Predicting UFC Fight Results

Max Brown and Jackson Seo

2023-12-11

Introduction

The UFC began as a professional mixed martial arts organization in 1993 serving as an alternative hand-to-hand combat sport that combined traditional boxing with wrestling, karate, kickboxing, and jiu-jistu fighting techniques. The entity was acquired by a group led by Dana White in 2001, who has served as the President for over two decades. Since establishing control, Dana White has exponentially grown the reach of UFC's product while also creating more structure and sanction to the sport of MMA. UFC currently has over 60 global broadcasting partners and is able to be accessed in over 165 different countries. With a traveling, tour-like model, the UFC has been able to sell-out many arenas across the world as equally become a highly-touted event to attend similar to boxing matches with well-known participants involved.

In many other North-american based sports, organizations have invested and founded their own analytics departments. These departments are responsible for using data to acquire and develop the right talent that will lead to on-field success and improve the team's product. Since the UFC's participants are individual fighters that often follow their own training regiment, there is a smaller focus on analytics within the sport.

The purpose of our project is to create a model that maximizes predictive accuracy for the purposes of assisting sports bettors in finding potential opportunities for value not seen by the public. Dana White and the UFC have fully embraced the recent popularity of sports betting, forming sponsorships with companies such as bet365 in the UK, DraftKings in the US, as well as many others located worldwide. There is an established market for sports betting in the UFC, and we hope to create a model that provides an estimation of a winner between two fighters along with some form of uncertainty that allows a sports-better to determine if the predicted odds over or under-estimate a fighter's chance of winning compared to the sportsbook odds given to the public.

Our data is each UFC fight from 1994-2021, containing each fighter's names, physical information (age, height, weight), the amount of wins and losses in their UFC history, and various fighting data. The fighting data includes the average amount of attempted and landed attacks over their UFC career, as well as the frequency of different types of attacks they have faced from their previous opponents. We plan to fit an elastic net model that incorporates the standardization of a ridge regression model, and the variable selection of a lasso regression model, to determine which predictors in our data set are most influential. Our response variable will be the winner in each fight, with that value randomized dependent on the color of the corner assigned during the fight (red or blue). The estimated probability values from the elastic net model can be compared against the moneyline odds to determine estimated value for sports bettors.

Data Description and EDA

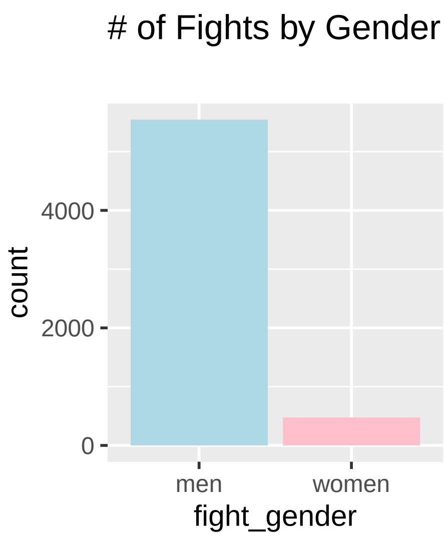
The data set used for modeling was initially sourced from ufcstats.com, where the data was processed and published on kaggle.com. The data set contains 6,012 unique fights over a 27 year span, with each row containing data on fighters in the red and blue corner, a universal classification system used throughout all rows. Since a given fighter could be in the red corner for one fight and in the blue corner for another, we don't place any value in the classification. The historical data for a given strike is split into four different columns. There is a separate column for the average amount of attempt and the average amount of strikes

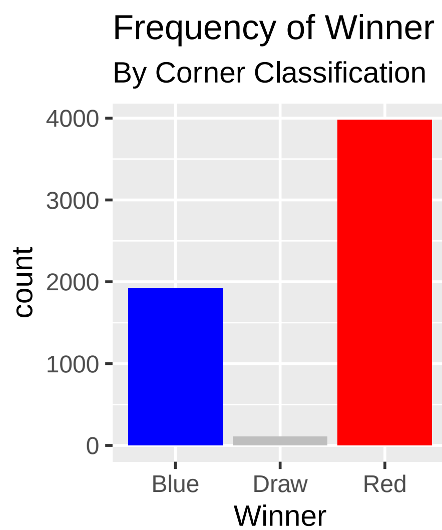
Table 1: Instances of a Fighter's 1st Fight

Var1	Freq
no	4316
yes	1696

landed for a fighter's prior fight. Additionally, for that given strike there includes identical metrics for the averages of their prior opponents. For example, if Conor McGregor had one prior fight in which his opponent attempted 20 strikes to Conor's head and connected on 8, the row for Conor's second fight would display avg_opp_HEAD_att equal to 20 and avg_opp_HEAD_landed to 8.

In the UFC, fighters can win in a variety of different ways. If both fighters are still standing at the end of the fighter, the judges will issue a decision that is either unanimous, split, or a majority decision. The data set includes the amount of victories by each form of decision, as well as by knockout or by the on-hand doctor stopping the fight. Our data set also provides the current winning or losing streak for each fighter, which can be useful as it can reflect the momentum and confidence a fighter may possess.





The plots above provide a better idea on the breakdown of the amount of male and female fights in UFC's history, as well as the results based on which corner was victorious. Similar to men's and women's lacrosse or men's and women's soccer, we believe that men's and women's UFC fights should be considered different sports given the difference in fighting style. Men's MMA is centered around wrestling, while women's MMA is centered around jiu-jitsu and judo with a strong preference for striking than grappling seen more commonly on the men's side. Since roughly 92% of the data are men's UFC fights, we will remove the observations in which the two fighters are female. The right plot above reveals that about 2/3 of fights with a winner are assigned to the red corner. Since the corners are simply a classifier and we are concerned that our models will be biased towards the red corner, we will randomize the fighters within each fight and reassign their corresponding statistics if necessary.

Another limitation within our data is that for a fighter's first career UFC fight, they have no historical data and thus their respective columns are N/A in our data set. Since our modeling techniques require clean data without missing data, we will need to remove instances of a fighter's first fight.

