UFC Fight winner prediction

Instructor: Prof. Ying Lin

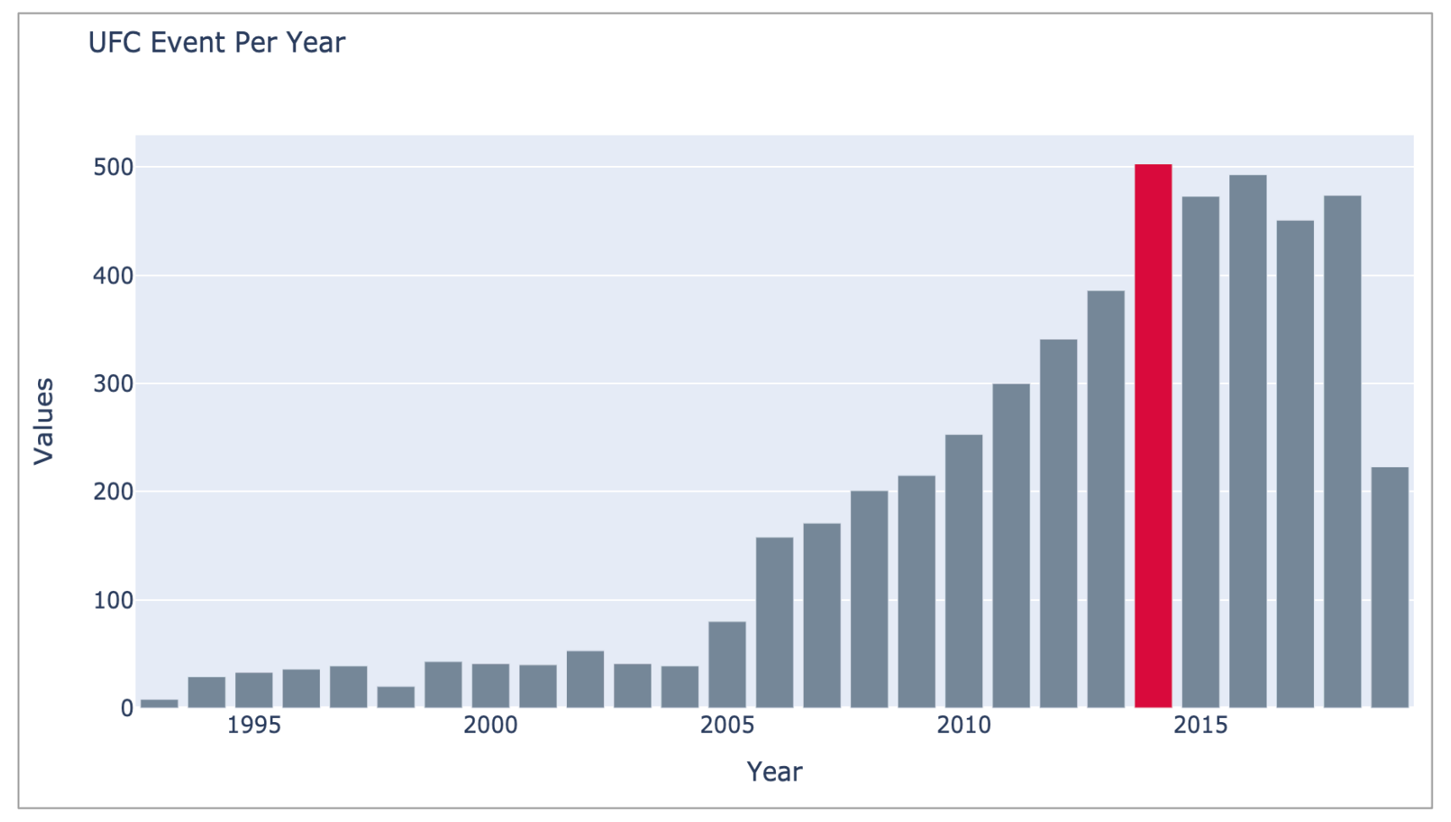
Team: Laxman Kumar, Saheb Singh, Abhiraj Singh. VIDEO LINK (https://drive.google.com/file/d/1kRpH7e9cptPaxJ_FzeExPHeRTyR5P9ag/view?usp=sharing)

