Final Project Report

UFC Betting Odds and Fighter Stats

Introduction

Mixed Martial Arts (MMA) has gained significant global attention, particularly through organizations like the Ultimate Fighting Championship (UFC). With its increasing popularity, betting on UFC fights has become a big part of the sport. Oddsmakers use fighter statistics, historical fight data, and subjective insights to set betting lines. Despite the sophisticated process behind these odds, questions remain how accurately these lines predict fight outcomes.

This project investigates the relationship between UFC fighter statistics and betting odds, with the goal of identifying trends, inefficiencies, and insights that could help bettors, analysts, and UFC fans better understand how odds are set. Specifically, I will analyze the effect of fighter attributes, such as reach, striking accuracy, and takedown defense, on the odds and outcomes of title fights. The objective is to explore if certain statistics are undervalued by sportsbooks and whether specific fighter attributes can predict fight results.

Data

To perform this analysis, two datasets were utilized:

2.1 UFC Fighter & Fight Statistics

I collected data from UFCStats.com, which contained height, reach, stance, striking accuracy, and takedown defense. This data was scraped on 209 fighter profile pages. I wrote a script to collect the fighter stats across these pages. This included the given fighters striking accuracy, takedown defense, height, reach, and stance. I created a dataframe called df_stats which included this scraped data and is in the ufc.ipynb file in the project folder.

This data required some cleaning. The columns needed to be renamed to Fighter, StrikingAcc, TakedownDef, Height, Reach, and Stance. The StrikingAcc and TakedownDef included words with the value which needed to be removed. The Height, Reach, and Stance columns needed the same thing as well as having a quotation removed from the end indicating inches. I created a

function to convert these strings to inches and applied. This was all done in the ufc.ipynb file in the project folder.

2.2 Betting Odds Data

I collected data from Kaggle (https://www.kaggle.com/datasets/mdabbert/ultimate-ufc-dataset?resource=download), which contained the fighter names, betting odds, fight results, and additional fighter statistics such as wins. This data was checked at random on sites such as ESPN to ensure legitimacy and accuracy with no incorrect values found. I wrote this data into a dataframe called df_fights from the fighter_data.csv downloaded from Kaggle. This data did not require any cleaning at this point as it needed to still be merged with the fighter statistics.

2.3 Combining Stats and Odds

Since both data sets included the fighter names, I had to merge based on the specific fighter name. To do this I used the red fighter to merge the statistics with the odds. This merge was a left merge on RedFighter to ensure the correct fighter was merged to red. This caused a challenge because some fighters who were RedFighter in one bout were BlueFighter in another, so I had to merge on a suffix using '_Red'. I dropped the initial "Fighter" column to clean the data. StrikingAcc, TakedownDef, Height, Reach, and Stance were all changed to match the red fighter (RedStrikingAcc, RedReach, etc.).

The same thing was done with the blue fighter, merging left on BlueFighter using the suffix '_Blue' to correctly merge to the blue fighter. The initial "Fighter" column was dropped again to avoid repeating values and to keep the data clean. I wrote and merged these to a dataframe called df_merged and then to a csv called fights_stats.csv. This is all done in the ufc.ipynb file in the project folder.

Table 1 Data Dictionary

Column	Type	Source	Description

RedFighter	Text	Both	Name of red corner fighter
BlueFighter	Text	Both	Name of blue corner fighter
RedOdds	Numeric	Kaggle	Red fighter betting odds
BlueOdds	Numeric	Kaggle	Blue fighter betting odds
Winner	Text	Kaggle	Winner of fight (Red or Blue)
RedStrikingAcc	Numeric	UFCStats	Career striking accuracy percentage for red fighter
RedTakedownDef	Numeric	UFCStats	Career takedown defense percentage for red fighter
RedHeight	Numeric	UFCStats	Height of red fighter (in inches)
RedReach	Numeric	UFCStats	Reach of red fighter (in inches)
BlueStrikingAcc	Numeric	UFCStats	Career striking accuracy percentage for blue fighter
BlueTakedownDef	Numeric	UFCStats	Career takedown defense percentage for blue fighter
BlueHeight	Numeric	UFCStats	Height of blue fighter (in inches)
BlueReach	Numeric	UFCStats	Reach of blue fighter (in inches)

