Paul Suh (phs92) ORIE 4741
Kenny Liang (ql75) ORIE 5741
Aditya Kompella (apk74) ORIE 5741

UFC Analysis

1 Introduction

1.1 Background

The UFC is among the world's most popular mixed martial arts organizations. In a mixed martial arts fight, two fighters, one who fights out of the red corner and another who fights out of the blue corner, are allowed to punch, kick, elbow, wrestle, and grapple their opponents. Traditionally, the red corner is given to the fighter who has a higher ranking and the blue corner is given to the fighter with a lower ranking. A standard UFC Fight has 3 rounds, each of which lasts 5 minutes. UFC Fights can end in either a win, a draw, or a loss. A fighter can win by knocking the opponent out with a strike (KO), submitting the opponent with a grappling technique (SUB), or getting a decision victory on the judges' scorecards (DEC). A fight can end in a draw if a fight lasts the full duration and the majority of the judges score the bout to be a draw.

1.2 Problem

Throughout this project, we want to uncover which features and fighting styles are most conducive to winning. The UFC is a multi-billion dollar sports company with their top-end fighters getting paid in the tens of millions of dollars per fight. Thus, if we are able to uncover certain fighting techniques or other features that can predict the winner of a match, our analysis would be of significant business value to fighters and coaches who are training to maximize their chances of winning. In addition to fight analysis, we want to develop a betting model that is profitable on real-world test data. If our model is able to generate a profit using the live odds, sports betting companies like DraftKings could use our analysis to readjust their odds and maximize profitability.

2 Data Analysis

2.1 Datasets

The *UFC-Fight Historical Data* is a comprehensive collection (~ 6,000 rows) of data, representing every UFC fight in the history of the organization from 1993 to 2021. It includes detailed information about past fights, fighter statistics, and match outcomes. Across 144 columns, some important features include average significant strike landed, fighter ranks, submissions attempted, and other useful data. We used another dataset, the *Ultimate UFC Datase*t, in order to get the live betting odds for each fight. The combination of these two datasets provides comprehensive and detailed information about each fight.

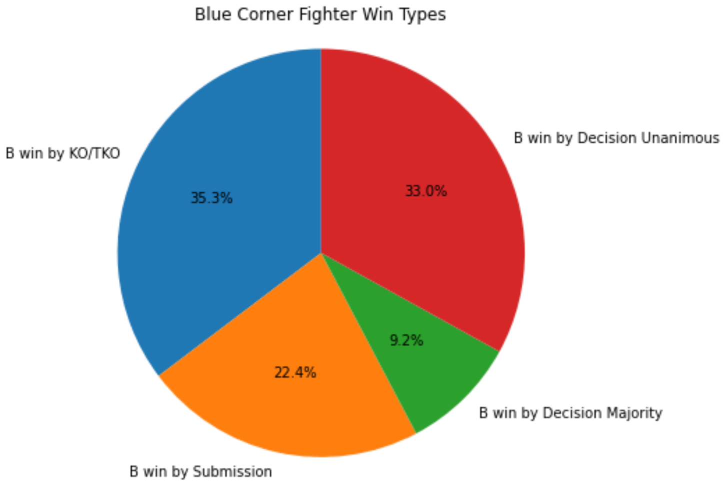
2.2 Data Processing

We mainly performed data analysis on the *UFC-Fight Historical Data* dataset, which contained many columns that required cleaning. First, we identified columns with missing values and created new features to indicate their absences (e.g., is_nan_<column_name>). We then fill the original value with 0 to maintain consistency, as well as suggesting to our models that this player is likely a rookie and therefore less competitive. For columns with categorical data, we performed one hot encoding to convert them into numerical features, allowing us to better capture patterns and relationships. Next, boolean columns simply require a conversion to 0 and 1s. Lastly, an important step was to perform label encoding on our target/prediction column, where we transformed the "Winner" column into a numerical feature: 0 for Blue Corner, 1 for Draw, and 2 for Red Corner.

