Generalized Linear Model approach to the Prediction of the outcome of Mixed Martial Arts Fights

by

Jackson Thurmond

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Master of Science
in Mathematics

at

The University of Wisconsin-Milwaukee
May 2025

ABSTRACT

Generalized Linear Model approach to the Prediction of the outcome of Mixed Martial Arts Fights

by

Jackson Thurmond

The University of Wisconsin-Milwaukee, 2025
Under the Supervision of Professor David Spade

Mixed martial arts is a complex combat sport that encompasses striking, grappling, and submissions. In a sport where fights can be won by finishing a fight or go to decision there is a multitude of factors that can influence the outcome of a fight. In an effort to determine which factors are statistically significant to a fight a generalized linear model approach was selected. Since mixed martial arts is a sport in which two competitors fight, and one is declared a winner, the result of a fight can be thought of a binary classification problem.

Table of Contents

List of Figures

List of Tables

Introduction

I.1 Motivation

The motivation of this research is to see what are the most important aspects to predicting fight outcomes in mixed martial arts (MMA). There is a long list of what is allowed in the sport, and this leads to a large, complex dataset of measurements being reported from each fight.

There has been previous work done on this same problem. Other research has been done using generalized linear models with more reduced datasets with respect to the number of possible predictor variables. In these other works generalized linear models were used to explore the tailored statistics created by the researchers (Jabbarov,2024)(Johnson,2009). The creation of tailored statistics is to see if there is a more efficient way of measuring a fighters performance. There has also been work done using a Bayesian framework (Apelgren,Eklund, 2024). Bayesian regression is a type of conditional modeling. The analysis of fighter performances to predict the outcome of fights and measure fighter performances is a growing field of interest in the sport.

Determining in which areas of the sport a fighter should train and place emphasis is a method popularized in sports such as baseball and basketball. The areas of research and development for sports teams is a growing industry and a market in MMA. One implementation of the models in this research would be to analyze many simulated fights in which two athletes fight with varying performances. Given some fixed statistics of a fighter and varying within some dataset of possible performances, a coach can determine a plan for the most likely outcome for success of a given fighter. Another possible application of a prediction model is in sports betting.

I.2 Problem Description

Our dataset consists of variables of several types. There are continuous, discrete, and indicator variables present in the dataset with a total of 5902 observations. The dataset we analyze in this research is a 5902*160 matrix. The overarching goal iss to find a model that could find which aspects of fighting were significant and to predict the outcome of fights. With this goal in mind, careful consideration is needed in choosing a function that could best describe the dataset. This determination will be justified later in this thesis. These choices are made to balance goals of maintaining the ability to interpret the model, prediction power, and the fit.

In the analysis of the UFC dataset there needs to be a method to analyze the data present. Observations in the dataset are labeled as red or blue indicating the winner. This is translated as 0, blue wins, and 1, red wins. Therefore, we are dealing with a binary response problem. There are 159 variables that could be important factors in determining what the most impactful are to the outcome of 0 or 1. This choice in changing the red and blue wins from categorical variables to binary variables is a natural choice since the same information is described by the variable, but we are able to handle them easier in analysis. This thesis will examine the application of logistic regression applied to the data set of the (UFC) up through 2021.

I.3 Literature Review

There has been increasing progress in the area of statistical analyses on the UFC datasets. As it is a new sport in comparison to baseball, football, and other popular sports. In a sport that is relatively new and constantly evolving there has been more interest in analyzing and studying fights. The UFC is a attractive to sports bettors, so this has also increased interest in the analysis of fights. The desire to create prediction model has increased and has expanded past the sports betting world and made its way into academic study as well. A few highlights of previous research are presented below.

Recently, a paper by Ismail Jabbarov (Jabbarov,2024) attempts to analyze the UFC dataset by utilizing a logistic regression model. The dataset uses the statistics provided by the UFC in its first release of fight results as well as some tailored statistics created by the authors in an attempt to offer a new perspective on the fights. Their dataset contains 5144 observations of fights compared to our 5902 observations. This discrepancy in number of fights comes from the more recent update of the dataset in which we got our dataset from the same Kaggle source ("UFC-Fight historical data from 1993 to 2021"). Thus this paper enables us to see where an earlier attempt at performing logistic regression to the dataset was done. This paper separates the distinction between what they consider first level variables such as punches thrown, knockdowns, take downs made, submission attempts, and positional advances. The second level variables are the ones they created which are ground effectiveness, power rating and punch ratio for example. These statistics were constructed to offer possible variables which they thought may offer a newer insight to the analysis of each fight observation.

In 2009, Jerimiah Douglas Johnson (Johnson,2009) attempts to analyze the UFC dataset by utilizing a logistic regression model. This is an example of generalized linear model. The dataset used uses count statistics as well as some tailored statistics created by the authors in an attempt to offer a new perspective on the fights. This paper, similarly to the (Jabbarov,2024) paper, uses the difference between cumulative statistics of the red and blue corner fighters.

Another more recent study, (Apelgren, Eklund- 2024) attempts to create two separate models to analyze the UFC data set to try to predict the outcome of a UFC fight. The two methods that were used are a logistic regression model and a Bayesian regression model. The Bayesian model is one that is so far the most notable difference between other papers in this grouping of papers before this project. This method requires using fighters with enough experience to have enough priors to obtain its distribution. This Bayesian model can be powerful if applied to fighters with enough fights in the UFC. This is a limitation to the Bayesian model that is not present in the logistic regression model built using the frequentist statistical framework. The Bayesian model incorporates the personal fighter information, by constructing the priors for the distribution. Our model does not use this information. Each observation in our model is assumed to be independent.

In another recent study (Walsh, 2022) attempts to predict the outcome of fights through a neural network and a random forest model. The random forest consists of decision trees making predictions based on majority vote. The neural network consists of an input layer, hidden layers, and an output layer. These two methods are compared in their prediction power. The random forest model has limitations in the dataset specifically since the ufc dataset is not well separable. Meaning that for each decision tree the a proper threshold of what constitutes a majority vote of being a win for red or blue has to be determined. In data that is not clearly separable, a random forest algorithm will iterate to overfit so that each node is singular, thus a majority vote threshold was placed in the algorithm.

Background

II.1 Mixed Martial Arts

MMA is a combat sport. Two fighters, indicated by their corner, either red or blue fight. Rounds are scheduled for 5 minutes long each, and there are 3 rounds total. There are also fights for a championship title that are 5 rounds total. Fights are scheduled in this way, but do not always go this long. A fight can be finished at any point in the scheduled time. If a fight goes for the full length scheduled, judges determine the winner of the fight. In mixed martial arts striking, grappling, and submissions are used.

Striking is what may be more familiar to most people. Striking is punching, kicking, knees, and elbows. Martial arts such as boxing, kickboxing, muay thai, karate, taekwondo and other striking martial arts are used. Each of these martial arts have their own techniques. Some of these techniques like a leg kick are applied differently. The distinction between types of kicks may seem pedantic, but applications of these techniques may lead to different outcomes. In striking, the goal is to hit the other fighter and not get hit yourself. As mixed martial arts is not based on a point system like fencing or sport karate, the focus in mixed martial arts is more similar to boxing in the sense that damage done to your opponent is the goal. You can either get a knockdown, knockout, or neither when hitting someone. A knockdown is when a fighter goes down to the ground from getting hit. A knockout is called when a fighter goes unconscious. A knockout occurring on the feet would be a hit on the feet that occurs in an unconscious opponent immediately. A technical knockout is when an opponent is knocked down and finished on the ground. Unlike boxing, when a knockdown occurs the fight is not over until a fighter is finished.

Grappling is a mixture of martial arts such as wrestling and jui-jujitsu for example. There are both standing components, such as take downs where a fighter takes their opponent to the ground and what they do on the ground by improving positions and submissions. This particular part of mixed martial arts is a place where new or casual fans have trouble understanding what is occurring. Striking is a more clear part of the sport in which one can see the fighters hitting each other. Actions such as having what is considered control in grappling over the opponent and improving positioning can be difficult to understand. It is important to discuss this piece of mixed martial arts since a large number of our variables include take downs, control time, and submission attempts for example. A submission is when an athlete puts there opponent in a hold that forces them to tap to the hold due to pain, threat of passing out, passing out, or a breaking of a limb.

In mixed martial arts, there is creativity and options open to all competitors on the techniques that can be used as long as they are not specifically illegal in the rule book. Essentially, mixed martial arts is fighting under a more open rule set compared to other combat sports. What started as a competition between different martial arts to see which was more effective eventually evolved into a sport where all were used to create more balanced and efficient fighters.

It must be addressed that the judging in MMA is subjective. Judges award points to fighters at the end of each round based on areas indicated by striking, grappling, and submissions as described above. There are also parts of fighting such as octagon control and aggression that are not only not measurable currently, but also highly subjective. This poses a more difficult aspect to consider. Fights that are not won by finish, are decided by subjective judging decisions. This background knowledge however, is even more motivation for analysis.

