UFC Visualization and Winner Prediction

Ishwa Obaid - 326869
Rameesha Gull - 327317
Muhammad Shoaib - 320124
Urooj Fatima - 329471

Abstract

In this report we have used the ultimate UFC dataset. There are three main parts of this report. The first one is Baseline analysis in which we have built a simple baseline model with basic data collection for data cleaning. Secondly, we did data visualization and used the power of visualization to gain a deeper insight into the data. We visualized some important insights using python language and specifically matplotlib library. Then, we did feature engineering and built a final model. Finally, we used feature engineering to strengthen our input and build a predictive model that gives upcoming fights predictions.

I. Introduction

UFC is one of the largest promotion companies, it stands for ultimate fighting championship. It is popular internationally as it proved fans with a spectacle, allowing them to witness the use of martial art techniques and skills of the fighters. There are very few rules and restrictions in UFC. For every bout there are several features that makes it a little difficult to foresee the outcome and winners. Placing bets on fights is a popular trend, but it is important to understand the data and predict odds of winning in a particular bout [1]. In this report, we presented easy to understand visualization to make it easier to interpret the data of bouts. Moreover, we have used a machine learning model to predict outcome of a bout for any two players.

II. Data Abstraction

The UFC dataset is taken from is work of Mdabbert and is taken from Kaggle [2]. It merges all public UFC datasets present on Kaggle. It contains data of all bouts from mid-2010 onwards to present.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Winner	title_bout	B_avg_KD	B_avg_op	B_avg_SIG	B_avg_op	B_avg_TD	B_avg_op	B_avg_SUI	B_avg_op	B_avg_RE	B_avg_op	B_avg_SIG
2	Red	FALSE	0	0	0.42	0.495	0.33	0.36	0.5	1	0	0	50
3	Red	FALSE	0.5	0	0.66	0.305	0.3	0.5	1.5	0	0	0	65.5
4	Red	FALSE	0.015625	0	0.45	0.4275	0.25	0.2	0.148468	0.098389	0	0	66.5
5	Blue	FALSE	0.015625	0	0.45	0.4275	0.25	0.2	0.148468	0.098389	0	0	66.5
6	Blue	FALSE	0.125	0	0.535625	0.57875	0.185	0.16625	0.125	0.1875	0.25	0	109.1875
7	Blue	FALSE	0	0	0.515	0.47375	0.435	0.25	1.25	0.625	0	0.25	86
8	Blue	FALSE	0.046875	0.125	0.459277	0.404688	0.322188	0.068867	0.189453	0	0.283203	0	130.8184
9	Red	FALSE	0.15625	0	0.791582	0.850586	0.269199	0.393398	0.029297	0.080078	0.023438	0.041016	39.21094
10	Red	FALSE	0.015625	0	0.45	0.4275	0.25	0.2	0.148468	0.098389	0	0	66.5
11	Red	FALSE	0.015625	0	0.45	0.4275	0.25	0.2	0.148468	0.098389	0	0	66.5
12	Blue	FALSE	1	0.625	0.47375	0.24125	0.28125	0	0.875	0	0	0	71.25
13	Blue	FALSE	1.0625	0	0.473125	0.663125	0.60125	0.105	0.5	0.0625	0.0625	0.5	49.3125
14	Red	FALSE	0.015625	0	0.45	0.4275	0.25	0.2	0.148468	0.098389	0	0	66.5
15	Red	FALSE	0.015625	0	0.496641	0.616719	0.327187	0.090938	1.59375	0.148438	0.179688	0.640625	75.90625
16	Blue	FALSE	0.5	0.625	0.464375	0.398125	0.48375	0.048125	0.5	0.34375	0.28125	1.0625	115.0313
17	Red	FALSE	0	0.3125	0.529375	0.4075	0.5	0.160625	1.3125	0.3125	0.375	0.25	158.8125

Figure 1: UFC dataset

As shown in figure 1, dataset includes complete history of UFC fighter rankings. There is data of about 4567 bouts with 137 features each. The dataset comprises of quantitative data. The data is in a tabular form where row represent items and column shows attributes. Here a bout is an item having both quantitative and categorical attributes. Each item has several attributes such as winner of bout, stats of bout, locations etc. Moreover, detail description of every fighter is also given in the dataset. A sample can be seen in figure 2. The two fighters are represented according to their team red or blue.