Winner_Red	Numeric	Merged	New column with a value of 1 if red fighter won and 0 if blue fighter won

Analysis

The analysis was conducted using a combination of descriptive statistics, correlation analysis, visualizations, and machine learning models. The goal was to assess how fighter statistics correlate with betting odds and fight outcomes, as well as to explore potential inefficiencies in the odds-setting process. All of the analysis was done in the ufc_analysis.ipynb file.

3.1 Research Questions

The primary goal of this project was to evaluate the relationship between fighter statistics and betting odds. The analysis was guided by the following research questions:

1. Do certain fighter attributes influence betting odds more than others?
2. Are there specific betting trends or inefficiencies that can be exploited based on historical fight data?
3. Do sportsbooks overvalue or undervalue certain fighter statistics when setting odds?

There was a question regarding how accurate betting odds are across weight classes, but this was dropped due to a lack of data to properly analyze this question in a way that would give any useful insight.

These questions aim to assess both the behavior of sportsbooks and the potential for analytical strategies to identify market opportunities.

3.2 Question 1

Do certain fighter attributed influence betting odds more than the others?

To answer this question, I used a correlation heatmap to show relationships between fighter statistics and betting odds. This is shown in *Figure 1* below. The strongest positive correlation found (aside from direct correlation) is between Reach and Height. This makes sense and shows the correlation matrix is working because someone who is tall is likely to have a longer reach, and vise-versa. RedOdds and BlueOdds have an almost perfectly inverse correlation, which makes sense as the fighters odds are opposite for the most part.

The statistic with the strongest positive correlation to BlueOdds is RedStrikingAcc. This shows that as RedStrikingAcc goes up, BlueOdds also go up. The strongest inverse relationship to BlueOdds other than RedOdds is BlueReach. This shows that as BlueReach goes up, BlueOdds goes down. For RedOdds, the strongest positive correlation is BlueReach. This shows that as BlueReach goes up, so does RedOdds. The strongest inverse relationship to RedOdds is RedStrikingAccuracy. This shows that as RedStriking accuracy goes up, RedOdds goes down.

This information can be used when determining which fighter statistics have the biggest part in determining odds. This has shown that for RedOdds, BlueReach is the biggest decider, and for BlueOdds, RedStrikingAccuracy is the strongest decider. BlueStrikingAccuracy is also a similarly strong correlation to RedOdds showing that there is likely a similar process in creating odds for red or blue fighters.