II.2 Statistical

Fixed effects models such as logistic regression are powerful statistical models that can be used in binary classification problems. We must first build the intuition of a linear model before an example of a generalized linear model. Multiple linear regression is a model with normally distributed errors, with mean 0 and fixed variance, which is useful for understanding the relationship between variables that appear linearly related to an effect.

In matrix notation, multiple linear regression can be represented as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where:

- $\mathbf{Y}$ is an $n \times 1$ vector of the dependent variable.
- $\mathbf{X}$ is an $n \times (p+1)$ matrix of the independent variables, including a column of ones for the intercept.
- $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of the coefficients.
- $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of the error terms.
- $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0,\sigma^2)$

The matrices are defined as follows:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The hat matrix, often denoted as $\mathbf{H}$, is defined as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Where:

- $\mathbf{X}$ is the matrix of independent variables.
- $\mathbf{X}^T$ is the transpose of $\mathbf{X}$.
- $(\mathbf{X}^T\mathbf{X})^{-1}$ is the inverse of the matrix product $\mathbf{X}^T\mathbf{X}$.

The hat matrix has several important properties:

- Symmetry: $\mathbf{H}$ is symmetric, meaning $\mathbf{H} = \mathbf{H}^T$.
- Idempotency: $\mathbf{H}$ is idempotent, meaning $\mathbf{H}\mathbf{H} = \mathbf{H}$
- Projection: $\mathbf{H}$ projects the vector of observed values $\mathbf{Y}$ onto the vector of fitted values $\hat{\mathbf{Y}}$.

Using the hat matrix, the fitted values $\hat{\mathbf{Y}}$ can be expressed as:

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

This relationship shows how the hat matrix transforms the observed values into the predicted values. The residuals, which are the differences between the observed and fitted values, can be expressed as:

$$\boldsymbol{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Where $\mathbf{I}$ is the identity matrix.

Multiple linear regression can be used for many different applications, however, we will show why a generalized linear model suits the binary classification problem more appropriately. Logistic regression models are powerful tools in statistics, enabling the prediction of a binary dependent variable based on the values of one or more independent variables. Unlike linear regression, logistic regression is used when the dependent variable is discrete, typically taking values 0 or 1. In this case 1 indicates a win for the red fighter and 0 indicates a win for the blue fighter. We assume that the response variable y_i is a Bernoulli random variable with probability distribution function listed below:

$$P(y_i = 1) = \pi_i$$

$$P(y_i = 0) = 1 - \pi_i$$

We also assume that the error terms ϵ_i have the property that $E(\epsilon_i) = 0$, this implies that the expected value of the response variable y_i is

$$E(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i$$

Thus, the expected value of the response variable is the probability that the value takes on 1. In the context of our situation the probability that red will win. The variance will also not be constant as shown below.

$$E[y_i - E(y_i)]^2 = (1 - \pi_i)^2\pi_i + (0 - \pi_i)^2(1 - \pi_i) = \pi_i(1 - \pi_i)$$

This presents the motivation for the choice in model for this data set. We are also dealing with a model that should produce the probability of achieving the value 1. A linear model would not be bounded between the values 0 and 1. This implies that a model used should be a non decreasing function that is sigmoid shape bounded between 0 and 1.

In logistic regression, the function relates the probability of the dependent variable y being 1 to the independent variables $\mathbf{x}$ through the logistic function:

The logistic regression model can be written in matrix notation as:

$$\tilde{\boldsymbol{\pi}} = \frac{1}{1 + \exp(-\mathbf{X}\boldsymbol{\beta})}$$

where:

- $\tilde{\boldsymbol{\pi}}$ is the vector of predicted probabilities.
- $\mathbf{X}$ is the matrix of independent variables.
- $\boldsymbol{\beta}$ is the vector of parameters.

From this logistic regression model it can be seen that the $\boldsymbol{\beta}$ parameters are linear in relation to $\mathbf{X}$. Thus, the logistic regression is categorized as a generalized linear model.The main objective is to then estimate the parameters $\boldsymbol{\beta}$. This is commonly done by maximizing the likelihood function L, which is a function of the parameters for sample point x. Each observation follows a Bernoulli distribution thus we may say that the probability distribution follows the form:

$$f_i(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

Since the observations are independent we may say that the likelihood functions follows:

$$L(\beta_1, \beta_2, ..., \beta_k | y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

It is more convenient to use the log likelihood function for our purposes so, since the natural log function is one to one and will preserve the data we will write:

$$\ln L(\tilde{\beta}|\tilde{y}) = \sum_{i=1}^{n} y_i \ln(\pi_i) + \sum_{i=1}^{n} (n_i - y_i) \ln(1 - \pi_i)$$

Where $\tilde{y}$ is the vector of response y_i variables.

Maximizing this log-likelihood function with respect to $\tilde{\beta}$ provides the maximum likelihood estimators for the parameters. The fixed effects approach as listed above is the method used to estimate the β parameters of the model. The method above is simply how the model is made when the variable selection is complete. However, it is impractical to assume that all variables available in a dataset are actually statistically significant to the model. Thus, the choice made was to use backwards selection method to test for the removal of not statistically significant variables from the model.

Backwards regression is a method that tests the statistical significance of a given predictors by testing the difference of the variable included in the model versus the model excluding that variable. For the fixed model method the values that will be used to test the significance of the variable are Akaike information criterion (AIC) and Bayesian information criterion (BIC).

$$AIC = -2 * \ln L(\hat{\beta}|y) + (2 * k)$$

$$BIC = -2 * \ln L(\hat{\beta}|y) + (k * \ln N)$$

where:

- k is the number of predicted estimated parameters.
- N is the number of observations.
- $\hat{\beta}$ is the maximized estimated parameters.

Binned residuals are a tool to evaluate the fit of logistic regression models. They will be the primary tool for the evaluation of model assumptions in this work. They help identify patterns or deviations that might indicate issues with the model.

Given a logistic regression model, the residuals are calculated as follows:

$$\epsilon_i = \pi_i - \hat{\pi}_i$$

where π_i is the observed value and $\hat{\pi}_i$ is the predicted probability.

To create binned residuals:

1. Sort the observations by their predicted values $\hat{\pi}_i$.

2. Divide the sorted observations into bins of equal size.

3. Calculate the mean predicted value and mean residual for each bin.

Binned residuals are used to:

- Assess model fit.

- Identify possible influential points.

- Identify possible trends.

Mixed regression models is a type of regression model that separates fixed components in regression, and has random components in regression. This method tends to be used to separate groups within a sample. One example would be a sample population from a hospital, with the groups within that hospital being in different wings. When construction a mixed regression model there are three possibilities, having a random intercept, random slope, or both. This determination is dependent on the dataset being analyzed. The equation for the mixed regression model is presented below for a mixed regression model for a linear regression. It would have to be transmored to be made a mixed regression model for logistic regression.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + \epsilon_{ij}$$

where:

- y_{ij} is the dependent variable for the j-th observation in the i-th group.

- β_0 is the fixed intercept.

- β_1 is the fixed slope coefficient.

- x_{ij} is the independent variable for the j-th observation in the i-th group.

- u_i is the random effect for the i-th group.

- ϵ_{ij} is the residual error for the j-th observation in the i-th group.

Methods of Model Building

III.1 Data Preparation

This thesis examines the application of logistic regression applied to the Mixed Martial arts data set of the Ultimate Fighting Championship up through 2021. Logistic regression is an extension of linear regression to binary response variables, in this case either a win for Red or a win for Blue. In this thesis, the use of this generalized linear model is applied to 5902 observations, corresponding to 5902 separate fights.

It is necessary to observe the correlation matrix between variables in order to see the relationships between the predictor variables that will be investigated. This step does not inform in particular which variables to remove to avoid redundant information, however we can see their correlation. It is useful to take steps to understand the particular dataset before applying models and methodologies. This is considered one of the few steps in exploratory data analysis (EDA). The correlation matrix of the data is one of the tools that will help in this part of the analysis. If the dataset were to have fewer possible predictor variables utilizing plots to see the relationships between predictors and the observation results is a common first step. However, in the case of this dataset's size viewing all possible combinations of the vectors in the data matrix would not be efficient. This method still produces a 160 by 160 matrix. Here, defining a threshold to view the highly correlated pairs is necessary. Scanning the matrix component by component would be inefficient here. Thus, a threshold of .95 correlation is determined to be that which would indicate high evidence