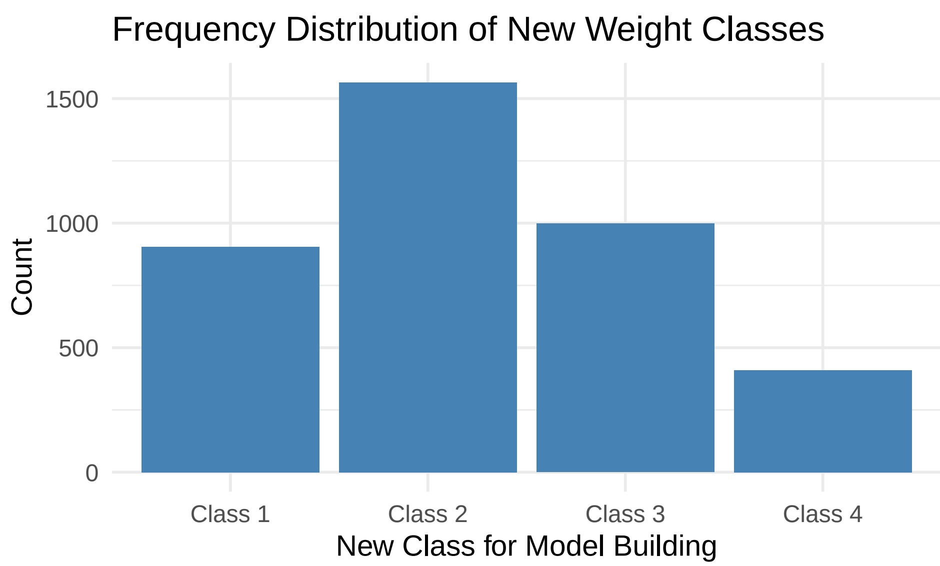
After shuffling the data, we see a much more even distribution of fight winners between the red and blue corner. Similar to comparing the difference between men's and women's MMA, we believe the fighting styles begin to differ as fighters increase in weight class. Therefore, rather than creating nine different models for each class, we will group weight classes together as seen below:

Table 2: Frequency of Winner by Corner

Var1	Freq
Blue	1970
Red	1908

Table 3: Frequency of Weight Class Fights

new_class	weight_class	max_weight	count
Class 1	Flyweight	125	173
Class 1	Bantamweight	135	331
Class 1	Featherweight	145	401
Class 2	Lightweight	155	782
Class 2	Welterweight	170	783
Class 3	Middleweight	185	582
Class 3	LightHeavyweight	205	416
Class 4	Heavyweight	265	381
Class 4	OpenWeight	300	29



The plot above displays the new distribution of observations by the new weight classes. While we were unable to create a completely even split of observations, we believe we have enough fights within each new segmentation to proceed in fitting four models. The table below shows the average attempts of different types of strikes for one classification of the fighters. We see that numbers tend to decrease for standing strikes such as the head, body, and leg as the weight class increases. This could be because heavier fighters prefer to spend more of the fight wrestling on the floor, or heavier fights typically lasting less time, with such cases not mutually exclusive.

Variable Transformations

Before constructing our four models, we first want to remove variables within our data set that are either redundant or will not provide useful information in predicting a winner. We will remove many of the multi-class variables, such as the red and blue fighter's names and the referee's name, that would significantly increase the complexity of our models. We will also remove the amount of draws a previous fighter has in

Table 4: Avg Attempt of Different Style Attacks

new_class	Rmean_head	Rmean_body	Rmean_leg	Rmean_clinch	Rmean_ground	Rmean_ctrltime
Class 1	73.7	12.0	8.0	7.5	8.5	143.6
Class 2	62.8	10.0	6.9	7.8	8.3	150.5
Class 3	51.9	7.8	6.0	8.0	9.1	138.7
Class 4	43.5	6.3	4.8	6.6	8.9	105.7

their career as all rows equate to zero for both the red and blue fighters. Elastic net model creation requires that the data is clean and does not contain missing values. Although we removed many of the initial missing values for when it was a fighter's UFC debut, there are other cases in our data set in which a fighter's age or reach is not recorded. We considered imputing the data using averages, but ultimately decided to remove these rows altogether to avoid potential biases that may arise from imputation. We also believed our sample size was large enough to construct our models.

To create the four separate models for the weight class segmentation, we have to create four separate data frames from which we can create training and test splits. Additionally, since the structure of our data set includes the number of attempts and successes for a given strike, as well as current win streak and overall wins on a fighter's record, we suspect that many potential predictors will be extremely multicollinear. To combat this, we plan to use ridge regression to introduce bias that can lower the variance of the estimates. However, since our models contain several predictors and ridge regression does not perform variable selection, we want to use lasso regression as well to assign estimates of irrelevant predictors to zero. The combination of both methods is called elastic net regression a regression model that includes both the L1 penalty of Lasso and the L2 penalty of Ridge regression. Elastic Net combines both L2 and L1 penalties of ridge regression and lasso. It controls the mixing of the two penalties through a parameter (lambda).

In looking at the correlation coefficients between predictors for each class, we notice that coefficients for the same type of strike exceed 0.95 (such has avg strikes attempted and landed to the opponent's head). Correlation coefficients are also extremely high for significant strike predictors and other forms of strikes, such as head, body, and leg. Therefore, once we fit each of the models, we will also construct likelihood ratio tests to determine if the inclusion of interactions among the selected variables help improve the model.

Since we plan to use elastic net that contain, we standardized the numerical predictors in our data. In doing so, we first created our training and test splits. From there, we scaled each split using the mean and standard deviation of the training data so there would not be any information leakage into our test set as we need to consider it as establishing new data. Elastic net also does not allow for multi-class categorical variables, so we need to transform the categorical variables (including our response variable) into numerical format so they can be used. We can do this by using dummy variables. We will not include a dummy variable indicating a Winner for the blue corner so our models are only in the context of Red winning or losing (1 = Red win, 0 = Red lose). Using a 70/30 training and test split, we can use cross validation for each alpha value to find the optimal lambda. Alpha is the mixing parameter between lasso (alpha = 1) and ridge (alpha = 0) regression. We will find the combination that gives the best performance (e.g., the lowest deviance), fit the elastic net model using the optimal alpha and lambda values, and finally evaluate the model on test set to check its performance.

