Octagon Insight Analysis

Group members: Deep Bhimani, Ritik Yadav, Jiachen Sun (Jason)

Data Source: Ultimate UFC Dataset from kaggle.com

Introduction

In the world of combat sports, the Ultimate Fighting Championship (UFC) stands to be at the top by captivating audiences worldwide with its thrilling matches. As enthusiasts of both data science and the UFC, our team is drawn to dive into the undiscovered insights lying in the adrenaline adrenaline-fueled arena.

The UFC was founded in 1993 and has become a global phenomenon by having elite fighters from all around of world fight recklessly. UFC has become the center of attention for talent, attracting fighters from all round of work with their unique fighting style and background. Did you know that the fastest knockout in UFC history occurred a mere 5 seconds into the fight? Or that the UFC octagon, measures 30 feet in diameter?

Goal/Aim

As we delved deeper into the world of UFC, we uncovered fascinating information that raised our curiosity and fueled our passion for data-driven exploration. Our exploration led us to formulate the following three questions for answers:

1) Which factor exerts the most significant influence on a fighter's chances of victory? Could it be their height, fighting stance, age, weight, country of origin, etc?
2) Using a machine learning model, can we accurately predict the outcome of a UFC fight when presented with information about two competing fighters?
3) How do various attributes of fighters evolve over the years? Are there discernible trends or patterns that emerge from analyzing these changes?

Data Gathering and Cleaning

After scouring the internet for datasets on UFC, we initially stumbled upon three separate datasets, each promising to fulfill our data needs. Our strategy was to clean these datasets individually and then merge them. However, a significant roadblock emerged when we realized that upon joining the tables, almost half of the rows would be lost due to inconsistencies in fighter data, rendering the merged dataset unreliable. Undeterred, we scoured the internet once again and fortuitously discovered the perfect dataset: fight_data.csv. This dataset comprehensively encapsulated fighter information from 2010 onwards, filling the gap left by the previous tables.

With the right dataset in hand, our next task was to meticulously clean it as seen in the Cleaning_data.ipynb. We carefully went through each of the 119 columns to ascertain their relevance. We judiciously removed 46 columns such as R_odds, B_odds, R_ev, B_ev,

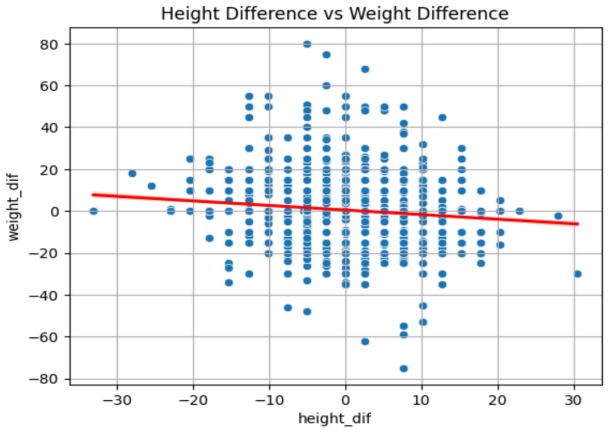
B_current_lose_streak, B_current_win_streak, etc. Recognizing their lack of relevance to our analysis. Addressing the issue of empty rows, we opted for a strategy of imputation, replacing missing values with the mean of their respective columns to maintain data accuracy. This approach was applied to columns like B_avg_SIG_STR_landed, B_avg_SIG_STR_pct, B_avg_SUB_ATT, etc.

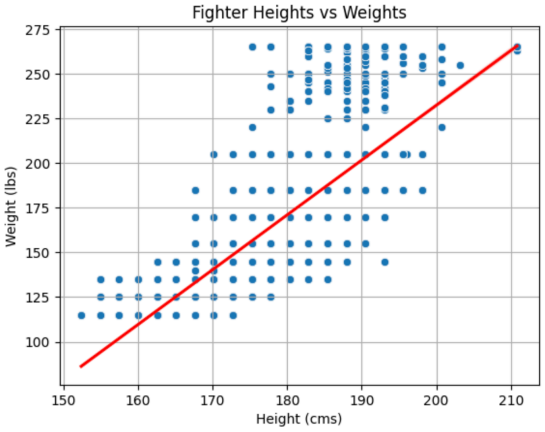
In the case of column B_Stance, which exhibited only two missing values, we conducted thorough research online to ascertain the correct stance for the fighters, ensuring data integrity. For pivotal columns such as finish_round and finish, conventional imputation methods like mean calculation were deemed inappropriate. Instead, we employed a method based on the distribution of non-empty values within these columns, filling in missing entries to align with the established distribution. For columns such as finish_details and finish_round_time, where the data held significant contextual importance, we introduced static values informed by our understanding of UFC fights. Additionally, we performed straightforward calculations to convert the total fight time from its original format into seconds for the total_fight_time_secs and the Weight_dif columns.