Var1	Var2	Correlation Value
B_avg_HEAD_att	B_avg_SIG_STR_att	0.979420032
B_avg_DISTANCE_att	B_avg_SIG_STR_att	0.967771913
B_avg_opp_HEAD_att	B_avg_opp_SIG_STR_att	0.981877269
B_avg_opp_DISTANCE_att	B_avg_SIG_STR_att	0.96915106
B_avg_DISTANCE_att	B_avg_HEAD_att	0.951405499
B_avg_opp_DISTANCE_att	B_avg_opp_HEAD_att	0.953194034
B_avg_BODY_landed	B_avg_BODY_att	0.957068766
B_avg_opp_BODY_landed	B_avg_opp_BODY_att	0.95406908
B_avg_LEG_landed	B_avg_LEG_att	0.981113508
B_avg_opp_LEG_landed	B_avg_opp_LEG_att	0.980329286
R_avg_HEAD_att	R_avg_SIG_STR_att	0.978203636
R_avg_DISTANCE_att	R_avg_SIG_STR_att	0.96676307
R_avg_opp_HEAD_att	R_avg_opp_SIG_STR_att	0.983137904
R_avg_opp_DISTANCE_att	R_avg_SIG_STR_att	0.972036916
R_avg_opp_DISTANCE_att	R_avg_opp_HEAD_att	0.956582802
R_avg_BODY_landed	R_avg_BODY_att	0.957846932
R_avg_opp_BODY_landed	R_avg_opp_BODY_att	0.954080939
R_avg_LEG_landed	R_avg_LEG_att	0.979821266
R_avg_opp_LEG_landed	R_avg_opp_LEG_att	0.975542578
R_avg_CLINCH_landed	R_avg_CLINCH_att	0.96978242
R_avg_opp_CLINCH_landed	R_avg_opp_CLINCH_att	0.964103986
R_avg_GROUND_landed	R_avg_GROUND_att	0.965950751
R_avg_opp_GROUND_landed	R_avg_opp_GROUND_att	0.966231506

Table III.1: Highly Correlated Pairs

of causing multicollinearity in the final model. Multicollinearity occurs when two or more predictor variables are highly correlated, meaning they provide redundant information about the response variable. This can make it difficult to determine the individual effect of each predictor on the outcome.

One factor that could explain the high correlation between particular pairs of variables is that they are linear combinations of one another. In the creation of this dataset there are indeed variables that are linear combinations of others. The removal of these linearly dependent vectors is a requirement for regression so that the hat matrix is full column rank. When the matrix is not full column rank there can be multiple solutions to the regression coefficients.

Once the linearly dependent variables are removed, the next step is removing variables with a high variance inflation factors to decrease the size of the initial full model before use. High variance inflation factors may indicate a possibility of mutlicollinearity. Limiting the possibility of multicollinearity early in the model through the removal of variables with high VIF's can help mitigate problems in the evaluation of the model later on. The equation for Variance Inflation Factors (VIF) is given by:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

$$R_i^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2/(y_i - \bar{y})^2$$

Where:

- y_i is the ith observation.
- $\bar{y}$ is the mean observation.
- $\hat{y}_i$ is the fitted value of the model.

VIF is preferred instead of correlation between predictors when investigating multicollinearity because correlation alone can be misleading. It should be noted that this would not be possible regressing on the indicator variables, so indicator variables are removed, so that variance inflation factors were only calculated by regressing on discrete and continuous variables. Indicator variables are replaced back into the dataset after the removal of the discrete and continuous variables that had variance inflation factors above 3. This is not yet the dataset that will be used to construct the initial model. The current indicator variables for weight class have an issue that for the lighter weight classes for men and women they have the same requirements for weight and the only defining feature that differentiates the fighters is sex. We may then remove the indicator variables for weight class and then create a variable for sex. In this way we can avoid the issue in which there is a variable for weight class and weight. We will instead have a variables for the weight of each fighter and an indicator variable for sex.

Weight Class	Gender	Weight (lbs)	Weight (kg)
Strawweight	Women	115 lbs	52.2 kg
Flyweight	Men and Women	125 lbs	56.7 kg
Bantamweight	Men and Women	135 lbs	61.2 kg
Featherweight	Men and Women	145 lbs	65.8 kg
Lightweight	Men	155 lbs	70.3 kg
Welterweight	Men	170 lbs	77.1 kg
Middleweight	Men	185 lbs	83.9 kg
Light Heavyweight	Men	205 lbs	93.0 kg
Heavyweight	Men	265 lbs	120.2 kg

Table III.2: UFC Weight Classes for Men and Women

We now have a dataset in which we have independence of each variable, variables with a high variance inflation factor were removed, and the creation of the sex variable. The dataset is now a matrix of 1 response variable and 114 possible predictor variables. We may now proceed with the model building process with this dataset.

III.2 Logistic regression using backwards elimination

The previous section describes the methods used to ensure full column rank of the hat matrix. Now the 5902 by 115 matrix can be utilized in the creation of the initial model.

$$\tilde{\pi} = \frac{1}{1 + \exp(-\mathbf{X}\boldsymbol{\beta})}$$

The following assumptions are made in the logistic regression analysis:

- The dependent variable Y is binary.
- There is a linear relationship between the logit of the outcome and the independent variables.
- Observations are independent.

After the stating of this model, before backwards regression can be performed, an initial model must be fit as described in the background section.

The table, F1 in appendix, gives the values of each of the predictors, their standard error, the calculated t statistic value, and estimated p-value. The full model is statistically significant. However, it can be seen from the table that in this large full model there are some high p-values, which indicate that they are not significant statistically significant to the model. Therefore, for this fixed model, utilizing backwards regression based on the BIC will be performed. This is done for the purposes of removing variables to simplify the model and improve fit. At each step we compute the BIC for the model. By testing the removal of each variable we identify the variable whose removal results in the largest reduction in BIC. We continue removing the variable that leads to the largest decrease in BIC. In this iterative process, we stop when the removal of any variable will lead to an increase in BIC.

The F2 table in the appendix are the values of each of the predictors their standard error, the calculated t statistic value, and estimated p-value. This new fit model is statistically significant. Observing the p-values in this table it can be seen that almost all p-values are low indicating that the after performing backwards regression based on the BIC the variables left are significant to the model. There are some variables that are above .05 but this can be explained due to the first initial model having 114 variables. We were still able to remove 46 variables.

For the next model we will perform backwards regression based on the AIC. The stated and fit full model is the same to be able to compare the resulting model after backwards regression based on the AIC and BIC. Here we find that the model that is built using AIC is the exact same model as the one built using BIC. This result gives support to the claim that this is the best model using backwards regression that could be built using this dataset. One explanation for this result is that the different criteria converge to the same solution because the data has a clear, optimal pattern that is robust to variations in modeling techniques. Another explanation that could explain what occurred is that since the algorithms are the

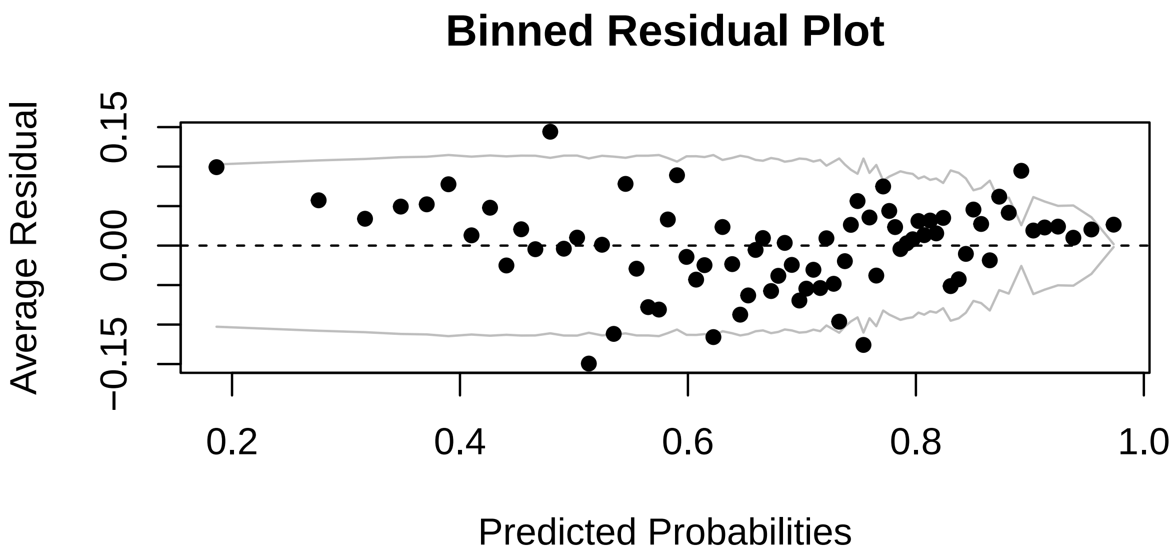

Figure III.1: Binned residual plot

same except for criteria used so a similar process could give the same results. Further research in this area would include an analysis of the variables removed at each iteration to see if the variables removed at each step were the same or if the algorithms got to the same model in different ways.

One important note is that the variable sex was removed in this model building process. This implies that there is not a significant difference between the result of a fight based on sex. As stated previously in the introduction to this section the creation of the sex variable was to avoid multicolinearity issues with having both a variable for weight and weight class. The creation of the sex variable ensured that both weight and sex were still considered in the creation of the initial model. It should be noted as well that the interpretation for sex being removed is not that sex does not matter, it is that in this dataset, where there are far more male fights than female fights, statistically sex is not a significant factor to the model. A further exploration of this factor will be considered in the piecewise model section.