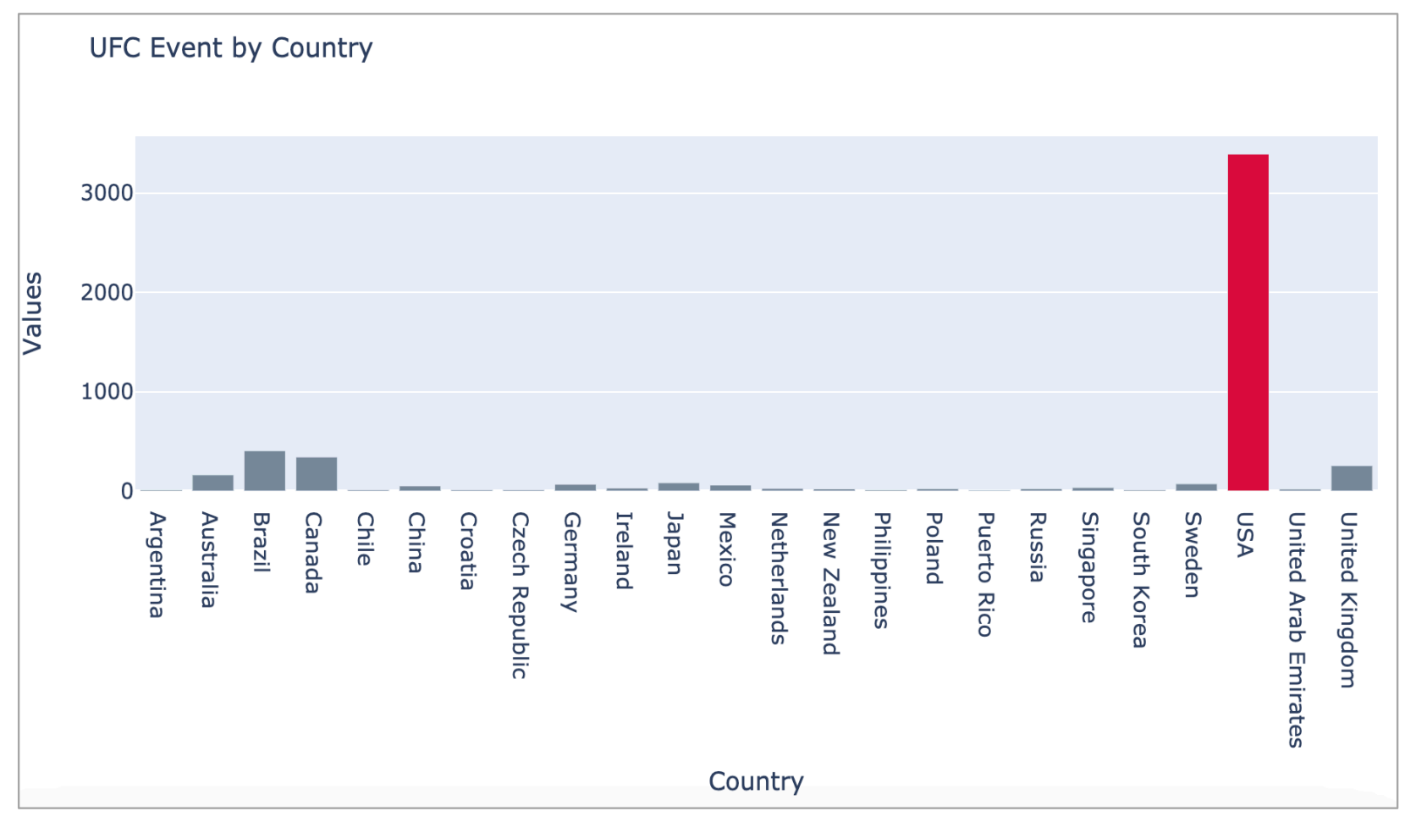
Problem Statement

- The goal of this project is to predict the outcome of UFC fights based on each players and groups fight statistics.
- An accurate prediction model could both inform the best placed bets (and potential risk associated) for each fight, but also could provide insight to coaches when accepting fights to begin with, simply by looking at the opponent's statistics relative to their fighter.
- It could also be used to help to identify which features are most significant in this prediction.