Lastly, we addressed the issue of messy numerical values in columns like Height_dif, Reach_dif, Sig_str_dif, etc by standardizing them to two decimal places, ensuring consistency and ease of interpretation in our dataset. In the country column, there were 2 values for the USA, one with an extra space that had to be removed. This comprehensive data-cleaning process took the most amount of time but laid a solid foundation for our analysis and interpretation as seen in fight_data_cleaned.csv.

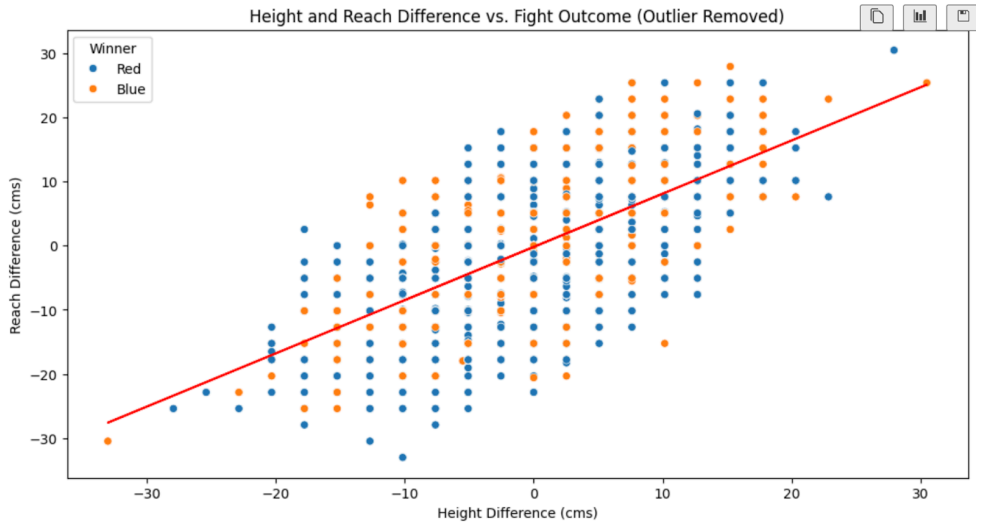
Statistical analysis

To understand the data a bit further to help us work towards our goal/aim questions, we created basic graphs in the statistical_graphs.ipynb file. Basic graphs include the distribution of fight time, average significant strike rates, men-to-women ratio, winners by stance, fight outcomes by weight class or country, Average Significant Strikes Landed by Weight Class, etc. These graphs laid a foundation for understanding the data.