The next step after finding the backwards regression model we need to validate model assumptions particularity using residual analysis. The binned residual plot in figure III.1

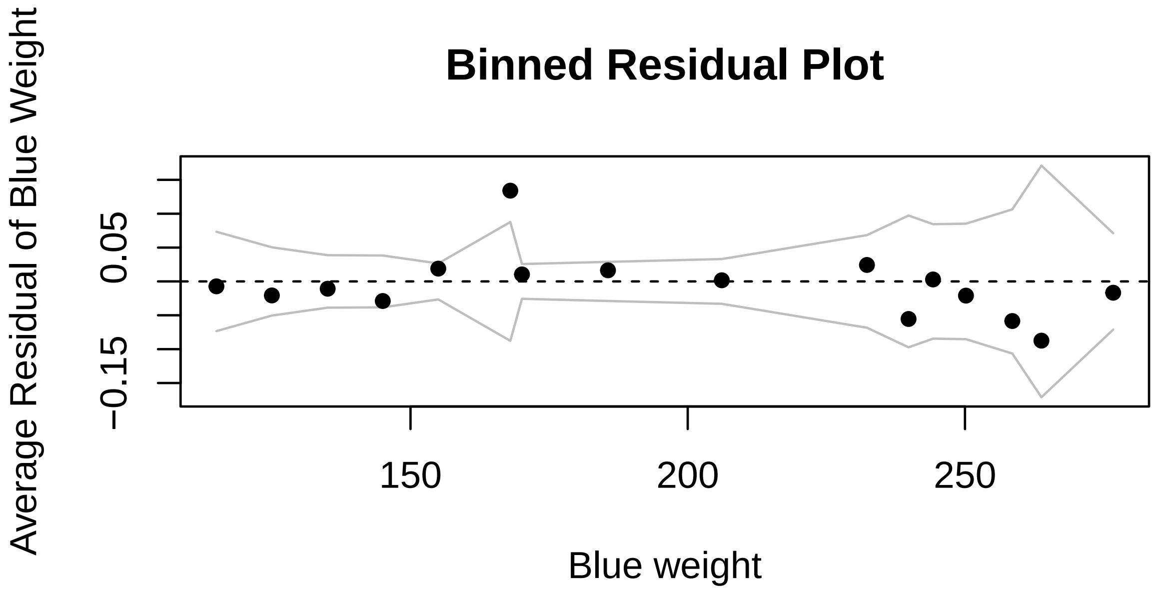

Figure III.2: Binned residual plot

presents random scatter about zero. There is no clear pattern present. Due to the bias in the dataset there are more observations of wins for the red corner which shows that there are more predicted probabilities for red winning. This is to be expected. What would be more concerning is that if residual plot did not represent this fact of the data. There are some values that fall outside the bounds of the confidence bounds but nothing extreme to cause for concern.

The next plots will represent the residuals for a variable that is continuous, and one that is for an indicator variable. For discrete and continuous variables the plot were all similar and the same goes for the indicator variables. The same analysis of all 68 residual plots were done in the same manner. For the sake of saving space not all plots will be shared in this paper as there would be 68 binned residual plots to share and discuss.

From the figure III.2, there is no apparent trend. The is random scatter about zero, and there is not significant deviation away from the from the confidence bounds of the plot. This implies that the error terms behave as to be expected and show that the fit of the model is generally acceptable. The same goes for the analysis of the indicator variable plots residual

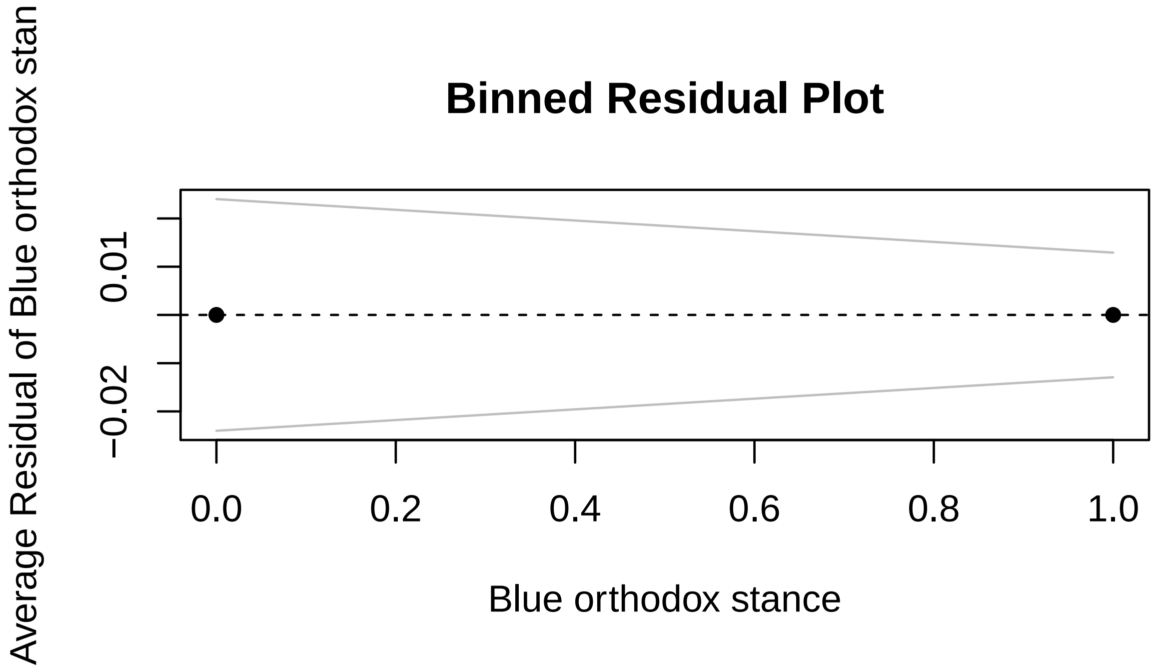

Figure III.3: Binned residual plot

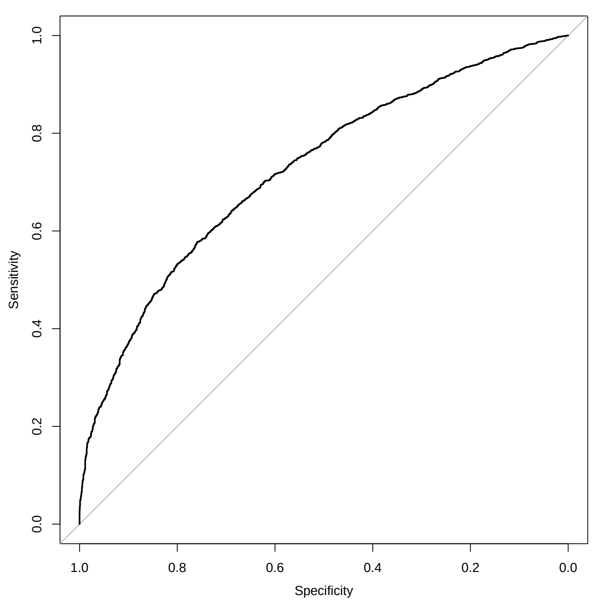

Figure III.4: Maximized Threshold of 0.714 from ROC analysis. For the residual plot for an indicator variable in figure III.3 is assesed there is not as much telling information here as the indicator is either 0 or 1.

The next step in the analysis of the backwards regression model is the model validation step. Here we will be analyzing the performance of the model on real data. In this step the data is split into a training and testing data set. Here 20 percent of the data set is randomly selected and removed form the data set. The model is then trained on the 80 percent of the dataset remaining. This data is the training set. From this model the ROC curve is maximized through specificity and sensitivity to choose the proper threshold in which π_i is determined to be either a 0 or 1. A receiver operating characteristic (ROC) curve is a graphical representation used to evaluate the performance of a binary classification model. It plots the specificity,a true negative rate, against the sensitivity, a true positive rate, at various threshold settings. From the maximization of the ROC curve a value of .714 was the threshold used to determine a win for red. The remaining data, the tesing data, is then imputed into the model and using the threshold is determined to be 0 or 1. It is then compared to the actual observations of the dataset for those particular fights. The results in the tables III. 3 and 4 are from this analysis.

	Reference	
Prediction	0	1
0	249	342
1	100	489

Table III.3: Confusion Matrix

The F1 score is the primary metric used for comparison in this paper because it provides a better reflection of model performance than accuracy alone. In this binary classification problem, the goal is to accurately predict when the red fighter will win and when the blue fighter will win.

The F1 score is the harmonic mean of precision and sensitivity (recall). By using the F1 score and the AUC (Area Under the ROC Curve), we gain a better understanding of model performance. It is important to have a reliable model that can predict the winner of a fight

Metric	Value
Accuracy	0.6254
95% CI	(0.5971, 0.6531)
Sensitivity	0.7135
Specificity	0.5884
Pos Pred Value	0.4213
Neg Pred Value	0.8302
Precision	0.4213
Recall	0.7135
F1	0.5298
Balanced Accuracy	0.6510

Table III.4: Model Validation

properly.

The current model shows better performance in predicting when the red fighter will win. However, there may be other models or methods that can improve predictions for both red and blue fighters. The AUC for the ROC curve is 0.7294, and this value will be used to compare with other models.

