**IST 707 – Final Project Report**

**UFC – Fight-By-Fight Winner Prediction**

**Shreyas Raghavan Sadagopan & Sourabh Ghosh**

**Contents**

**Module 1: Introduction and Problem Statement 2**

**Module 2: Data Assimilation and Understanding 2**

**Module 3: Cleaning and Manipulation 3**

**Module 4: Exploratory Data Analysis 4**

**Module 5: Feature Engineering 5**

**Module 6: Feature Importance study 6**

**Module 7: Predictive Models 7**

**Module 8: Business/ Statistical Inference 8**

**Module 9: Conclusion 9**

**Module 10: References 9**

[**Appendix**](#appendix)

**Module 1 Introduction and Problem Statement**

Mixed martial arts (MMA) is a full-contact combat sport that allows striking and grappling, both standing and, on the ground, using techniques from other combat sports and martial arts. Based in the United States, the UFC produces events worldwide that showcase twelve weight divisions and abide by the Unified Rules of Mixed Martial Arts. It has many real-life applications such as gambling, journalism and improvement of player performance by scrutinizing the advantages of the other opponents and refining his own performance in accordance with that data. Our Plan is to develop a comprehensive predictive model to predict the winners of UFC Fights from each fighters' statistics prior to the fight. We were intrigued by the highly unpredictable nature of the sport which drove us to analyze and look out for any patterns from the statistics of the fighters. We would be predicting the winner of a given brawl with the information of the two fighters and their prior records.

**Major data Questions:**

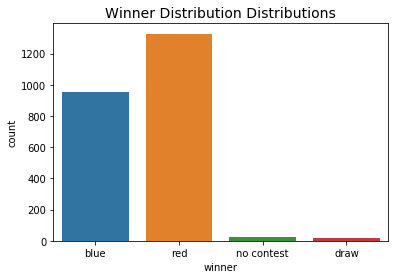
Our primary objective is to answer the following question through this project:

* Predicting the outcome of a match given the details of the fighters.
* What variables/factors have the most telling impact on the outcome of the match?
* How are demographic variables like Age/Height related to the outcome?
* What are the most popular locations in UFC?
* What is the most popular way to win the fight?
* Comparing techniques used by fighters

**Module 2: Data Assimilation and Problem Understanding**

* Dataset contains a list of all UFC fights since 2013 with summed up entries of each fighter's round by round record preceding that fight.
* Each row represents a single fight - with each fighter's previous records summed up prior to the fight. Blank stats mean it’s the fighter's first fight since 2013 which is where granular data for UFC fights.
* There are about 894 columns for almost 2318 unique fights
* Target class: The Y variables defines the outcome of the fight – The ‘Winner’ Variable
* We have conceived this a classification problem. In order to predict results of future fights (our prediction will be in one of four classes for each game win, loss, draw or no contest), we will be evaluating multiple models ranging from neural networks to decision trees

**The Dependent Variable has 4 levels**: **Broad Categories of the Explanatory Variables**



Demographic Variables like:

Age

Height

Location

Fighter Technique Variables like:

Punches

Kicks

Fight Detail Variables like:

Date

Winby : How did the fighter win the fight

Apart from this, the dataset contains all the techniques attempted and landed by the fighters in each round.

**Module 3: Cleaning and Manipulation**

1. **Ambiguity in the maximum number of rounds:**

* The maximum number of rounds possible should be 3 or 5. But in some cases, it was found that there were 4 rounds in 27 fights. It is an anomaly and should be rectified.
* Most of the 3-round fights were won by decision (as opposed to KO/submission) - this means that there was no opportunity for a 4th round. We have replaced all the 4 round fights to 3 round fights

1. **Age is not dynamic:**

* Age of the fighters was not dynamic and was calculated from a fixed date.
* A fighter fighting today and fighting one year later, would have the same age , which is a major anomaly.
* The Age field represents the fighter's age at the time of dataset creation, not at the time of the fight.