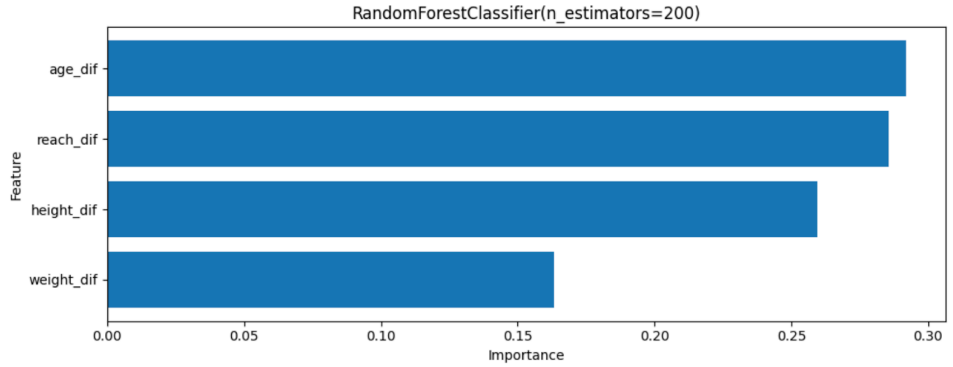


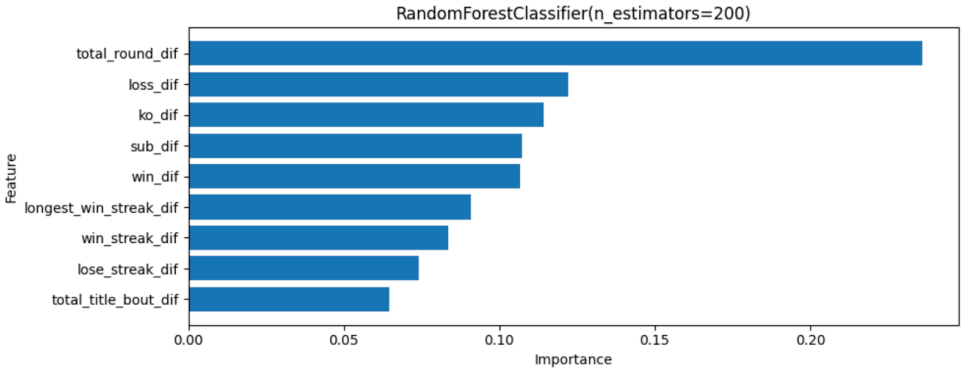
In the same statistical_graphs.ipynb file, to dive deeper, we made a height difference vs weight difference graph and found the Correlation Coefficient (r-value): -0.148. This shows a weak negative linear relationship. However, we calculated a Correlation Coefficient (r-value) of 0.80 for the fighter heights vs weights showing a strong positive linear relationship. The taller the fighter is the more he weights. Since we see negative weak linear relationships, this shows that in UFC the matches are often fought with fighters with similar heights/weights.



To further analyze the physical aspects, we created a height and reach difference vs fight outcome graph. For example, if the height is negative this means that the red fighter is taller than the blue. The red line shows a Correlation coefficient (r): 0.64, being a strong positive correlation. We can depict that more than height, reach plays a bigger role in terms of winning the match. Throughout the x-axis, we can see that almost equal amounts of blue and red win however in terms of the y-axis, we can see that more blue wins as they start reaching reach advantage over the red fighters.

Now to find answers to the question of which feature is important to win the fight, we created feature_imporance.ipynb file. We built machine-learning models with the features we wanted. We made two analyses, one with the fighter's physical attributes and the second one with the fighter's experience. Our test group will be (height_dif, reach_dif, age_dif, weight_dif) for the physical attribute table and (lose_streak_dif, loss_dif, win_dif, win_streak_dif, longest_win_streak_dif, total_round_dif, total_title_bout_dif, ko_dif, sub_dif) for the fighter experience table. The result group is only the Winner which is either Red or Blue.





The aim of machine learning with stature data here is not to predict the result of a fight only based on the columns, but we want to find out each feature's importance in influencing the winning of a fight, we chose a random forest model and decision model as our main machine learning methods, since they are both tree-based models, and it's easy to utilize their built-in feature_importance__ attribute. The accuracy of the two models for both charts is around 51%, which demonstrates that it's hard to determine winners based on their builds for fights. From the first chart, we can see age has the most importance around 0.28 followed by reach, height, and weight. In the second chart, of the player's experience we can see, that the number of rounds a fighter has an importance of around 0.30, followed by amount of losses, amount of K.Os, etc. These insights are fascinating to see because they give behind-the-scenes insights into the UFC fights.

Model Development and Comparision of Models

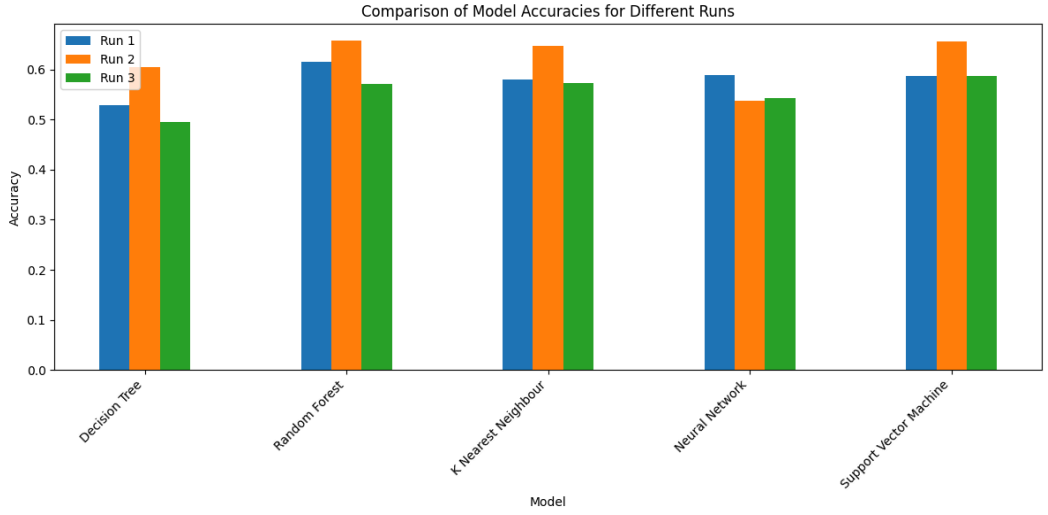
In this section, we used the cleaned data from the first section to develop models for predicting the outcome of the match. We used five models for comparison which are DecisionTreeClassifier, RandomForest Classifier, KNNClassifier, MLPClassifier, and Support Vector Classifier. This model development is done in 3 iterations.