The values of the F1 score, accuracy, and AUC are not meaningful in isolation. These measurements of performance require comparison with other models to provide context. For example, an accuracy and F1 score above 0.5 are considered good for a model predicting fight outcomes due to the complex nature of fighting, where a single punch can change the result.

III.3 Mixed logistic regression model

As noted earlier in the data preparation section, there is a possible multicollinearity issue present in the indicator variables of weight class. The weight class variable indicates both weight class and sex. In the fixed regression model the sex variable was removed in backwards regression using the AIC and BIC method. Since sex is removed in this way, there may be a possibility that sex has a random effects component not found in the first model. We will now create a mixed regression model using the construction found in the statistical background section. We again will use the the logistic function to have a generalized mixed regression model.

In this model the factor that will separate groups for random effects is sex. When calculating the fit for this model the random effects coefficients are calculated to be zero with minimal variance. This implies that the random effects model only added a zero term to the model and did not improved model performance. This increases BIC as it incorporates the number of predictors in the model, however performance of the model does not improve. Therefore, the usage of a mixed effects model using sex to separate groups was not an effective solution to improve the performance of the model.

III.4 Piecewise logistic regression model

The last model that we will construct is a piecewise logistic regression model. This is another way other than a mixed regression model to view the differences between groups. This separation of the dataset into two groups allows backwards regression based on the BIC to be applied to each group separately. This could not only show what factors are significant for the outcome of a female fight and a male fight. This also improves the prediction power of the piecewise model compared to the model that is not separated based on sex.

$$\tilde{\pi_{\mathbf{m}}} = \frac{1}{1 + \exp(-\mathbf{X_m}\boldsymbol{\beta_m})}$$

$$\tilde{\pi_{\mathbf{f}}} = \frac{1}{1 + \exp(-\mathbf{X_f}\boldsymbol{\beta_f})}$$

Where

- X_m is the data for the male fights only.
- X_f is the data for the female fights only.
- β_m is the fitted predictors for the male fights.

- β_f is the fitted predictors for the female fights.
- $\tilde{\pi_m}$ is the predicted probabilities of the male fights.
- $\tilde{\pi_f}$ is the predicted probabilities of the female fights.

The following assumptions are made in the logistic regression analysis:

- The dependent variable Y is binary.
- There is a linear relationship between the logit of the outcome and the independent variables.
- Observations are independent.

After the stating of this model, before backwards regression can be performed, an initial model must be fit. This initial model fit is shown below for both the male and female groups. As the dataset is now split for each group the full initial model will not be the same for each in terms of fit and statistical significance.

After the full model is fit for each backwards regression based on BIC will be performed on each. The fit for each found in the appendix F.3 and F.4 respectively.

For each model we must complete residual analysis the same as the as we did with both fixed regression model that was not separated based on sex. This full residual analysis must be completed the same for each the male and female fighter model. Again for the sake of the reader the 100+ plots will not be placed in this paper for the sake of the reader. Each binned residual plot was analyzed.

The binned residual plots are evenly distributed around zero. This indicates that the model's predictions are unbiased and that the errors are randomly distributed. Most residuals fall within the confidence bands. This implies that there are not significant influential points that would indicate model misspecification or outliers. There are no clear trends or patters in the residuals. Such patterns could suggest issues such as nonlinearity of the log odds.The

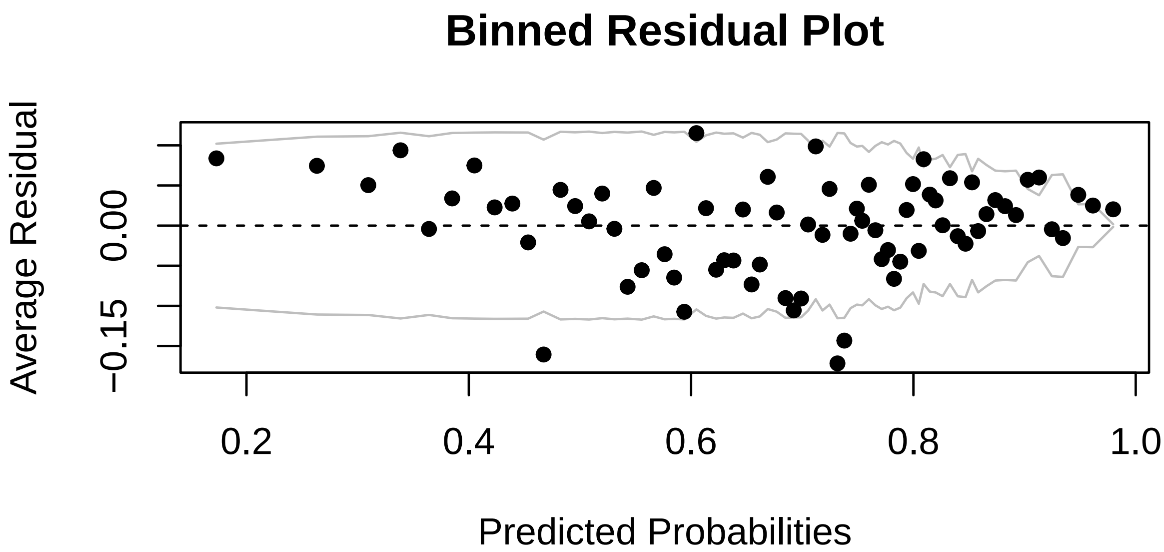

Figure III.5: Binned residual plot male

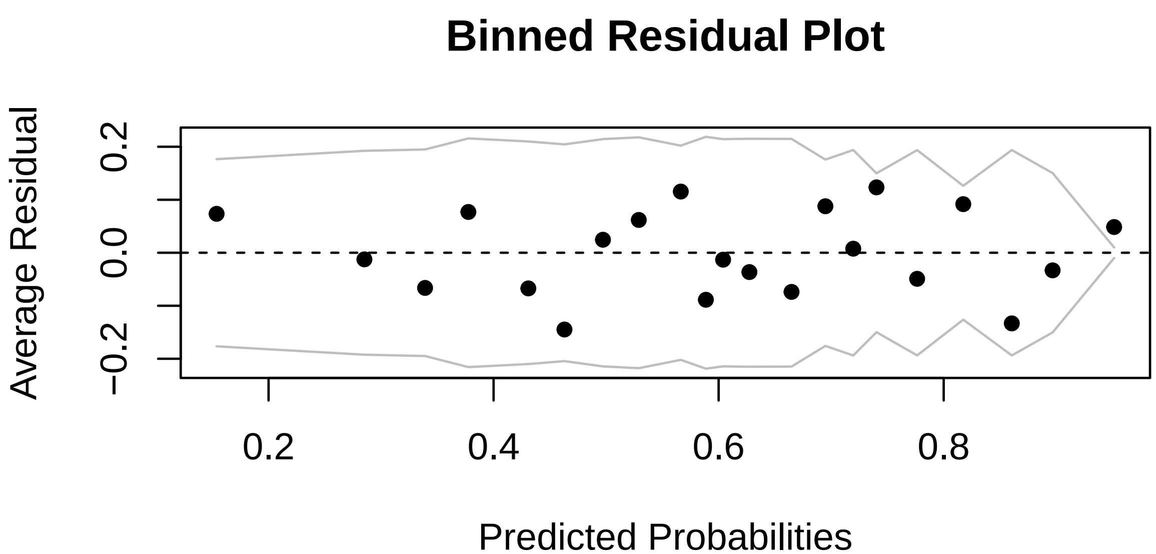

Figure III.6: Binned residual plot female

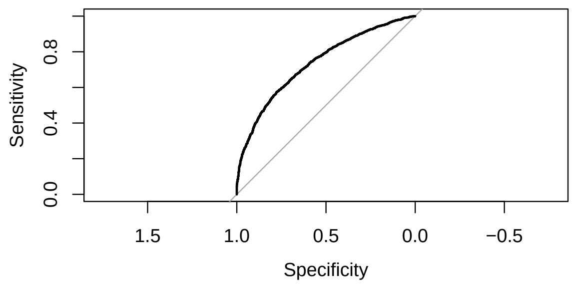

Figure III.7: Maximized Threshold value of .7296

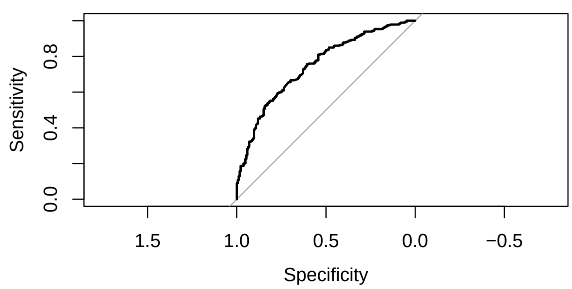

Figure III.8: Maximized Threshold value of .6

female model has less bias towards predicted probabilities above 0.5 indicating that the data set of female fights has a more balance winnings of blue and red corner fighters.