Accuracy: proportion of the total number of predictions that were correct. Precision: ratio of correctly predicted positive observations to the total predicted positives. Recall (Sensitivity): ratio of correctly predicted positive observations to all observations in the actual class. F1-Score: weighted average of Precision and Recall.

Class 1 Model (Flyweight, Bantamweight, Featherweight)

Methodology

For class 1 our optimal alpha value is 0.55, which results in the variable selection of 23 predictors. Given the high coefficient values we found earlier in our methodology, we attempted to include three interaction terms. In a new model with the selected variables, we added the interaction of the average amount of reversal strikes landed for the red fighter and the average amount of significant strikes landed by the red fighter, the interaction of the average amount of body strikes landed and the average amount of significant strikes landed by the Red Fighter, and the average amount of take downs and the average amount of significant strikes previous opponents have landed on the Red fighter. Additionally, we also included a quadratic effect for the average amount of significant strikes landed under the assumption that as a fighter delivers more damage to their opponent the probability of winning would increase exponentially. However, we failed to reject the null hypothesis in favor for the base model when each of the added terms were included. Given our elastic net model for class 1 received a 0.55 alpha value, we hypothesized that some of our predictors would still be multicollinear since less of a ridge penalty was used. However, in running a VIF test on our chosen predictors we did not spot any that would signify high correlation with each other which would justify either removing the predictor or testing further interactions.

Since the use of interactions and the VIF did not change our model, we decided to look at the statistically significant predictors and evaluate the model's performance with just these predictors chosen. We discovered that our model performs better with less predictors, and given that our goal is to maximize predictive accuracy, we proceeded with it as our best model. We then compared this p-value variable selection with best subset selection, both stemmed from the elastic net model. With best subset, using five of the 23 predictors minimized the BIC. However, the nine statistically significant predictors from the elastic net model, satisfying a 0.1 significance level, performed better than best subset with five predictors chosen.

$$log(P(Winner = Red)/(1 - P(Winner = Red))) = B_0 + B_1 * B\_avg\_opp\_SIG\_STR\_pct+$$
$$B_2 * B\_win\_by\_Decision\_Unanimous + B_3 * R\_avg\_opp\_SIG\_STR\_pct + B_4 * R\_avg\_BODY\_landed+$$
$$B_5 * R\_avg\_opp\_CLINCH\_att + B_6 * R\_current\_win\_streak+$$
$$B_7 * B\_age + B_8 * R\_age + B_9 * B\_StanceSwitch$$

In summary, for the class 1 model we used elastic net to select 23 of the existing 140 variables. We then used likelihood ratio tests to determine that none of our hypothesized interactions or inclusion of quadratic effects were significant enough to add to the model. From there, we also concluded through a VIF test that none of the selected predictors contained high levels of multicollinearity that would justify the variable to be removed. Finally, we compared selecting the nine statistically significant predictors against the best subset selection and concluded that the nine statistically significant predictors better predicted our testing data, choosing that model as our final selection. The roc curve and AUC value of 0.59 shown in the appendix signify that our model could be improved, but we will proceed to our other models to determine if the variability in UFC fight winners may place an upper bound in our model's performance.

Inference

Looking at the final model above, we notice that significant strike percentage was an important predictor for a red fighter and their facing opponent. Holding all else constant, a standard deviation increase in the red fighter's average faced significant strike percentage from previous opponents would decrease the red fighter's log odds of winning by 0.327. Moreover, a standard deviation increase in the Blue Fighter's average faced significant strike percentage from previous opponents would increase the red fighter's log odds of winning by 0.372. Momentum is also an important factor for the red fighter, as a standard deviation in the red fighter's current win streak increases their log odds of winning by 0.295. One major limitation within this particular model is that we did not explore all potential nonlinear effects, meaning the true relationship between one of

Table 5: Coefficient Estimates for Class 1 Model

	x
(Intercept)	-0.029
B_avg_opp_SIG_STR_pct	0.217
B_win_by_Decision_Unanimous	-0.419
R_avg_opp_SIG_STR_pct	-0.327
R_avg_BODY_landed	0.372
R_avg_opp_CLINCH_att	-0.214
R_current_win_streak	0.295
B_age	0.383
R_age	-0.322
B_StanceSwitch	-0.678

Table 6: Class 1 Elastic Net –> Sig. Predictors Model Accuracy

	Values
Accuracy	0.581
Precision	0.558
Recall	0.549
F1	0.554

these predictors and the response variable may not truly be positive. Additionally, we notice that the switch stance classifier has the highest coefficient, but that may be due to the lack of observations with that stance in our training and test data, as switch stances are much less common than orthodox or southpaw. Most importantly, our model has a predictive accuracy of only 58.1% on the test data, meaning when sports bettors input data into our model they should consider the predicted probability to determine how much confidence they should have in their betting decision.

Class 2 Model (Lightweight, Welterweight)

Methodology

For class 2, we calculated an optimal alpha value of 0.65 which resulted in the variable selection of 29 predictors. We then used likelihood ratio tests to explore the inclusion of interactions and quadratic effects and discovered that including the quadratic effect for R_avg_HEAD_landed and R_avg_opp_CLINCH_att would be significant (at an alpha = 0.05 level) for our model. We then looked at the VIF values of our model and determined that none needed to be removed due to the potential presence of significant multicollinearity. Similar to class 1, we chose the 14 predictors that were statistically significant within the elastic net model to test if such reduction would improve the model's predictive accuracy. Unlike the class 1 model, such reduction decreased our model's predictive accuracy. Finally, we used best subset on the elastic net model in which the BIC value was minimized with 8 predictors. However, the full elastic net model still had the best predictive accuracy, so we will proceed with it as our final model.