1. **Missing Weights and Demog Details.**

* The Weights for 12 unique players were missing. They were represented by NAs,
* We replaced those NAs with present values available on the internet

1. ***Performed the Univariates Test: Please find the image in the appendix -*** [***(Screenshot 1)***](#bookmark=id.iq3s44f5yxo)
2. **Handling ID, Date and the Location variables:**

* We will be removing the ID, Date and the Location columns at a later stage before modelling. We will be creating Level 2 variables using these columns and hence saving this step for a later stage.

1. **Outlier Treatment:**

* All the outliers in the continuous columns have been identified and capped using a 5% window on either side (0.05 and 0.95 percentile). This step is done specifically before handling the NAs so as to not miss the essence of the raw and miscalculate the interpolations.

1. **NULL Value Interpolation:**

* We found that there are a few missing values in our data. Age and Height are important features in any combat sport, and there were 17 and 12 missing values out of 2318 rows. We can simply delete rows with missing values, but usually we would want to take advantage of as many data points as possible. Replacing missing values with zeros would not be a good idea - as age 0 will have actual meanings and that would change our data. So, we have used the "fillna" function replaces every NaN (not a number) entry with the mean of the column

**Module 3: Exploratory Data Analysis**

1. Red Side seems to win slightly more than blue (867/1477 = 58.7%)
2. Max-3-round fights are ~2.3 rounds on average, while max-5-round fights are ~3 rounds on average.
3. There are more fighters fighting debut fights. This statistic however could be skewed by the fact that our data set assumes the debuts of every fighter in 2013.
4. Submission wins tend to happen a little after halfway through the fight: average 1.7 rounds in a max-3-round fight, 2.6 in a max-5-round.The match duration affects the distribution of win type. KOs are most likely in a max-5-round match, while decisions are most common for max-3-rounders. Max-3-round matches often end in decision, meaning a clear winner wasn't declared by the end of the fight. However, max-5-round matches end in KO or submission 2/3 of the time! [(Illustration 1)](#bookmark=id.k0pbq2rkqfdc)
5. The weight classes follow a relatively normal distribution with a peak at 70kg / 154 lbs. [(Illustration 2)](#bookmark=id.on47duzif2ih)

* Wins by decision are most common with lighter fighters, and wins by KO are most common with the heaviest fighters
* The likelihood of a match ending by submission seems independent of weight class

1. Heavier fighters are more likely to win; however, the weight difference is likely not a strong estimator of victory. [(Illustration 3)](#bookmark=id.cke2a31u7wqa)
2. Age is a big factor in any sport, moreover in MMA where you must have a combination of strength, agility and speed (among other skills). Found that most fights have been won by fighters in their late 20’s through early 30’s as they peak during this time and then lose strength, quickness and cardiovascular. Younger fighters do not develop peak strength till 27-28 while older fighters are usually slower and more likely to lose. Interestingly, most fighters are below 35. Age matters, and youth is a clear advantage. fighters aged 25-35 fight ~2x more than older fighters AND ~2x more than the youngest (age 19-24) fighters. The newcomers are likely to take longer recovery times and lower viewer demand for older fighters. The latter possibly due to being busy fighting in other MMA leagues, or having a harder time getting fights as a newcomer. [(Illustration 4)](#bookmark=id.4mijnaeklj6u)
3. The Most common hometown and training location for fighters is Rei De Janeiro, Brazil.[(Illustration 5)](#bookmark=id.34jdqpka34x)
4. Height is also a major advantage in MMA as it means more the height more is the reach. Taller fighters

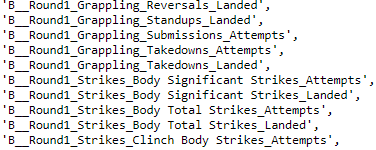
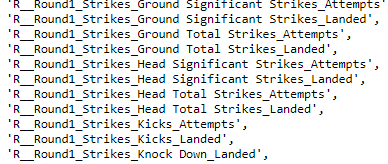
have an advantage and, on average, win.

