“Predicting fight outcomes in UFC”
Final Report

Owner: Usman Mahmood

Contents

1 Context **2**

2 Problem Statement **3**

3 Data Collection and Pre-processing **3**

3.1 Data Source 3

3.2 Data Pre-processing 3

3.3 Feature Selection 3

3.4 Missing Data 4

3.5 Feature Engineering 4

3.6 Train-Test set split 4

3.7 Data Standardization 5

4 Methodology **5**

5 Evaluation and Final Results **5**

5.1 Fight Predictions 5

5.2 Key Findings 7

5.3 Hypothesis Testing 7

6 Reference: **8**

1 Context

The Ultimate Fighting Championship (UFC) is a rapidly growing and successful sports organization that operates in the realm of Mixed Martial Arts (MMA). In contrast with other mainstream sports of the world, MMA is relatively new. Yet, its popularity is not confined to a particular geography or culture making it a truly global sport albeit in its infancy. Fighters in the UFC hail from all over the world and compete in the UFC events all year round.

As the name suggests, the sport combines a variety of different combat sports or martial arts into a single unified format. The scoring system takes multiple aspects of fighting into account and is intended to not favor one or more types of martial arts over others. Similarly, fights can be ended unilaterally if a fighter is able to knock out or submit the opponent.

Because of the physically taxing nature of the sport, injuries abound and fighters only tend to compete 2 to 3 times a year on average. This results in limited data about fighters and their skill sets available for analysis. Furthermore, fighters undergo rigorous training sessions and often tend to improve on new disciplines and techniques in between consecutive fights. This means that unlike traditional combat sports, such as Boxing, Kickboxing, Judo, Wrestling, etc., MMA fights are notoriously difficult to predict even for the most experienced analysts and professionals in the sport. This unpredictability also makes the sport more exciting for fans and fighters alike.

As a fan of the sport, there are many nuances that I have come across from elite athletes, commentators, and professional analysts, such as the "**intangibles**". Experts believe that the intangibles are latent and often ignored factors that tend to be better predictors of fights than typical descriptive statistics, such as fighters' weight, height, reach, win record, striking accuracy, defence accuracy, etc. that are the basis of most predictions and fan chatter. Some of these intangible factors pointed out by various experts are:

1. **Pressure**
   A highly active fighter who dictates the pace of the fight with constant attacks and forward (attacking) movement drains or exhausts the opponent both mentally and physically ultimately leading to their defeat. Is this true?

2. **Damage absorbed in previous fights**
   One of the ubiquitous comments in the sport often heard is that fighters "leave a piece of themselves in the octagon" with each fight. This alludes to the fact that once a fighter takes significant damage during a fight, it negatively affects his/her future performance either physically or psychologically or both regardless of the outcome of the fight. Is this true?

3. **Hunger (desire) to win**
   It is often believed that a hungry or motivated fighter is a more lethal fighter and one favored to win more often than not. The "hunger" can be in the form of career earnings, type of fight (i.e. Title vs. Non-title), losing streak (i.e. threat of being fired from the organization prematurely), etc. Does hunger translate to greater success?

4. **Jack-of-all vs. Master-of-few**
   Experts often emphasize that fighters in this sport should be well-rounded, i.e. they should be skilled in multiple martial art forms. However, I have often seen fighters who are exceptionally talented in "Wrestling" or "Brazilian Jiu Jitsu" and less than average in other martial art forms dominate their opponents consistently. Is it better to be a jack-of-all as opposed to a master-of-few? In the case of the latter, does mastery of certain disciplines lead to more success than that attained from mastery of others?

Over the past couple of years, these are some of the beliefs or claims that I have come across that remain untested and unproven absent objective data analysis.

2 Problem Statement

- Validate the intangible factors as either significantly true or debunk them as mere myths based on statistical data analysis.
- Determine the weights (importance) of the intangible factors towards win/loss probabilities.
- Build a Machine Learning model that can reasonably accurately predict the outcome of fights based on a combination of descriptive and newly engineered custom features that capture the intangible factors listed earlier.