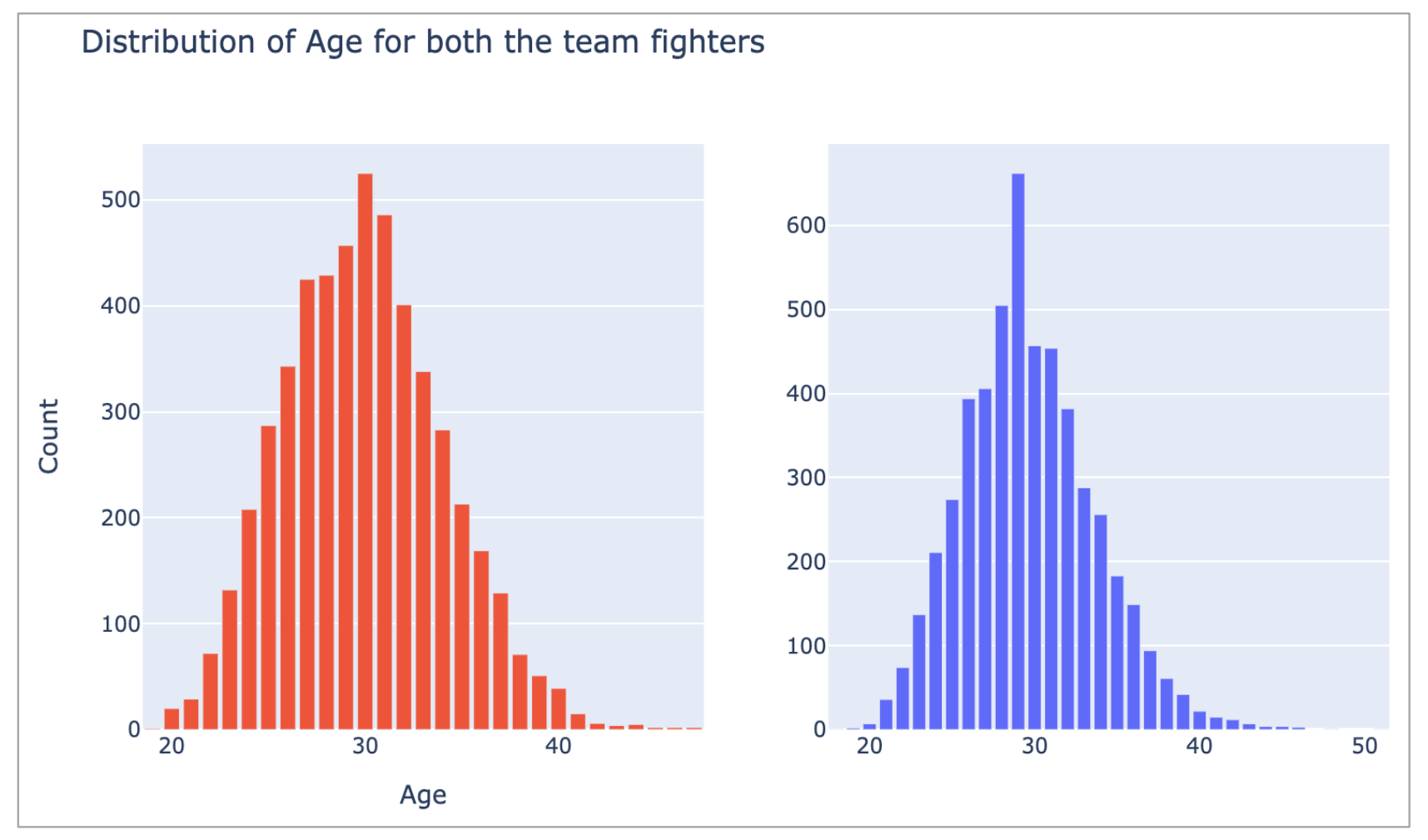
Dataset Description

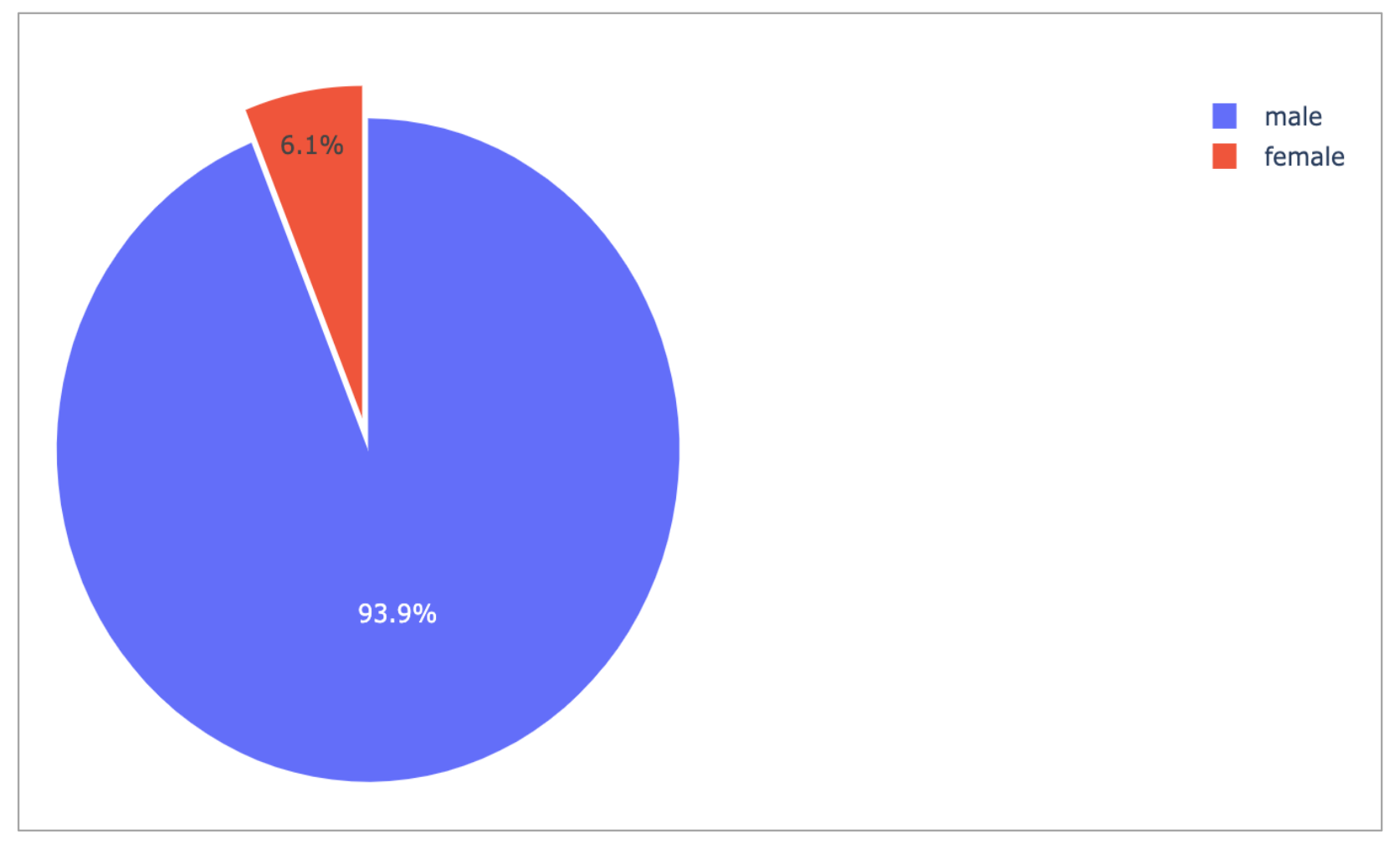
- Dataset for this project is taken from Kaggle named " ".
- There are over one hundred different fighter statistics on UFC Stats for each of the fights in UFC record from 1993 to 2019.
- For each player information included are their personal information such as height, weight, reach, stance and their fight statistics such as win streaks, strike percentage, guard passes and strikes landed by location.
- Dataset contains multiple csv files
  - **raw_fighter_details.csv :** This file contains the information about the fighter. The attributes are fighter name, height, weight, stance, reach and date of birth.
  - **raw_total_fight_data.csv :** This file consist of information about each of the fight. The information is divided in teams R and B. Details for each fight includes the the fighter individual details, number of rounds, date and the winner.
- From the data, UFC had conducted maximum number of events in 2014 (503 events) and United States is the country where most of the UFC events are conducted 3392 events.