1. There was a linear relationship between the number of offensive strokes and the score. More offensive attempts you make should mean more land to your opponent (and with the right skills and power - more chance you must win the fight). The bigger the risk, the greater the reward.
2. We compared the Round 1 and Round 5’s fighting stats. We found that grappling reversals increase for fighters between weight 80-90, while takedowns decrease in the lighter weight groups. Fighters prefer to land more head strikes during round 1. By Round 5, fighters (who are now worn-out) are hardly landing any leg and body strike. They are still landing a good amount of Head strikes. This makes sense as the fight is coming to an end and instead of depending on the judges, they want to go for a Knockout. ([Illustration 7](#bookmark=id.cbjfvxt1g0gw) and [Illustration 8](#bookmark=id.eib3vg60fcbz))

**Module 5: Feature Engineering**

1. **Variable Manipulation**

* For a given fight our raw data contains fight metric related variables for each of the fighters, Red and Blue.
* To illustrate, we had variables as the ones shown below. B represents the Blue fighter and R represents Red fighter. The data was available for each of the 5 rounds

* Using the above variables, we have created **ratio variables** depicting Blue vs Red fighting metric for each metric for a round for each of the fights. For Example, two separate variables ‘B-Round1\_Grappling\_Landed” and ‘R-Round1\_Grappling\_Landed” were transformed as ‘B-Round1\_Grappling\_Landed”/ ‘R-Round1\_Grappling\_Landed” which is nothing but the ratio of Grappling landed by B vs No of Grappling Landed by R.

|  |  |  |
| --- | --- | --- |
|  | Number of Columns | Number of rows |
| Before Transformation | 883 | 2318 |
| After Transformation | 448 | 2318 |

1. **Creating Level 1 variables:**

* Using the transformed variables from the previous step, we have created a cumulative score for each of the rounds in a fight. We have assigned weights ranging from 1 to 5 for Round 1 to Round 5 respectively.
* As a fighter moves and gets tired during a fight, we are increasing the weights for the achievements/ fight scores as he progresses through rounds during a fight . We have arrived at a cumulative score for each of the fight IDs which represent the ratio of fight metrics with due weightage to the rounds too.

|  |  |  |
| --- | --- | --- |
|  | **Number of Columns** | **Number of Rows** |
| **Before Transformation** | 448 | 2318 |
| **After Transformation** | 224 | 2318 |

1. **Creating Level 2 variables:**

* Using the variables from the previous step, we further created variables which tell us the Aggression quotient of the fighter and his hit rate.
* We created Number of attempts vs Number of strikes ratio variables to extract theri hit rate and their aggression quotient. This would be really helpful in our model as we saw that the hit rate was major influencer for a win in our EDA process

|  |  |  |
| --- | --- | --- |
|  | **Number of Column**s | **Number of Rows** |
| **Before Transformation** | 224 | 2318 |
| **After Transformation** | 112 | 2318 |

1. **Variable Reduction:**

* There are a whopping 112variables in our dataset. There is no way we would be able to use all of these variables for our predictive models.
  + Dropping Variables with more than 50% values missing

1. Removed all the variables which have more than 50 % missing values.
2. There were a total of 9 variables with more than 50 % missing values
   * Dropping variables with very less variance
3. Used a 10 % threshold for variance.
4. Dropped variables which are below 10 %
5. There were a total of 27 variables below 10 %
   * Dropping variables using correlation
6. Dropped variables which have a correlation coefficient of more than 0.7
7. Dropped a total of 35 variables

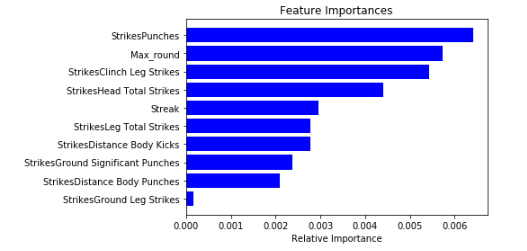
|  |  |  |
| --- | --- | --- |
|  | **Number of Column**s | **Number of Rows** |
| **Before Transformation** | 112 | 2318 |
| **After Transformation** | 41 | 2318 |

**Module 6: Feature Selection and Importance study**

**Feature ranking with Recursive Feature Elimination**

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the model’s coef\_ or feature\_importances\_ attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model.

