The average accuracy for both rural and urban projects are 4.61% and 4.81%, respectively. Urban projects in the dataset depicted a higher variability between the engineer’s estimate and low bid than rural projects.
Cost Distribution:

The cost distribution is portrayed in Figure 22. Most projects fell under $50M in value, however, a few projects exceeded that threshold with some costing higher than $300M. The summary statistics on the dataset cost distribution is presented in Table 7 below.

![Cost Distribution in Dataset](image)

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$4,855,038.15</td>
</tr>
<tr>
<td>Median</td>
<td>$1,783,130.50</td>
</tr>
<tr>
<td>Max</td>
<td>$345,252,917.75</td>
</tr>
<tr>
<td>Min</td>
<td>$7,150.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>$656,746.60</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>$4,356,765.56</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$15,539,942.59</td>
</tr>
</tbody>
</table>

*Table 7: Summary Statistics for Project Distribution*

The mean and median are measures of location that identify where our data is centered. The mean is the average of all the datapoints in the dataset, while the median represents the cost of the project in the middle of the dataset. The min, max, first quartile, and third quartiles are all measures of spread. The dataset contained projects with a minimum cost of $7,150.00 and maximum cost of $345,252,917.75. The cost of the projects at the 25% percentile is $656,746.60 and the 75% percentile is $4,356,765.56, as
portrayed by the first quartile and third quartile respectively. The standard deviation measures the dispersion of data according to the mean and is defined by Equation 3 below.

\[
\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}} \tag{Equation 3}
\]

Where \(x_i\) is the value of the \(i\)th point in the data set, \(\bar{x}\) is the mean value of the data set and \(n\) is the number of data points in the dataset. Our dataset showed a standard deviation equal to $15,539,942.59, which indicated that there is a large variance in the observations obtained by the research team.

![Figure 23: Cost Distribution by Location Type](image)

The cost distribution of projects located in rural or urban locations exhibited similar results as shown in Figure 23 above and Table 8 below.

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Rural</th>
<th>Urban</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$ 4,483,055.80</td>
<td>$ 4,743,592.40</td>
</tr>
<tr>
<td>Median</td>
<td>$ 1,792,381.90</td>
<td>$ 1,550,464.00</td>
</tr>
<tr>
<td>Max</td>
<td>$ 209,905,586.51</td>
<td>$ 161,252,864.47</td>
</tr>
<tr>
<td>Min</td>
<td>$ 7,150.00</td>
<td>$ 23,920.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>$ 652,272.03</td>
<td>$ 618,509.43</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>$ 4,425,016.05</td>
<td>$ 4,157,168.70</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$ 12,527,720.26</td>
<td>$ 14,297,298.25</td>
</tr>
</tbody>
</table>
Table 8: Summary Statistics for Cost Distribution of Projects by Location Type

The summary statistics for both rural and urban location types exhibit similar trends. Urban locations have a higher variability than rural locations since the standard deviation was recorded as $14,297,298.25 for urban locations vs $12,527,720.26 for rural locations.

Project Duration:

![Pie chart showing data distribution by forecasted duration range]

**Figure 24: Data Distribution by Forecasted Duration Range**

The data distribution by forecasted duration range is illustrated in Figure 24. Projects with durations between 30 to 99 days (level 2) were the leading projects at 961 observations (55%). There was almost half the number of projects within the duration range 100 to 299 days (level 3) as there was for projects within the duration range 30 to 99 days, and projects of more than 299 days (level 4) were found to be the least common among the observations in the dataset at 151 observations (9%).
Figure 25 provides insight on the cost distribution by duration. The trend among all levels for the forecasted duration range showed that projects with higher cost take a longer time to complete.

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>1: &lt;30 days</th>
<th>2: 30 ~ 99 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$537,396.29</td>
<td>$2,098,595.46</td>
</tr>
<tr>
<td>Median</td>
<td>$388,994.09</td>
<td>$1,309,810.00</td>
</tr>
<tr>
<td>Max</td>
<td>$5,616,410.71</td>
<td>$15,348,561.42</td>
</tr>
<tr>
<td>Min</td>
<td>$7,150.00</td>
<td>$81,638.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>$183,064.05</td>
<td>$594,179.97</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>$684,740.41</td>
<td>$2,925,909.50</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$574,977.97</td>
<td>$2,122,923.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>3: 100 ~ 299 days</th>
<th>4: &gt; 299 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$5,394,336.22</td>
<td>$26,448,489.00</td>
</tr>
<tr>
<td>Median</td>
<td>$4,057,902.28</td>
<td>$10,762,846.41</td>
</tr>
<tr>
<td>Max</td>
<td>$35,836,904.84</td>
<td>$345,252,917.75</td>
</tr>
<tr>
<td>Min</td>
<td>$194,428.73</td>
<td>$332,291.32</td>
</tr>
<tr>
<td>First Quartile</td>
<td>$2,163,784.93</td>
<td>$4,865,740.81</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>$7,557,470.80</td>
<td>$27,509,250.37</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$4,640,921.05</td>
<td>$46,861,957.38</td>
</tr>
</tbody>
</table>

Table 9: Summary Statistics for Cost Distribution by Forecasted Duration Range
Table 9 contains the summary statistics for the cost distribution by duration range. The statistics all proved that the project cost grows proportional to the project duration.

**Number of Bidders:**

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.37</td>
</tr>
<tr>
<td>Median</td>
<td>3.00</td>
</tr>
<tr>
<td>Max</td>
<td>13.00</td>
</tr>
<tr>
<td>Min</td>
<td>1.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>2.00</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>4.00</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.74</td>
</tr>
</tbody>
</table>

*Table 10: Summary Statistics of Distribution of Number of Bidders*
Type of Work:

Figure 27 represents the data distribution by project type. The most common project type among the state DOTs is roadway redesign at 23%. With the increase in traffic, highway expansion is a priority. Resurfacing projects (20%), maintenance or minor upgrades (19%), and safety and traffic control (18%) closely follow. Earthwork (7%), bridge replacement (4%), new bridge construction (2%), utilities (2%), environmental mitigation (2%), bicycle/pedestrian (2%), and road or culvert replacement (2%) are less common.
Figure 28 shows the distribution of cost estimates per project type. The summary statistics for each project type are shown in Table 11 below.

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Resurfacing</th>
<th>Maintenance or minor upgrades</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$4,111,897.30</td>
<td>$4,185,644.62</td>
</tr>
<tr>
<td>Median</td>
<td>$2,976,758.39</td>
<td>$1,078,949.00</td>
</tr>
<tr>
<td>Max</td>
<td>$45,834,511.20</td>
<td>$161,252,864.47</td>
</tr>
<tr>
<td>Min</td>
<td>$139,452.54</td>
<td>$23,920.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>$1,323,571.69</td>
<td>$424,905.88</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>$5,938,981.46</td>
<td>$3,051,041.52</td>
</tr>
<tr>
<td></td>
<td>Earthwork</td>
<td>Bicycle/pedestrian</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td><strong>Standard Deviation</strong></td>
<td><strong>4,344,979.82</strong></td>
<td><strong>12,177,972.17</strong></td>
</tr>
<tr>
<td>Mean</td>
<td>$5,083,795.08</td>
<td>$4,424,172.56</td>
</tr>
<tr>
<td>Median</td>
<td>$3,286,386.69</td>
<td>$412,132.50</td>
</tr>
<tr>
<td>Max</td>
<td>$35,836,904.84</td>
<td>$68,584,834.08</td>
</tr>
<tr>
<td>Min</td>
<td>$75,195.00</td>
<td>$39,397.20</td>
</tr>
<tr>
<td>First Quartile</td>
<td>$1,002,407.58</td>
<td>$203,375.75</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>$8,526,717.83</td>
<td>$546,796.50</td>
</tr>
<tr>
<td>Deviation</td>
<td>$5,380,071.61</td>
<td>Deviation</td>
</tr>
<tr>
<td><strong>Bridge replacement</strong></td>
<td><strong>Environmental mitigation</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>$15,900,191.87</td>
<td>$3,477,180.07</td>
</tr>
<tr>
<td>Median</td>
<td>$2,206,357.34</td>
<td>$1,605,137.35</td>
</tr>
<tr>
<td>Max</td>
<td>$209,905,586.51</td>
<td>$13,061,000.00</td>
</tr>
<tr>
<td>Min</td>
<td>$194,428.73</td>
<td>$103,000.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>$1,002,048.40</td>
<td>$1,202,677.67</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>$6,344,725.53</td>
<td>$5,100,245.00</td>
</tr>
<tr>
<td>Deviation</td>
<td>$44,354,209.85</td>
<td>Deviation</td>
</tr>
<tr>
<td><strong>Utilities</strong></td>
<td><strong>Safety and traffic control</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>$1,263,311.74</td>
<td>$3,575,428.31</td>
</tr>
<tr>
<td>Median</td>
<td>$638,113.00</td>
<td>$631,431.43</td>
</tr>
<tr>
<td>Max</td>
<td>$9,088,769.20</td>
<td>$345,252,917.75</td>
</tr>
<tr>
<td>Min</td>
<td>$75,000.00</td>
<td>$44,080.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>$357,793.77</td>
<td>$380,188.04</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>$1,096,222.50</td>
<td>$1,574,540.72</td>
</tr>
<tr>
<td>Deviation</td>
<td>$1,761,596.68</td>
<td>Deviation</td>
</tr>
<tr>
<td><strong>Roadway redesign</strong></td>
<td><strong>New bridge construction</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>$5,450,672.71</td>
<td>$4,861,048.51</td>
</tr>
<tr>
<td>Median</td>
<td>$3,212,304.60</td>
<td>$1,514,207.28</td>
</tr>
<tr>
<td>Max</td>
<td>$155,410,996.00</td>
<td>$32,035,819.90</td>
</tr>
<tr>
<td>Min</td>
<td>$7,150.00</td>
<td>$101,365.75</td>
</tr>
<tr>
<td>First Quartile</td>
<td>$1,554,064.28</td>
<td>$867,763.96</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>$5,013,212.85</td>
<td>$4,202,881.29</td>
</tr>
<tr>
<td>Deviation</td>
<td>$11,671,838.93</td>
<td>Deviation</td>
</tr>
<tr>
<td><strong>Road or culvert replacement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>$9,413,403.35</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>$3,968,694.42</td>
<td></td>
</tr>
</tbody>
</table>
Max $ 44,694,935.90  
Min $ 205,596.50  
First Quartile $ 975,088.19  
Third Quartile $ 10,182,525.01  
Standard Deviation $ 12,820,760.08  

Table 11: Summary Statistics for Project Distribution by Project Type

Bridge replacement projects in our dataset were found to have the highest mean cost estimate at $15,900,191.87, while utility projects had the lowest mean cost at $1,263,311.74. The highest median cost was found in projects related to road or culvert replacement at $3,968,694.42, and the lowest was found in bicycle/pedestrian projects at $412,132.50. Bridge replacement projects were also found to be the most variable in cost at a standard deviation of $44,354,209.85, and utility project were the least variable at a standard deviation of $1,761,596.68.

Figure 29: Distribution of Number of Bidders by project Type
The distribution of number of bidders by project type demonstrates a variation for each project type as shown in Figure 29. The summary of statistics related to the distribution of number of bidders by project type is shown in Table 12.

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Resurfacing</th>
<th>Maintenance or minor upgrades</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.21</td>
<td>3.43</td>
</tr>
<tr>
<td>Median</td>
<td>3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Max</td>
<td>9.00</td>
<td>10.00</td>
</tr>
<tr>
<td>Min</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>2.00</td>
<td>First Quartile</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>4.00</td>
<td>Third Quartile</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.27</td>
<td>Standard Deviation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Earthwork</th>
<th>Bicycle/pedestrian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4.12</td>
<td>3.43</td>
</tr>
<tr>
<td>Median</td>
<td>4.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Max</td>
<td>11.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Min</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>3.00</td>
<td>First Quartile</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>5.00</td>
<td>Third Quartile</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.80</td>
<td>Standard Deviation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Bridge replacement</th>
<th>Environmental mitigation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.76</td>
<td>5.64</td>
</tr>
<tr>
<td>Median</td>
<td>3.00</td>
<td>5.00</td>
</tr>
<tr>
<td>Max</td>
<td>13.00</td>
<td>11.00</td>
</tr>
<tr>
<td>Min</td>
<td>1.00</td>
<td>2.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>2.50</td>
<td>First Quartile</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>5.00</td>
<td>Third Quartile</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.80</td>
<td>Standard Deviation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Utilities</th>
<th>Safety and traffic control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.71</td>
<td>3.87</td>
</tr>
<tr>
<td>Median</td>
<td>3.00</td>
<td>4.00</td>
</tr>
<tr>
<td>Max</td>
<td>7.00</td>
<td>12.00</td>
</tr>
<tr>
<td>Min</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>First Quartile</td>
<td>1.50</td>
<td>First Quartile</td>
</tr>
<tr>
<td>Third Quartile</td>
<td>3.50</td>
<td>Third Quartile</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.54</td>
<td>Standard Deviation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Roadway redesign</th>
<th>New bridge construction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.53</td>
<td>3.73</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>Max</td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>70</td>
<td>2.00</td>
<td>9.00</td>
</tr>
<tr>
<td>100</td>
<td>4.00</td>
<td>7.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Max</th>
<th>Min</th>
<th>First Quartile</th>
<th>Third Quartile</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Road or culvert replacement</td>
<td>4.31</td>
<td>4.00</td>
<td>9.00</td>
<td>1.00</td>
<td>3.00</td>
<td>6.00</td>
<td>2.28</td>
</tr>
</tbody>
</table>

Table 12: Summary Statistics for Project Distribution by Project Type

Environmental mitigation projects were found to have the highest mean number of bidders at 5.64, and a median of 5 bidders, but it had the highest variability at a standard deviation of 2.28. On the other hand, roadway redesign projects had the lowest number of bidders and variability, with a mean of 2.53, a median of 2 and a standard deviation of 1.27.
4.4 Feature Engineering

Feature engineering is a machine learning technique that leverages data to simplify and speed up data transformations such that a high model accuracy can be obtained. The research team employed feature engineering to transform, modify, and/or create features that are usable in the prediction model (Boutilier 2020). This section is dedicated to the steps taken to ensure the dataset was in the appropriate usable form for the research. Figure 30 details the components of feature engineering.