From the maximization of the ROC curve a value of 0.7296 was the threshold used to determine a win for red for the male model. From the maximization of the ROC curve a value of 0.6 was the threshold used to determine a win for red for the female model. These thresholds determined for the maximization on sensitivity and specificity for both models can be found in plots III. 7 and 8.

The next step is to validate each piece of the piecewise model separately. This will help to give a better understanding of its performance as a whole and compare to the fixed logistic regression model that did not separate based on sex. As stated in the fixed model's validation the values such as accuracy and F1 score hold meaning in comparison to other models rather than as stand alone values for evaluating performance. Another method of comparison that has been a key value in this project has been BIC. Again BIC has been used in as a measure in backwards regression and at each step removal of a variable based on BIC is used to compare each model. This is utilizing all of the data. In this model validation step we will again for each model split the data randomly into 80 percent and 20 percent of the data by randomly selecting 20 percent and removing it. This step is the same as for the fixed model and is done instead on both parts of the piecewise model.

	Reference	
Prediction	0	1
0	256	325
1	78	428

Table III.5: Confusion Matrix Male Model

Metric	**Value**
Accuracy	0.6293
95% CI	(0.5998, 0.658)
Sensitivity	0.7665
Specificity	0.5684
Pos Pred Value	0.4406
Neg Pred Value	0.8458
Precision	0.4406
Recall	0.7665
F1	0.5596
Balanced Accuracy	0.6674

Table III.6: Model Validation Male Model

	Reference	
Prediction	0	1
0	28	17
1	10	38

Table III.7: Confusion Matrix Female Model

The female model performs the best compared to the male model and the first model. It classifies wins for red and wins for blue correctly compared to the other models. This piecewise model as a whole performs better than the model not separated by sex. Both the male model and female model classify wins more accurately and in F1 score. The male model has and AUC value of 0.694 and the female model has and AUC value of 0.6553. The AUC

Metric	**Value**
Accuracy	0.7097
95% CI	(0.6064, 0.7992)
Sensitivity	0.7368
Specificity	0.6909
Pos Pred Value	0.6222
Neg Pred Value	0.7917
Precision	0.6222
Recall	0.7368
F1	0.6747
Balanced Accuracy	0.7139

Table III.8: Model Validation Female Model

value considers the model's performance across all possible thresholds. Therefore, a model might perform better at a particular threshold (improving the F1 score) but not as well across all thresholds (resulting in a lower AUC value). Therefore, through the maximization of a threshold for each model individually we found that we could improve F1 score for the piecewise model to have better performance, but across all thresholds the first model performs better. Using both AUC value and F1 score we can have a better perspective on the trade offs of choosing one model over another.

Conclusion

In this research we construct 3 different models. The backwards elimination model, the mixed regression model, and the piecewise regression model. The first model is constructed first by running backwards elimination based on BIC. Then constructing a model running backwards elimination based on AIC. From these two separate criterion we come to the same final model. We then construct the mixed regression model where the random effects coefficients were calculated to be 0 implying that we are only punished for adding random effects coefficients that do not impact the performance of the model. We then construct a piecewise model by splitting the data into male fights and female fights, then performing backwards regression based on BIC. Each of these models performed better than the first model that we developed. Specifically the female fighter model having an F1 score above .7 is a significant improvement compared to all other models in this research. Overall, these models provide a framework for understanding and predicting the outcome of fights. They offer insights into which variables are significant, and gender-specific differences in fights from the piecewise model. An area to apply these models would be in a coaching scenario to prepare an athlete. With fixed fighter information, different fight scenarios can be simulated to provide coaches with crucial information to support coaching choices. Exploring neural networks in future research may further improve prediction accuracy.

Bibliography

Apelgren, S., and Eklund, C. (2024). Predicting UFC matches using regression models. Degree Project in Technology, First cycle, 15 credits.

Berthet, V. (2023). FightTracker: Real-time predictive analytics for Mixed Martial Arts bouts

Jabbarov, I. (2024). Predicting Mixed Martial Arts fight outcomes using multilinear regression analysis. Degree Project in Applied Mathematics and Industrial Economics, First cycle, 15 credits.

Warrier, R. (2021). UFC-Fight historical data from 1993 to 2021. Kaggle. Retrieved from https://www.kaggle.com/datasets/rajeevw/ufcdata.

Walsh, G. (2022). Predictive Analysis of UFC Fights. National College of Ireland.

Appendix

Table F.1: Variable list names

Variable	Name
y	y
x2	B_avg_SIG_STR_pct
x3	B_avg_opp_SIG_STR_pct
x4	B_avg_TD_pct
x5	B_avg_opp_TD_pct
x6	B_avg_SIG_STR_att
x7	B_avg_SIG_STR_landed
x8	B_avg_opp_SIG_STR_att
x9	B_avg_opp_SIG_STR_landed
x10	B_avg_TOTAL_STR_att
x11	B_avg_TOTAL_STR_landed
x12	B_avg_opp_TOTAL_STR_att
x13	B_avg_opp_TOTAL_STR_landed
x14	B_avg_TD_att
x15	B_avg_TD_landed
x16	B_avg_opp_TD_att
x17	B_avg_opp_TD_landed
x18	B_avg_HEAD_att
x19	B_avg_HEAD_landed
x20	B_avg_opp_HEAD_att
x21	B_avg_opp_HEAD_landed
x22	B_avg_BODY_att
x23	B_avg_BODY_landed
x24	B_avg_opp_BODY_att
x25	B_avg_opp_BODY_landed
x26	B_avg_LEG_att
x27	B_avg_DISTANCE_att
x28	B_avg_DISTANCE_landed
x29	B_avg_opp_DISTANCE_att
x30	B_avg_opp_DISTANCE_landed
x31	B_avg_CLINCH_att
x32	B_avg_CLINCH_landed

x33	B_avg_opp_CLINCH_att
x34	B_avg_opp_CLINCH_landed
x35	B_avg_CTRL_timeseconds
x36	B_avg_opp_CTRL_timeseconds
x37	B_total_time_fought_seconds
x38	B_total_rounds_fought
x39	B_total_title_bouts
x40	B_current_win_streak
x41	B_current_lose_streak
x42	B_longest_win_streak
x43	B_wins
x44	B_losses
x45	B_win_by_Decision_Majority
x46	B_win_by_Decision_Split
x47	B_win_by_Decision_Unanimous
x48	B_win_by_KOTKO
x49	B_win_by_Submission
x50	B_win_by_TKO_Doctor_Stoppage
x51	B_Height_cms
x52	B_Reach_cms
x53	B_Weight_lbs
x54	R_avg_SIG_STR_pct
x55	R_avg_opp_SIG_STR_pct
x56	R_avg_TD_pct
x57	R_avg_opp_TD_pct
x58	R_avg_SIG_STR_att
x59	R_avg_SIG_STR_landed
x60	R_avg_opp_SIG_STR_att
x61	R_avg_opp_SIG_STR_landed
x62	R_avg_TOTAL_STR_att
x63	R_avg_TOTAL_STR_landed
x64	R_avg_opp_TOTAL_STR_att
x65	R_avg_opp_TOTAL_STR_landed
x66	R_avg_TD_att
x67	R_avg_TD_landed
x68	R_avg_opp_TD_att
x69	R_avg_opp_TD_landed
x70	R_avg_HEAD_att
x71	R_avg_HEAD_landed
x72	R_avg_opp_HEAD_att
x73	R_avg_opp_HEAD_landed
x74	R_avg_BODY_att
x75	R_avg_BODY_landed
x76	R_avg_opp_BODY_att
x77	R_avg_opp_BODY_landed

x78	R_avg_LEG_att
x79	R_avg_DISTANCE_att
x80	R_avg_DISTANCE_landed
x81	R_avg_opp_DISTANCE_att
x82	R_avg_opp_DISTANCE_landed
x83	R_avg_CLINCH_att
x84	R_avg_CLINCH_landed
x85	R_avg_opp_CLINCH_att
x86	R_avg_opp_CLINCH_landed
x87	R_avg_CTRL_timeseconds
x88	R_avg_opp_CTRL_timeseconds
x89	R_total_time_fought_seconds
x90	R_total_rounds_fought
x91	R_total_title_bouts
x92	R_current_win_streak
x93	R_current_lose_streak
x94	R_longest_win_streak
x95	R_wins
x96	R_losses
x97	R_win_by_Decision_Majority
x98	R_win_by_Decision_Split
x99	R_win_by_Decision_Unanimous
x100	R_win_by_KOTKO
x101	R_win_by_Submission
x102	R_win_by_TKO_Doctor_Stoppage
x103	R_Height_cms
x104	R_Reach_cms
x105	R_Weight_lbs
x106	R_age
x107	B_Stance_Open_Stance
x108	B_Stance_Orthodox
x109	B_Stance_Sideways
x110	B_Stance_Southpaw
x111	R_Stance_Open_Stance
x112	R_Stance_Orthodox
x113	R_Stance_Sideways
x114	R_Stance_Southpaw
x115	sex