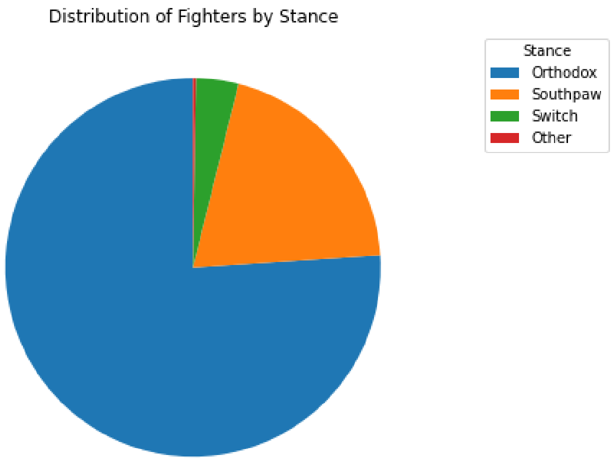
2.3 Data Visualization

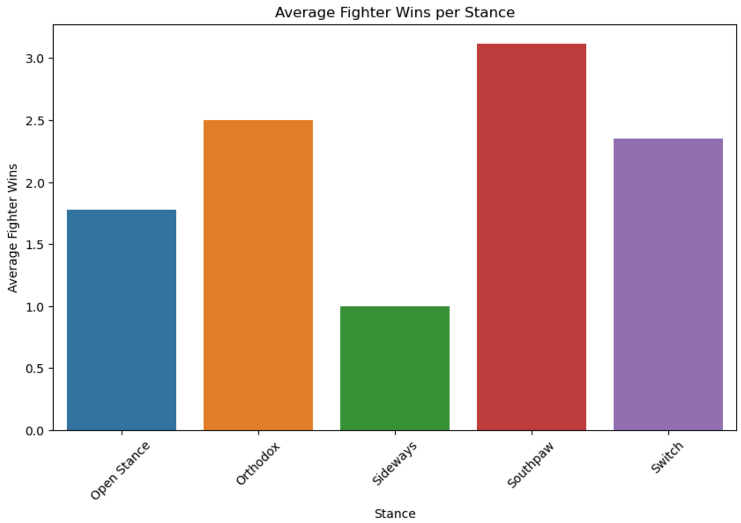
2.3.1 Win Types

In the pie chart to the right, we can see the 4 main ways that a fight can be won: by decision, by knockout, by submission, or by majority decision. A decision victory is when a fight is not ended early due to a KO or submission and instead, the judges determine who won the fight by points. Although decision victories account for the largest fraction of wins, no category has an overwhelming majority. Mixed martial arts is evidently a multidimensional sport with some fighters obtaining a decision victory with overwhelming cardio, other fights getting a knockout by trading on their feet, and others winning through their grappling technique. Fighters must be well-versed in all of these unique disciplines to be successful in the cage.