- There are only 6.1% female fighters in UFC compare to male fighters which is 93.6%. Also both the teams have mean age value as 30.





Evaluation Metrics

- We are trying to predict which team will win the match based on the player statistics and their fight data and accuracy is one of the best metric which can provide us this information.
- Another factor in determining the metrics is the imbalance in target values.
- Hence the best metric(s) to do that would be use **Recall & Accuracy.**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

Data Preprocessing

Data preprocessing is the important step for any machine learning process. Following were the steps followed to preprocess the data.

- Null values and duplicated removal
- Splitting the multiple values from single column to multiple column
- Converting the data from string to numerical and fraction to percentage
- Replacing the null value in winner column with "draw"
- Replacing the winner name to winner team
- Merging the fighter_data and fighter_details dataframe into one

Modeling

1. SVM Classifier

Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier . An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

2. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. It is a predictive algorithm using independent variables to predict the dependent variable, just like Linear Regression, but with a difference that the dependent variable should be categorical variable.

3. Naïve Bayes

Naive Bayes algorithm is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to the presence of any other feature.

4. Random Forest Classifier

The random forest is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which we could call a "forest"), this model uses two key concepts that gives it the name *random*: Random sampling of training data points when building trees. Random subsets of features considered when splitting nodes.

5. Gradient Boosting Classifier

Gradient Boosting Classifier builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

6. Extra Tree Classifier

Extra Trees is an ensemble machine learning algorithm that combines the predictions from many decision trees. It is related to the widely used random forest algorithm. It can often achieve as-good or better performance than the random forest algorithm, although it uses a simpler algorithm to construct the decision trees used as members of the ensemble.

7. K Means

K Means is a clustering algorithm which divides observations into k clusters. Since we can dictate the amount of clusters, it can be easily used in classification where we divide data into clusters which can be equal to or more than the number of classes.

8. Deep Neural Network

Artificial Neural Networks (ANN) are multi-layer fully-connected neural nets. They consist of an input layer, multiple hidden layers, and an output layer. Every node in one layer is connected to every other node in the next layer. We can make the network deeper by increasing the number of hidden layers.

Experiment Result

Below are the results of the models.

Model Name	Accuracy	Recall
SVM-C	68.88	86.10
Logistic Regression	69.53	72.00
Naïve Bayes	69.00	70.00
Random Forest Classifier	70.11	96.00
Gradient Boosting Classifier	70.66	95.00
Extra tree classifier	69.88	96.00
K-Mean	53.16	65.00
Deep Neural Network	71.04	94.00

- Out of all the algorithms tested, for 2-class Neural network has performed best with accuracy 71.04 and recall rate as 94%.
- Random Forest Classifier and Extra Tree classifier gives the highest recall with 96%.

Further Discussions

- Continuing the work on the application of Principle Component Analysis on the datasets.
- Further research to be done on Feature Engineering.
- Work on the application of Artificial Neural Networks.

References

- https://www.kaggle.com/rajeevw/ufcdata
- https://dash.plotly.com/

Video Link

https://drive.google.com/file/d/1kRpH7e9cptPaxJ_FzeExPHeRTyR5P9ag/view?usp=sharing