Caveat:

Because of the limited data available about fighters and bout statistics, it may not be possible to evaluate all intangible factors listed in this study, but I'm confident I'll be able to test some of them.

Future Potential:

I'm not aware whether such intangibles are factored in to the models deployed by sports betting organizations as specified in their odds prior to each fight. If the model proves reliable and if the win/loss probabilities from the model do not perfectly align with the betting odds, there may exist arbitrage opportunities for bettors.

3 Data Collection and Pre-processing

3.1 Data Source

The data set was acquired from Kaggle[1], that included statistics about fights since 2010, gambling odds for each fight, as well as UFC official fighter rankings. Fights prior to 2010 were recorded improperly and are thus excluded from this partially pre-processed data set. Nevertheless, the initial data set includes fight statistics for 4,566 fights (bouts), approximately 134 features, and 1,667 unique fighters.

3.2 Data Pre-processing

Most feature columns were numeric and were set to float type. Fighter "Stance" was the only categorical variable that was encoded using One-Hot-Encoding technique. Binary encoding was performed on other categorical variables, e.g. "Winner" (response variable), "Gender", "Title Bout". A number of other corrections were also made such as misspelled fighter names, trimming of columns, etc as needed.

3.3 Feature Selection

The data set included a number of features with very high null counts, such as the ranking columns as well as certain bout specific stats that have recently become available to public. Additionally, some features were irrelevant to the analysis, such as "Location" of the fight, "Country" fighters hail from, "Empty Arena". Moreover, certain features were redundant, e.g. information about the fight "Weight Class" is also captured by the fighter "Weight" feature, "Expected Value" from a $100 bet on a fighter is directly related to the betting "Odds" for that fighter. Those features were thus excluded from the analysis.

Subsequently, the data set was evaluated for the presence of multi-collinearity among the remaining set of features. Figure 1 shows the heat plot of the final set of features.

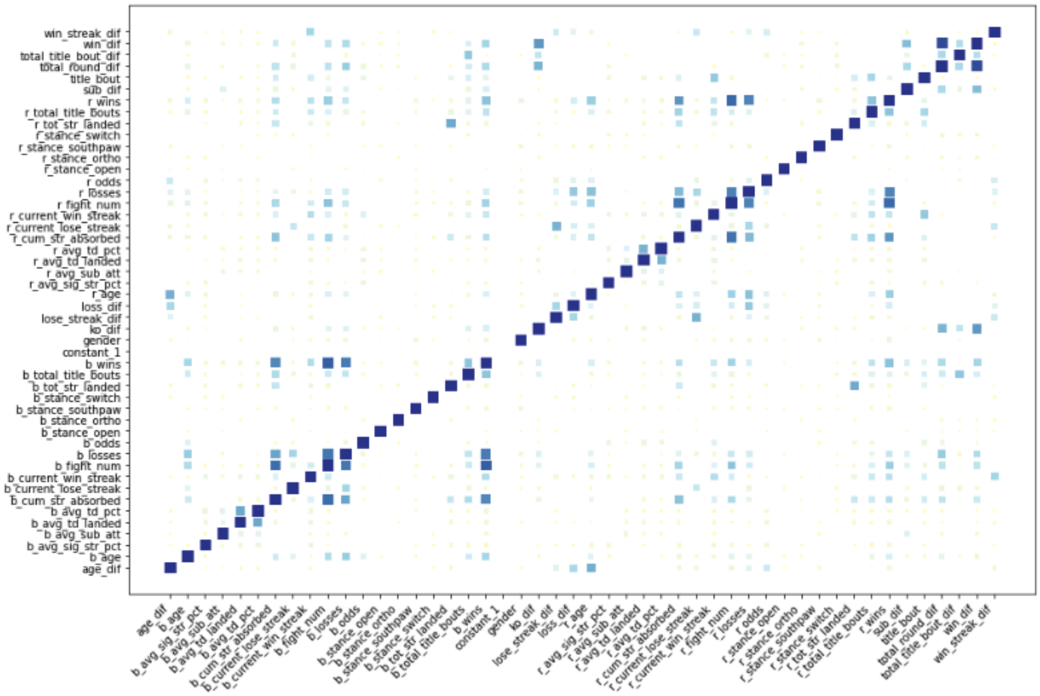


Figure 1: Heat plot of Features correlation

As evident from the heat plot, the input features were not highly correlated with one another.