![Figure 30: Feature Engineering Components](image)

The detailed feature engineering process is outlined in Table 13 below.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Reflection</th>
<th>Corrective Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal Categorical</td>
<td>The Location Type feature has only 2 categorical classes. It is considered</td>
<td>Designated 0 for rural and 1 for urban</td>
</tr>
<tr>
<td></td>
<td>a binary categorical feature</td>
<td></td>
</tr>
<tr>
<td>Nominal Categorical</td>
<td>The Type of Work feature has 11 categorical classes of no natural order</td>
<td>Performed one-hot or ordinal encoding of all 11 classes</td>
</tr>
<tr>
<td>Ordinal Categorical</td>
<td>The Forecasted Duration Range feature has 4 categorical classes with a</td>
<td>Replaced Level 1: &lt; 30 days with “1”</td>
</tr>
<tr>
<td></td>
<td>logical order</td>
<td>Replaced Level 2: 30 ~ 99 days with “2”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replaced Level 3: 100 ~ 299 days with “3”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replaced Level 4: &gt; 299 days with “4”</td>
</tr>
</tbody>
</table>
## Ordinal Categorical

The Project Size feature has 8 categorical classes of no natural order

<table>
<thead>
<tr>
<th>Level</th>
<th>Replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1: &lt; $10,000</td>
<td>1</td>
</tr>
<tr>
<td>Level 2: $10,000 ~ $49,999</td>
<td>2</td>
</tr>
<tr>
<td>Level 3: $50,000 ~ $99,999</td>
<td>3</td>
</tr>
<tr>
<td>Level 4: $100,000 ~ $499,999</td>
<td>4</td>
</tr>
<tr>
<td>Level 5: $500,000 ~ $999,999</td>
<td>5</td>
</tr>
<tr>
<td>Level 6: $100,000,000 ~ $999,999,999</td>
<td>6</td>
</tr>
<tr>
<td>Level 7: $100,000,000 ~ $249,999,999</td>
<td>7</td>
</tr>
<tr>
<td>Level 8: &gt; $249,999,999</td>
<td>8</td>
</tr>
</tbody>
</table>

## Ordinal Categorical

The % of Cost Based Estimating Used is the target of this research, and it is binary ordinal

Designated 0 for historical bid-based estimating and 1 for 5% and 10% combination estimating

## Feature Normalization

Our feature set includes features with very different scales, so feature normalization is required

Used the rescaling method of feature normalization

---

*Table 13: Feature Engineering Process*
CHAPTER 5: DEVELOPMENT OF COST ESTIMATING PREDICTION MODELS

5.1 Introduction

Machine Learning (ML) is a subdivision under Artificial Intelligence (AI) that has taken over in the field of data analysis due to the ease of programming and automated enhancement, and it’s regarded as the most popular technology in the fourth industrial revolution (Kotsiantis 2006, Sarker 2021). ML algorithms can be classified into 5 types (Boutilier 2020, Sarker 2021):

1. Supervised learning: Requires the user to input observations that include both the features and targets
2. Unsupervised learning: Requires the user to input observations that include features alone, and the target is interpreted by the AI model
3. Semi-supervised learning: Requires the user to input mixed observations, with some including both the features and target and others only including the features
4. Reinforced learning: No user input of pre-existing data. Instead, the model learns through trial and error
5. Deep learning: The user is not required to assign a target variable but allows the algorithm to follow the human process of learning

ML problems are also categorized according to the type of target feature being addressed in the problem. Regression targets are of a continuous nature, while classification targets are discreet and depend on the number of classes (Boutilier 2020). In this research, the target variable has 2 classes, i.e., it is binary in nature. The target is set for the % of the cost-based estimating feature, such that a “0” represents 0% and a “1” represents 5-10%.

The goal of the model also highly affects the assumptions and their validation. ML models can either address:
(1) Prediction problems: Where the goal is to predict the target value as accurately as possible for future observations, or

(2) Explanation problems: Where the goal is to explain the relationship between the target and the features. In this case, we need to interpret the regression coefficients and care is needed with respect to model assumptions, otherwise the information we extract from the regression coefficients may be misleading.

The ML models explored in this research depend on the supervised learning approach. There is an abundance of studies on the impact of the features selected on the engineer’s estimate accuracy in the past body of literature. To better address the engineer’s estimate accuracy problem, the research team decided to use a prediction model and focus on the accuracy of the model to predict the most suitable method of estimation with some regard to the impact of the different features on the target.

Since the target feature in ML problems is a binary feature, the models were trained to choose the class “0” or “1” depending on a probability threshold. This threshold is sometimes called a rounding threshold such that projects with probabilities above the threshold are classified as “1”, in this case meaning that 5-10% of cost-based estimating is more likely to yield an engineer’s estimate that would vary from low bid by 10% or less. Projects under the threshold are classified as “0” i.e., only using historical bid-based estimating is sufficient. The threshold is typically identified by experts in the field of study depending on the anticipated or accepted level of error. The threshold rule is denoted by $T(\hat{y}_i)$ and given by:

$$T(\hat{y}_i) = \begin{cases} 1, & \text{if } \hat{y}_i \geq \text{threshold coeff} \\ 0, & \text{if } \hat{y}_i < \text{threshold coeff} \end{cases}$$

The following measures are instrumental in ML problems to cross-check a model (Boutilier 2020):

(1) True positive rate (TPR): This is a measure of the sensitivity of the model. It identifies the proportion of positives that are correctly identified. Equation 4 represents its mathematical formulation.

$$TPR = \frac{TP}{TP + FN}$$  \hfill (Equation 4)
(1) True negative rate (TNR): This is a measure of the specificity of the model. It identifies the proportion of negatives that are correctly identified. Equation 5 represents its mathematical formulation.

\[
TNR = \frac{TN}{TN + FP}
\]

(1) False positive rate (FPR): This is a measure of a Type I error, which is based on the proportion of negatives that are incorrectly identified. Equation 6 represents its mathematical formulation.

\[
FPR = \frac{FP}{TN + FP}
\]

(2) False negative rate (FNR): This is a measure of a Type II error, which is based on the proportion of positives that are incorrectly identified. Equation 7 represents its mathematical formulation.

\[
FNR = \frac{FN}{TP + FN}
\]

Where TP stands for the number of true positives, TN stands for the number of true negatives, FP stands for the number of false positives, and FN stands for the number of false negatives.

Classification error preference is highly dependent on the type of machine learning problem. For the purpose of this thesis study, it is preferred that the FNR is higher than the FPR since the use of cost-based estimating exhausts the resources of STAs. This means that the preference lies with the prediction of the use of historical bid-based estimating, when the optimum approach would have been combination estimating with 5-10%. The research team decided on a goal TPR of 0.85. A confusion matrix was used to visualize the true positives, true negatives, false positives, and false negatives.

The performance measure used to compare the performance of the models was the area under the receiver operating curve (AUC). The receiver operating curve (ROC) visualizes the trade-off between the TPR and FPR. In particular, the ROC illustrates the TPR and FPR as the discriminant threshold is varied. Figure 31 from Boutilier (2020) illustrates the components of an ROC. The performance of a random classifier (i.e., guessing) is given by the red dotted line. Note the direction of classifier improvement.
The area under the ROC curve (AUC) is the most commonly used performance metric for classification problems. AUC measures the area under the ROC curve. An AUC of 1 indicates a perfect classifier, while an AUC of 0.5 indicates random guessing. AUCs below 0.5 imply that the model is worse than guessing. AUC is commonly used because it is a single metric that accounts for the trade-off between true positives and false positives (Boutilier 2020).

All models were constructed using 4-fold randomized cross validation to reduce their variance and bias.

5.2 Multiple Linear Regression Model

5.2.1 Background

Regression analysis is a statistical analysis that is used as a means to study the relationship between two or more quantitative variables in both prediction and explanation problems. MLR in particular refers to regression models with one dependent variable, called the target variable, and more than one independent variable, called the feature variables (Boutilier 2020). Binary MLR analysis is used when the target variable is represented by either a “0” or “1”.
For a given data point, let $y$ represent the target variable and $x_1, x_2, \ldots, x_F$ represent $F$ feature variables. Our aim was to determine the best values for the regression coefficients, $\beta_0, \beta_1, \ldots, \beta_F$ such that our predicted target values for the MLR problem, outlined in Equation 8 (Boutilier 2020, Jobson 1991)

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_F x_F$$

Equation 8

were as close as possible to the true target values (denoted $y$). To do this, we defined the error (or residual) for a given observation to be $(\hat{y} - y)$. Then, we chose our regression coefficients to minimize the mean squared error (MSE) between the true target values and the predicted target values across all data points. For example, suppose we observe $n$ data points of the form $(y_i, x_{i1}, x_{i2}, \ldots, x_{iF})$, $i = 1, \ldots, n$. We could minimize the mean squared error (MSE) as follows in Equation 9:

$$\min_{\beta_0, \beta_1, \ldots, \beta_F} \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_F x_{iF} - y_i)^2$$

Equation 9

This type of MLR is called the least squares problem. The regression coefficients that minimize are denoted $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_F$ (Boutilier 2020, Jobson 1991).

5.2.2 Verification of Assumptions

One downfall of MLR models is their reliance on strong assumptions for their success. These assumptions include:

1. Weak exogeneity: The feature values are known exactly and not subject to error. This must be assumed by the modeler.
2. Linearity: The target can be written as a linear combination of features.
3. Constant variance or homoscedasticity: The variance of the error terms is constant across all samples.
4. No autocorrelation: The error terms are not correlated with each other. Note that this assumption is satisfied if the targets are assumed to be independent from each other.
(5) Normality (or other distribution information): The error terms can, but are not required, to be normally distributed. If they are, the regression coefficients are also normally distributed. This assumption impacts the statistical tests used for the regression coefficients and other distributional assumptions are possible.

Observations in the dataset were directly obtained from the STAs, so assumption (1) related to weak exogeneity is fulfilled. Assumption (4) was also checked using the Pearson correlation, so there is an unlikely chance that the features are highly correlated. The research team was unable to verify assumptions (2) and (3) because they require the use of residual analysis, which is visualized through a scatter plot. The data is discrete and so are the residuals. As a result, plots of raw residuals were inconclusive on the assumptions. Assumption (5) was studied using QQ Plots as shown in Figure 32.

![QQ Plot of Linear Regression Model](image)

*Figure 32: QQ Plot of Linear Regression Model*

The normality assumption was not verified. The dataset is closer to a uniform distribution with a prolonged small right and left tail where the majority of values in the plot fall closer to the center. In a normal distribution, the theoretical values will fall 2 and -2 standard deviations from the center. This was also verified using a Shapiro Wilk test, which resulted in a p-value of 0.0, and a statistic of 0.2341828346 as shown in Figure 33.
5.2.3 Model Training and Results

The total 1,761 observations were split between training and testing datasets for the MLR regression model using a random ratio of 80:20 respectively (Zhang 2017). The model was trained using the observations in the training set with the variables outlined in Table 6 serving as the model features, and the MLR regression model was successfully executed with an R-squared of 0.327. The predicted coefficients for the model are shown in the mathematical representation in Equation 10.

\[
\hat{y} = 0.2606 - 0.1182 \times \text{Location Type} + 0.0004 \times \text{Forecasted Duration Range} + 0.0913 \\
- 0.0016 \times \% \text{Diff between engineer's estimate and} - 0.0101 \\
- 1.027e^{-10} \times \text{Construction Spending} - 11.8611 \\
+ 0.0078 \times \text{Crude Oil Prices} - 0.0836 \\
+ 0.0864 \times \text{Earthwork} + 0.5388 \\
+ 0.4574 \times \text{Environmental Mitigation} \\
+ 0.5190 \times \text{New Bridge Construction} + 0.1787 \times \text{Resurfacing} + 0.1635 \\
+ 0.3046 \times \text{Road or Culvert Replacement} + 0.4671 \times \text{Roadway Redesign} + 0.6571 \\
+ 0.4669 \times \text{Utilities}
\]

\[\text{Equation 10}\]
The model summary is outlined in Figure 34.

The “coef” column displays the regression coefficients \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_F \). The “\( P > |t| \)” column displays the p-value corresponding to each regression coefficient and the following two columns display the 95% confidence interval. The impact of these features is determined by the regression coefficients. The larger the absolute value of the coefficient, the more impactful it is on the target variable. A regression coefficient is said to be statistically significant from zero when the value displayed in the “\( P > |t| \)” column is less than 0.05. The features are displayed in Table 14 in order of most impactful and most statistically significant to least impactful and least statistically significant.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>Statistically significant from zero?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prime Loan Rate (PLR)</td>
<td>-11.8611</td>
<td>Yes</td>
</tr>
<tr>
<td>Safety and Traffic Control</td>
<td>0.6571</td>
<td>Yes</td>
</tr>
<tr>
<td>Environmental Mitigation</td>
<td>0.5388</td>
<td>Yes</td>
</tr>
<tr>
<td>Feature</td>
<td>Coefficient</td>
<td>Statistical Significance</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>--------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Utilities</td>
<td>0.4669</td>
<td>Yes</td>
</tr>
<tr>
<td>Maintenance or Minor Upgrades</td>
<td>0.4574</td>
<td>Yes</td>
</tr>
<tr>
<td>Roadway Redesign</td>
<td>0.3046</td>
<td>Yes</td>
</tr>
<tr>
<td>Constant</td>
<td>0.2606</td>
<td>Yes</td>
</tr>
<tr>
<td>New Bridge Construction</td>
<td>-0.1906</td>
<td>Yes</td>
</tr>
<tr>
<td>Resurfacing</td>
<td>0.1787</td>
<td>Yes</td>
</tr>
<tr>
<td>Location Type</td>
<td>-0.1182</td>
<td>Yes</td>
</tr>
<tr>
<td>Project Size</td>
<td>0.0913</td>
<td>Yes</td>
</tr>
<tr>
<td>Crude Oil Prices</td>
<td>0.0078</td>
<td>Yes</td>
</tr>
<tr>
<td>Road or Culvert Replacement</td>
<td>0.1635</td>
<td>No</td>
</tr>
<tr>
<td>Earthwork</td>
<td>0.0864</td>
<td>No</td>
</tr>
<tr>
<td>Bridge Replacement</td>
<td>-0.0836</td>
<td>No</td>
</tr>
<tr>
<td>Number of bidders</td>
<td>-0.0101</td>
<td>No</td>
</tr>
<tr>
<td>% Difference between the Engineer's Estimate and Low Bid</td>
<td>-0.0016</td>
<td>No</td>
</tr>
<tr>
<td>Forecasted Duration Range</td>
<td>0.0004</td>
<td>No</td>
</tr>
<tr>
<td>Construction Spending (CS)</td>
<td>1.03E-10</td>
<td>No</td>
</tr>
</tbody>
</table>

*Table 14: Feature Coefficients and Statistical Significance from 0*

From Figure 34 and Table 14, the research team was able to establish that the features with the highest impact on the chosen method of estimation are the prime loan rate, project category, location type, and project size. The lowest impact was related to the construction spending during the year of construction, forecasted duration range, and to the research team’s surprise, the number of bidders.