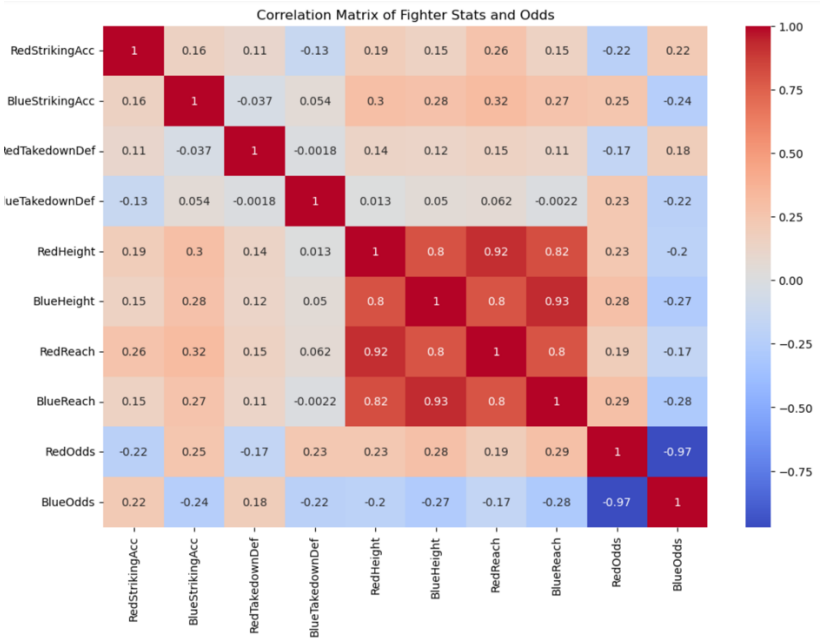

Figure1 Correlation Matrix of Fighter Stats and Odds

I have also included a bar plot of median red fighter betting odds by reach quartile to visually assess how a physical attribute may relate to odds. This can be seen in *Figure 2* below. This shows that Red Fighters in the second reach quartile are likely to be a higher favorite than those in the 1st, 3rd, or 4th. This indicates that they are perceived as stronger favorites by sportsbooks. This could indicate a potential inefficiency or bias in how reach is factored into odds.

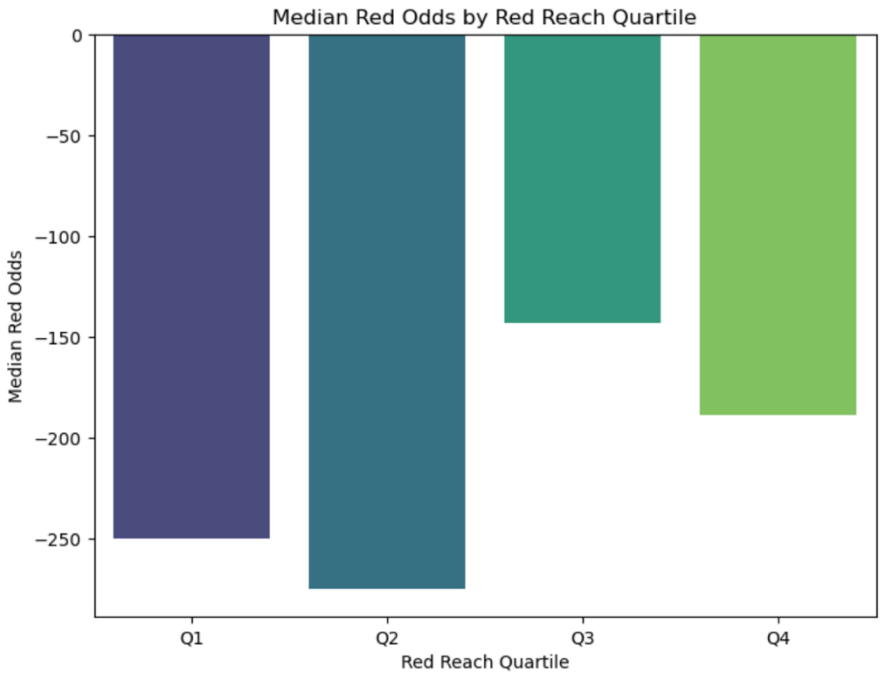

Figure 2 Median Red Odds by Red Reach Quartile

3.3 Question 2

Are there specific betting trends or inefficiencies that can be exploited based on historical fight data?

To answer this question, I used a logistic regression model trained on fighter attributes to test whether attributes can predict winners better than odds alone. The accuracy on this regression came out to 0.67. The confusion matrix shown below in *Figure 3* shows that 25 fights were predicted 1 and were actually 1 (True Positive), 3 fights were predicted 0 and actually 1 (False

Negative), 11 fights were predicted 1 and actually 0 (False Positive), and 4 fights were predicted 0 and actually 0 (True Negative). This is pretty accurate as there is more true positives than all other categories combined.