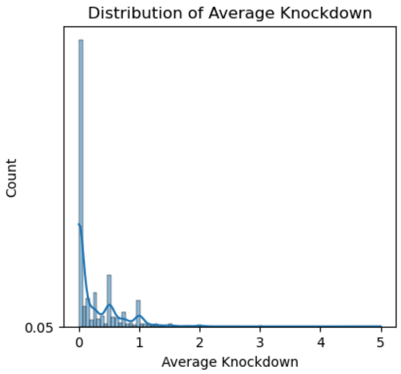
2.3.2 Fighter Stances

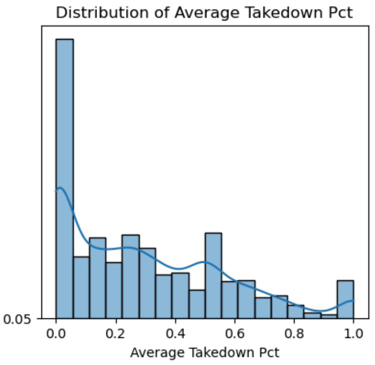


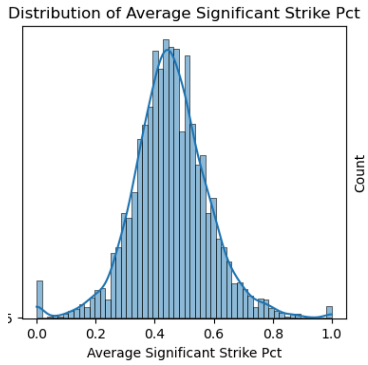


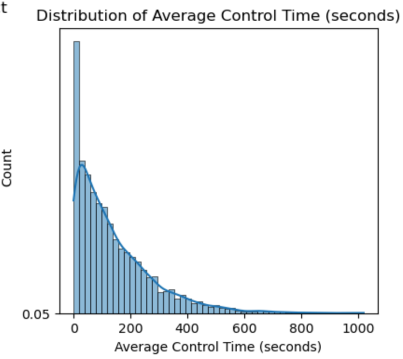
In the visualizations above, we explore the different stances that fighters can employ. There are 3 main stances: orthodox where the fighter puts their left foot forward, southpaw where the fighter puts their right foot forward, and switch where the fighter alternates between orthodox and southpaw. As can be seen in the pie chart, the most popular fighting stance is orthodox. However, when examining the average fighter wins per stance. It can be seen that Southpaw fighters seem to have the highest number of average wins. This can be attributed to the fact that Southpaw fighters are significantly rarer so Orthodox fighters often do not have much training or fighting experience against Southpaws. Thus, trainers and fighters may consider adopting more unique stances due to the advantages they provide.

2.3.3 Fighting Tactics



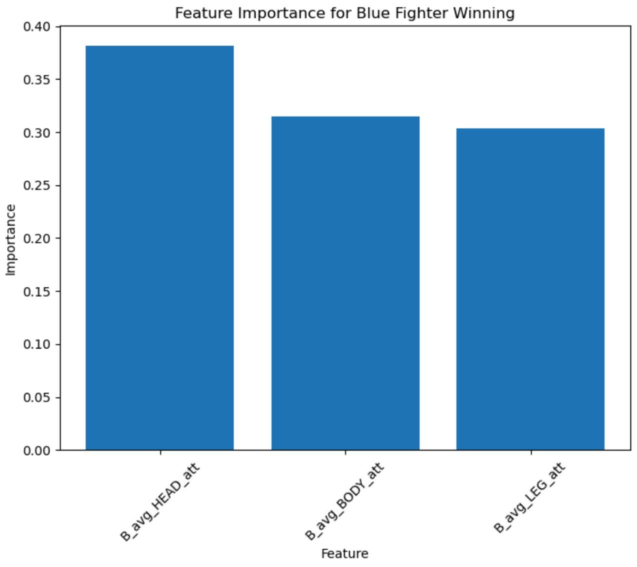






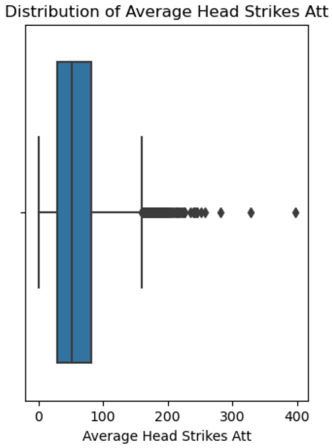
As discussed earlier, fighters employ a variety of different tactics. Some fighters are pressure fighters who take control of the ring while other fights are strong wrestlers who take their opponents down. In the KDE graphs to the left, we graphed the distribution of 4 features that epitomize different fighting styles. Fighters can use these graphs to benchmark their skills and visualize their strengths/weaknesses as compared to the rest of the UFC fighters.

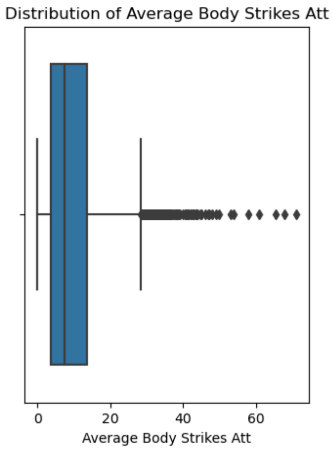
2.3.4 Striking

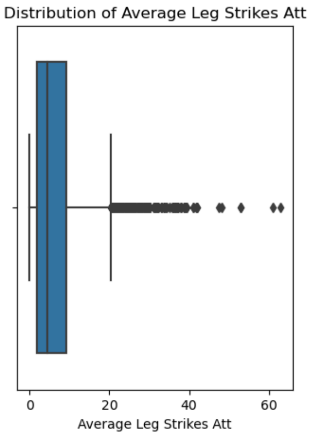


There are 3 main targets for a strike: the body, the leg, and the head. In order to generate the graph on the left, a random forest was trained based on just 3 features: the number of head strikes attempted, the number of body strikes attempted, and the number of leg strikes attempted. Then, the feature importance was measured. The 3 box and whisker plots down below show the distributions of body, leg, and head strikes. We can see that head strikes are much more frequently attempted than body and leg strikes.