Table F.2: Full model summary

Term	**Estimate**	**Std. Error**	**Statistic**	**P-value**
(Intercept)	5.63904	1.27198	4.43326	1e-05
x2	-0.29759	0.38639	-0.77018	0.44119
x3	0.16546	0.36734	0.45043	0.6524
x4	0.36801	0.17968	2.04811	0.04055
x5	0.78605	0.19292	4.07451	5e-05
x6	-0.02856	0.06068	-0.47062	0.63791
x7	0.07354	0.04608	1.59592	0.11051
x8	0.11881	0.03866	3.07319	0.00212
x9	-0.12597	0.04883	-2.57999	0.00988
x10	0.00415	0.01536	0.27015	0.78704
x11	0.00191	0.01723	0.1108	0.91177
x12	-0.05794	0.01715	-3.37903	0.00073
x13	0.06953	0.01939	3.58658	0.00034
x14	-0.06936	0.02395	-2.89595	0.00378
x15	-0.13055	0.06083	-2.14616	0.03186
x16	-0.01103	0.02278	-0.48407	0.62833
x17	-0.15563	0.05946	-2.6172	0.00887
x18	0.01002	0.05133	0.1952	0.84524
x19	-0.06659	0.03425	-1.94455	0.05183
x20	-0.0244	0.02868	-0.85085	0.39485
x21	0.02227	0.03581	0.62196	0.53397
x22	0.00309	0.05413	0.05713	0.95444
x23	-0.05703	0.04101	-1.39072	0.16431
x24	-0.06875	0.0357	-1.92558	0.05416
x25	0.07371	0.04401	1.67484	0.09397
x26	-0.03255	0.04193	-0.77627	0.43759
x27	0.01363	0.01592	0.85577	0.39213
x28	-0.0217	0.02401	-0.90371	0.36615
x29	-0.03131	0.01707	-1.83484	0.06653
x30	0.03516	0.02446	1.43723	0.15065
x31	0.01267	0.02572	0.49239	0.62244
x32	-0.01591	0.03612	-0.4404	0.65965
x33	-0.00558	0.02846	-0.19616	0.84448
x34	0.01027	0.03913	0.26239	0.79302
x35	0.00083	0.00062	1.32586	0.18488
x36	5e-05	0.00064	0.08399	0.93306
x37	-0.00065	0.00038	-1.71233	0.08683
x38	-0.01001	0.01871	-0.53485	0.59276
x39	0.16869	0.04599	3.66831	0.00024
x40	-0.00188	0.03571	-0.05271	0.95797
x41	-0.04578	0.06602	-0.69344	0.48804

Term	Estimate	Std. Error	Statistic	P-value
x42	-0.00924	0.04802	-0.19252	0.84733
x43	0.49916	0.52038	0.95921	0.33745
x44	0.05422	0.05648	0.95995	0.33708
x45	-0.13977	0.61514	-0.22722	0.82025
x46	-0.35829	0.5284	-0.67807	0.49773
x47	-0.5089	0.5265	-0.96658	0.33375
x48	-0.52705	0.52383	-1.00614	0.31435
x49	-0.52007	0.52398	-0.99253	0.32094
x50	-0.3093	0.54516	-0.56735	0.57047
x51	0.02198	0.00859	2.55742	0.01055
x52	-0.04842	0.00677	-7.15513	0
x53	0.00146	0.00271	0.53908	0.58983
x54	0.8846	0.36258	2.43973	0.0147
x55	-0.85209	0.35999	-2.367	0.01793
x56	0.37481	0.17849	2.09985	0.03574
x57	0.26164	0.18377	1.42369	0.15454
x58	0.13531	0.06982	1.93793	0.05263
x59	-0.11112	0.04319	-2.57271	0.01009
x60	0.06937	0.03423	2.02663	0.0427
x61	-0.08451	0.04425	-1.91009	0.05612
x62	-0.05155	0.01598	-3.22487	0.00126
x63	0.06038	0.01795	3.36444	0.00077
x64	-0.03923	0.01608	-2.43986	0.01469
x65	0.04998	0.01806	2.76715	0.00565
x66	0.04712	0.0221	2.13209	0.033
x67	-0.00362	0.05956	-0.06081	0.95151
x68	0.0484	0.0228	2.12236	0.03381
x69	-0.00424	0.05934	-0.07139	0.94309
x70	-0.04169	0.06126	-0.68054	0.49616
x71	0.00029	0.03009	0.00957	0.99237
x72	0.02725	0.02414	1.12914	0.25884
x73	-0.04318	0.03094	-1.39568	0.16281
x74	-0.08904	0.06317	-1.40951	0.15868
x75	0.04463	0.0382	1.16815	0.24275
x76	0.03395	0.03093	1.09782	0.27228
x77	-0.05078	0.03996	-1.27094	0.20375
x78	-0.03809	0.05523	-0.68975	0.49035
x79	-0.04461	0.01611	-2.76919	0.00562
x80	0.05841	0.02325	2.51193	0.01201
x81	-0.05717	0.01617	-3.53562	0.00041
x82	0.06672	0.02389	2.79231	0.00523
x83	-0.00191	0.02648	-0.07226	0.9424
x84	-0.00263	0.03684	-0.07142	0.94306
x85	-0.07788	0.02573	-3.02755	0.00247

Term	Estimate	Std. Error	Statistic	P-value
x86	0.08809	0.03641	2.41952	0.01554
x87	-0.00163	0.00061	-2.68696	0.00721
x88	-0.00298	0.00062	-4.83844	0
x89	0.00075	0.00035	2.11604	0.03434
x90	0.01788	0.01463	1.22173	0.22181
x91	0.03974	0.03003	1.32366	0.18561
x92	0.01232	0.0283	0.43526	0.66337
x93	-0.08321	0.06186	-1.3451	0.17859
x94	0.08057	0.03676	2.19157	0.02841
x95	-0.33924	0.35204	-0.96364	0.33523
x96	-0.05874	0.04611	-1.27406	0.20264
x97	0.24452	0.41005	0.59632	0.55096
x98	0.03281	0.36055	0.09099	0.9275
x99	0.27731	0.35793	0.77475	0.43849
x100	0.26475	0.35393	0.74805	0.45443
x101	0.31235	0.35585	0.87777	0.38007
x102	0.38179	0.36845	1.03621	0.3001
x103	-0.01398	0.00864	-1.61874	0.1055
x104	0.01572	0.00684	2.29818	0.02155
x105	0.00823	0.00278	2.9633	0.00304
x106	-0.07014	0.00908	-7.72581	0
x107	0.75455	1.09052	0.69191	0.48899
x108	0.1096	0.15357	0.71369	0.47542
x109	11.66031	265.91012	0.04385	0.96502
x110	-0.0471	0.16424	-0.2868	0.77426
x111	0.60253	0.69183	0.87092	0.3838
x112	0.02377	0.1765	0.13465	0.89289
x113	11.12711	376.98436	0.02952	0.97645
x114	0.2305	0.18753	1.22912	0.21903
x115	-0.17683	0.141	-1.25415	0.20979

Table F.3: Backwards regression summary

Term	Estimate	Std. Error	Statistic	P-value
(Intercept)	4.78613	1.12547	4.25256	2e-05
x4	0.33321	0.17731	1.87919	0.06022
x5	0.82049	0.17803	4.60876	0
x6	-0.00728	0.00328	-2.22288	0.02622
x7	0.05246	0.02341	2.24136	0.025
x8	0.09302	0.0216	4.30727	2e-05
x9	-0.09752	0.02689	-3.62653	0.00029
x10	0.00603	0.0023	2.618	0.00884
x12	-0.06049	0.01624	-3.72483	2e-04
x13	0.07196	0.01841	3.90978	9e-05
x14	-0.06786	0.02308	-2.93987	0.00328
x15	-0.12515	0.05998	-2.08648	0.03694
x17	-0.17571	0.04074	-4.31333	2e-05
x19	-0.06493	0.02456	-2.64423	0.00819
x23	-0.06318	0.02473	-2.55454	0.01063
x24	-0.04288	0.01734	-2.47241	0.01342
x25	0.05008	0.02356	2.1256	0.03354
x26	-0.04124	0.01907	-2.16262	0.03057
x29	-0.02804	0.01231	-2.27728	0.02277
x30	0.0287	0.01804	1.59079	0.11166
x35	0.00074	0.00052	1.42578	0.15393
x37	-0.00061	0.00029	-2.09784	0.03592
x39	0.15451	0.04084	3.78384	0.00015
x43	0.15336	0.05482	2.79747	0.00515
x47	-0.18358	0.06657	-2.75771	0.00582
x48	-0.18595	0.06382	-2.91388	0.00357
x49	-0.17481	0.06269	-2.78871	0.00529
x51	0.02336	0.00833	2.80435	0.00504
x52	-0.04699	0.00663	-7.08679	0
x54	0.89618	0.35725	2.50858	0.01212
x55	-0.86109	0.35595	-2.41912	0.01556
x56	0.35401	0.14875	2.37983	0.01732
x57	0.23319	0.15109	1.54342	0.12273
x58	0.08946	0.02025	4.41806	1e-05
x59	-0.10742	0.02548	-4.21659	2e-05
x60	0.09764	0.02372	4.11645	4e-05
x61	-0.12822	0.03099	-4.13825	3e-05
x62	-0.04794	0.01482	-3.23352	0.00122
x63	0.05645	0.01672	3.37535	0.00074
x64	-0.03825	0.01567	-2.44164	0.01462
x65	0.04843	0.01761	2.74935	0.00597