$$log(P(Winner = Red)/(1 - P(Winner = Red))) = B_0 + B_1 * B\_avg\_opp\_SIG\_STR\_pct +$$
$$B_2 * B\_avg\_opp\_SUB\_ATT + B_3 * B\_avg\_REV + B_4 * B\_avg\_TD\_att + B_5 * B\_avg\_TD\_landed +$$
$$B_6 * B\_avg\_CLINCH\_landed + B_7 * B\_avg\_CTRL\_time.seconds. + B_8 * B\_total\_time\_fought.seconds. +$$
$$B_9 * B\_total\_title\_bouts + B_{10} * B\_current\_lose\_streak + B_{11} * B\_longest\_win\_streak +$$
$$B_{12} * B\_win\_by\_Submission + B_{13} * B\_win\_by\_TKO\_Doctor\_Stoppage + B_{14} * R\_avg\_SIG\_STR\_pct +$$

$$+B_{15} * R\_avg\_SUB\_ATT + B_{16} * R\_avg\_HEAD\_landed + B_{17} * R\_avg\_HEAD\_landed^2 +$$
$$B_{18} * R\_avg\_opp\_CLINCH\_att + B_{19} * R\_avg\_opp\_CLINCH\_att^2 + B_{20} * R\_avg\_GROUND\_att +$$
$$B_{21} * R\_avg\_opp\_CTRL\_time.seconds. + B_{22} * R\_current\_win\_streak + B_{23} * R\_longest\_win\_streak$$
$$B_{24} * R\_win\_by\_Decision\_Split + B_{25} * R\_win\_by\_Decision\_Unanimous + B_{26} * R\_win\_by\_Submission$$
$$B_{27} * B\_age + B_{28} * R\_age + B_{29} * B\_StanceOrthodox + B_{30} * B\_StanceSouthpaw$$

The model above results in a 0.64 AUC value, which shows improvement compared to class 1, yet has room for improvement. We will continue to monitor how the AUC values change based on the class used to construct the model.

Inference

Table 7: Coefficients for Class 2 Most Important Predictors

	x		x
poly(R_avg_opp_CLINCH_att, 2)2	5.067	poly(R_avg_HEAD_landed, 2)2	-11.406
poly(R_avg_HEAD_landed, 2)1	3.088	poly(R_avg_opp_CLINCH_att, 2)1	-6.561
B_age	0.442	R_win_by_Decision_Split	-0.306
B_avg_opp_SIG_STR_pct	0.166	B_StanceSouthpaw	-0.299
B_avg_REV	0.164	R_age	-0.261
B_win_by_TKO_Doctor_Stoppage	0.147	B_win_by_Submission	-0.188
R_avg_SUB_ATT	0.135	B_total_title_bouts	-0.172
R_longest_win_streak	0.132	R_avg_opp_CTRL_time.seconds.	-0.169

Table 8: Class 2 Elastic Net Model Accuracy

	Values
Accuracy	0.601
Precision	0.609
Recall	0.601
F1	0.605

Above are the most positive and negative estimates towards a class 2 red fighter's log odds of victory. The most notable observation among the coefficients is the concave-down relationship between the average amount of red fighter's head strikes landed on their opponent. This means that holding all else constant and given our training data, the log odds the red fighter wins decreases as the amount of head strikes landed reaches a large amount. This shows signs towards an added importance of submissions for this class, which can be supported by a positive coefficient for the average amount of submission attempts for a red fighter. One limitation within our model selection process is that we did not explore a higher degree than two for potential nonlinear effects, meaning that for R_avg_HEAD_landed the true relationship may not be quadratic. Nonetheless, the model provides an accuracy measure of 0.601 on our testing data which was the strongest of any model selection technique we explored for this class.

Class 3 Model (Middleweight, LightHeavyweight)

Methodology

Our optimal alpha value of 0.85 resulted in the variable selection of 46 predictors. Using contextual knowledge, we again considered the inclusion of various interactions and discovered that the interaction of B_avg_GROUND_att and B_avg_CTRL_time.seconds., and R_avg_BODY_landed and

R_avg_SIG_STR_pct improved our model through the likelihood ratio test. With these interactions included, we then took the 22 terms with a p-value less than 0.1 and compared the model accuracy measures. The full elastic net model provided better accuracy results, so we finally compared the model against the BIC backwards selection subset model as best subset was not computationally possible starting with a high number of predictors. While the BIC backwards selection model resulted in only six significant predictors, the full elastic net model was still more accurate on the test data. Given our biggest priority is predictive accuracy, the full elastic model is our final selection.

$$log(P(Winner = Red)/(1 - P(Winner = Red))) = B_0 + B_1 * B\_avg\_opp\_SIG\_STR\_pct +$$
$$B_2 * B\_avg\_opp\_TD\_pct + B_3 * B\_avg\_REV + B_4 * B\_avg\_opp\_REV + B_5 * B\_avg\_SIG\_STR\_landed$$

$$B_6 * B\_avg\_opp\_TOTAL\_STR\_landed + B_7 * B\_avg\_TD\_att + B_8 * B\_avg\_opp\_TD\_att +$$
$$B_9 * B\_avg\_LEG\_landed + B_{10} * B\_avg\_opp\_LEG\_att + B_{11} * B\_avg\_opp\_CLINCH\_landed +$$
$$B_{12} * B\_avg\_GROUND\_att + B_{13} * B\_avg\_opp\_GROUND\_landed + B_{14} * B\_avg\_opp\_GROUND\_landed +$$
$$B_{15} * B\_avg\_CTRL\_time.seconds. + B_{16} * B\_current\_win\_streak +$$
$$B_{17} * B\_current\_lose\_streak + B_{18} * B\_longest\_win\_streak + B_{19} * B\_win\_by\_Decision\_Split +$$
$$B_{20} * B\_win\_by\_Decision\_Unanimous + B_{21} * B\_win\_by\_TKO\_Doctor\_Stoppage +$$
$$B_{22} * B\_Height\_cms + B_{23} * B\_Reach\_cms + B_{24} * B\_Weight\_lbs + B_{25} * R\_avg\_KD +$$
$$B_{26} * R\_avg\_SIG\_STR\_pct + B_{27} * R\_avg\_opp\_TD\_pct + B_{28} * R\_avg\_opp\_SUB\_ATT +$$
$$B_{29} * R\_avg\_REV + B_{30} * R\_avg\_opp\_TOTAL\_STR\_landed + B_{31} * R\_avg\_TD\_landed +$$
$$B_{32} * R\_avg\_opp\_TD\_att + B_{33} * R\_avg\_opp\_HEAD\_att + B_{34} * R\_avg\_BODY\_landed +$$
$$B_{35} * R\_avg\_opp\_LEG\_att + B_{36} * R\_longest\_win\_streak + B_{37} * R\_losses +$$
$$B_{38} * R\_win\_by\_Decision\_Split + B_{39} * R\_win\_by\_KO.TKO + B_{40} * R\_Height\_cms +$$
$$B_{41} * R\_Reach\_cms + B_{42} * R\_age + B_{43} * B\_StanceOpenStance +$$
$$B_{44} * B\_StanceOrthodox + B_{45} * R\_StanceSouthpaw + B_{46} * R\_StanceSwitch +$$
$$B_{47} * B\_avg\_GROUND\_att * B\_avg\_CTRL\_time.seconds. +$$
$$B_{48} * R\_avg\_BODY\_landed * R\_avg\_SIG\_STR\_pct$$