The performance of a multiple linear regression problem is usually measured by the coefficient of determination $R^2$. The goodness of fit is proportional to the value of the $R^2$ coefficient, and this MLR model yielded a very low $R^2$ coefficient, which is an indicator of poor performance. One possible reason for poor performance of a MLR problem is that the assumptions of linearity was not met and that it is not well suited for categorical problems since it primarily deals with continuous target values, and it is sensitive to imbalanced data (Boutilier 2020, Jobson 1991).

### 5.3 Logistic Regression Model

#### 5.3.1 Background

Logistic regression (LOGIT) is a classification algorithm used to assign observations to a discrete set of classes. In linear regression, the outcome is usually
continuous and can be any possible value. However, in the case of logistic regression, the predicted outcome is discrete and restricted to the class values. It works by predicting the probability of occurrence of a target variable. Logistic regression was suitable for this MLR problem since it deals with a target variable with a set of two classes as previously explained (Boutilier 2020, Jobson 1991).

LOGIT uses a more complex prediction function, which is defined as the ‘sigmoid function’ or ‘logistic function’. The sigmoid function maps the predictions to probabilities between “0” and “1” for binary problems (i.e., \( \hat{y}_i \in \{0, 1\} \)). It is given by Equation 11.

\[
\hat{y}_i = \frac{1}{1 + e^{-z}}
\]

Equation 11

For each data point i, let \( y_i \in \{0, 1\} \) represent the target variable and \( x_{i1}, x_{i2}, \ldots, x_{iF} \) represent F feature variables. Multiple features in a LOGIT model operate similar to MLR regression using a linear combination of the features with feature coefficients (i.e., \( \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_F x_{iF} \)). Logistic regression combines the sigmoid function and the linear combination of features to create Equation 12, which represents the logistic function (Boutilier 2020, Jobson 1991).

\[
\hat{y}_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_F x_{iF})}}
\]

Equation 12

When Equation 12 is converted to the logistic scale, it represents linear regression on the logarithm of the odds (logit), as portrayed by Equation 13.

\[
\text{logit}(\hat{y}) = \frac{\hat{y}_i}{1 - \hat{y}_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_F x_{iF}
\]

Equation 13

The coefficients \( \beta_0, \beta_1, \ldots, \beta_F \) are interpreted as the change in the log odds, or the percentage change in the odds.

Linear regression uses the least squares equation to create a closed form solution for the regression coefficients. This is not applicable to LOGIT, which relies on the use of the maximum likelihood estimation (MLE). The MLE problem requires the use of numerical methods such as the stochastic gradient descent or Newton’s method (Boutilier 2020, Jobson 1991).
A threshold rule is required for all classification problems that rely on the probability, so it also applies to logistic regression models. The selected threshold corresponds to an FPR of 0.85 as selected by the research team for all ML models. A visual representation of the logistic function is shown in Figure 35 below, as obtained from Alam’s “Implementation of Logistic Regression Using Python”.

![Figure 35: Sigmoid Function with Threshold](image)

### 5.3.2 Verification of Assumptions

The success of a binary logistic regression function is dependent on the fulfillment of its assumptions for goodness of fit (Jobson 1991). These assumptions can be summarized into:

1. **Dichotomous target**: The predicted outcome is strictly binary of values “0” and “1”
2. **No autocorrelation**: The error terms are not correlated with each other. Note that this assumption is satisfied if the targets are assumed to be independent from each other.
3. **Linearity**: The target can be written as a linear log odds combination of the continuous features
4. **Sample size**: Large sample sizes

Through feature engineering, the research team designated “0” for historical bid-based estimating and “1” for 5% and 10% combination estimating. All observations in the dataset were doublechecked for erroneous values outside of “0” and “1” and assumption
(1) was verified as such. Using Pearson’s correlation, assumption (2) was also verified by the research team. Assumption (4) was verified using Equation 2 in the EDA section. The research team collected 1,761 observations and the minimum was calculated as 40.

One method used to test that the continuous features are linearly related to the log odds of the target feature, assumption (3), is through plotting the linear log odds for the numerical features such that they represent either an S-shaped curve or U-shaped curve. Since the shape is usually not obvious, the focus is shifted to the following characteristics: flat top and bottom and an increasing or decreasing middle. The log odds for the continuous numerical features are shown in Figures 36 to 40.

![Figure 36: % Diff between the Engineer’s Estimate and Low Bids Log Odds Linear Plot](image)

Figure 36 provides insight on the % Difference between the Engineer’s Estimate and Low Bids feature. The distribution of data resembles an S-curve distribution. The tails are more distributed around the fitted line with flat top and bottom and a decreasing slope towards the midpoint of the plot.
The log odds linear plot for the number of bidders is outlined in Figure 37. The distribution of data resembles an S-curve distribution. The tails are more distributed around the fitted line with flat top and bottom and an increasing slope towards the midpoint of the plot.

Construction spending also fulfilled the log odds linearity assumption. Figure 38 presents a very clear S-curve data distribution. The tails are more distributed around the
fitted line with flat top and bottom and an increasing slope towards the midpoint of the plot.

Figure 39: Prime Loan Rate Log Odds Linear Plot

Figure 39 portrays the Prime Loan Rate feature. The distribution of data resembles an S-curve distribution. The tails are more distributed around the fitted line with flat top and bottom and a decreasing slope towards the midpoint of the plot.

Figure 40: Crude Oil Prices Log Odds Linear Plot
The Crude Oil Prices feature also fulfilled the log odds linearity assumption. Figure 40 presents an S-curve data distribution. The tails are more distributed around the fitted line with flat top and bottom and an increasing slope towards the midpoint of the plot.

The use of a LOGIT model was deemed appropriate after the assumptions were verified.

5.3.3 Model Training and Results

Logistic regression for prediction problems relies on the use of regularization. Regularization is a method of reducing overfitting by penalizing the regression coefficients. In regularized regression, we add an additional term, $\lambda \sum_{k=1}^{K} |\beta_k|^p$, to the objective of the logit function, where $p$ is a number that denotes the type of regularization and $\lambda$ is a hyperparameter. The $\lambda$ parameter controls the degree of regularization and can be chosen during a 4-fold randomized cross-validation procedure. Regularization is useful because it limits the size of the coefficients, preventing over-fitting to the training data.

There are three common types of regularization:

1. L1-regularization: Known as lasso regression, it helps select the most important features by penalizing irrelevant features such that they become equivalent to “0”.
2. L2-regularization: Known as ridge regression, it helps reduce the correlation between the features in a dataset.
3. Elastic Net Regularization: Combined approach of L1 and L2 regularization, it helps reduce the unimportant features but doesn’t equate them to “0”.

The total 1,761 observations were split between training and testing dataset for the logistic regression model using a random ratio of 80:20 respectively (Zhang 2017). The model was trained using the observations in the training set with the variables outlined in Table 6 serving as the model features. Three models were trained using no regularization, L1-regularization, and L2-regularization, respectively. The performance of the models was evaluated using the AUC score to select the model with highest performance.
5.3.3.1 Base Model Results

The logistic regression model with no regularization was successfully executed with an R-squared of 0.3954. The Prime Loan Rate feature had the highest predicted coefficient at -226.2853. The coefficient is more than forty times the coefficient for the next ranking feature, Safety and Traffic Control. The mathematical formula outlining all coefficients in the logarithmic scale is shown in Equation 14, as shown below.

\[
Z = -3.0477 - 0.9347 \times \text{Location Type} - 0.0345 \times \text{Forecasted Duration Range} + 0.5488 \\
- 0.0253 \times \text{Project Size} - 0.9347 \times \% \text{Diff between engineer’s estimate and} - 0.1083 \\
- 2.003e^{-9} \times \text{Construction Spending} - 226.2853 \\
- 0.1906 \times \text{Crude Oil Prices} - 0.5422 \\
- 0.0345 \times \text{Prime Loan Rate} + 0.1906 \times \text{Crude Oil Prices} - 0.5422 \\
- 0.0345 \times \text{Prime Loan Rate} + 0.1906 \times \text{Crude Oil Prices} - 0.5422 \\
- 0.0345 \times \text{Bridge Replacement} + 0.5693 \times \text{Earthwork} + 3.9026 \\
- 2.1770 \times \text{New Bridge Construction} + 0.7937 \times \text{Resurfacing} + 0.7946 \\
- 1.6675 \times \text{Roadway Redesign} + 5.6219 \\
+ 3.1473 \times \text{Utilities}
\]

Equation 14

To convert the score from the logarithmic scale to the probability of the binary value “0” or “1”, the research team used Equation 15, as shown below.

\[
f(z) = \frac{1}{1 + e^{-z}}
\]

Equation 15
The model summary is outlined in Figure 41, as shown above. An LL-Null model refers to an objective function of maximum likelihood with no features. The obtained log-likelihood is higher than that of the LL-Null model, which indicates that the model is successful.

The features are displayed in Table 15 below in order of most impactful and most statistically significant to least impactful and least statistically significant.

![Logit Regression Results](image)

<table>
<thead>
<tr>
<th>Dep. Variable</th>
<th>No. Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of cost based estimating used</td>
<td>1408</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model:</th>
<th>Logit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method:</td>
<td>MLE</td>
</tr>
<tr>
<td>DF Residuals:</td>
<td>1389</td>
</tr>
<tr>
<td>Date:</td>
<td>Tue, 28 Jun 2022</td>
</tr>
<tr>
<td>Time:</td>
<td>22:59:47</td>
</tr>
<tr>
<td>converged:</td>
<td>True</td>
</tr>
<tr>
<td>Pseudo R-squ.:</td>
<td>0.3954</td>
</tr>
<tr>
<td>Log-Likelihood:</td>
<td>-531.11</td>
</tr>
<tr>
<td>LL-Null:</td>
<td>-870.39</td>
</tr>
<tr>
<td>Covariance Type:</td>
<td>nonrobust</td>
</tr>
<tr>
<td>LLR p-value:</td>
<td>8.071e-136</td>
</tr>
</tbody>
</table>

| Feature                          | coef  | std err | z     | P>|z|  | [0.025| 0.975|
|----------------------------------|-------|---------|-------|-----|------|------|
| const                            | -3.0477 | 1.065  | -2.862 | 0.004  | -5.135 | -0.961 |
| Location type                    | -0.9547 | 0.206  | -4.688 | 0.000  | -1.339 | -0.531 |
| Forecasted duration range        | -0.0045 | 0.111  | -0.312 | 0.755  | -0.251 | 0.162 |
| Project size                     | 0.5466 | 0.283  | 1.942  | 0.052  | -0.005 | 1.103 |
| % difference between the engineer estimate low bid | 0.0253 | 0.037  | 0.847  | 0.343  | -0.027 | 0.078 |
| Number of bidders                | -0.1063 | 0.049  | -2.203 | 0.028  | -0.205 | -0.012 |
| Construction Spending ($)        | 2.003e-09 | 7.3e-10 | 2.671 | 0.008  | 5.34e-10 | 3.47e-09 |
| Prime Loan Rate                  | -226.2853 | 20.105 | -11.255 | 0.000  | -265.689 | -186.681 |
| Crude oil prices ($ per barrel)  | 0.1906 | 0.018  | 10.633 | 0.000  | 0.156  | 0.225 |
| Bridge replacement               | -0.5422 | 0.740  | -0.733 | 0.464  | -1.983 | 0.096 |
| Earthwork                        | 0.5693 | 0.645  | 0.883  | 0.377  | -0.695 | 1.834 |
| Environmental mitigation         | 3.9026 | 1.031  | 3.787  | 0.000  | 1.883  | 5.523 |
| Maintenance or minor upgrades    | 2.6639 | 0.620  | 4.327  | 0.000  | 1.468  | 3.900 |
| New bridge construction          | -2.1770 | 1.208  | -1.805 | 0.071  | -4.544 | 0.190 |
| Resurfacing                      | 0.7937 | 0.615  | 1.290  | 0.197  | -0.413 | 2.000 |
| Road or culvert replacement      | 0.7946 | 0.845  | 0.941  | 0.347  | -0.661 | 2.450 |
| Roadway redesign                 | 1.6675 | 0.611  | 2.728  | 0.006  | 0.470  | 2.685 |
| Safety and traffic control       | 5.6219 | 0.786  | 7.133  | 0.000  | 4.077  | 7.167 |
| Utilities                        | 3.1473 | 0.863  | 3.645  | 0.000  | 1.495  | 4.840 |

*Figure 41: Logistic Regression Results and Weights with no Regularization*
<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>Statistically significant from zero?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prime Loan Rate (PLR)</td>
<td>-226.2853</td>
<td>Yes</td>
</tr>
<tr>
<td>Safety and Traffic Control</td>
<td>5.6219</td>
<td>Yes</td>
</tr>
<tr>
<td>Environmental Mitigation</td>
<td>3.9026</td>
<td>Yes</td>
</tr>
<tr>
<td>Utilities</td>
<td>3.1473</td>
<td>Yes</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.0477</td>
<td>Yes</td>
</tr>
<tr>
<td>Maintenance or Minor Upgrades</td>
<td>2.6839</td>
<td>Yes</td>
</tr>
<tr>
<td>Roadway Redesign</td>
<td>1.6675</td>
<td>Yes</td>
</tr>
<tr>
<td>Location Type</td>
<td>-0.9347</td>
<td>Yes</td>
</tr>
<tr>
<td>Crude Oil Prices</td>
<td>0.1906</td>
<td>Yes</td>
</tr>
<tr>
<td>Number of bidders</td>
<td>-0.1083</td>
<td>Yes</td>
</tr>
<tr>
<td>Construction Spending (CS)</td>
<td>2.00E-09</td>
<td>Yes</td>
</tr>
<tr>
<td>New Bridge Construction</td>
<td>-2.177</td>
<td>No</td>
</tr>
<tr>
<td>Road or Culvert Replacement</td>
<td>0.7946</td>
<td>No</td>
</tr>
<tr>
<td>Resurfacing</td>
<td>0.7937</td>
<td>No</td>
</tr>
<tr>
<td>Project Size</td>
<td>0.5488</td>
<td>No</td>
</tr>
<tr>
<td>Earthwork</td>
<td>0.5693</td>
<td>No</td>
</tr>
<tr>
<td>Bridge Replacement</td>
<td>-0.5422</td>
<td>No</td>
</tr>
<tr>
<td>% Difference between the Engineer's Estimate and Low Bid</td>
<td>0.0253</td>
<td>No</td>
</tr>
<tr>
<td>Forecasted Duration Range</td>
<td>-0.0345</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 15: Feature Coefficients and Statistical Significance from 0

The model performance accuracy was evaluated using the AUC. The score obtained for both the training and the testing sets was 0.89, which indicates a strong performance of the model in correctly predicting the preferred estimation method. The ROC is shown in Figure 42.