3.4 Missing Data

For a number of fights, some of the fighter's stats such as "Average Significant Strikes Landed", "Average Significant Strikes Percent", "Average Take-downs Landed", "Average Take-downs Percent", and "Average Submission Attempted" had null values especially for their first fight in UFC. In order to retain those data points, missing data for a fighter was imputed based on the average of those features for all other fights of that fighter. The final data set contained 4,051 data points and 56 features.

3.5 Feature Engineering

In order to test one of the hypotheses as well as to evaluate the predictive power of whether cumulative damage absorbed by fighters over their career affects their performance or fight outcomes, I computed a new feature, "Cumulative Strikes Absorbed". As the name suggests, the new feature is the accumulated total significant strikes absorbed by a fighter up to each fight and was computed based on "Average Significant Strikes Landed per minute" (by their opponent) and "Total Fight Time in secs" data available for each fight.

3.6 Train-Test set split

The data was split in to 80% Training set and 20% Test set. The test set comprised of all the recent fights, while the training set comprised of older data. This choice of the chronological split was intentional as I wanted to make sure that all the past data for the fighters was accounted for when making future fight predictions. It was also important because due to the nature of the sport, each fighter only has a limited number of fights (data points) and randomizing selection was probably not the preferred option. 5-fold cross validation was also performed on the training set when fitting various models to the data.

3.7 Data Standardization

As a number of models tested are distance measure based, features were standardized prior to model fit. This was important because the scale and units of measurement varied significantly across the features and standardizing features would prevent unwanted bias.

4 Methodology

Following machine learning classification algorithms were used to predict the Win/Loss outcomes for the available data:

- Logistic Regression
- Logistic Regression with Regularization (Lasso and Elastic Net)
- K Nearest Neighbors
- Support Vector Machines (Linear)
- Support Vector Machines (Kernel based, non-linear)
- Decision Tree
- Random Forest
- Neural Network
- AdaBoost
- Naïve Bayes

First, the above models were fit using full set of features (i.e. 56 in total) obtained following the data preparation phase.

Second, Principal Component Analysis (PCA) was used to reduce the number of features and extract the top set of components that explained the most variability in the response variable. Top components were selected based on eigenvalues ≥ 1. Data with reduced dimensions (principal components) was then fit across all the models and their performance was evaluated.

Third, top features were extracted based on importance scores from the Random Forest Model. Also, top features were found and compared with those from Regularized Logistic Regression (using L2 penalty, i.e. Lasso). Data with those set of features was fit across all the models and their performance was evaluated.

5 Evaluation and Final Results

5.1 Fight Predictions

As mentioned earlier, cross validation was used for all the models listed above. Also, following metrics were used to evaluate performance of the models on both training and test sets:

- **Classification Accuracy**
  For each model, the Classification accuracy score from Cross Validation was based on average (1 - error rate) across all 5 folds of the training set. Similarly, the Classification accuracy score for Training set was based on (1 - error rate) from the model fitted using the full training set and evaluated against the Training set. Lastly, the Classification accuracy score for Test set was based on (1 - error rate) from the model fitted using the full training set and evaluated against the 'unseen' Test set.
- **Precision**
  For each model, weighted Precision scores were obtained by evaluating fit on Training and Test sets separately.

- **Recall**
  For each model, weighted Recall scores were obtained by evaluating fit on Training and Test sets separately.
- **F1-Score**
  For each model, weighted F1-Scores were obtained by evaluating fit on Training and Test sets separately.