From these two graphs, it appears that body and leg strikes are vastly underutilized by fighters – they are thrown with very low frequency relative to head strikes despite the fact that body and leg strikes have a high feature importance.



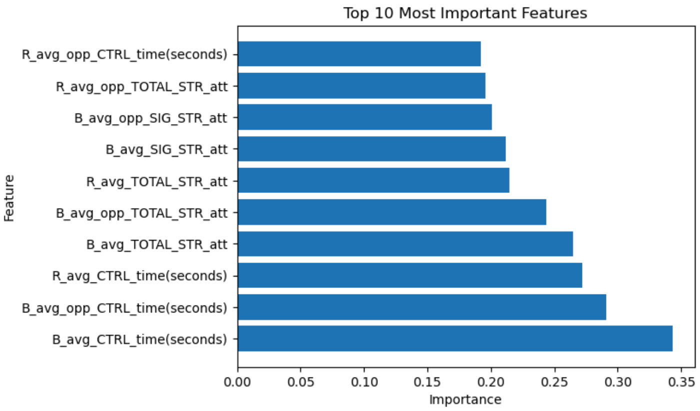




3. Algorithms

3.1 Principal Component Analysis

The dataset currently contains a large number of features. Given that we only have about 6000 rows, it is very likely our model will overfit if we use all of the features. Thus, we want to choose a subset of these features when training our models to improve generalization as well as make our models more interpretable. To do this, we utilized Principal Component Analysis (PCA) to compute the principal components. Consequently, we can reduce the dimensionality of the dataset while retaining the most significant features that contribute to the variance in the data. This can be particularly useful for reducing noise, eliminating redundant or irrelevant features, and improving the interpretability and computational efficiency of training models.

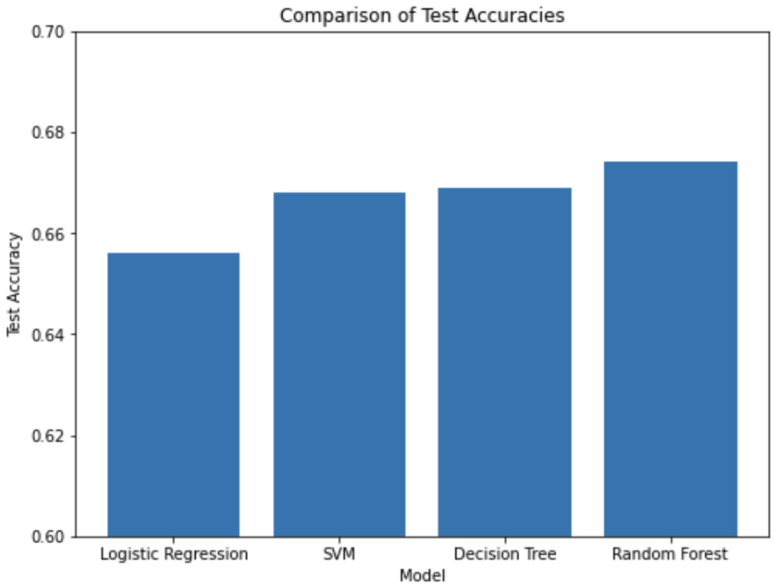


Results: After applying the data preprocessing techniques mentioned earlier, we identified the top ten most important features through our PCA analysis. The results suggest that the average control time is an important stat to keep in mind. It shows that landing significant strikes alone is not always the key to winning. Considering our previous analysis, since matches often don't end in submissions or knockouts, having control of the ring can be a decisive factor in winning through decisions.