![Figure 42: AUC Score for Testing and Training Sets for Base Logistic Regression Model](image-url)
The false positives outlined in the confusion matrix for the testing set in Figure 43 reported that only 5 of 353 projects in the testing set falsely identified cost-based estimating as the optimal estimation method. The confusion matrix shows that the model is predicting historical bid-based estimating rather than combination estimating, which indicates bias.

![Figure 43: Confusion Matrix of Testing Set in Base Logistic Regression Model](image)

### 5.3.3.2 L1-regularization Model Results

The logistic regression model with Lasso regularization was not successfully executed with no model convergence. An R-squared of 0.3947 was calculated from the model. The Prime Loan Rate feature had the highest predicted coefficient at -224.698. The coefficient is more than forty times the coefficient for the next ranking feature, Safety and
Traffic Control. The mathematical formula outlining all coefficients in the logarithmic scale is shown in Equation 16.

\[ Z = -2.9109 - 0.9329 \times \text{Location Type} - 0.0346 \times \text{Forecasted Duration Range} + 0.5498 \]

* Project Size - 0.0253 * % Diff between engineer’s estimate and \(- 0.1083\)
* number of bidders + 1.872 \(e^{-9}\) * \(\text{Construction Spending} - 224.698\)
* \(\text{Prime Loan Rate} + 0.1894 \times \text{Crude Oil Prices} - 0.5459\)
* \(\text{Bridge Replacement} + 0.5664 \times \text{Earthwork} + 3.8525\)
* \(\text{Environmental Mitigation} + 2.6805 \times \text{Maintenance or Minor Upgrades} - 2.1782 \times \text{New Bridge Construction} + 0.7904 \times \text{Resurfacing} + 0.7910\)
* \(\text{Road or Culvert Replacement} + 1.6645 \times \text{Roadway Redesign} + 5.6146\)
* \(\text{Safety and Traffic Control} + 3.138 \times \text{Utilities}\)

To convert the score from the logarithmic scale to the probability of the binary value “0” or “1”, the research team used Equation 17, as shown below.

\[ f(z) = \frac{1}{1 + e^{-z}} \]

The features are displayed in Table 16 below in order of most impactful and most statistically significant to least impactful and least statistically significant.
The model performance accuracy was evaluated using the AUC. The score obtained for both the training and the testing sets was 0.89, which indicates a strong performance of the model in correctly predicting the preferred estimation method. The ROC is shown in Figure 44.

![Figure 44: AUC Score for Testing and Training Sets for L1-Regularization Logistic Regression Model](image)

The false positives outlined in the confusion matrix for the testing set in Figure 45 reported that 24 of 353 projects in the testing set falsely identified cost-based estimating as the optimal estimation method.
5.3.3.3  **L2-regularization Model Results**

The logistic regression model with ridge regularization was successfully executed with no model convergence. An R-squared of 0.373 was calculated from the model. The Prime Loan Rate feature had the highest predicted coefficient at -84.9338. The coefficient is more than seventeen times the coefficient for the next ranking feature, Safety and Traffic Control. The mathematical formula outlining all coefficients in the logarithmic scale is shown in Equation 18.
To convert the score from the logarithmic scale to the probability of the binary value “0” or “1”, the research team used Equation 19, as shown below.

\[ f(z) = \frac{1}{1 + e^{-z}} \]  

The features are displayed in Table 17 below in order of most impactful and most statistically significant to least impactful and least statistically significant.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>Statistically significant from zero?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prime Loan Rate (PLR)</td>
<td>-84.9338</td>
<td>Yes</td>
</tr>
<tr>
<td>Safety and Traffic Control</td>
<td>4.907</td>
<td>Yes</td>
</tr>
<tr>
<td>Environmental Mitigation</td>
<td>2.845</td>
<td>Yes</td>
</tr>
<tr>
<td>Utilities</td>
<td>2.2874</td>
<td>Yes</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.3529</td>
<td>Yes</td>
</tr>
<tr>
<td>Maintenance or Minor Upgrades</td>
<td>2.0997</td>
<td>Yes</td>
</tr>
<tr>
<td>Roadway Redesign</td>
<td>1.102</td>
<td>Yes</td>
</tr>
<tr>
<td>Location Type</td>
<td>-0.8627</td>
<td>Yes</td>
</tr>
<tr>
<td>Crude Oil Prices</td>
<td>0.06966</td>
<td>Yes</td>
</tr>
<tr>
<td>Number of bidders</td>
<td>-0.0646</td>
<td>Yes</td>
</tr>
<tr>
<td>Construction Spending (CS)</td>
<td>3.38E-10</td>
<td>Yes</td>
</tr>
<tr>
<td>New Bridge Construction</td>
<td>-2.3386</td>
<td>No</td>
</tr>
<tr>
<td>Road or Culvert Replacement</td>
<td>0.3848</td>
<td>No</td>
</tr>
<tr>
<td>Resurfacing</td>
<td>0.45403</td>
<td>No</td>
</tr>
<tr>
<td>Project Size</td>
<td>0.4995</td>
<td>No</td>
</tr>
</tbody>
</table>
The model performance accuracy was evaluated using the AUC. The score obtained for both the training and the testing sets was 0.89 for the testing set and 0.88 for the training set, which indicates a strong performance of the model in correctly predicting the preferred estimation method. The ROC is shown in Figure 46.

The false positives outlined in the confusion matrix for the testing set in Figure 47 reported that 24 of 353 projects in the testing set falsely identified cost-based estimating as the optimal estimation method.
5.3.3.4 Model Selection

Logistic regression is one of the most popular models for classification problems due to its ability to handle discreet data. However, the model is susceptible to overfitting such that the prediction may exhibit high bias.

The base model resulted in the same model accuracy at an AUC score of 0.89, however, the use of regularization is essential to limit the probability of overfitting and bias. Model 2 using L1-regularization did not affect the accuracy or impact the model enough to justify use. The model coefficients and feature importance followed the same pattern. On the other hand, L2-regularization impacts the bias of a model by limiting feature correlation. The Prime Loan Rate feature coefficient went down to -84.8338 from -226.2853 in the base model. While the coefficient is more than ten times more than the coefficients for the other features, its reduction using L2-regularization provides for a more

Figure 47: Confusion Matrix of Testing Set in Ridge Logistic Regression Model
uniform model that still considers the prime loan rate to be the leading feature in determining the appropriate method of estimation.

## 5.4 Classification and Regression Trees Model

### 5.4.1 Background

Classification and regression trees (CART) are a type of decision tree algorithm that can be used with versatility for classification or regression problems and continuous or categorical targets (Boutilier 2020, Timofeev 2004).

CART models outperform logistic regression models and MLR models when proper hyperparameter tuning is done. The algorithm is non-parametric, which gives it the ability to capture non-linear relationships and high-order interactions between features and the target (De’Ath and Fabricius 2002, Speybroeck 2012). Including all interaction terms may lead to dimensionality issues and over-fitting, and selecting the best interaction terms is a challenging feature selection problem for CART models.

A decision tree algorithm relies on dividing the model’s target prediction across multiple distinct predictions related to each of the feature spaces. A visualization of the decision tree and the splitting nodes can be seen in Figure 48 below.

![Decision Tree Visualization](image)

**Figure 48: Decision Tree Visualization**

Figure 48 pertains to a decision tree for a binary classification problem where the prediction was partitioned across three split levels. The nodes and splits of a decision tree give indication on the feature importance. There are a total of 4 different nodes that can be identified in a decision tree (Timofeev 2004):
(1) Root node: The first node in the tree
(2) Leaf node: A terminal node in the tree
(3) Parent node: The node directly above a reference node
(4) Child node: The node directly below a reference node

One of a CART models main characteristics is that it provides the user with an easily interpretable interface to understand the outcome of a prediction model. Users follow the splits in the tree until they reach a leaf node, which corresponds to a computed probability based on the proportion of each class in the leaf. The probability can be converted to the binary classification ("0" or "1") using the threshold identified by the research team (Boutilier 2020, De’Ath and Fabricius 2002).

The target of ML problems using a CART model is to determine the splitting sequence that maximizes the prediction accuracy. CART models are susceptible to overfitting when no limitations are applied on the decision tree. In turn, they have multiple hyperparameters that can be tuned to obtain the optimum performance of a model without overfitting. The hyperparameters include:

(1) Splitting criteria:

Finding the best splitting sequence can be obtained by either considering all the possible features in the model or a subset of said features (Speybroeck 2012).

(2) Stopping criteria:

This is the most important stage in training a decision tree because no criteria on when to stop equates to an overfit decision tree. Stopping can be determined using:

- maximum depth
- maximum number of observations per leaf
- minimum number of observations per split

(3) Number of features to consider for each split (Pruning):
The idea behind pruning is to remove portions of the tree that do not contribute to out-of-sample prediction accuracy. This is done by replacing the split with a leaf node (that corresponds to a prediction) and observing the out-of-sample accuracy. Pruning is usually done in a bottom-up fashion until no further improvements are observed in the model.

The research team used two metrics for fitting the CART model. The first metric we used for evaluating the different feature–split combinations was the Gini impurity metric, which measures the likelihood that a randomly selected observation will be incorrectly classified by calculating the weighted average of the child nodes in a split as a single metric to represent the value of that split. This metric follows the same logic of the FPR and the FNR (Timofeev 2004). The other metric - information gain - is attributed to the measurement of the purity of the node split (Boutilier 2020).

5.4.2 Verification of Assumptions

CART models have no formal distributional assumptions. They are non-parametric and can thus handle skewed and multi-modal data, as well as categorical data that is ordinal or non-ordinal.

5.4.3 Model Training and Results

Unlike ML and logistic regression, CART models are not sensitive to data assumptions. This makes CART an ideal algorithm to identify the optimum estimation method for STAs since the QQ Plots identify a non-linear relationship.

The total 1,761 observations were split between training and testing datasets for the logistic regression model using a random ratio of 80:20 respectively (Zhang 2017). A model was trained using an optimized hyperparameter tuning using a 4-fold randomized cross-validation. The hyperparameters were separately investigated to identify trends for each hyperparameter.
- Maximum depth:

![Figure 49: Maximum Depth Hyperparameter Effect on the CART AUC Score](image)

Figure 49 represents the trends in the AUC score of the CART model by tuning the maximum depth hyperparameter. The AUC score for testing and training sets follows the same trend. It increases at a constant rate up to a maximum depth of 9. Beyond that, the score slightly fluctuates in both the training and testing sets. To reduce the probability of overfitting, the research team decided to stop at a maximum depth of 9.

![Figure 50: Minimum Number of Samples per Leaf Hyperparameter Effect on the CART AUC Score](image)
The minimum samples per leaf hyperparameter follows an opposing trend to the maximum depth, as shown in Figure 50. The AUC score of the testing and training sets starts between 0.9 and 1.0, and then decreases gradually to 0.5 at around 700 samples per leaf. The chosen minimum number of samples by the cross validation was 4 samples per leaf.

![Figure 51: Minimum Number of Samples per Split Hyperparameter Effect on the CART AUC Score](image)

Figure 51 represents the trends in the AUC score of the CART model by tuning the minimum number of samples per split. The AUC score for testing and training sets follows the same trend. It decreases at an almost constant rate until around 700 samples and then the score remains constant at 0.5. The cross-validation identified the optimum minimum number of samples per split as 8.

The features are displayed in Table 18 below in order of importance, as indicated by the feature coefficients.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintenance or Minor Upgrades</td>
<td>2.09970</td>
</tr>
<tr>
<td>Crude Oil Prices</td>
<td>0.75859</td>
</tr>
<tr>
<td>Road or Culvert Replacement</td>
<td>0.38480</td>
</tr>
<tr>
<td>Safety and Traffic Control</td>
<td>0.13180</td>
</tr>
<tr>
<td>Prime Loan Rate (PLR)</td>
<td>0.09190</td>
</tr>
<tr>
<td>% Difference between the Engineer’s Estimate and Low Bid</td>
<td>0.01005</td>
</tr>
<tr>
<td>Number of bidders</td>
<td>0.00765</td>
</tr>
<tr>
<td>Environmental Mitigation</td>
<td>0.00000</td>
</tr>
</tbody>
</table>
A visualization of the features splitting in the CART model is another method used to identify the features with the highest importance. Figure 52 represents the first 3 splits in the CART model. The Crude Oil Prices feature, Safety and Traffic Control feature, and Number of Bidders feature were identified as the most important features using the CART algorithm.