The model above provides an AUC value of 0.59, which is approximately similar to classes 1 and 2. It is becoming more conclusive that we are unable to reach a significantly high value due to the unpredictability of fight winners regardless of the classification threshold we set.

Inference

In evaluating the class 3 model coefficients, we noticed that younger red fighters have a greater chance of winning, and the log odds increase by 0.372 for every standard deviation increase in the blue fighter's height (cms), holding all else constant. We also spotted that take down (TD) statistics are prevalent among the most important predictors, both positively and negatively. The average amount of take downs landed increases the red fighter's chances, but so does the average amount of take down attempts they have had to defend. Overall, with our 0.5 predicted probability threshold, our model predicted the test data correctly at a 56% rate. The biggest limitation within this model is that the amount of predictors suggests a possibility of overfitting, meaning there may be a simpler model we did not consider that would have greater predictive accuracy.

Table 9: Most Important Predictors for Class 3

	x		x
B_avg_opp_TOTAL_STR_landed	0.511	R_StanceSwitch	-1.459
B_Height_cms	0.372	B_StanceOpenStance	-1.043
R_avg_TD_landed	0.253	B_avg_opp_TD_att	-0.443
R_avg_opp_TD_att	0.243	B_avg_TD_att	-0.405
B_win_by_Decision_Split	0.241	R_age	-0.400
R_avg_KD	0.211	B_Weight_lbs	-0.363
B_avg_opp_TD_pct	0.202	R_avg_SIG_STR_pct:R_avg_BODY_landed	-0.316
R_win_by_KO.TKO	0.198	B_avg_opp_LEG_att	-0.242

Table 10: Class 3 Elastic Net Model Accuracy

	Values
Accuracy	0.564
Precision	0.578
Recall	0.599
F1	0.588

Class 4 Model (Heavyweight, Openweight)

Methodology

For class 4, the optimal alpha value was 0.1, resulting in a variable selection with 19 predictors. Using likelihood ratio tests, we found strong significance in including the quadratic effect for B_avg_opp_SIG_STR_pct, or how well Red's current opponent has defended against significant strikes, and the interaction between B_avg_REV and B_avg_opp_REV, which are the average amount of reversals Red's current opponent, Blue, has landed and received from their prior opponents. In evaluating the vif values of this model, we noticed that R_avg_opp_LEG_att and R_avg_opp_LEG_landed each had a generalized value greater than 5, which we deemed was large enough to remove from the model. From there, we selected the seven predictors that were statistically significant and compared the accuracy measures to the full elastic net model. The full elastic net model with the interactions and the quadratic effect included performed better on the test data. Finally, we used best subset and created a new model with the three predictors chosen from the elastic net model. Once again, the elastic net model performed better in predictive accuracy, thus selecting it as our final model to proceed.

$$log(P(Winner = Red)/(1 - P(Winner = Red))) = B_0 + B_1 * B\_avg\_opp\_KD +$$
$$B_2 * B\_avg\_opp\_SIG\_STR\_pct + B_3 * B\_avg\_opp\_SIG\_STR\_pct^2 +$$
$$B_4 * B\_avg\_opp\_TD\_pct + B_5 * B\_avg\_REV + B_6 * B\_avg\_opp\_REV + B_7 * B\_losses +$$
$$B_8 * B\_win\_by\_Decision\_Majority + B_9 * B\_win\_by\_Submission + B_{10} * B\_Reach\_cms +$$
$$B_{11} * R\_avg\_opp\_TOTAL\_STR\_landed + B_{12} * R\_avg\_opp\_LEG\_att + B_{13} * R\_avg\_opp\_LEG\_landed +$$
$$B_{14} * R\_win\_by\_Decision\_Majority + B_{15} * R\_win\_by\_TKO\_Doctor\_Stoppage + B_{16} * B\_age +$$
$$B_{17} * B\_StanceOrthodox + B_{18} * B\_StanceSouthpaw + B_{19} * B\_avg\_REV * B\_avg\_opp\_REV$$

The AUC value for the model above is 0.63, similar to the prior AUC values calculated in the models for the other classes.

Inference

For the class 4 model, there is not a noticeable prevalence in the types of predictors. We spot that reach is important, as a standard deviation increase in the blue fighter's reach decreases the red fighter's log odds of winning by 0.671, holding all else constant. Additionally, sports bettors should be extremely aware of the blue fighter's stance, as Orthodox and Southpaw decrease red's predictive chances relative to a Switch stance. The accuracy measure is the strongest of the four classes of which we have fitted models, predicting the test data winner at a rate of 62%. One limitation in this model, however, is the presence of only 93 observations in our test data, which may inflate our predictive accuracy. We are interested to see how our prediction rate would chance as UFC hosts more fights for the Heavyweight and Openweight classes.