3.2 Model Selection

After conducting feature selection, we represented our data using only the most important features. We proceeded to train four distinct models on the preprocessed dataset. To ensure unbiased evaluation, we employed an 80/20 train-test split and trained each model on the training subset. Since our objective involved classification, we opted for a diverse set of classification algorithms, encompassing both simple linear models and more sophisticated nonlinear methods.

3.2.1 Logistic Regression with L2 Regularization



Result: We began by using Logistic Regression, a simple yet effective linear model, to establish a baseline for our predictions. To prevent overfitting, we employed L2 regularization, which adds a penalty term to the loss function to discourage large weights. This technique is particularly useful when dealing with high-dimensional data, as it helps to reduce the impact of noisy or irrelevant features. By applying PCA to reduce the dimensionality of our dataset to around 20 features, we further mitigated the risk of overfitting. The resulting model achieved a test accuracy of approximately 65.5%, significantly better than randomly guessing the winner (50% accuracy). This suggests that our model was able to identify useful patterns in the data that are indicative of UFC match outcomes.

3.2.2 Support Vector Machine with a regularization parameter C of 1 and an RBF Kernel

Result: Next, we turned to the Support Vector Machine (SVM) algorithm, which aims to maximize the margin between classes. By using an RBF kernel, we enabled the SVM to capture non-linear relationships between features. The SVM's ability to find the optimal hyperplane in the feature space makes it well-suited for datasets with complex boundaries. In our case, the SVM achieved a test accuracy of around 67%, outperforming the Logistic Regression model. This improvement can be attributed to the SVM's ability to handle non-linear relationships and its robustness to noise and outliers.

3.2.3 Decision Tree with a Gini Impurity criterion and a maximum depth of 5

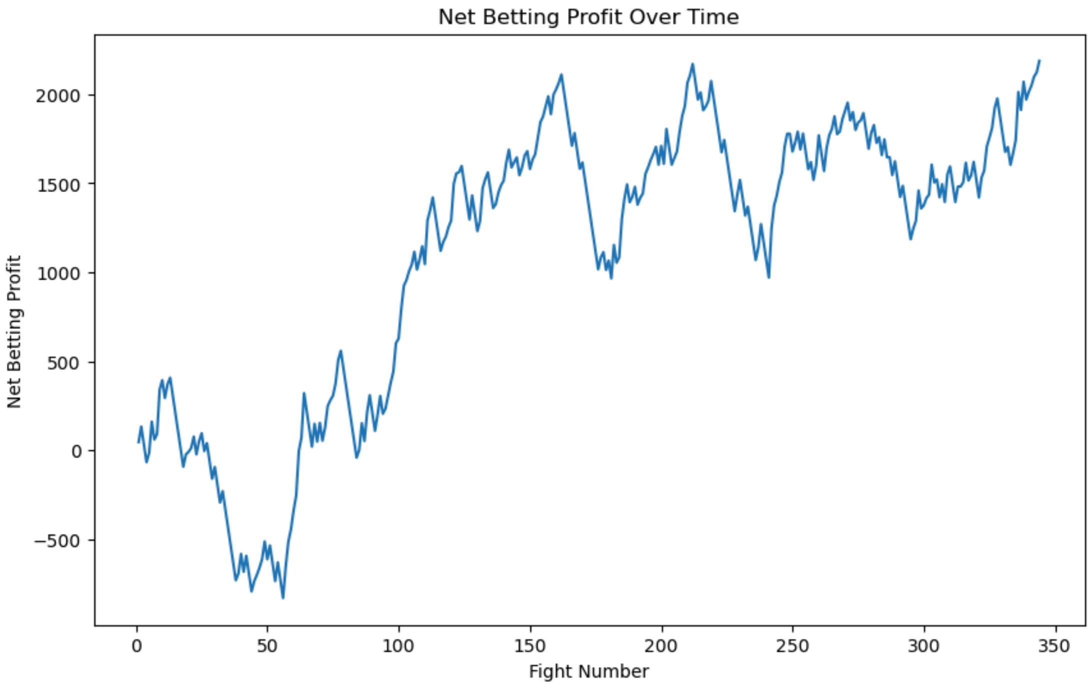
Result: We chose the Decision Tree algorithm for its ability to handle non-linear relationships and identify important features. By using the Gini impurity criterion, we ensured that the tree was optimized for classification tasks. The maximum depth of 5 was set to prevent overfitting and promote generalizability. The Decision Tree model achieved an accuracy of approximately 67.5%, slightly improving upon the linear models. This suggests that there are non-linear connections between features that are important for predicting UFC match outcomes. The relatively shallow tree depth also implies that the most important features are likely to be those that are most strongly correlated with the outcome.