Table 18: Feature Importance Ranking

<table>
<thead>
<tr>
<th>Feature</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilities</td>
<td>0.00000</td>
</tr>
<tr>
<td>Roadway Redesign</td>
<td>0.00000</td>
</tr>
<tr>
<td>Location Type</td>
<td>0.00000</td>
</tr>
<tr>
<td>Construction Spending (CS)</td>
<td>0.00000</td>
</tr>
<tr>
<td>New Bridge Construction</td>
<td>0.00000</td>
</tr>
<tr>
<td>Resurfacing</td>
<td>0.00000</td>
</tr>
<tr>
<td>Project Size</td>
<td>0.00000</td>
</tr>
<tr>
<td>Earthwork</td>
<td>0.00000</td>
</tr>
<tr>
<td>Bridge Replacement</td>
<td>0.00000</td>
</tr>
<tr>
<td>Forecasted Duration Range</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

Figure 52: Split Level 3 Tree in CART Model
The model performance accuracy was evaluated using the AUC. The score obtained for both the training and the testing sets was 0.99, which indicates that even after the hyperparameter tuning, the CART model overfitted the data. The ROC is shown below in Figure 53.

The false positives outlined in the confusion matrix for the testing set also reiterated that the model overfit. There were no reported false positives or false negatives. All 353 observations in the test set were correctly identified.
Depending on the data characteristics, CART tended to continue splitting, such that every split corresponds to one observation. If that happens, the accuracy is 100%, but it is highly overfitted. In this case, the cross validation identified a highly overfit model as the optimum model. Since the relationship between the hyperparameters and the AUC score is reflective of a linear relationship, the research team was unable to tune the hyperparameters with enough confidence that the model would not be overfit. Thus, they concluded that a CART model is not optimal for the prediction of the method of estimation.

### 5.5 Random Forests Model

#### 5.5.1 Background

The base Random Forest (RF) algorithm was first invented in 1995 by Dr. Tin Kam Ho – a researcher from Bell Labs. However, in 2001, L. Breiman proposed the RF algorithm version we use commonly today, as a solution to the CART model’s tendency to overfit. It is characterized by a combination of tree predictors such that each tree depends
on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The mutually exclusive predictor results are then averaged to obtain the final model performance (Biau and Scornet 2016, Breiman 2001).

Similar to CART, random forests can be used for both classification and regression problems. They are optimum for this machine learning problem because they were invented with prediction problems in mind, but they can also be used for explanation problems.

Random forests are one of the most popular ensemble learners that make use of bootstrap aggregation to reduce model variance and sensitivity to a particular choice of dataset. Bootstrap aggregation is commonly known as the bagging algorithm, which relies on sampling with replacement. Given a training set of \( n \) observations denoted \( T \), \( M \) new training sets are created \( (T_1, \ldots, T_M) \) by randomly sampling \( n \) observations from \( T \) with replacement. Practically, this means that \( T_1, \ldots, T_M \) may have repeated observations. Note that each \( T_1, \ldots, T_M \) is the same size as the original training set \( T \). Breiman (2001) mathematically proved that for a large \( n \), the proportion of unique observations in each \( T_1, \ldots, T_M \) is \( 1 - \frac{1}{e} \approx 63\% \). That means 37\% of the observations are duplicates. \( M \) CART models are then trained, one for each \( T_1, \ldots, T_M \). Each model provides a predicted target value that is then used to compute the proportion of 1s / majority vote for classification problems (Boutilier 2020, Qi 2012).

Random forests introduce another risk of errors to machine learning models. If the individual CART models are trained using the same training data \( T \), the resulting trees will be highly correlated (and many trees may be identical). Correlated trees are known to increase the bias of the predictions. Bagging attempts to avoid this problem by training each tree on a different training set (Boutilier 2020, Breiman 2001).

Extended bagging applies the same theory of bagging for random forests by only considering a random subset of features at each split. The default value of features at each split is calculated using Equation 20.

\[
\text{# of features per split} = \sqrt{F}
\]

Equation 20
Like CART, RF uses the same hyperparameters for the tuning of each tree (splitting criteria, stopping criteria, and number of features to consider at each split). Another addition to the list of hyperparameters is the number of trees in the model.

The bagging procedure incorporates the out-of-bag error to measure model performance. Each sampled training set $T_i, \ldots, T_M$ is called a bag. To obtain an out-of-bag error estimate, each observation is run $x_i$ through the random forest using only the trees that were not trained with the observation. The prediction and corresponding error for each observation is calculated and averaged over all observations (Biau 2016, Boutilier 2020).

To make a prediction in random forest, the prediction for each bag / CART model is first calculated. For classification problems, the prediction is denoted as $\hat{y}_{im} \in \{0, 1\}$ where $i$ corresponds to the observation and $m$ corresponds to the bag. Note that $\hat{y}_{im}$ has already been converted from a probability to a binary value by the CART model. The final random forest classification prediction is calculated using Equation 21.

$$\hat{p}_i = \frac{\sum_{m=1}^{M} \hat{y}_{im}}{M}$$  

where $\hat{p}_i$ is the probability that $\hat{y}_i = 1$. The binary prediction is then determined using a rounding threshold. The research team chose a threshold corresponding to a goal TPR of 0.85.

For explanation problems, random forests provide some information about feature importance using the following approaches (Boutilier 2020):

1. Mean decrease in impurity / mean improvement in accuracy: Each split improves the accuracy of the prediction. Each leaf node has a corresponding criteria value and each splitting of the parent node into two new child nodes, each of those new child nodes should improve the criteria value. For each feature, the improvement in accuracy is measured every time the feature is split. A weighted average is computed based on assigning weights to the trees according to their depth. The feature importance values are then scaled between “0” and “1” to obtain the relative measure of feature importance.
Permutation importance: The performance metric value of the entire random forest model is calculated using the out-of-bag errors. To obtain the importance of a particular feature, the entries for that feature are randomly permuted. In other words, the feature value for each observation is replaced with a random number. The model accuracy is then recomputed, and the difference is used to represent feature importance. If there is no change in model accuracy, then the feature has little importance. Similar to above, we can scale all feature importance values to [0, 1].

5.5.2 Verification of Assumptions

Random forests have no formal distributional assumptions. They are non-parametric and can thus handle skewed and multi-modal data, as well as categorical data that are ordinal or non-ordinal.

5.5.3 Model Training and Results

The total 1,761 observations were split between training and testing datasets for the logistic regression model using a random ratio of 80:20 respectively (Zhang 2017). A model was trained using an optimized hyperparameter tuning using a 4-fold randomized cross-validation. The cross validation identified a maximum depth of 2, minimum samples per leaf of 8, minimum samples per split of 9, and number of trees of 219. Each hyperparameter was individually explored for trend identification as shown in Figures 55 to 58.
• Maximum depth:

![Figure 55: Maximum Depth Hyperparameter Effect on the RF AUC Score](image)

Figure 55 represents the trends in the AUC score of the RF model by tuning the maximum depth hyperparameter. The AUC score for testing and training sets follows the same trend. It increases at almost a constant rate up to a maximum depth of 2, beyond which the model overfits.

![Figure 56: Minimum Number of Samples per Leaf Hyperparameter Effect on the RF AUC](image)

The minimum samples per leaf hyperparameter follow an opposing trend to the maximum depth, as shown in Figure 56. The AUC score of the testing and training sets
starts between 0.9 and 1.0 and then decreases gradually to 0.5 at around 450 samples per leaf. The chosen minimum number of samples by the cross validation was 8 samples per leaf.

![Minimum Number of Samples per Split Hyperparameter Effect on the CART AUC Score](image)

**Figure 57: Minimum Number of Samples per Split Hyperparameter Effect on the CART AUC Score**

Figure 57 represents the trends in the AUC score of the RF model by tuning the minimum number of samples per split. The AUC score for testing and training sets follows the same trend. It decreases at an almost constant rate until around 450 samples and then the score remains constant at 0.5. The cross-validation identified the optimum minimum number of samples per split as 9.

![Impact of Number of Trees in Decision Tree Model on AUC Score](image)

**Figure 58: Impact of Number of Trees in Decision Tree Model on AUC Score**
The number of trees hyperparameter increases up to around 50 trees before remaining constant at an AUC score of around 1, which means that the model is susceptible to overfitting. The chosen number of trees hyperparameter by the cross validation was 219 trees.

The features are displayed in Table 19 below in order of most important as indicated by the feature coefficients.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude Oil Prices</td>
<td>0.30196</td>
</tr>
<tr>
<td>Safety and Traffic Control</td>
<td>0.21973</td>
</tr>
<tr>
<td>Prime Loan Rate (PLR)</td>
<td>0.16098</td>
</tr>
<tr>
<td>Construction Spending (CS)</td>
<td>0.12297</td>
</tr>
<tr>
<td>New Bridge Construction</td>
<td>0.05352</td>
</tr>
<tr>
<td>Bridge Replacement</td>
<td>0.03260</td>
</tr>
<tr>
<td>Earthwork</td>
<td>0.02791</td>
</tr>
<tr>
<td>Location Type</td>
<td>0.01887</td>
</tr>
<tr>
<td>Resurfacing</td>
<td>0.01493</td>
</tr>
<tr>
<td>Maintenance or Minor Upgrades</td>
<td>0.01314</td>
</tr>
<tr>
<td>Forecasted Duration Range</td>
<td>0.00804</td>
</tr>
<tr>
<td>Number of bidders</td>
<td>0.00785</td>
</tr>
<tr>
<td>% Difference between the Engineer’s Estimate and Low Bid</td>
<td>0.00742</td>
</tr>
<tr>
<td>Roadway Redesign</td>
<td>0.00584</td>
</tr>
<tr>
<td>Environmental Mitigation</td>
<td>0.00289</td>
</tr>
<tr>
<td>Utilities</td>
<td>0.00079</td>
</tr>
<tr>
<td>Project Size</td>
<td>0.00052</td>
</tr>
<tr>
<td>Road or Culvert Replacement</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

*Table 19: Feature Importance Ranking*

The features with the highest importance were the Crude Oil Prices, Safety and Traffic Control, Prime Loan Rate, and the Construction Spending features. All of these features correspond to the state of the economic market. They were followed by the features related to the project type; however, this indicates that the choice of method of estimation is highly reliant on the market conditions.
The model performance accuracy was evaluated using the AUC. The score obtained for both the training and the testing sets was 0.90, which is a strong performance of the model in correctly predicting the preferred estimation method. The ROC is shown in Figure 59.

![Figure 59: AUC Score for Testing and Training Sets for RF Model](image)

The false positives outlined in the confusion matrix for the testing set in Figure 60 reported that 20 of 353 projects in the testing set falsely identified cost-based estimating as the optimal estimation method.
The RF model exhibited the highest AUC score without risk of overfitting. This is because it combines the advantages of CART models related to the flexibility of the data assumptions and good generalization across project types, as well as the advantage of a logistic regression model related to a low probability of overfitting.

5.6 Model Comparison

5.6.1 Algorithm Performance Accuracy

The research team employed the use of multiple machine learning algorithms to obtain the model with the highest performance accuracy in relation to the prediction of the optimal method of estimation. A total of four algorithms were trained and tested to account for the tradeoff between the advantages and disadvantages of each algorithm.

Linear regression exhibited poor performance due to the non-linear relationship between the features and the target used in the model. The research team determined that it should be excluded from the analysis. Three logistic regression models were evaluated in terms of performance and errors due to variance. Ridge regularization resulted in an AUC score of 0.87, demonstrating a trade-off between high performance and no overfitting. A
CART model was built using optimized hyperparameters, which unfortunately proved to be incompatible with the machine learning problem being evaluated due to its tendency to overfit. Random forests were explored as a solution to the CART model’s overfitting, and the model performance was measured by an AUC score of 0.9, proving to be the most suitable prediction model of the aforementioned algorithms. Figure 61 presents the summary AUC scores in a combined ROC of the models.

![ROC Comparison of Different Models](image.png)

**Figure 61: Comparison of AUC Algorithm Scores for Estimate Predictions**

### 5.6.2 Feature Importance

The objective of the models is not extended to establishing an explanation between the features and the model. The body of literature and the estimating peer exchange provided extensive evidence on the impact of the features chosen in the model. However, this section is dedicated to comparing the feature importance in each of the models and drawing similarities between expert opinion in the estimating peer exchange, literature review, and the machine learning models.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Ridge Logistic Regression</th>
<th>Classification and Regression Trees</th>
<th>Random Forests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude Oil Prices</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Safety and Traffic Control</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Prime Loan Rate (PLR)</td>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 20: Feature Importance Ranking by Model

Table 20 outlines the feature ranking per machine learning algorithm. Some features exhibited similar levels of importance across all three algorithms. The Prime Loan Rate feature and Safety and Traffic Control feature both ranked within the top 5 features in all of the models. The prime loan rate is an economic indicator that pertains to the interest rate per year granted to the STAs, which is one indicator that relates to the funding available. Safety and traffic control projects consistently affected the prediction across the model, which means that they impacted the decision on the method of estimation. In logistic regression, they had a positive coefficient, which is indicative that combination estimating is preferred for safety and traffic control projects.

The number of bidders was one of the most commonly identified features by STAs, as well as the body of literature. However, the feature was ranked between the top 7 to 15 projects in the models. The features related to the economic state consistently outweighed the importance of the project characteristics, namely the project’s size, location type, and forecasted duration. The project types ranked differently in terms of importance, but generally ranked after the economic indicators.
CHAPTER 6: DISCUSSION AND CONCLUSIONS

6.1 Conclusions

The process of cost estimating follows that of a highway project’s lifecycle: (1) planning, (2) project development (preliminary design), (3) final design (4) right-of-way acquisition, (5) construction, and (6) operation and maintenance. This research explored the engineer’s estimate prepared during the final design phase, which is used in comparison to contractor bids. Inconsistent cost estimates of projects in the transportation industry have been the subject of study since the 1980s. This paper built from the work of previous studies on the accuracy of the engineer’s estimate with a primary goal of examining the accuracy of the engineer’s estimate in relation to the method of cost estimation used by the estimating team for different project types. The team’s research findings have been split between the qualitative and quantitative results obtained from the estimating peer exchange, and the analysis of four different data-driven machine learning algorithms (MLR, LOGIT, CART, RF) that were developed for the prediction of the most suitable method of cost estimation depending on project and economic factors.