Term	Estimate	Std. Error	Statistic	P-value
x66	0.04544	0.01618	2.80808	0.00498
x68	0.04724	0.01736	2.72167	0.0065
x74	-0.04434	0.01504	-2.94726	0.00321
x75	0.0374	0.02175	1.71916	0.08559
x79	-0.04478	0.0123	-3.64119	0.00027
x80	0.06018	0.01752	3.43422	0.00059
x81	-0.05973	0.01599	-3.73536	0.00019
x82	0.07154	0.02367	3.02214	0.00251
x85	-0.08443	0.02461	-3.43048	6e-04
x86	0.09502	0.03495	2.71878	0.00655
x87	-0.00182	0.00053	-3.41587	0.00064
x88	-0.00314	0.00056	-5.62802	0
x89	0.00105	3e-04	3.52094	0.00043
x91	0.05588	0.02613	2.13855	0.03247
x93	-0.11716	0.04795	-2.44359	0.01454
x94	0.0871	0.0297	2.93293	0.00336
x95	-0.25094	0.04978	-5.0413	0
x99	0.23341	0.05764	4.0497	5e-05
x100	0.19352	0.05489	3.52582	0.00042
x101	0.24259	0.05333	4.54924	1e-05
x102	0.30992	0.12827	2.41627	0.01568
x103	-0.01278	0.00853	-1.49745	0.13428
x104	0.01663	0.00669	2.48426	0.01298
x105	0.00909	0.00178	5.12246	0
x106	-0.07105	0.00873	-8.13964	0
x108	0.14214	0.07048	2.01674	0.04372
x114	0.20779	0.07715	2.69342	0.00707

Table F.4: Male model summary

Term	Estimate	Std. Error	Statistic	P-value
(Intercept)	5.65578	1.24404	4.5463	1e-05
x4	0.38895	0.18837	2.06484	0.03894
x5	0.79024	0.18897	4.18183	3e-05
x6	-0.01435	0.00447	-3.20938	0.00133
x8	0.10355	0.02309	4.48522	1e-05
x9	-0.10608	0.02891	-3.66944	0.00024
x11	0.00722	0.00239	3.0199	0.00253
x12	-0.06371	0.01696	-3.75681	0.00017
x13	0.07657	0.0193	3.96702	7e-05
x14	-0.07271	0.02358	-3.08401	0.00204
x15	-0.12086	0.05993	-2.01678	0.04372
x17	-0.16649	0.04852	-3.43126	6e-04
x18	0.0131	0.0055	2.38036	0.0173
x19	-0.01984	0.00677	-2.93071	0.00338
x24	-0.06588	0.01866	-3.52986	0.00042
x25	0.07191	0.02522	2.85154	0.00435
x29	-0.03728	0.01335	-2.79161	0.00524
x30	0.03422	0.01962	1.74404	0.08115
x36	-0.00107	0.00054	-1.98464	0.04718
x39	0.1715	0.04353	3.93969	8e-05
x42	-0.05599	0.03827	-1.46324	0.1434
x43	0.19428	0.0593	3.27602	0.00105
x47	-0.22054	0.06996	-3.15257	0.00162
x48	-0.20127	0.06637	-3.03233	0.00243
x49	-0.19444	0.06526	-2.97925	0.00289
x51	0.02906	0.00875	3.31945	9e-04
x52	-0.05076	0.00697	-7.28011	0
x54	0.80434	0.37532	2.14305	0.03211
x55	-0.7459	0.37021	-2.01482	0.04392
x56	0.40919	0.15917	2.57073	0.01015
x57	0.23518	0.16124	1.45855	0.14469
x58	0.09431	0.02179	4.32769	2e-05
x59	-0.11549	0.02741	-4.21426	3e-05
x60	0.11617	0.02621	4.43261	1e-05
x61	-0.15079	0.03402	-4.43301	1e-05
x62	-0.04894	0.01613	-3.03347	0.00242
x63	0.0577	0.01821	3.1682	0.00153
x64	-0.04662	0.01723	-2.70579	0.00681
x65	0.05822	0.01933	3.01158	0.0026
x66	0.04636	0.01719	2.69704	0.007
x68	0.03979	0.01848	2.15274	0.03134

Term	Estimate	Std. Error	Statistic	P-value
x74	-0.04736	0.01597	-2.96568	0.00302
x75	0.04486	0.02312	1.94038	0.05233
x79	-0.05107	0.01315	-3.88337	1e-04
x80	0.07056	0.01881	3.75171	0.00018
x81	-0.06953	0.0177	-3.9271	9e-05
x82	0.084	0.02597	3.23421	0.00122
x85	-0.09242	0.02722	-3.39497	0.00069
x86	0.10027	0.03855	2.60098	0.0093
x87	-0.00195	0.00058	-3.39087	7e-04
x88	-0.00323	6e-04	-5.34227	0
x89	0.00106	0.00032	3.29259	0.00099
x90	0.02554	0.01316	1.94038	0.05233
x93	-0.09753	0.05603	-1.74067	0.08174
x94	0.07667	0.03428	2.2369	0.02529
x95	-0.29946	0.06839	-4.37875	1e-05
x96	-0.09371	0.04234	-2.21323	0.02688
x99	0.22488	0.06049	3.71753	2e-04
x100	0.23395	0.05883	3.97679	7e-05
x101	0.27893	0.05774	4.83106	0
x102	0.38121	0.12976	2.93782	0.00331
x103	-0.01803	0.00893	-2.01772	0.04362
x104	0.01661	0.00704	2.35907	0.01832
x105	0.00905	0.00184	4.92027	0
x106	-0.07593	0.0095	-7.98999	0
x108	0.16539	0.07341	2.25282	0.02427
x114	0.19862	0.08012	2.47903	0.01317

Table F.5: Female model summary

Term	Estimate	Std. Error	Statistic	P-value
(Intercept)	0.38735	2.21959	0.17451	0.86146
x3	-2.59703	1.78707	-1.45324	0.14616
x4	-1.05187	0.54296	-1.93729	0.05271
x5	1.79638	0.66673	2.69434	0.00705
x6	0.58804	0.15621	3.76443	0.00017
x12	-0.10257	0.03707	-2.76698	0.00566
x13	0.10889	0.04325	2.51778	0.01181
x17	-0.62903	0.22844	-2.75361	0.00589
x18	-0.59988	0.15605	-3.84401	0.00012
x21	0.04301	0.02093	2.05457	0.03992
x22	-0.59945	0.15513	-3.8641	0.00011
x25	0.0888	0.03863	2.29893	0.02151
x26	-0.56125	0.1537	-3.65156	0.00026
x29	0.11514	0.03734	3.08349	0.00205
x30	-0.13399	0.04895	-2.73712	0.0062
x31	0.07987	0.05406	1.47752	0.13954
x32	-0.10707	0.06795	-1.57582	0.11507
x33	0.16048	0.08279	1.93844	0.05257
x34	-0.24377	0.1063	-2.29329	0.02183
x35	0.0058	0.00192	3.02148	0.00252
x36	0.00841	0.0023	3.66236	0.00025
x37	-0.00363	0.00161	-2.25836	0.02392
x38	0.07639	0.03866	1.97593	0.04816
x39	-0.32508	0.18947	-1.71568	0.08622
x42	0.8877	0.24084	3.68585	0.00023
x43	-0.64781	0.23079	-2.80697	0.005
x45	-16.06093	535.41158	-0.03	0.97607
x53	-0.05971	0.02018	-2.9587	0.00309
x54	2.5381	1.7264	1.47017	0.14152
x58	0.09493	0.04458	2.12948	0.03321
x59	-0.11532	0.05264	-2.19086	0.02846
x60	-0.15688	0.07016	-2.2362	0.02534
x61	0.26069	0.09214	2.8292	0.00467
x62	-0.08298	0.04322	-1.91985	0.05488
x63	0.09662	0.04926	1.96152	0.04982
x64	0.08643	0.05076	1.70277	0.08861
x65	-0.09281	0.05709	-1.62566	0.10402
x68	0.1692	0.06674	2.5352	0.01124
x73	-0.06735	0.02405	-2.79993	0.00511
x77	-0.07853	0.03958	-1.9842	0.04723
x81	0.07901	0.03697	2.13693	0.0326

Term	**Estimate**	**Std. Error**	**Statistic**	**P-value**
x82	-0.14618	0.05545	-2.63614	0.00839
x88	-0.00462	0.00153	-3.0203	0.00253
x91	0.36286	0.13462	2.69548	0.00703
x98	-0.45646	0.22046	-2.07054	0.0384
x105	0.05473	0.01927	2.83935	0.00452