Table 11: Model Coefficients for Class 4 Red Fighter

	x		x
poly(B_avg_opp_SIG_STR_pct, 2)1	5.956	poly(B_avg_opp_SIG_STR_pct, 2)2	-7.298
(Intercept)	2.002	B_StanceOrthodox	-1.618
B_avg_REV	0.682	B_StanceSouthpaw	-1.041
R_win_by_TKO_Doctor_Stoppage	0.498	B_Reach_cms	-0.671
B_avg_opp_REV	0.410	B_avg_REV:B_avg_opp_REV	-0.446
B_avg_opp_KD	0.291	R_win_by_Decision_Majority	-0.316
B_win_by_Submission	0.278	R_avg_opp_LEG_landed	-0.296
B_win_by_Decision_Majority	0.259	R_avg_opp_TOTAL_STR_landed	-0.191
B_avg_opp_TD_pct	0.215	R_avg_opp_LEG_att	-0.016
B_age	0.051	B_losses	0.016

Table 12: Class 4 Elastic Net Model Accuracy

	Values
Accuracy	0.624
Precision	0.684
Recall	0.531
F1	0.598

Conclusion

Despite recognizing there is room for improvement for the predictive accuracy for each of the classes, we are satisfied with the efficiency of our model selection processes. Sports bettors can research historical data of the fighters in an upcoming fight and based on the fight's weight class, import the data into the appropriate model to receive a prediction probability for the red fighter to win. A sports bettor could consider using the red probability's complement to determine the blue fighter's probability of winning, but must be aware that the probability of a draw, although small, is included in that value. Assuming two fighters are evenly matched, sportsbooks typically place the "moneyline" of a fighter winning at -110, meaning a bettor is profitable if they bet correctly at least 55% of the time. Each of our models eclipse that rate, but sportsbooks do not provide -110 lines for every single fight. Therefore, if we were to create our models again, we would create one model for all classes and focus our attention on cases in which the predictive probability was high for a losing fighter, and then try to identify a pattern for which "favorites" in our model were not chosen correctly. Additionally, we could compare our odds of prior fights to the sportsbook's implied odds based on the lines they set, allowing us to spot cases where significant differences are present.

Appendix

Data Dictionary

R_ and B_: Prefix signifies red and blue corner fighter stats respectively

opp: Containing in columns is the action done by the opponent on the fighter

fighter: Name of the fighter

Referee: Referee/On-Hand Doctor of the fight. They are responsible for ending a fight if they believe a fighter is unable to continue

date: Date of the fight

location: Location of the fight

Winner: The corner of the winning fighter. We will turn this into a dummy variable and will serve as our model's response variable

title_bout: True/False indicator of a championship fight for a weight class

weight_class: Categorical variable indicating the division of the fight. There are nine male divisions and four female divisions. We will reassign the male divisions to create our models

KD: the number of knockdowns

SIG_STR: the of significant strikes 'landed of attempted'

SIG_STR_pct: significant strikes percentage

TOTAL_STR: total strikes 'landed of attempted'

TD: number of take downs

TD_pct: take down percentages

SUB_ATT: number of submission attempts

REV: number of reversals landed

HEAD: number significant strikes to the head (att = attempted, landed = successful attempts)

BODY: number of significant strikes to the body (att = attempted, landed = successful attempts)

LEG: number of significant strikes to the leg (att = attempted, landed = successful attempts)

CLINCH: number of significant strikes in the clinch, also known as close quarters (att = attempted, landed = successful attempts)

GROUND: number of significant strikes on the ground (att = attempted, landed = successful attempts)

Stance: the stance of the fighter (orthodox, southpaw, etc.)

Height_cms: the height of the fighter in centimeters

Reach_cms: the reach of the fighter (arm span) in centimeters

Weight_lbs: the weight of the fighter in pounds (lbs)

age: the age of the fighter

current_lose_streak: the amount of consecutive previous fights the fighter has lost (0 if they won their previous fight)

current_win_streak: the amount of consecutive previous fights the fighter has won (0 if they lost their previous fight)

draw: the number of draws in the fighter's ufc career wins: the number of wins in the fighter's ufc career

losses: the number of losses in the fighter's ufc career

total_rounds_fought: the average of total rounds fought by the fighter

total_time_fought(seconds): the count of total time spent fighting in seconds

total_title_bouts: the total number of title bouts taken part in by the fighter

win_by_Decision_Majority: the number of wins by majority judges decision in the fighter's ufc career (often 2-0 with one judge deciding a draw)

win_by_Decision_Split: the number of wins by split judges decision in the fighter's ufc career (often 2-1 in favor of one fighter)

win_by_Decision_Unanimous: the number of wins by unanimous judges decision in the fighter's ufc career

win_by_KO/TKO: the number of wins by knockout in the fighter's ufc career

win_by_Submission: the number of wins by submission in the fighter's ufc career

win_by_TKO_Doctor_Stoppage: the number of wins by doctor stoppage in the fighter's ufc career

Table 13: VIF Values for Class 1 Elastic Net Model

	x
B_avg_KD	1.21
B_avg_opp_SIG_STR_pct	1.15
B_avg_TD_att	1.37
B_avg_opp_TD_att	1.34
B_avg_opp_HEAD_landed	1.21
B_avg_CLINCH_att	1.36
B_win_by_Decision_Majority	1.00
B_win_by_Decision_Unanimous	1.47
B_win_by_Submission	1.12
R_avg_KD	1.22
R_avg_opp_SIG_STR_pct	1.15
R_avg_opp_TD_pct	1.19
R_avg_opp_SUB_ATT	1.18
R_avg_REV	1.16
R_avg_SIG_STR_landed	2.95
R_avg_BODY_landed	2.36
R_avg_LEG_att	1.54
R_avg_opp_CLINCH_att	1.28
R_win_by_KO.TKO	1.41
R_current_win_streak	1.18
B_age	1.28
R_age	1.15
B_StanceSwitch	1.10