3.2.4 Random Forest with a Gini Impurity criterion, 500 trees, and a maximum depth of 9

Result: Finally, we employed the Random Forest algorithm, which combines multiple Decision Trees to improve generalizability and reduce overfitting. By using a bagging approach, Random Forest can handle datasets with limited samples, such as ours, and provide more robust predictions. With 500 trees and a maximum depth of 9, our Random Forest model achieved the highest test accuracy of around 68%. This improvement can be attributed to the ensemble effect, where the collective wisdom of multiple trees leads to more accurate predictions. The Random Forest's ability to handle complex interactions between features and its robustness to noise and outliers make it a suitable choice for our dataset.

3.3 Real-Life Performance

In order to assess the performance of our models in a financial context, we sought to apply them to sports betting scenarios. Initially, we filtered the training set to exclusively encompass fights that occurred from 1993-2020, and trained a model on these years. Our test set was composed of fights in the year 2021. The reasoning behind this approach is to have a model that learns from past data and predicts on data in the “future”, giving us an unbiased estimate of how our model will perform on real-world data. We then performed a join/merge operation of our test dataset with another dataset to acquire the live betting odds for each fight. When using our model on the test set, we executed a betting strategy whereby we wagered $100 on the fighter predicted by the model to win, be it the Red or Blue-side fighter. Finally, we recorded the net profit generated by our algorithm over the course of the betting year using our model.



Result: Our model generated a net profit of $2186 over 350 fights. The graph shows how the net profit changed over time while taking into account the $100 we spent each day on betting. On average, our model is quite accurate, allowing us to make significant profits over a long period of time. This translates to an average profit of 6.25% per game, which is impressive considering the unpredictable nature of UFC matches, including lucky knockouts and decisions. These results demonstrate the successful application of our model in financial contexts.

4 Conclusion

4.1 Key Takeaways

Based on our models and data analysis, it is clear that some techniques and features seem to correlate strongly with winning. We explored how fighting stance, control time, body and leg strikes, and more all correlate with a fighter's chance of winning. Trainers can adopt some of the analytical insights we provide to optimize their fighter's training and better prepare them for the ring. Secondly, since our betting model was able to generate a non-trivial profit, we were successful at creating a model that was more accurate than the "crowd" or the betting odds. Sports betting companies may find our models and data analysis useful to set more accurate odds.

4.2 Confidence and Improvements

We performed multiple rounds of testing and validation; each step confirming that our model is roughly 65% accurate and significantly better than guessing. However, before deploying to a production setting, we would want to rescrape a more recent data of the UFC, particularly the fights that occurred throughout the years 2021-2024, and retrain our model.

4.3 Fairness and Ethics

The outcome we are attempting to predict (the winner of a fight) is very measurable and we are training our models against roughly 6000 fights. Moreover, we use sound experimental design methods with train/test splits. We obtained similar prediction accuracies with multiple different models, further validating our results. We advise all users of our predictive model to be aware of the inherent risks and volatility related to sports gambling. We do not believe our model is a Weapon of Math Destruction (WMD) and we believe our models will benefit both the sports betting community, UFC trainers, and fight fans.

4.4 Contributions

We pair-programmed the majority of the project together and were very collaborative. However, Aditya Kompella took more ownership of the model training, whereas Kenny Liang and Paul Suh took more ownership of the data visualizations and data pre-processing.

4.5 References

https://www.kaggle.com/datasets/rajeevw/ufcdata
https://www.kaggle.com/datasets/mdabbert/ultimate-ufc-dataset

4.5 Source Code

https://github.com/KennyLiang2302/ufc_prediction