The estimating peer exchange was primarily used to determine the state of practice of the STAs in preparing the engineer’s estimate and to derive patterns regarding the engineer’s estimate accuracy. Participants in the estimating peer exchange were the Departments of Transportation for the states of Illinois (IDOT), Iowa (IDOT), Kentucky (KTC), Michigan (MDOT), Montana (MDT), Minnesota (MnDOT), Nebraska (NDOT), North Dakota (NDDOT), Ohio (ODOT) and South Dakota (SDDOT). The peer exchange dataset was supplemented with data from four other Departments of Transportation for the states of Arkansas (ARDOT), California (Caltrans), Washington (WSDOT), and Tennessee (TDOT). The participating states uncovered the main governing factors behind the fluctuation of the price of projects in the highway and infrastructure sector. STAs concurrently agreed on the following list of most impactful factors; low bid competition, geographic location type, the limitations behind outdated data used in historical bid-based estimating, shortage of skilled workers and time, and market volatility. Most of the states (fourteen of the total sixteen) also reported tracking the engineer’s estimate accuracy for
the purpose of yearly evaluation, and constant improvement to their respective estimating practices.

The estimating peer exchange also served as an exchange of the state practices related to estimate preparation. The estimating protocol was found to alternate between a centralized and decentralized approach. Eight of the sixteen STAs use a decentralized by project staff approach for estimate preparation, while four STAs each reported using both a centralized and decentralized by designated staff approach. In addition, STAs identified two main goals behind the value of the engineer’s estimate. 70% of the states reported that they target an engineer’s estimate that is close to the lowest responsible bidder, while 30% aim for a “fair and reasonable estimate” that more closely resembles the second or third lowest bidder. The state participants also provided information regarding the primary methods of cost estimation. Similar to findings from the literature review, historical bid-based estimating was the most common method used by the STAs in the peer exchange, with 8 out of the 13 states who responded to this question choosing it over combination estimating and cost-based estimating. The percentage of estimate line items calculated using cost-based estimating ranged between 0% to 70% for the STAs. 43% of the 8 states also reported using an index to adjust the unit price of the historical data used in historical bid-based estimating, while 57% reported using the same values within a three-year timeframe. Since the majority of states prioritize historical bid-based estimating, 7 of the 14 states reported an average of 0% cost-based estimating used. The remaining 7 were separated into groups: three with an average of 5%, two with an average of 10%, and one each for 60% and 70% estimate line items calculated using cost-based estimating.

Moreover, low bid competition was reported to be the most prominent area of concern by STAs. The average number of bidders fluctuated around the three bidders mark for most states, with the highest average reported in 2018 at 3.49 and the lowest average reported in 2019 at 3.01. The number varied between 2.4 and 3.7 for each individual STA. Of those bidders, the STAs reported an average of 48%, 50%, and 48% of the bids falling within ±10% of the engineer’s estimate. The average percentage by project type was then evaluated, and there was a strong indication that the accuracy of the engineer’s estimate increases proportionally to the percentage of line items calculated using cost-based
estimating. Historical bid-based estimating was found to have the lowest accuracy at 47% of projects’ bids falling within ±10% of the engineer’s estimate items. Combination estimating closely followed at 48%, but cost-based estimating was deemed to have the highest average percentage with 53% of projects within ±10% of the engineer’s estimate items. Although this indicated that a shift in focus towards cost-based estimating might be warranted for a higher number of projects to fall within the ±10% mark, the dataset used did not consider the impact of the multiple project and economic factors identified by STAs and throughout the literature review.

Accordingly, this research study was extended to examine the methods of cost estimating used in the transportation industry using factors related to the project characteristics including: type of work, location type, forecasted duration range, project size, % difference between the engineer’s estimate and low bid, and number of bidders, as well as factors related to the economic conditions such as; construction spending (CS), prime loan rate (PLR), and crude oil prices. The research team used data-driven approaches to predict the optimum method of estimation. A second dataset was obtained from six STAs: Montana Department of Transportation (MDT), Nebraska Department of Transportation (NDOT), North Dakota Department of Transportation (NDDOT), Tennessee Department of Transportation (TDOT), Washington State Department of Transportation (WSDOT), and Wisconsin Department of Transportation (WisDOT). Unfortunately, the data obtained only included observations for projects estimated using historical bid-based estimating and combination estimating with a 5-10% average of line items estimated using cost-based estimating. States did not track the exact percentage of line items estimated using cost-based estimating, so the metric was adjusted to represent the average percentage used by each of the STAs. A total number of 8,635 datapoints were initially obtained, but only projects where the percentage difference between low bid and the engineer’s estimate was less than or equal to ±10% were considered. After EDA and feature engineering, the dataset was collectively reduced to 1,761 datapoints.

In this research, the target variable was set as the % of cost-based estimating. It was converted to a binary state since the data was split between two classes, such that a “0” represents 0% and a “1” represents 5-10% of cost-based estimating line items within the
estimate. Four data-driven techniques to assist STAs in selecting the method of estimation for transportation projects were used: multiple linear regression (MLR), logistic regression (LOGIT), classification and regression trees (CART), and random forests (RF). Model performance was evaluated using the area under the receiver operating curve (AUC) which indicates the diagnostic ability of a binary classifier system by exhibiting a value between “0” and “1”. The models developed to predict the optimum method of estimation in this study serve two functions: first, they can be used to identify projects that require the allocation of funds associated with performing cost-based estimating. The model predictions can be used to establish budgets for their long-term transportation plans and program management; second, the developed model can be used to evaluate which projects take precedent over others when time and cost are of the essence.

MLR was used as a benchmark statistic technique to evaluate the need for a more complex machine learning algorithm. The statistical assumptions concerning a constant variance and normal distribution of the residuals were violated, thus undermining the consistency and reliability of the MLR model predictions. As a result, the MLR model does not fit to serve long-term purposes of this research and was discounted from the list of data-driven approaches that were compared in terms of accuracy and generalization.

The LOGIT regression function followed suit with three different models. The statistical assumptions in logistic regression models are less strict with only a focus on a linear log odds relationship of the features and the target and a large sample size. A base model with no regularization was first trained, followed by a logistic regression model where 11 regularization was applied to reduce the dimensionality of the model and a 12 regularization logistic regression model was applied to reduce the feature correlation in the model. All three models obtained a classification accuracy of 89%. The regularized logistic regression model using 12 regularization surpassed the logistic regression models with the same AUC score and a lower risk of model bias. The base model exhibited skewed predictions and systematically predicted historical bid-based estimating, rather than combination estimating.

Furthermore, a CART model was trained using optimized hyperparameters to determine the best model using the CART algorithm. Using hyperparameter tuning, a
maximum model accuracy of 0.99 was obtained using a maximum depth of 9, minimum samples per leaf of 4, and minimum samples per split of 8. While the AUC score increased, the model exhibited high variance and it overfit, which deemed it an unfit algorithm for the prediction of the method of estimating. As a result, an RF model was developed due to its capability to remedy the susceptibility of the CART model to overfit. The model exhibited the highest performance among all the data-driven approaches at an AUC score of 0.9 without presenting any evidence of overfitting. This means that the RF model is the most reliable for use for the prediction of the method of estimation in transportation projects.

It is important to note that the scope of this research is not extended to establish an explanatory relationship between the features and the target variables since their importance was verified by the STAs as well as the literature review. Nevertheless, the degree of influence of the factors in the models was still explored to identify the most impactful sector that determines the method of estimation. It was ascertained that the economic indicators consistently surpassed the impact of the project indicators across all the models. The primary economic factors included the prime loan rate, and crude oil prices. The influence of the prime loan rate feature ranked first, fifth, and third in the 12-regularized LOGIT model, CART model, and RF model respectively. On the other hand, the crude oil price feature ranked eighth, second, and first respectively. Economic data is seldom considered when drafting estimates using historic bid-based estimating, however, developing an adjustment factor for the state of the economy is highly suggested.

The project type feature was the most significant among the project related variables. Project types classified under safety and traffic control, maintenance and minor upgrades, environmental mitigation, and roadway redesign are most influential on the method of cost estimation to be used. The aforementioned project types combined with road or culvert replacement, earthwork, and resurfacing project types perceived the use of combination estimating as an advantage, while new bridge construction and bridge replacement favored historical bid-based estimating.
6.2 Limitations & Recommendations

This thesis study was developed in order to understand the features with the highest impact on the cost of transportation projects, as well as provide STAs with a clear data-driven directive on the use of the different cost estimating methodologies. Although the research presents STAs with a data-driven medium to determine the optimum method of estimation, it encounters three major limitations:

(1) The models developed in this study are most reliable for the STAs that provided data because the data used for model development were retrieved from their database. While the research team reached out to all 50 state DOTs, only 6 responded with data. To increase the reliability and consistency of model performance, other DOTs can duplicate the modeling procedure employed in this study to build a new optimal model using their own data.

(2) This study does not consider the primary use of cost-based estimating, and only examines projects that use 5-10% cost-based estimating or projects that use historical bid-based estimating. The lack of involvement of states who primarily rely on cost-based estimating limited the dimensionality of the research and its ability to capture information on all three methods of cost estimating. This also limited the reliability of the extracted information on the feature importance in the model, thus, a prediction model was pursued rather than an explanation model.

(3) STAs do not track the percentage of cost-based estimating used per project, but instead only provided the research team with average percentages that were used as the baseline for the model.

STAs allocate a significant portion of their funds to data collection for the constant development of the database for historical bid-based estimating. The literature review outlines the plethora of data that is not synthesized due to the lack of manpower. It is recommended that STAs relocate funds spent on data collection for historical bid-based estimating to data analysis and synthesis and work towards training the cost estimation employees on cost-based estimating. With the unstable state of the economy, reliance on historical estimates is inaccurate, and the need for skills to accurately estimate cost should
be prioritized in collaboration with the models developed by the research team. STAs can make use of the trends of economic variation among the past years to further solidify the guidelines behind the use of historical bid-based estimating.

The question around the engineer’s estimate accuracy will remain constant if mechanisms are not established for data sharing to public researchers that can prioritize the exploration of the tradeoff between the economic state and the impact on the engineer’s estimate. While most STAs possess some form of data, the fear of confidentiality was a limiting factor that affected the full advancement of this research. Once datasets from more state DOTs are obtained, this research can be expanded to benefit a larger body of audience. State DOTs may choose to individually replicate the RF model created in this research in the event that data sharing cannot be achieved. The dataset used in this research includes sufficient observations for the success of the models. Thus, the most suitable models to be used if more data were to become available are the LOGIT and RF models.

The proposed data-driven approach should be formatted into a user-friendly layout, so that STAs can easily output results on the method of estimation related to their projects depending on the characteristics. The codes are programed in python, which might not be a common interface among the public. A recommendation for an interface includes the Visual Basic Application (VBA), which is compatible with macro enabled Microsoft Excel, a very common software program.
APPENDICES

A: References


http://digital.library.wisc.edu/1793/54136


https://www fhwa dot gov/programadmin/contracts/cacc pdf

https://www fhwa dot gov/environment/publications/flexibility/flexibility pdf

https://www fhwa dot gov/publications/research/infrastructure/pavements/CostEstimate DevelopmentAndOversight/CostEstimateDevelopmentAndOversight pdf


https://www fhwa dot gov/majorprojects/cost_estimating/guidance.cfm/


“Highway Infrastructure Programs, Federal Highway Administration, Transportation.” USAspending.gov, Department of the Treasury, 17 Apr. 2022, https://www.usaspending.gov/federal_account/069-0548


Engineers Specialty Conference on Leadership and Management in Construction, ASCE, Hilton Head SC, 120-128.


https://www.researchgate.net/publication/242419834_COST_ESTIMATION_OF_HIGHWAY_PROJECTS_IN_DEVELOPING_COUNTRIES_ARTIFICIAL_NEURAL_NETWORK_WORK_APPROACH


B: Estimating Peer Exchange Survey

Q-1. Does your state have estimate documentation?

☐ Yes
☐ No

Q-2. What is the primary goal of your preliminary engineer’s estimate?

☐ Close to low bid
☐ Fair and reasonable estimate

Q-3. What is the primary approach used by your DOT for developing the preliminary or initial engineer’s estimate?

☐ Cost Based (material costs, labor costs, overhead, etc. are estimated)
☐ Historical Bid Based Unit Cost
☐ Combination
☐ Other   If other, please briefly describe

Q-4. What is the primary approach used by your DOT for developing the final engineer’s estimate?

☐ Cost Based (material costs, labor costs, overhead, etc. are estimated)
☐ Historical Bid Based Unit Cost
☐ Combination
☐ Other   If other, please briefly describe

Q-5. Do you use an index to adjust historical bid prices to present day pricing?

☐ Yes
☐ No
Q-6. How is the preliminary or initial engineer’s estimate prepared?

☐ Centralized

☐ Decentralized (District/Region) by designated staff

☐ Decentralized by project staff

Q-7. What percentage of your estimate is based on cost-based estimating?

_______

Q-8. How is the final engineer’s estimate prepared?

☐ Centralized

☐ Decentralized (District/Region) by designated staff

☐ Decentralized by project staff

Q-9. Is the process different based on project size, contracting type, etc.?

☐ Yes    ☐ No   If yes, please briefly describe

Q-10. Do you have mobilization maximums in your estimate?

☐ Yes

☐ No

☐ Payment Cap

Q-11. Do you track or update the preliminary engineer’s estimate throughout the design process?

☐ Yes    ☐ No   If yes, please briefly describe

Q-12. Do you utilize an engineer’s estimate review process prior to soliciting bids?

☐ Yes    ☐ No   If yes, please briefly describe

Q-13. What percentage of the time has the low bid fallen within plus or minus 10% of your final engineer’s estimate for the last 3 years?
Q-14. Do you have other methods of measuring accuracy of the final engineer’s estimate other than determining the percentage difference from the low bid?

☐ Yes  ☐ No  If yes, please briefly describe

Q-15. What are the average number of bidders per project for the last 3 years?