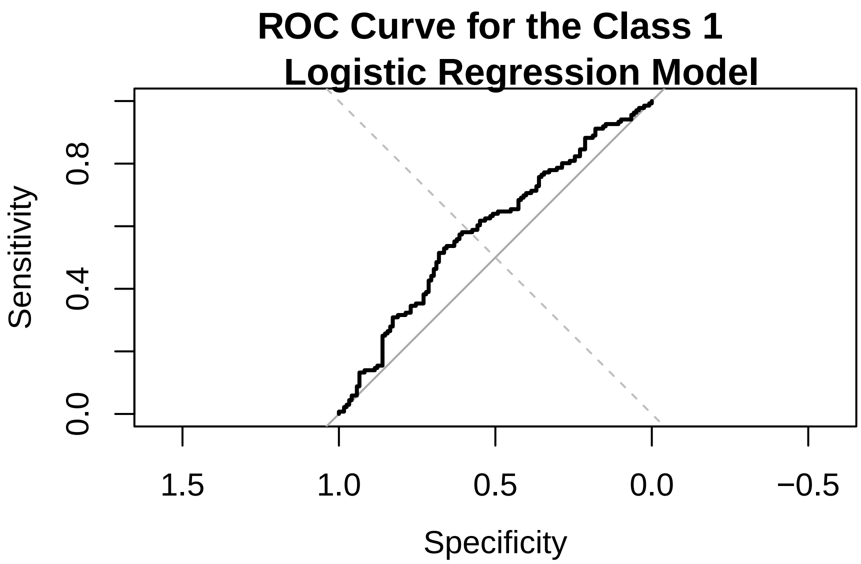
```
## function (x)  .Primitive("length")
```

Table 14: Class 1 Elastic Net Model Accuracy

	Values
Accuracy	0.531
Precision	0.504
Recall	0.516
F1	0.510

Table 15: Class 1 Elastic Net –> Best Subset Model Accuracy

	Values
Accuracy	0.558
Precision	0.532
Recall	0.541
F1	0.537




```
## AUC: 0.5906461
```

Table 16: VIF Values for Class 2 Elastic Net Model

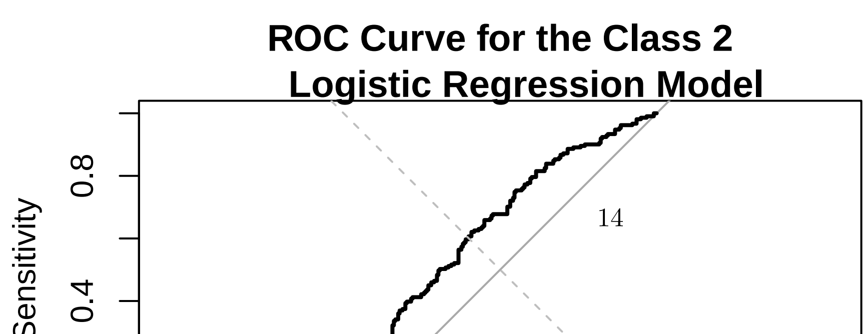
	GVIF	Df	GVIF^(1/(2*Df))
B_avg_opp_SIG_STR_pct	1.14	1	1.07
B_avg_opp_SUB_ATT	1.19	1	1.09
B_avg_REV	1.20	1	1.09
B_avg_TD_att	3.19	1	1.79
B_avg_TD_landed	4.25	1	2.06
B_avg_CLINCH_landed	1.22	1	1.10
B_avg_CTRL_time.seconds.	3.37	1	1.84
B_total_time_fought.seconds.	1.56	1	1.25
B_total_title_bouts	1.32	1	1.15
B_current_lose_streak	1.11	1	1.05
B_longest_win_streak	1.79	1	1.34
B_win_by_Submission	1.42	1	1.19
B_win_by_TKO_Doctor_Stoppage	1.24	1	1.11
R_avg_SIG_STR_pct	1.14	1	1.07
R_avg_SUB_ATT	1.30	1	1.14
poly(R_avg_HEAD_landed, 2)	1.47	2	1.10
poly(R_avg_opp_CLINCH_att, 2)	1.30	2	1.07
R_avg_GROUND_att	1.35	1	1.16
R_avg_opp_CTRL_time.seconds.	1.22	1	1.10
R_current_win_streak	1.44	1	1.20
R_longest_win_streak	3.06	1	1.75
R_win_by_Decision_Split	1.25	1	1.12
R_win_by_Decision_Unanimous	2.18	1	1.48
R_win_by_Submission	1.51	1	1.23
B_age	1.25	1	1.12
R_age	1.36	1	1.16
B_StanceOrthodox	6.48	1	2.55
B_StanceSouthpaw	6.42	1	2.53

Table 17: Class 2 Elastic Net –> P-Value Sig. Predictors Model

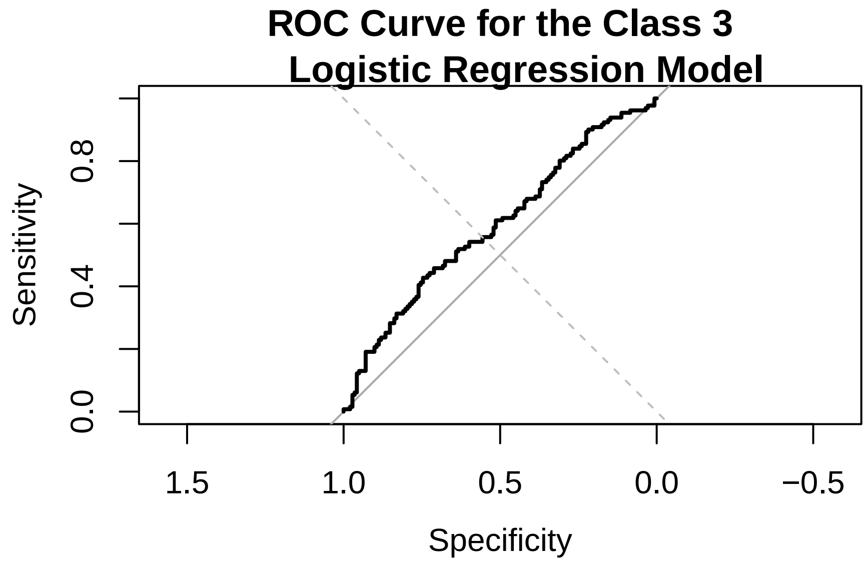
	Values
Accuracy	0.564
Precision	0.573
Recall	0.555
F1	0.564