	A	B	C	D	E	F	G	H	I	J
1	fighter_na	Height	Weight	Reach	Stance	DOB	SLpM	Str_Acc	SApM	Str_Def
2	Tom Aaron		155 lbs.			Jul 13, 197	0	0%	0	0%
3	Papy Abe	5' 11"	185 lbs.		Southpaw	Jun 30, 19	2.8	55%	3.15	48%
4	Shamil Ab	6' 3"	235 lbs.	76"	Orthodox	Sep 02, 19	2.45	44%	2.45	58%
5	Danny Ab	5' 11"	155 lbs.		Orthodox	Jul 03, 198	3.29	38%	4.41	57%
6	Hiroyuki A	5' 6"	145 lbs.		Orthodox		1.71	36%	3.11	63%
7	Ricardo Al	5' 11"	185 lbs.		Orthodox	Apr 27, 19	3.79	31%	3.98	68%
8	Daichi Ab	5' 11"	170 lbs.	71"	Orthodox	Nov 27, 19	3.8	33%	4.49	56%
9	David Abb	6' 0"	265 lbs.		Switch		1.35	30%	3.55	38%
10	Klidson Al	6' 0"	205 lbs.	74"	Orthodox	Dec 24, 19	2.05	40%	2.9	55%
11	Daniel Ac	5' 8"	180 lbs.		Orthodox	Dec 27, 19	3.52	36%	2.85	62%
12	Scott Adar	6' 0"	225 lbs.		Southpaw		0	0%	0	0%
13	Juan Adan	6' 5"	265 lbs.	80"	Orthodox	Jan 16, 19	7.09	55%	4.06	34%
14	Anthony A	6' 1"	185 lbs.	76"	Orthodox	Jan 13, 19	3.17	41%	5.93	44%

Figure 2: Fighter information in the dataset

III. Task Abstraction

The data can be used for several tasks but the ones that we focused on a few important ones. Most important is analyzing the dataset to find trends in the data. Moreover, it is used to compare information of bouts and fighters and identify outliers in various features such as locations, weight classes etc. Furthermore, the winner of a bout can be predicted on the basis of multiple attributes.

IV. Implementation

We have visualized Ultimate fighting dataset using python language and at the end developed a predictive model for UFC data and our model is predicting the winning percentage of red corner and blue corner. There are more than 130 features in this dataset, and we can even create more than 100 visualizations, but we have chosen most important part of the data.

A. Language and libraries used for visualization

Python language is used for these visualizations and panadas library used for basic data preprocessing and for the visualization purpose we used matplotlib pyplot and patches libraries.

B. Visulaizations

We made 3 visualizations to present the data effectively:

- Bar charts to visualize winner teams for the bout.

- Bar plot to visualize the weight class feature in the dataset.
- Pie chart to present numbers of fights in particular locations.

1) Bout's winner distribution

a) Viusalization

We used bar graph to show the overall winners of the bouts. Y-axis show the number of bouts and x-axis show the players (Red and Blue teams). We used this visualization as it represents the data in effective and easy to understand form. This form of representation makes it easier to know which team has better odds of winning without extensive study of dataset. Moreover, the information is presented in a form that reduces the cognitive load and is easy to remember.

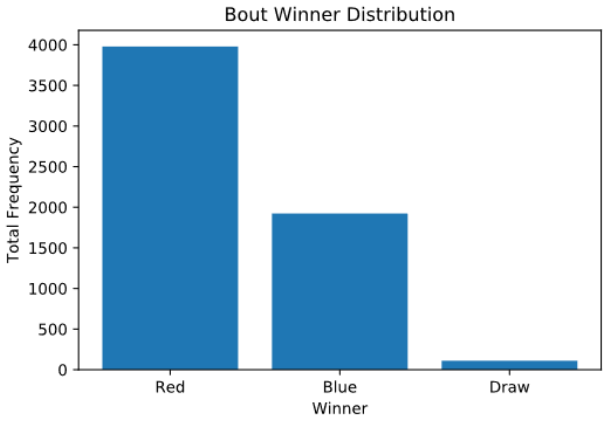

Figure 3:Bout Winner Distribution Bar Graph

b) Evalutation:

The purpose of the graph is shown by its title and the axes convey the data being used. The graph is simple and conveys one information i.e. frequency of wins. Grouping by length makes the graph easy to read. Here magnitude channels (length of bars and horizontal position on aligned scale) are used because data is quantitative. This graph is effective because position and length are used to present attributes.