2018  _____
2019  _____
2020  _____

Q-16. Do you have any initiatives to increase bidding competition on projects?

☐ Yes  ☐ No  If yes, please briefly describe

Q-17. Does your state have guidelines or manuals on how to prepare the cost estimate?

☐ Yes  ☐ No

Q-18. Does your DOT have criteria for rejecting bids that overrun the engineer’s estimate?

☐ Yes  ☐ No  If yes, please list them

Q-19. Does your DOT analyze bids that appear to be mathematically or materially unbalanced?

☐ Yes  ☐ No  If yes, please briefly describe how this analysis is performed.

Q-20. Does your DOT have a process for dealing with penny unit bids?

☐ Yes  ☐ No  If yes, please briefly describe

Q-21. What are your DOT’s top 3 issues regarding preparation of the engineer’s estimate?

Q-22. Would your DOT be willing to participate in a virtual Peer Exchange in late July or early August focusing on developing the engineer’s estimate?
Q-23. If you were to participate in a Peer Exchange, would you be willing to present details on your DOT’s processes and procedures?

☐ Yes  ☐ No

Q-24. If you were to participate in a virtual Peer Exchange in late July or early August, how long would you like it to be?

☐ 2 half-day sessions

☐ 3 half-day sessions

☐ 4 half-day sessions

Q-25. Please provide your contact information

State DOT: ________________________

Responder’s Name: __________________

Responder’s Position: ________________

Responder’s E-Mail: __________________

Responder’s Phone Number: ____________
C: Multiple Linear Regression Python Code

```python
#Partition the data into training and testing
X = df.drop(['% of cost based estimating used'], axis=1)
y = df['% of cost based estimating used']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=21)
regr = linear_model.LinearRegression()
regr.fit(X, y)
print('Intercept: 
', regr.intercept_)
print('Coefficients: 
', regr.coef_)
X = sm.add_constant(X) # adding a constant

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

print_model = model.summary()
print(print_model)
```

D: Logistic Regression Python Code

```python
#Partition the data into training and testing
x = df.drop(['% of cost based estimating used'], axis=1)
y = df['% of cost based estimating used']

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=21)

#Model Summary
Model = sm.Logit(y_train, sm.add_constant(x_train))
Result = Model.fit()
Result.summary()

#Compute ROC curves and AUC
yhat_Test = Model.predict(Result.params,sm.add_constant(x_test))
yhat_Train = Model.predict(Result.params,sm.add_constant(x_train))

#Get the FPR, TPR corresponding to all possible rounding thresholds
fprQ, tprQ, threshQ = roc_curve(y_test, yhat_Test)
fpr2, tpr2, thresh2 = roc_curve(y_train, yhat_Train)

#Determine the AUC
roc_auc_Test= roc_auc_score(y_test, yhat_Test)
roc_auc_Train= roc_auc_score(y_train, yhat_Train)
```
# Plot the ROC curve

```python
fig = figure(figsize=(10, 6)) # Create the figure space
plt.plot([0, 1], [0, 1], linestyle='--') # Add the dashed line for guessing
plt.plot(fpr0, tpr0, label='Base Testing Set (AUC = 0.25)', color='r') # Print the Test ROC curve and AUC
plt.plot(fpr2, tpr2, label='Base Training Set (AUC = 0.32)', color='g') # Print the Test ROC curve and AUC
plt.xlim([0.0, 1.0]) # Set x-axis
plt.ylim([0.0, 1.0]) # Set y-axis
plt.xlabel('False Positive Rate') # Set x-axis label
plt.ylabel('True Positive Rate') # Set y-axis label
plt.legend(loc='lower right') # Set legend location
plt.show() # Display plot
```

# Round the predictions according to the threshold with a TPR of 0.85

```python
# Base Model
pred_round = [] # List to store binary outcomes
thresh = thresholds[find([x[i] for x in enumerate(tpr2) if x[1] > 0.85])]
for i in y_test: # Loop over all predictions
    if i > thresh:
        pred_round.append(1)
    else:
        pred_round.append(0)
```

# Create a confusion matrix for Base model

```python
# Compute confusion matrix values
cmatrix = confusion_matrix(y_true = y_test, y_pred = pred_round)

# Plot
fig = figure(figsize=(10, 6)) # Create figure space
plt.imshow(cmatrix, cmap=plt.cm.Blues) # Plot CM
plt.text(0, 0, '0', )
plt.text(0, 1, '1')
plt.text(1, 0, '0')
plt.text(1, 1, '1')
plt.xlabel('True label')
plt.ylabel('Predicted label')
plt.title("Base Confusion Matrix")
plt.show()
```
# Lasso Regression

```python
LogReg = LogisticRegression(penalty='l1', solver='liblinear', max_iter=1200)
Logparams = {"C": [0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000, 70000, 1000000]}
clfLogL1 = GridSearchCV(LogReg, Logparams, cv=4, scoring='roc_auc')
clfLogL1.fit(X_train, y_train)

# Look at the parameters for the best model
clfLogL1.best_estimator_
```

```python
# Use the best model to make predictions
predL1 = clfLogL1.best_estimator_.predict_proba(X_test).T[1]

from sklearn.metrics import r2_score
r2_score(y_test, predL1)

print(clfLogL1.best_estimator_.coef_, clfLogL1.best_estimator_.intercept_)
```

# Plot the ROC curve for L1-Regularization

```python
# Training set
clfLogL1_train = clfLogL1.best_estimator_.predict_proba(X_train).T[1]
fpr1, tpr1, thresh1 = roc_curve(y_train, clfLogL1_train)
roc_auc_train = roc_auc_score(y_train, clfLogL1_train)

# Testing set
fpr2, tpr2, thresh2 = roc_curve(y_test, clfLogL1_test)
roc_auc_test = roc_auc_score(y_test, clfLogL1_test)

# Plot the ROC curves
fig = plt.figure(figsize=(10, 6))
plt.plot([0, 1], [0, 1], linestyle='--')
plt.plot(fpr1, tpr1, label='Training set (AUC = %0.2f)' % roc_auc_train)
plt.plot(fpr2, tpr2, label='Testing set (AUC = %0.2f)' % roc_auc_test)
plt.xlim([-0.01, 1.0])
plt.ylim([0.0, 1.0])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend(loc='lower right');
plt.show()
```
# Round the predictions according to the threshold with a TPR of 0.85

# L1 model
pred_l1round = [] # List to store binary outcomes
L1threshold = thres[[X[0] for x in enumerate(tpr2) if x[1] > 0.85][0]]

for i in predl1: # Loop over all predictions
    # If the prediction is larger than the threshold value corresponding to a TPR of 0.85 then it's a 1
    if i > L1threshold:
        pred_l1round.append(1)
    else:
        pred_l1round.append(0)

# Create a confusion matrix for L1 model

# Compute confusion matrix values
cMatrix = confusion_matrix(y_true = y_test, y_pred = pred_l1round)

# Plot
fig = figure(figsize=(10, 6)) # Create figure space
plt.imshow(cMatrix, cmap=pl.cm.Blues) # Plot CM

# Add numbers to plot
plt.text(0, 0, '{:.0f}'.format(cMatrix[0, 0]), horizontalalignment='center', fontsize='xx-large')
plt.text(0, 1, '{:.0f}'.format(cMatrix[0, 1]), horizontalalignment='center', fontsize='xx-large')
plt.text(1, 0, '{:.0f}'.format(cMatrix[1, 0]), horizontalalignment='center', fontsize='xx-large')
plt.text(1, 1, '{:.0f}'.format(cMatrix[1, 1]), horizontalalignment='center', fontsize='xx-large')

# Add Health and Heart Disease labels to each axis
labels = ['Historical Bid-Based Estimating', 'Combination Estimating']
plt.xticks(tick_marks, labels, rotation=90, fontsize='x-large')
plt.yticks(tick_marks, labels, fontsize='x-large')

# Add axis labels
plt.xlabel('True label', fontsize='xx-large')
plt.ylabel('Predicted label', fontsize='xx-large')
plt.title('L1 Confusion Matrix')
plt.show()

# Ridge Regression
LogReg = LogisticRegression(penalty='l2', solver='liblinear', max_iter=1200)

Logparams = {'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000, 70000, 1000000]}

clfLogL2 = GridSearchCV(LogReg, Logparams, cv=4, scoring='roc_auc')
clfLogL2.fit(X_train, y_train)

# Look at the parameters for the best model
clfLogL2.best_estimator_

# Use the best model to make predictions
from sklearn.metrics import r2_score
r2_score(y_test, predL2)

print(clfLogL2.best_estimator_.coef_, clfLogL2.best_estimator_.intercept_)

#Plot the ROC curve for L2-Regularization

#Training set
clfLogL2_train = clfLogL2.best_estimator_.predict_proba(X_train).T[1]
fpr1, tpr1, thresh1 = roc_curve(y_train, clfLogL2_train)
roc_auc_train = roc_auc_score(y_train, clfLogL2_train)

#Testing set
fpr2, tpr2, thresh2 = roc_curve(y_test, clfLogL2_test)
roc_auc_test = roc_auc_score(y_test, clfLogL2_test)

# Plot the ROC curves
fig = plt.figure(figsize=(10, 5))
plt.plot([0, 1], [0, 1], linestyle='--')
plt.plot(fpr1, tpr1, label='Training set (AUC = %0.2f)' % roc_auc_train)
plt.plot(fpr2, tpr2, label='Testing set (AUC = %0.2f)' % roc_auc_test)
plt.xlim([-0.01, 1.0])
plt.ylim([0.0, 1.01])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend(loc='lower right')
plt.show()

#Round the predictions according to the threshold with a TPR of 0.85

#L2 model
pred_L2round = [] #list to store binary outcomes
L2thresh = thresh2[[x[0] for x in enumerate(tpr2) if x[1] > 0.85][0]]

for i in predL2: #loop over all predictions
    if the prediction is larger than the threshold value corresponding to a TPR of 0.85 then it's a 1
    if i > L2thresh:
        pred_L2round.append(1)
    else:
        pred_L2round.append(0)
# Create a confusion matrix for L2 model

# Compute confusion matrix values
cMatrix = confusion_matrix(y_true = y_test, y_pred = pred_l2round)

# Plot
fig = figure(figsize=(10, 6)) # Create figure space
plt.imshow(cMatrix, cmap=plt.cm.Blues) # Plot CM

# Add numbers to plot
plt.text(0, 0, '{}'.format(cMatrix[0, 0]), horizontalalignment='center', fontsize = 'xx-large')
plt.text(0, 1, '{}'.format(cMatrix[1, 0]), horizontalalignment='center', fontsize = 'xx-large')
plt.text(1, 0, '{}'.format(cMatrix[0, 1]), horizontalalignment='center', fontsize = 'xx-large')
plt.text(1, 1, '{}'.format(cMatrix[1, 1]), horizontalalignment='center', fontsize = 'xx-large')

# Add Health and Heart Disease labels to each axis
tick_marks = [0,1]
labels = ['Use 10% Cost-based Estimating', 'Only use Historical Bid-based Estimating']
plt.xticks(tick_marks, labels, rotation=90, fontsize = 'x-large')
plt.yticks(tick_marks, labels,fontsize = 'x-large')

# Add axis labels
plt.ylabel('True label', fontsize = 'xx-large')
plt.xlabel('Predicted label', fontsize = 'xx-large')
plt.title("L2 Confusion Matrix")
plt.show()

---

**E: CART Python Code**

# Define the model
CARTmod = DecisionTreeClassifier(random_state=0)

# Choose some hyperparameter values
CARTparams = {'splitter':['random', 'best'], 'max_features':['auto', 'None'], 'min_samples_split':randint(1,10), 'max_depth':randint(1,10), 'min_samples_leaf':randint(1,10), 'class_weight':[None, 'balanced'], 'criterion':['gini', 'entropy']}

# Run the random search
clfCART = RandomizedSearchCV(CARTmod,CARTparams,#model and parameters
cv=4,#number of cross validation folds
scoring='roc_auc',#accuracy metric
n_iter=100)#number of random parameter combinations
clfCART.fit(X_train,y_train)

# Look at the parameters for the best model
clfCART.best_estimator_
# Determine the feature importances using the built in method
importanceDT = clfCART.best_estimator_.feature_importances_
print(importanceDT)

# Print out the corresponding features
feats = X_train.columns
print(feats)

# Organize the output
print(''
for i in range(len(importanceDT)):
    print(feats[i])
    print(importanceDT[i])

# We'll start with max depth

# Choose a bunch of depths (1 to 15 in this case)
max_depths = np.linspace(1, 15, endpoint=True)

# Storage arrays
train_results = []
test_results = []

# Loop over each depth
for max_depth in max_depths:

    # Train a CART model
    dt1 = DecisionTreeClassifier(max_depth=max_depth)
dt1.fit(X_train, y_train)

    # Get training set AUC
    preds_train = dt1.predict_proba(X_train).T[1]
    roc_auc_train = roc_auc_score(y_train, preds_train)
    train_results.append(roc_auc_train)

    # Get testing set AUC
    preds_test = dt1.predict_proba(X_test).T[1]
    roc_auc_test = roc_auc_score(y_test, preds_test)
test_results.append(roc_auc_test)
# Now let's consider the minimum number of samples per Split

# Choose a bunch of min samples per Split (1 to 1000 in this case)
min_samps_per_split = np.linspace(1, 1000, endpoint=True)

# Storage arrays
train_results = []
test_results = []

# Loop over each depth
for min_samp_per_split in min_samps_per_split:

    # Train a CART model
    dt1 = DecisionTreeClassifier(min_samples_leaf=int(min_samp_per_split))
    dt1.fit(X_train, y_train)

    # Get training set AUC
    preds_train = dt1.predict_proba(X_train).T[1]
    roc_auc_train = roc_auc_score(y_train, preds_train)
    train_results.append(roc_auc_train)

    # Get testing set AUC
    preds_test = dt1.predict_proba(X_test).T[1]
    roc_auc_test = roc_auc_score(y_test, preds_test)
    test_results.append(roc_auc_test)