In the first iteration, we use the data cleaned from our intuition in the first section. We use a StandardScaler and LabelEncoder to preprocess the numerical and categorical columns respectively. We then create a test train split of a 25:75 ratio. The model list consists of a RandomForest Classifier tuned in with the following parameters: n_estimators = 400, max_depth = 12, random_state = 2. We agreed to use it after training the data multiple rounds with different values for these parameters. For the KNN classifier, we have used

n_neighbours at 100. For the MLPCLassifier, we used internal layers of size(64,128,64) and a max_iter of 300.

The second approach involved using the raw data from the start and sorting the columns based on their correlation value with the result('Winner') column. First, we discarded the columns where more than 40 percent of the data was missing. Then, we found the correlation for each column with the label and discarded all the numerical columns with a correlation below 5 percent. Next, we used the process of scaling and labeling the data described in process 1 and used the same set of models with the same parameters for training the data.

The third approach was about using PCA to reduce the dimensionality of data. We started with 119 columns. We then dropped the columns where we had more than 40 percent of the data missing. Then, we performed the scaling and labeling operations. Next, the process was to use PCA. Before using it we had the size of data as (4698,88). We wanted to preserve 99 percent of the variance and the result was surprising. We ended up with a feature matrix of size just (4698,3). We then used this data from PCA to train on the same set of models and plotted a chart comparing the models and iterations which is shown in the figure below



Now, We ran the code multiple times and there were minor changes in the result each time. However, the results were consistent with the pattern shown above. The best prediction came at close to 66 percent by RandomForest and SVM using approach 2. There was not a huge change in prediction percentage across the iterations but we do see that Method 2 was the most successful one (except for the case of MLP). On closer inspection, we see that the Neural Network model (MLP) demonstrated the least susceptibility to the method used. In fact, it was the only model to perform best with the first approach. We could also say that RandomForest Classifier performed best on average across all iterations and SVM and KNN came a close second to it.

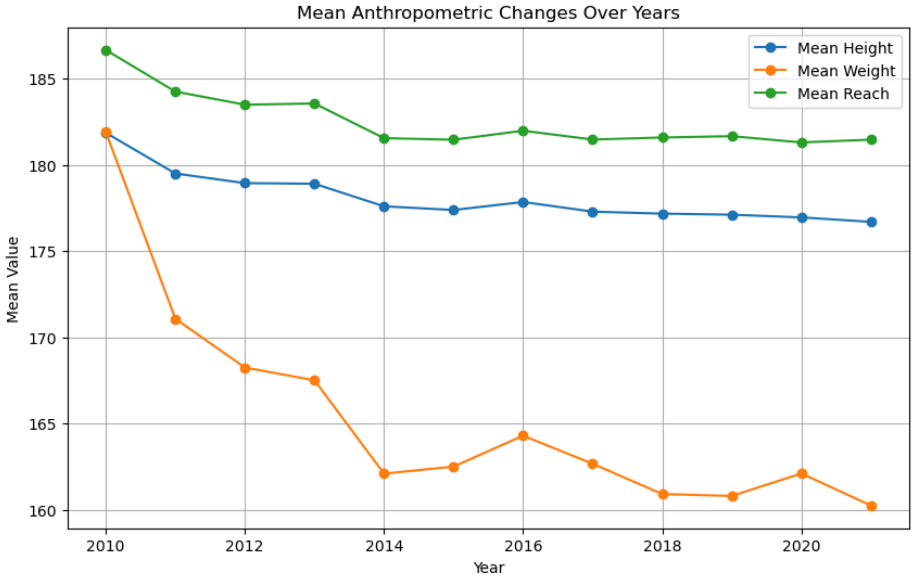
However, we had guessed that the results from methods 2 and 3 would be closer to each other because both of them rely indirectly on using correlation techniques. However, from the

diagram, we can say that our assumption was wrong. We are not too sure why it happened but we can guess Method 2 relies on correlation between features and the target variable to select relevant features. However, correlation does not always capture the full relationship between variables. On the other hand, method 3 directly reduces dimensionality without considering the target variable and thus, potentially removing features that may not be highly correlated with the result but still contribute to the prediction.