2) Weight Class Distribution

a) Viusalization

- This is a bar graph that shows the distribution of players into different weight classes.
- The length of the bar represents the number of players in a particular weight class.
- The color of the bar represents the gender to which the weight class belongs.

We used bar graph to show the distribution of players into different weight classes. Y-axis show the number of players in a particular class and x-axis show the weight classes. We used this visualization as it represents the data in effective and easy to understand form. This form of representation makes it easier to identify outlier weight classes. It also aids in comparison of different weight classes. Moreover, it makes it easier to compare the male weight classes and female weight classes.

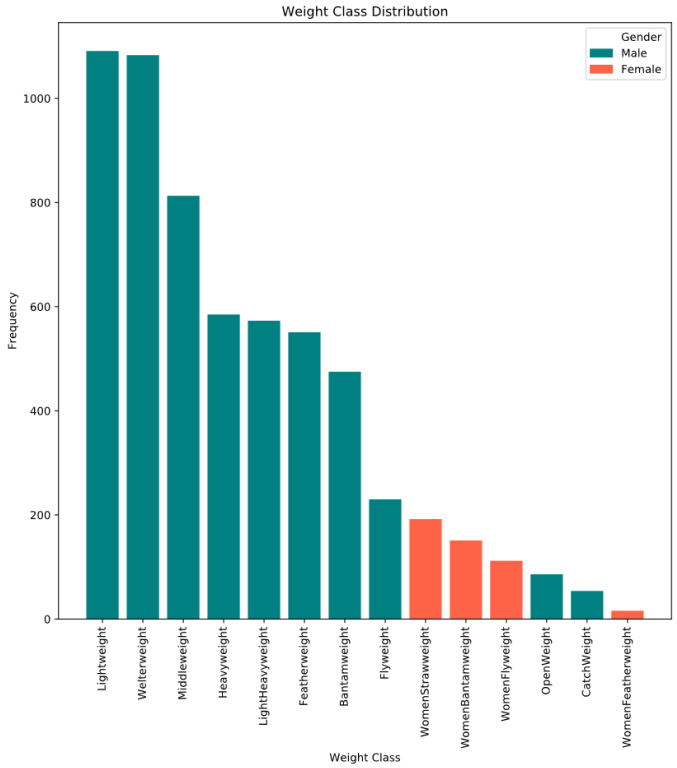

Figure 4:Wright Class Distribution Bar Graph

b) Evalutation:

The purpose of the graph is shown by its title and the axes convey the data being used. This graph conveys two types of information in a very simple manner, frequency of a particular weight class and whether that class comprises of female fighters or male fighters. Grouping by length makes the graph easy to read. Here magnitude channels (length of bars and horizontal position on aligned scale) are used because data is quantitative. Hue is used as an identity channel to distinguish among the two categories: male and female. This graph is effective because position, length and hue are the highest ranked channels

3) Bout's location distribution

a) Viusalization

We used pie chart to show the frequency of bout occurring in a particular location. We selected this form of visualization as in is easy to interpret and the information is representation in an interesting and catchy yet simple way.

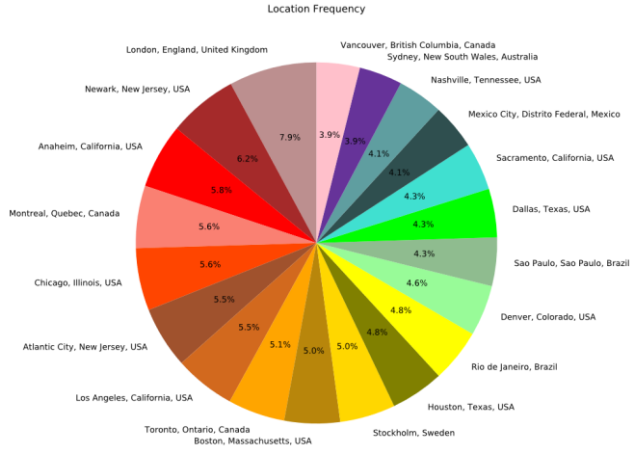

Figure 5:Bout location distribution pie chart

b) Evalutation:

The purpose of the graph is shown by its title and the radial slices convey the data being used. The graph is straightforward and conveys one information i.e. frequency of bouts in a particular location. Here magnitude channel is size of the radial slice and it displays quantitative data i.e. frequency of occurrence of bouts. Identity channel hue is used to show the categorical attribute i.e. location. The highest ranked channels are radial slices and hue.

4) Prediction Model

a) Implementation of model

We used machine learning technique Xgboost classifier to implement the prediction model because of the high speed of hyperparameter tuning and high-quality results.

We used the dataset to predict the future fight between any player and our model is giving us a very higher accuracy. User interface is very simple. Two players are selected one from red corner one from blue corner and after prediction our model shows us the winning percentage of red corner and blue corner individually and percentage or draw.

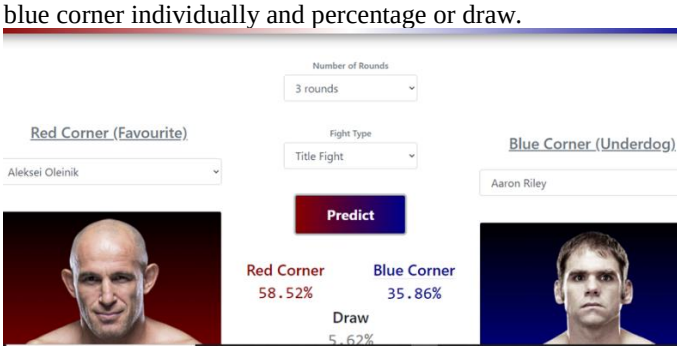

Figure 6:Prediction model employment

b) Deployment of model

We deployed the model as a webpage. Figure 3 shows the layout of the webpage. The model is available on the following link.

https://predator-ai-mma.herokuapp.com/

V. Disscusion

The ultimate UFC dataset comprises data of 4567 bouts with more than 130 features each. After visualizing the data, it is observed that there is a great difference in the number of bouts won by a team. It is clear that most bouts are won by red team and only a very few ended in a draw. Furthermore, it can be noted the division of players among weight classes is not similar for both genders. Male weight classes have a greater number of fighters as compared to the female ones. The number of female weight classes is also much smaller than male weight classes. From the visualization of bout frequency in different locations it can be seen which cities organize more bouts. If country wise frequency is viewed, it can be concluded that most bouts take place in USA. However, London UK has most bouts than any other city. It can also be seen that Canada and Australia have least number of bouts.

The visualizations have been designed to make sure that good visualization principles have been followed. Firstly, the gestalt principle of Continuity is followed in both the bar graphs. The data is displayed such that all the bars are present in a descending order. Moreover, similarity principle is also followed as same color is used to represent data belonging to same classes. There is also symmetry and order in all the visualizations as easier to interpret shapes (rectangles and circle) are used. Therefore, the visualizations are easier to interpret and understand. These visualizations are simple yet effective.

For further work, more features can be added to data. As these features are less for this problem. Using more features will aid in achieving a better prediction model. Moreover, we can use visualization tools such as tableau to make interactive dashboards to make it more appealing.

VI. Conclusion

We selected UFC dataset with data of every UFC bout from mind 2010 onwards till 2021. The dataset was very extensive with more than 100 features per bout. We used different visualizations to display important aspects of the data effectively. The visualizations were made using python programming language and libraries like matplotlib. The visualizations can be used for analyzing the data, making comparison of different features, identifying outliers etc. Furthermore, we implemented a prediction model to predict the outcome of the bout.

References

[1] M. McQuaide, “Applying Machine Learning Algorithms to Predict UFC Fight Outcomes”. Source: http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647731.pdf

[2] Dataset: https://www.kaggle.com/mdabbert/ultimate-ufc-dataset