Model/Metric	Accuracy			Precision		Recall		F1-Score	
	CV	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.66	0.67	0.62	0.66	0.61	0.67	0.62	0.66	0.61
Logistic Regression w/ Reg.	0.65	0.67	0.63	0.66	0.63	0.67	0.63	0.66	0.62
K-Nearest Neighbors	0.59	0.69	0.60	0.69	0.60	0.69	0.60	0.69	0.60
Linear SVM	0.65	0.67	0.63	0.66	0.63	0.67	0.63	0.66	0.63
RBF SVM	0.61	0.95	0.57	0.95	0.57	0.95	0.57	0.95	0.57
Decision Tree	0.57	1.00	0.54	1.00	0.54	1.00	0.54	1.00	0.54
Random Forest	0.64	1.00	0.62	1.00	0.62	1.00	0.62	1.00	0.62
Neural Network	0.62	0.82	0.57	0.82	0.57	0.82	0.57	0.82	0.57
AdaBoost	0.64	0.69	0.59	0.69	0.59	0.69	0.59	0.69	0.59
Naive Bayes	0.59	0.59	0.59	0.67	0.60	0.59	0.59	0.44	0.45

Table 1: Model Performance for data containing all 56 features

Model/Metric	Accuracy			Precision		Recall		F1-Score	
	CV	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.61	0.62	0.59	0.61	0.57	0.62	0.59	0.59	0.55
Logistic Regression w/ Reg.	0.60	0.62	0.59	0.61	0.57	0.62	0.59	0.59	0.55
K-Nearest Neighbors	0.56	0.72	0.56	0.72	0.54	0.72	0.56	0.72	0.55
Linear SVM	0.59	0.61	0.58	0.62	0.56	0.61	0.58	0.54	0.50
RBF SVM	0.58	1.00	0.58	1.00	0.34	1.00	0.58	1.00	0.43
Decision Tree	0.60	0.65	0.58	0.64	0.56	0.65	0.58	0.63	0.54
Random Forest	0.60	1.00	0.59	1.00	0.57	1.00	0.59	1.00	0.55
Neural Network	0.59	0.70	0.58	0.70	0.57	0.70	0.58	0.68	0.56
AdaBoost	0.60	0.67	0.59	0.66	0.58	0.67	0.59	0.66	0.57
Naive Bayes	0.58	0.58	0.56	0.54	0.49	0.58	0.56	0.49	0.47

Table 2: Model Performance for data containing Top 20 Principal Components

Model/Metric	Accuracy			Precision		Recall		F1-Score	
	CV	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.65	0.65	0.62	0.65	0.61	0.65	0.62	0.65	0.61
Logistic Regression w/ Reg.	0.65	0.66	0.62	0.65	0.62	0.66	0.62	0.65	0.61
K-Nearest Neighbors	0.60	0.75	0.58	0.74	0.57	0.75	0.58	0.74	0.57
Linear SVM	0.65	0.66	0.63	0.65	0.62	0.66	0.63	0.65	0.62
RBF SVM	0.58	0.99	0.58	0.99	0.54	0.99	0.58	0.99	0.48
Decision Tree	0.62	0.67	0.60	0.67	0.59	0.67	0.60	0.67	0.59
Random Forest	0.64	1.00	0.61	1.00	0.60	1.00	0.61	1.00	0.60
Neural Network	0.65	0.67	0.62	0.66	0.61	0.67	0.62	0.66	0.61
AdaBoost	0.64	0.68	0.60	0.68	0.59	0.68	0.60	0.68	0.59
Naive Bayes	0.65	0.65	0.61	0.65	0.61	0.65	0.61	0.65	0.61

Table 3: Model Performance for data containing Top 10 most important features

5.2 Key Findings

The accuracy scores obtained using the models listed above did not exceed 65%. Also, linear models tended to perform better than non-linear models. Few models such as the RBF Support Vector Machines, Decision Tree, and Random Forest often resulted in overfit models as evident from perfect accuracy, precision, recall and F1-scores when evaluated against the training data. The scores dropped significantly as those models were evaluated against the test set indicating low bias and high variance. The results did not vary significantly with fewer features or when PCA was performed to reduce dimensions.