RFE requires a specified number of features to keep, however it is often not known in advance how many features are valid. To find the optimal number of features, cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features. The RFECV visualizer plots the number of features in the model along with their cross-validated test score and variability and visualizes the selected number of features.

****

**Found out that all the top 41 variables, are highly important and produce the best accuracy in tandem in our train set.**  [**[ See Feature Selection Illustrations]**](#bookmark=id.am7vve2ognb)

**Module 7: Predictive Models**

**We have started building the predive models using the 41 optimal variables obtained from the Recursive feature elimination exercise. We have explained the outputs below:**

For building the classification model, we followed incremental techniques and built a base model in each of the cases and tuned them using ‘Hyperparameter tuning’ to arrive at the best parameter combinations.

**Logistic Regression :**

Selection of multiclass logistic regression technique as our base model in the engineered feature gave us an accuracy of 57.1%([Illustration 9](#bookmark=kix.d9tonjwms31h)  ). We performed hyper parameter tuning for model selection with “c” value ranging from 0.0001 to 1000, with regularization parameters as Ridge “L2” and Lasso “L1” and solver values as “liblinear” and “saga” resulting in accuracy of 57.8%.

**Random Forest Classifier :**

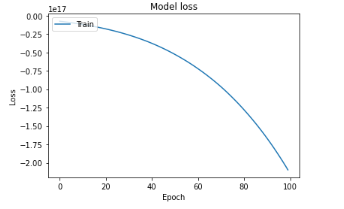
The classification model uses random forest classifiers leading to reduction of variance of large data and resulting in an accuracy score of 60%([Illustration 10](#bookmark=kix.o8q0vsa279f9) ) which was better than the logistic regression. The progressive modelling was implemented by performing hyperparameter tuning using tune grid with wide range of values, for this purpose a random grid search was performed for 100 iterations and 3 cross validations performing 300 fits to the data based on the random search a tune grid was created on a ranges of values for 'max\_depth': 80-110, 'max\_features': [2, 3], 'min\_samples\_leaf': [3, 4, 5], 'min\_samples\_split': [8, 10, 12], 'n\_estimators': [100, 200, 300, 1000].The best parameters resulted to be ‘n\_estimators': 200, 'min\_samples\_split': 5,'min\_samples\_leaf': 2, 'max\_features': 'sqrt', 'max\_depth': 10, 'bootstrap': True.The model built performed as expected resulting with an accuracy of 60%([Illustration 11](#bookmark=kix.4d6gny8yfy9m)  ) for the random search and 62%[(Illustration 12](#bookmark=kix.t1889t5fe025) ) for the tune grid. Since the data was a bit unbalanced, looking at the F1-Scores of the models also improved progressively from 60% - 64% - 73% for the base model, random grid and with best hyper parameters respectively.

**Gradient Boosting Machine :**

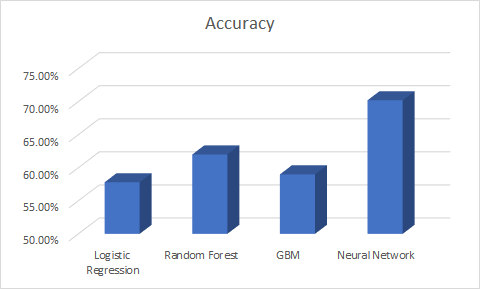
The Gradient Boosting Machine resulted in an accuracy score of 57.78%[(Illustration 13](#bookmark=kix.p1yckvn6b30u) ) for the base model.. The progressive modelling was implemented by performing hyperparameter tuning using tune grid with wide range of values, for this purpose a random grid search was performed for 288iterations and 3 cross validations performing 864fits to the data based on tune grid on a ranges of values for 'max\_depth': 80-110, 'max\_features': [2, 3], 'min\_samples\_leaf': [3, 4, 5], 'min\_samples\_split': [8, 12], 'n\_estimators': [100, 200, 300, 1000]. The best parameters resulted to be 'max\_depth': 90, 'max\_features': 3, 'min\_samples\_leaf': 3, 'min\_samples\_split': 8,'n\_estimators': 1000.The model built performed as expected resulting with an accuracy of 59%[(Illustration 14 )](#bookmark=kix.rakg466m7ymy) for the tune grid.