# Plot the results
fig = plt.figure(figsize=(10, 6))
line1, = plt.plot(min_samps_per_leaf, train_results, 'b')
line2, = plt.plot(min_samps_per_leaf, test_results, 'r')
plt.legend(['Training set', 'Testing set'])
plt.ylabel('AUC')
plt.xlabel('Minimum number of samples per split')
plt.show()
# Now let's consider the minimum number of samples per leaf

# Choose a bunch of min samples per leaf (1 to 1000 in this case)
min_samps_per_leaf = np.linspace(1, 1000, endpoint=True)

# Storage arrays
train_results = []
test_results = []

# Loop over each depth
for min_samp_per_leaf in min_samps_per_leaf:
    # Train a CART model
    dt1 = DecisionTreeClassifier(min_samples_leaf=int(min_samp_per_leaf))
    dt1.fit(X_train, y_train)

    # Get training set AUC
    preds_train = dt1.predict_proba(X_train)[:, 1]
    roc_auc_train = roc_auc_score(y_train, preds_train)
    train_results.append(roc_auc_train)

    # Get testing set AUC
    preds_test = dt1.predict_proba(X_test)[:, 1]
    roc_auc_test = roc_auc_score(y_test, preds_test)
    test_results.append(roc_auc_test)

# Plot the results
fig = plt.figure(figsize=(10, 6))
line1, = plt.plot(min_samps_per_leaf, train_results, 'b')
line2, = plt.plot(min_samps_per_leaf, test_results, 'r')
plt.legend(['Training set', 'Testing set'])
plt.ylabel('AUC')
plt.xlabel('Minimum number of samples per leaf')
plt.show()

text_representation = tree.export_text(clfCART.best_estimator_)
print(text_representation)

with open("decision_tree.log", "w") as fout:
    fout.write(text_representation)

fig = plt.figure(figsize=(20, 25))
_ = tree.plot_tree(clfCART.best_estimator_,
                   feature_names=X.columns,
                   rounded=True, proportion=False,
                   filled=True)
fig.savefig("CART best estimator", dpi=130)
# Compute the training and testing set ROC curves

clpreds_train = clfCART.best_estimator_.predict_proba(X_train).T[1]
fp1, tp1, thresh1 = roc_curve(y_train, clp nons_train)
roc_auc_train = roc_auc_score(y_train, clp nes_train)

clpreds_test = clfCART.best_estimator_.predict_proba(X_test).T[1]
fp2, tp2, thresh2 = roc_curve(y_test, clp nes_test)
roc_auc_test = roc_auc_score(y_test, clp nes_test)

# Plot the ROC curves
fig = plt.figure(figsize=(10, 6))
plt.plot([0, 1], [0, 1], linestyle='--')
plt.plot(fp1, tp1, label='Training set (AUC = %0.2f)' % roc_auc_train)
plt.plot(fp2, tp2, label='Testing set (AUC = %0.2f)' % roc_auc_test)
plt.xlim([-0.01, 1.0])
plt.ylim([0.0, 1.0])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend(loc='lower right');
plt.show()

# Use the best model to make predictions
predCART = clfCART.best_estimator_.predict_proba(X_test).T[1]

# Round the predictions according to the threshold with a TPR of 0.85

# CART model
predCARTround = [] # list to store binary outcomes
CARTthresh = thresh2[x[0] for x in enumerate(tp2) if x[1] > 0.85][8]

for i in predCART: # loop over all predictions

    # If the prediction is larger than the threshold value corresponding to a TPR of 0.85 then it's a 1
    if i > CARTthresh:
        predCARTround.append(1)
    else:
        predCARTround.append(0)
# Create a confusion matrix for CART model

cMatrix = confusion_matrix(y_true = y_test, y_pred = predCARTround)

# Plot
fig = figure(figsize=(10, 6))  # Create figure space
plt.imshow(cMatrix, cmap=pl.cm.Blues)  # Plot CM

# Add numbers to plot
plt.text(0, 0, '{:.1f}'.format(cMatrix[0, 0]), horizontalalignment='center', fontsize = 'xx-large')
plt.text(0, 1, '{:.1f}'.format(cMatrix[0, 1]), horizontalalignment='center', fontsize = 'xx-large')
plt.text(1, 0, '{:.1f}'.format(cMatrix[1, 0]), horizontalalignment='center', fontsize = 'xx-large')
plt.text(1, 1, '{:.1f}'.format(cMatrix[1, 1]), horizontalalignment='center', fontsize = 'xx-large')

# Add Health and Heart Disease labels to each axis

tick_marks = [0,1]
labels = ['Historical Bid-Based Estimating', 'Combination Estimating']
plt.xticks(tick_marks, labels, rotation=90, fontsize = 'x-large')
plt.ylim([-0.5,1.5])
plt.yticks(tick_marks, labels,fontsize = 'x-large')

# Add axis labels
plt.ylabel('True label', fontsize = 'xx-large')
plt.xlabel('Predicted label', fontsize = 'xx-large')
plt.title('CART Confusion Matrix')
plt.show()

---

**F: Random Forest Python Code**

```python
# Define the model
RFmod = RandomForestClassifier(random_state=0)

# Choose some hyperparameter values
RFparams={
'n_estimators':randint(10,250),
'max_features':['auto',None],
'min_samples_split':randint(1,10),
'max_depth':randint(1,10),
'min_samples_leaf':randint(1,10),
'class_weight':[None,'balanced']
}

# Look at the parameters for the best model
clfRF.best_estimator_

# Run the random search
clfRF = RandomizedSearchCV(RFmod,RFparams,# model and parameters
cv=4,#number of cross validation folds
scoring='roc_auc',#accuracy metric
n_iter=1)#number of random parameter combinations
clfRF.fit(X_train,y_train)
```
# We'll start with max depth

# Choose a bunch of depths (1 to 15 in this case)
max_depths = np.linspace(1, 15, endpoint=True)

# Storage arrays
train_results = []
test_results = []

# Loop over each depth
for max_depth in max_depths:

    # Train a RF model
    dt1 = RandomForestClassifier(max_depth=max_depth)
dt1.fit(X_train, y_train)

    # Get training set AUC
    preds_train = dt1.predict_proba(X_train).T[1]
    roc_auc_train = roc_auc_score(y_train, preds_train)
    train_results.append(roc_auc_train)

    # Get testing set AUC
    preds_test = dt1.predict_proba(X_test).T[1]
    roc_auc_test = roc_auc_score(y_test, preds_test)
    test_results.append(roc_auc_test)

# Plot the results
fig = plt.figure(figsize=(10, 6))
line1, = plt.plot(max_depths, train_results, 'b')
line2, = plt.plot(max_depths, test_results, 'r')
plt.legend({'Training set', 'Testing set'})
plt.ylabel('AUC')
plt.xlabel('Max depth')
plt.show()
# Now let's consider the minimum number of samples per Split

# Choose a bunch of min samples per Split (1 to 1000 in this case)
min_samps_per_split = np.linspace(1, 1000, endpoint=True)

# Storage arrays
train_results = []
test_results = []

# Loop over each depth
for min_samp_per_split in min_samps_per_split:
    # Train a RF model
    dt1 = RandomForestClassifier(min_samples_leaf=int(min_samp_per_split))
    dt1.fit(X_train, y_train)

    # Get training set AUC
    preds_train = dt1.predict_proba(X_train).T[1]
    roc_auc_train = roc_auc_score(y_train, preds_train)
    train_results.append(roc_auc_train)

    # Get testing set AUC
    preds_test = dt1.predict_proba(X_test).T[1]
    roc_auc_test = roc_auc_score(y_test, preds_test)
    test_results.append(roc_auc_test)

# Plot the results
fig = plt.figure(figsize=(10, 6))
line1, = plt.plot(min_samps_per_split, train_results, 'b')
line2, = plt.plot(min_samps_per_split, test_results, 'r')
plt.legend({'Training set', 'Testing set'})
plt.ylabel('AUC')
plt.xlabel('Minimum number of samples per split')
plt.show()
# Now let's consider the minimum number of samples per leaf

# Choose a bunch of min samples per leaf (1 to 1000 in this case)
min_samps_per_leaf = np.linspace(1, 1000, endpoint=True)

# Storage arrays
train_results = []
test_results = []

# Loop over each depth
for min_samp_per_leaf in min_samps_per_leaf:
    # Train a RF model
    dt1 = RandomForestClassifier(min_samples_leaf=int(min_samp_per_leaf))
    dt1.fit(X_train, y_train)

    # Get training set AUC
    preds_train = dt1.predict_proba(X_train).T[1]
    roc_auc_train = roc_auc_score(y_train, preds_train)
    train_results.append(roc_auc_train)

    # Get testing set AUC
    preds_test = dt1.predict_proba(X_test).T[1]
    roc_auc_test = roc_auc_score(y_test, preds_test)
    test_results.append(roc_auc_test)

# Plot the results
fig = plt.figure(figsize=(10, 6))
line1, = plt.plot(min_samps_per_leaf, train_results, 'b')
line2, = plt.plot(min_samps_per_leaf, test_results, 'r')
plt.legend({'Training set', 'Testing set'})
plt.ylabel('AUC')
plt.xlabel('Minimum number of samples per leaf')
plt.show()
# Now let’s consider the number of trees in the model

# Choose a bunch of n_estimators (1 to 1000 in this case)
n_estimators = np.linspace(1, 300, endpoint=True)

# Storage arrays
train_results = []
test_results = []

# Loop over each depth
for n_estimator in n_estimators:

    # Train a RF model
    dt1 = RandomForestClassifier(n_estimators=int(n_estimator))
    dt1.fit(X_train, y_train)

    # Get training set AUC
    preds_train = dt1.predict_proba(X_train).T[1]
    roc_auc_train = roc_auc_score(y_train, preds_train)
    train_results.append(roc_auc_train)

    # Get testing set AUC
    preds_test = dt1.predict_proba(X_test).T[1]
    roc_auc_test = roc_auc_score(y_test, preds_test)
    test_results.append(roc_auc_test)

# Plot the results
fig = plt.figure(figsize=(10, 6))
line1, = plt.plot(n_estimators, train_results, 'b')
line2, = plt.plot(n_estimators, test_results, 'r')
plt.legend(['Training set', 'Testing set'])
plt.ylabel('AUC')
plt.xlabel('Number of Trees')
plt.show()

# Determine the feature importances using the built in method
importanceDT = clfRF.best_estimator_.feature_importances_
print(importanceDT)

# Print out the corresponding features
feats = X_train.columns
print(feats)

# Organize the output
print('
for i in range(len(importanceDT)):
    print(feats[i])
    print(importanceDT[i])
# Compute the training and testing set ROC curves

clfpreds_train = clfRF.best_estimator_.predict_proba(X_train).T[1]
fpr1, tpr1, thresh1 = roc_curve(y_train, clfpreds_train)
roc_auc_train = roc_auc_score(y_train, clfpreds_train)

clfpreds_test = clfRF.best_estimator_.predict_proba(X_test).T[1]
fpr2, tpr2, thresh2 = roc_curve(y_test, clfpreds_test)
roc_auc_test = roc_auc_score(y_test, clfpreds_test)

# Plot the ROC curves
fig = plt.figure(figsize=(10, 6))
plt.plot([0, 1], [0, 1], linestyle='--')
plt.plot(fpr1, tpr1, label='Training set (AUC = %0.2f)' % roc_auc_train)
plt.plot(fpr2, tpr2, label='Testing set (AUC = %0.2f)' % roc_auc_test)
plt.xlim([-0.01, 1.0])
plt.ylim([0.0, 1.0])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend(loc='lower right')
plt.show()

# Use the best model to make predictions

edRF = clfRF.best_estimator_.predict_proba(X_test).T[1]

# Round the predictions according to the threshold with a TPR of 0.85

#RF model
predRFround = []  # list to store binary outcomes
RFthresh = thresh2[[x[0] for x in enumerate(tpr2) if x[1] > 0.85][0]]

for i in predRF:  # loop over all predictions
    # If the prediction is larger than the threshold value corresponding to a TPR of 0.85 then it's a 1
    if i > RFthresh:
        predRFround.append(1)
    else:
        predRFround.append(0)
G: Model Comparison Python Code

```python
# Create a confusion matrix for CART model

# Compute confusion matrix values
cMatrix = confusion_matrix(y_true = y_test, y_pred = predRFround)

# Plot
fig = figure(figsize=(10, 6)) # Create figure space
plt.imshow(cMatrix, cmap=pl.cm.Blues) # Plot CM

# Add numbers to plot
plt.text(0, 0, '{:.2f}'.format(cMatrix[0, 0]), horizontalalignment='center', fontsize = 'xx-large')
plt.text(0, 1, '{:.2f}'.format(cMatrix[1, 0]), horizontalalignment='center', fontsize = 'xx-large')
plt.text(1, 0, '{:.2f}'.format(cMatrix[0, 1]), horizontalalignment='center', fontsize = 'xx-large')
plt.text(1, 1, '{:.2f}'.format(cMatrix[1, 1]), horizontalalignment='center', fontsize = 'xx-large')

# Add Health and Heart Disease Labels to each axis
tick_marks = [0.1]
labels = ['Historical Bid-Based Estimating', 'Combination Estimating']
plt.xticks(tick_marks, labels, rotation=90, fontsize = 'x-large')
plt.yticks(tick_marks, fontsize = 'x-large')

# Add axes labels
plt.ylabel('True label', fontsize = 'xx-large')
plt.xlabel('Predicted label', fontsize = 'xx-large')
plt.title("RF Confusion Matrix")
plt.show()
```

```python
# Plotting ROC curve
fig = plt.figure(figsize=(10, 6))
plt.plot([0, 1], [0, 1], linestyle='--')
plt.plot(fprLogReg, tprLogReg, label='Logistic Regression (AUC = %0.2f)' % roc_auc_logReg)
plt.plot(fprCART, tprCART, label='CART (AUC = %0.2f)' % roc_auc_cart)
plt.plot(fprRF, tprRF, label='Random Forest (AUC = %0.2f)' % roc_auc_rf)
plt.xlim([0, 1.0])
plt.ylim([0, 1.0])
plt.title('ROC Comparison of Different Models', fontsize = 15)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend(loc='lower right');
plt.show()
```