For most models, cross validation scores were often similar to performance on the test set indicating that cross validation accuracy score is a reliable estimate of the accuracy on the test set.

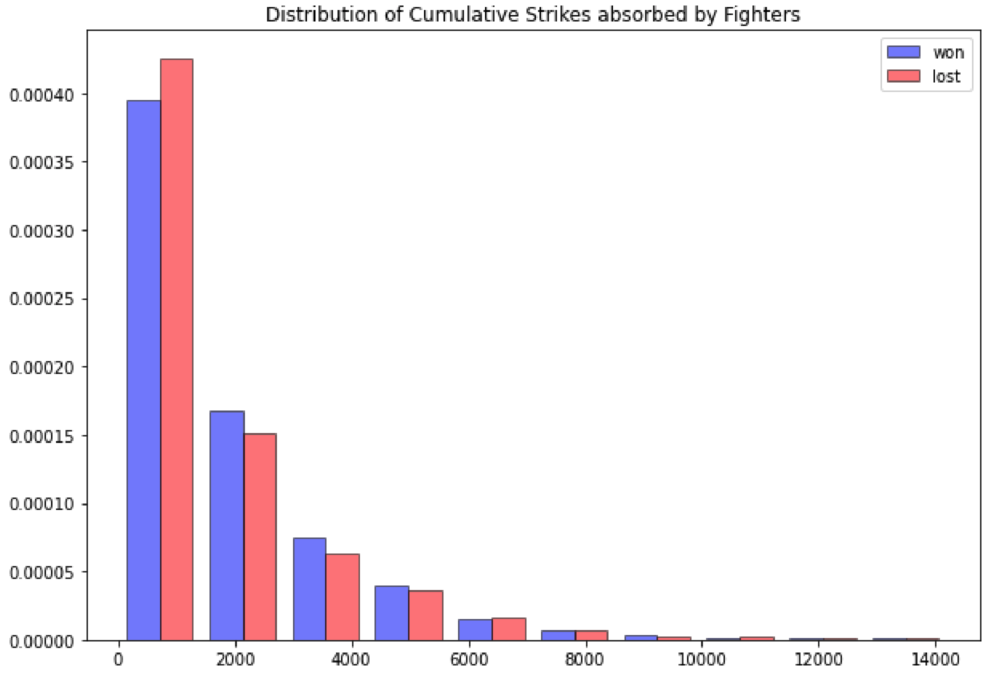
Top features by Importance score from Random Forest are: “Fighter Betting odds”, “Average Significant Strike Percent”, “Cumulative Strikes Absorbed”, “Age Differential”, “Total Strikes Landed”, “Significant Strike Differential”.

In summary, based on Tables 1 and 3, Linear Support Vector Machines and Logistic Regression performed best with relatively high Accuracy, Precision, Recall, and F1-Score in both cross validation and test set.

5.3 Hypothesis Testing

Hypothesis: One of the ubiquitous comments in the sport often heard is that fighters “leave a piece of themselves in the octagon” with each fight. This alludes to the fact that once a fighter takes significant damage during a fight, it negatively affects his/her future performance either physically or psychologically or both regardless of the outcome of the fight.

To test this hypothesis, “Cumulative Strikes Absorbed” feature was generated and its distribution was evaluated for both win/loss outcomes for comparison:



The above side-by-side density histogram indicates that both distributions are not significantly different from one another. In other words, the cumulative damage sustained by fighters over their careers does not seem to influence the outcome of their fights.

6 Reference:

1. https://www.kaggle.com/mdabbert/ultimate-ufc-dataset?select=ufc-master.csv