Table 18: Class 2 Elastic Net –> Best Subset Model Accuracy

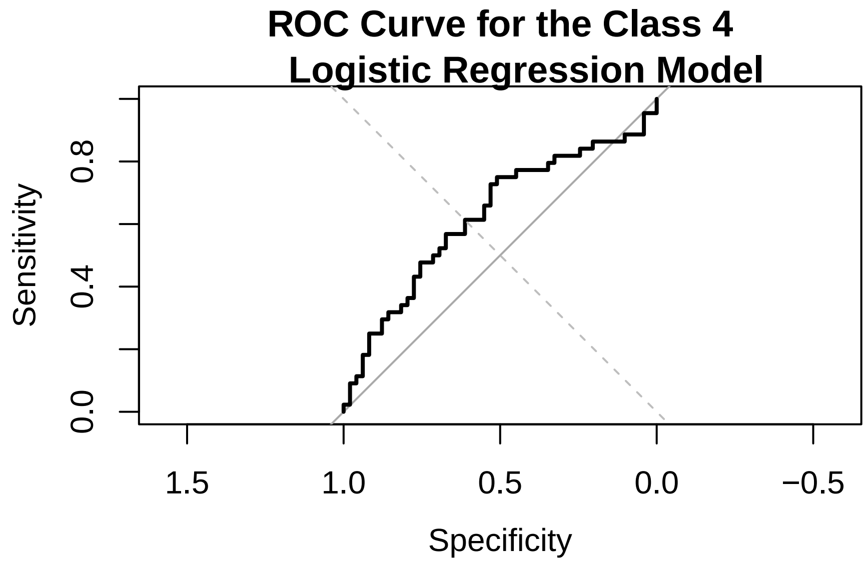
	Values
Accuracy	0.545
Precision	0.558
Recall	0.509
F1	0.532



```
## AUC: 0.6397235
```




```
## AUC: 0.5948285
```




```
## AUC: 0.6252319
```

Table 19: VIF Values for Class 3 Elastic Net Model

	x
B_avg_opp_SIG_STR_pct	1.45
B_avg_opp_TD_pct	1.52
B_avg_REV	1.28
B_avg_opp_REV	1.35
B_avg_SIG_STR_landed	3.31
B_avg_opp_TOTAL_STR_landed	4.13
B_avg_TD_att	2.58
B_avg_opp_TD_att	1.60
B_avg_LEG_landed	1.85
B_avg_opp_LEG_att	1.86
B_avg_opp_CLINCH_landed	1.95
B_avg_GROUND_att	2.39
B_avg_opp_GROUND_landed	2.70
B_avg_CTRL_time.seconds.	3.55
B_current_win_streak	3.46
B_current_lose_streak	1.72
B_longest_win_streak	3.74
B_win_by_Decision_Split	1.45
B_win_by_Decision_Unanimous	1.95
B_win_by_TKO_Doctor_Stoppage	1.15
B_Height_cms	2.09
B_Reach_cms	1.92
B_Weight_lbs	1.57
R_avg_KD	1.42
R_avg_SIG_STR_pct	1.58
R_avg_opp_TD_pct	1.47
R_avg_opp_SUB_ATT	1.26
R_avg_REV	1.24
R_avg_opp_TOTAL_STR_landed	2.57
R_avg_TD_landed	1.45
R_avg_opp_TD_att	1.57
R_avg_opp_HEAD_att	2.42
R_avg_BODY_landed	1.57
R_avg_opp_LEG_att	1.53
R_longest_win_streak	4.47
R_wins	10.95
R_losses	3.16
R_win_by_Decision_Split	1.56
R_win_by_KO.TKO	4.26
R_Height_cms	2.14
R_Reach_cms	2.14
R_age	1.55
B_StanceOpenStance	1.08
B_StanceOrthodox	1.26
R_StanceSouthpaw	1.23
R_StanceSwitch	1.12
B_avg_GROUND_att:B_avg_CTRL_time.seconds.	1.63
R_avg_SIG_STR_pct:R_avg_BODY_landed	1.36

Table 20: Class 2 Elastic Net –> P-Val Simplified Model Accuracy

	Values
Accuracy	0.546
Precision	0.556
Recall	0.634
F1	0.592

Table 21: Class 3 Elastic Net –> Backwards BIC Model Accuracy

	Values
Accuracy	0.524
Precision	0.534
Recall	0.662
F1	0.591

Table 22: VIF Values for Class 4 Elastic Net Model

	GVIF	Df	GVIF^(1/(2*Df))
B_avg_opp_KD	1.34	1	1.16
poly(B_avg_opp_SIG_STR_pct, 2)	1.50	2	1.11
B_avg_opp_TD_pct	1.24	1	1.11
B_avg_REV	3.07	1	1.75
B_avg_opp_REV	1.64	1	1.28
B_losses	2.00	1	1.41
B_win_by_Decision_Majority	1.28	1	1.13
B_win_by_Submission	1.40	1	1.18
B_Reach_cms	1.40	1	1.18
R_avg_opp_TOTAL_STR_landed	1.69	1	1.30
R_avg_opp_LEG_att	33.68	1	5.80
R_avg_opp_LEG_landed	34.80	1	5.90
R_avg_opp_GROUND_att	22.93	1	4.79
R_avg_opp_GROUND_landed	23.21	1	4.82
R_win_by_Decision_Majority	1.16	1	1.08
R_win_by_TKO_Doctor_Stoppage	1.20	1	1.09
B_age	1.68	1	1.30
B_StanceOrthodox	5.67	1	2.38
B_StanceSouthpaw	6.13	1	2.48
B_avg_REV:B_avg_opp_REV	3.06	1	1.75

Table 23: Class 4 Elastic Net –> Sig. P-Value Model Accuracy

	Values
Accuracy	0.591
Precision	0.649
Recall	0.490
F1	0.558

Table 24: Class 4 Elastic Net –> Best Subset Selection

	Values
Accuracy	0.548
Precision	0.595
Recall	0.449
F1	0.512