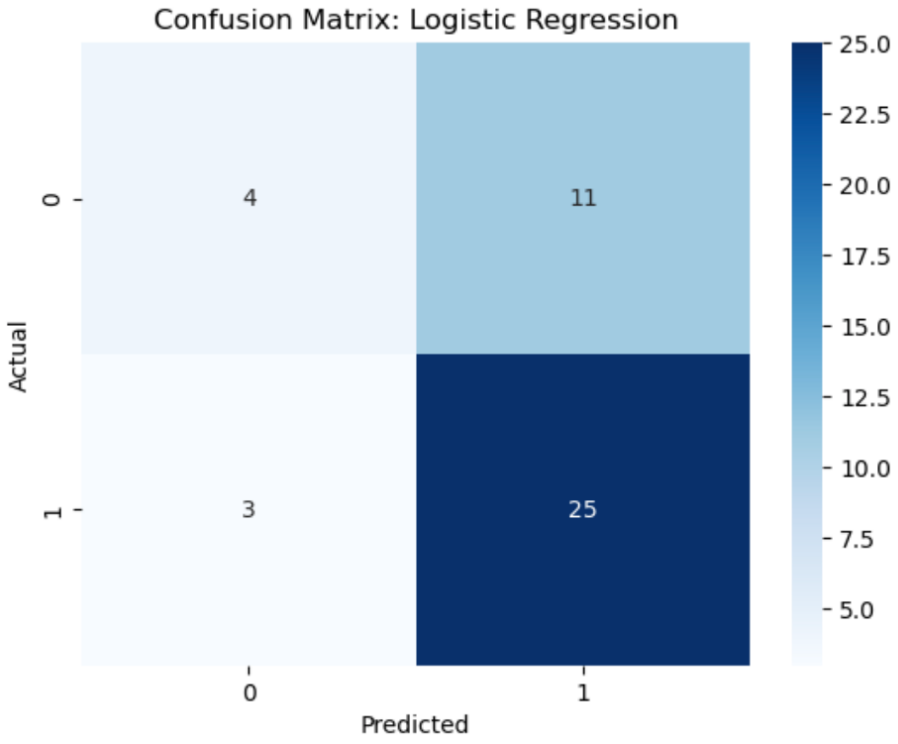

Figure 3 Confusion Matrix: Logistic Regression

Comparing this to a model with odds-only can allow insight nto potential inefficiencies or overlooked patterns in betting markets. For the odds-only model, the accuracy came out to 0.65. The confusion matrix shown below in *Figure 4* shows that 25 fights were predicted 1 and were actually 1 (True Positive), 3 fights were predicted 0 and actually 1 (False Negative), 12 fights were predicted 1 and actually 0 (False Positive), and 3 fights were predicted 0 and actually 0 (True Negative). This is also pretty accurate as there is more true positives than all other categories combined.

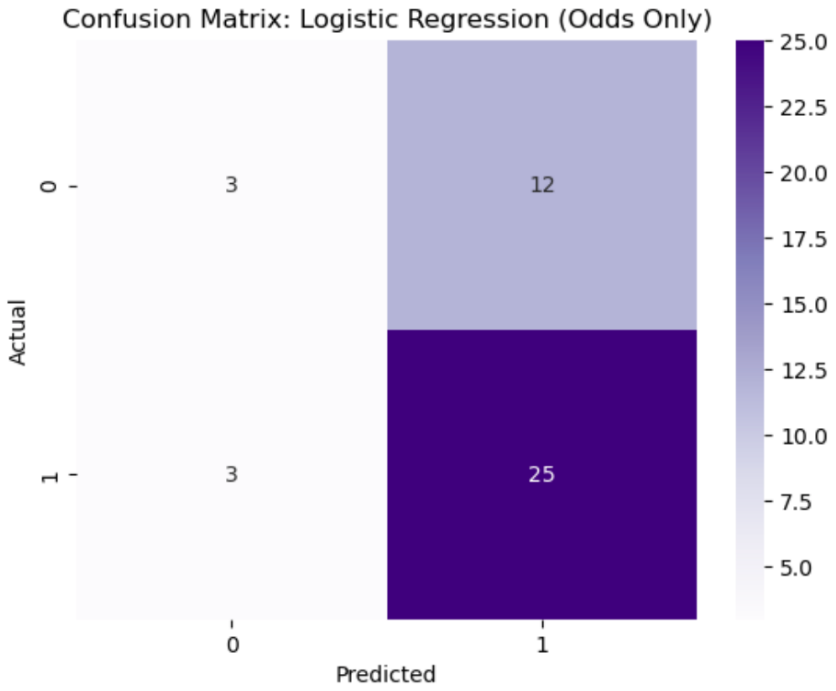

Figure 4 Confusion Matrix: Logistic Regression (Odds Only)