Stature Evolution Analysis

Moving on, to answer our 3rd question, the evolution of fighter builds through the 12 years (2010 - 2021). We divided data by the year of fights and put every fighter's weight, height, and reach in each year to make a comparison on certain year's fighters' data. The results might be surprising that fighters are turning shorter and shorter over the years. We noticed that the mean value of weight, height, and reach reduced till 2014, and after 2014 the value remained the same at least in the graph.

The year with the fighters who have the highest average weight, height, and reach data is 2010, and they are 181 cm in height, 181 lbs in weight, and 186 cm in reach. The weight dropped the most rapidly to 160 lbs by 2021. Average height also changed to 177 cm, and reach goes to 181 cm. This shows that UFC had more fighters joining in the past 10 years, but their weight and height requirements are not as high as they used to be 10 years ago.



Finally, we decided to perform a statistical test to check if the data from 2010 were the same as 2021 and if 2014's data were the same as 2021's. It involved performing equal variance testing and normality testing on each of the columns. For that, we used stats. normaltest and stats.Levene method respectively. If the data passes this test, we use t_test otherwise we perform the Mann-Whitney test. The value of alpha used is 0.05.

First, we compared 2010 and 2021 data. By testing the normality of each feature, we discovered that their reaches are normally distributed but their heights and weights are not following the normal distribution. Especially the weight of UFC fighters in 2021 compared to

the data in 2010, is severely skewed because more UFC fighters are trying to keep low weight to join the lower weight class fights. Also, none of the columns passed the normality and levene test. Therefore, we used the Mann-Whitney test for all three features: height, weight, and reach and for all of those tests, the p-values came <0.05 which suggests there is sufficient evidence that the mean of these values is not the same (which is pretty obvious from our graph).

Next, we did the same operation for 2014 and 2021 data (because they looked close to the same in the graph). The following was the result.

```
Column: Height_cms
Normality test p-value for 2014: 0.8725009334260156
Normality test p-value for 2021: 1.36271849228339e-05
Equal variance test p-value: 0.007436955675790429
Mann-Whitney U test p-value for Height_cms: 0.09182121220531972
Column: Reach_cms
Normality test p-value for 2014: 0.5435872833719692
Normality test p-value for 2021: 0.055599071379003016
Equal variance test p-value: 0.01578365915529554
Mann-Whitney U test p-value for Reach_cms: 0.9544420518099087
Column: Weight_lbs
Normality test p-value for 2014: 2.819848859203893e-26
Normality test p-value for 2021: 1.1348780880383478e-22
Equal variance test p-value: 0.00532769211493841
Mann-Whitney U test p-value for Weight_lbs: 0.023760774363718498
```

From this data, there is sufficient evidence to say that the mean height and reach were the same in 2014 as compared to 2021. However, the same cannot be implied for weight.

Limitations

The analysis presented in this report is subject to several limitations that warrant consideration. Firstly, despite efforts to obtain comprehensive UFC data, reliance on a single dataset from 2010 onwards may have introduced biases and missed historical trends, potentially limiting the generalizability of findings from 1993 to 2010 and 2021 to 2024. Additionally, challenges in data cleaning and imputation techniques, particularly the subjective nature of resolving missing data and potential biases introduced by imputing missing values with column means, may have impacted the accuracy of the analysis.

Moreover, the predictive models' modest accuracy rates around 66% highlight the inherent complexities of UFC matches, suggesting that other unaccounted factors beyond fighter attributes significantly influence fight outcomes. Furthermore, the simplicity of feature selection and model complexity may have overlooked critical predictors, such as fighting style and psychological factors, contributing to oversimplified models. Lastly, temporal trends analysis and assumptions regarding data normality may overlook historical shifts and introduce methodological constraints, impacting the completeness and robustness of insights gained.

Accomplishment Statements

Deep Bhimani

- Performed data cleaning to ensure the accuracy and reliability of the dataset.
- Utilized basic statistical models to derive fundamental insights from the data.
- Applied machine learning techniques to analyze feature importance.
- Authored the introduction, my findings, and outlined limitations in the report.

Ritik Yadav

- Worked on creating models for the prediction part of the project.
- Developed insights and visual representation of data from the accuracies of the trained model
- Recleaned parts of data for use in model prediction and helped in adding parts to the evolutionary data

Jiachen Sun (Jason)

- Utilized machine learning techniques to analyze evolutionary stature data.
- Revised and organized evolutionary data for enhanced visualization, employing data plotting techniques.
- Produced a part of the comprehensive reports and maintained updates on the GitHub page.