While the odds-only model also performs reasonably, it is slightly less accurate and misclassifies one more fight than the attributes-based model. Both models have the same number of true positives, but the attributes-based model has fewer false positives and more true negatives, suggesting a slightly better balance in prediction quality.

3.4 Question 3

Do sportsbooks overvalue or undervalue certain fighter statistics when setting odds?

By analyzing correlations between fighter statistics and odds we can identify which statistics appear strongly linked to betting lines. In *Figure 5*, it shows the correlation between Winner_Red and the statistics for the red fighter. Using this we can compare to *Figure 1* to see if a statistic is strongly correlated to winning, but weakly to odds (or vise-versa) to see if it may be under/over valued by sportsbooks.

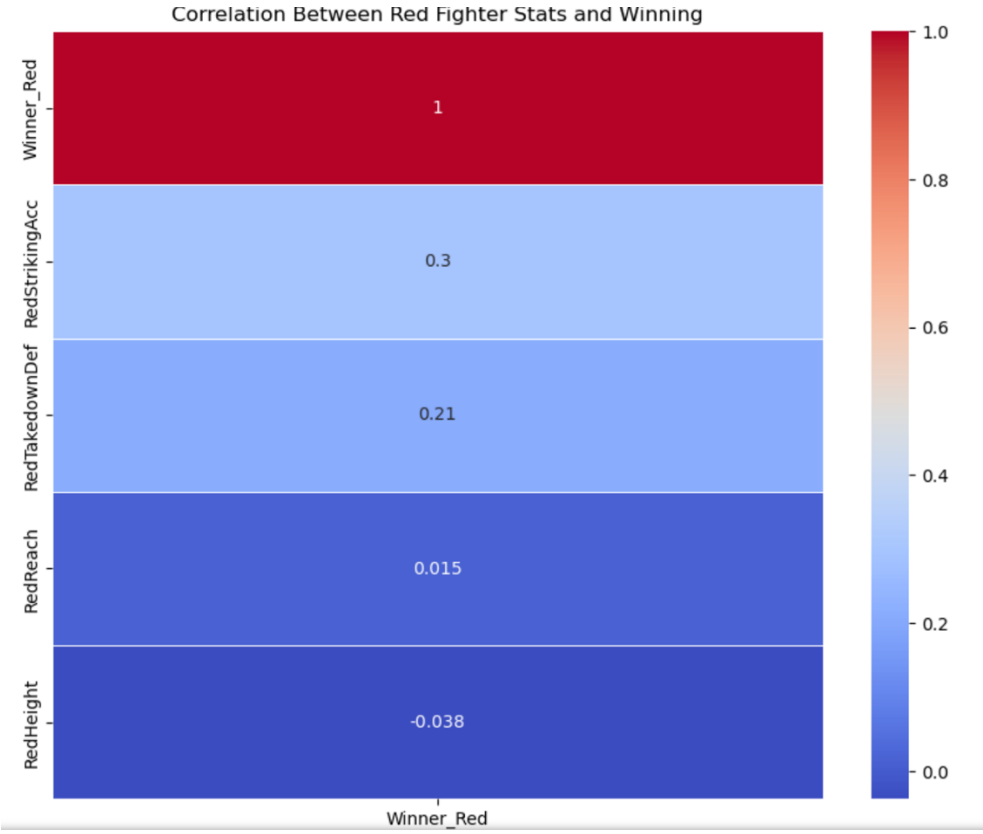

Figure 5 Correlation Between Red Fighter Stats and Winning

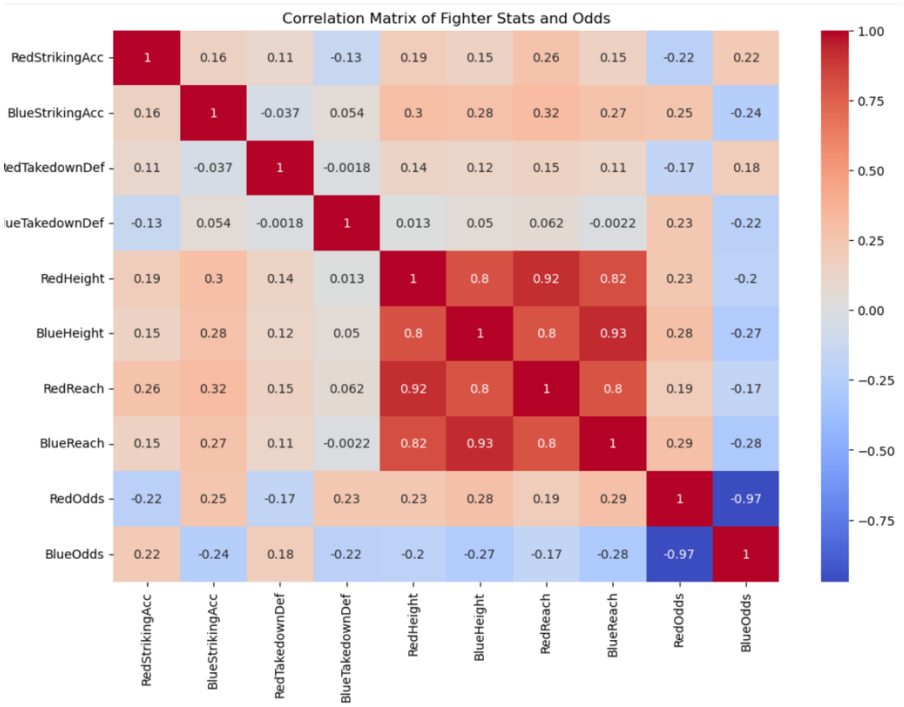

Figure 1 Correlation Matrix of Fighter Stats and Odds

We can see that the coefficients are similar between the two figures, showing that sportsbooks may not over or undervalue fighter statistics in a significant way. The coefficient of 0.25 is similar enough to 0.3 to conclude that the statistic is not really over or undervalued, and is more so related directly to odds.

Conclusion

This project explored the relationship between UFC fighter statistics and betting odds, looking to uncover trends and potential inefficiencies. The analysis revealed that certain fighter attributes, particularly striking accuracy and reach, have notable correlations with betting odds. For example, a higher RedFighter striking accuracy was associated with lower RedOdds, indicated greater favoritism. Additionally, reach was shown to affect perceptions of a fighter's advantage, although the relationship was non-linear across quartiles, suggesting potential biases in how sportsbooks factor physical attributes into their odds.

Through machine learning models such as logistic regression, the project demonstrated that fighter attributes alone could predict fight outcomes with moderate success (67% accuracy). This suggests that while betting odds remain a strong predictor of outcomes, there may be exploitable gaps, especially when certain statistics are undervalued or overlooked. Models using only betting odds provided a useful benchmark, and the comparison helped assess the relevance of raw fighter data versus market expectations.

Despite promising results, the project has several limitations. First, the dataset was limited to a sample of 209 fighters across 233 bouts, which restricts generalizability. Second, certain variables such as fight-specific context such as injuries were not included due to the data not being available, which may affect the accuracy of predictions. Finally, a dropped research question about the impact of weight class on betting accuracy highlights the need for more comprehensive data to explore additional dimensions of the sport.

Future work could expand the analysis by incorporating fight-specific metrics. Additionally, refining machine learning models could further improve predictive accuracy. As the UFC continues to grow in popularity and data availability increases, combining advanced analytics with domain knowledge may offer valuable tools for analysts and bettors.