**Neural Networks :**

We have performed training of our dataset into a neural network with the input , output and 2 hidden layers.The activation function was set as ‘sigmoid’ for the outer layer and ‘relu’ for others. The number of neurons as per manual tuning being 20,12,8,1 for all the 4 layers from input to output, the learning algorithm kept as “stochastic gradient descent”/”adam” for optimization. This baseline setting of the model efficiently improved the accuracy in the given feature engineered dataset. The accuracy increased to 70% while decreasing the learning rate on validation data with this neural network with 100 epoch as shown in the below figure.



Hyperparameter tuning was performed for parameters such as 'softmax', 'softplus', 'softsign', 'relu', 'tanh', 'sigmoid', 'hard\_sigmoid', 'linear' and dropout rates ranging from 0 to 0.9. The tuning on number of neurons was performed ranging from 1 to 30 neurons resulting in best hyper parameters to be “softmax” with dropout value of 0 and number of neurons to be 12.The resulting accuracy was 70.02%[(Illustration 15 )](#bookmark=kix.qzo1t1dvrq5o) .



**Module 8: Business/Statistical Inference:**

UFC being one of the most unpredictable games of today's generation involves a lot of investment by the organization and the sponsors. This task of predicting the winner of the game will lead to better decisions if opted by the business personals of the fighter managers and sponsors to help them evaluate which fighter could be on the top of their list of representatives. Furthermore, this being the type of game which inculcates the trend of betting on the fighters and hence the machine learning models could be used to determine the possibility of winning for a fighter helping them to manipulate and make large profits in the betting market.

The evaluation presented in this report is based on limited data and engineering of available data on models that could predict the chance of winning a fighter. We have used accuracy as an evaluation metric to predict the winner in a game. We are primarily worried about how accurate our predictions are and hence did not concentrate on the precision or the recall values. However, we used F1-score as a supporting evaluation metric as there were 4 different classes and the data was biased towards just 2 classes- win and loss.

**Module 9: Conclusion**

**To answer the major data questions in module 1, we can conclude that**

* For predicting the outcome of a match given the details of the fighters, we have come up with 41 best predictors after engineering and found Neural Networks to be the best model with 70.02% accuracy. We believe this is the best possible analysis given the amount of available data and its inherent noise.
* The most important features that made a telling impact on the match decisions were found to be , “Maximum Number of Rounds” played in a match, strategies such as “Strikes Punches”, “Clinch Leg/Head Strikes, Ground Punches , Ground Leg strikes and Body Distance . Looking at the research on how players fight these are the major traits that the trainers and fighters keep in mind while performing and training. Also, the past performances of the fighter were determined to play an important role since a better performing fighter is expected to perform better in future matches
* Also, the features such as player weight and age play an important role in win result for a red/blue fighter with weight value 70Kg and age range 30-34 years more probable to win in past scenario
* Geographically stating the United States and Brazil has most of the players which is justified as UFC is the second-best sport played in brazil.
* The most popular way to win a fight is by Decision (50%) and then KO(30%) and the least, Submission(20%)

**Module 10 : References :**

Hitkul, Karmanya Aggarwal, Neha Yadav, and Maheshwar Dwivedy. "A Comparative Study of Machine Learning Algorithms for Prior Prediction of UFC Fights." *Harmony Search and Nature Inspired Optimization Algorithms Advances in Intelligent Systems and Computing* (2018): 67-76.

Johnson, Jeremiah Douglas. *Predicting outcomes of mixed martial arts fights with novel fight variables*. Diss. University of Georgia, 2012.

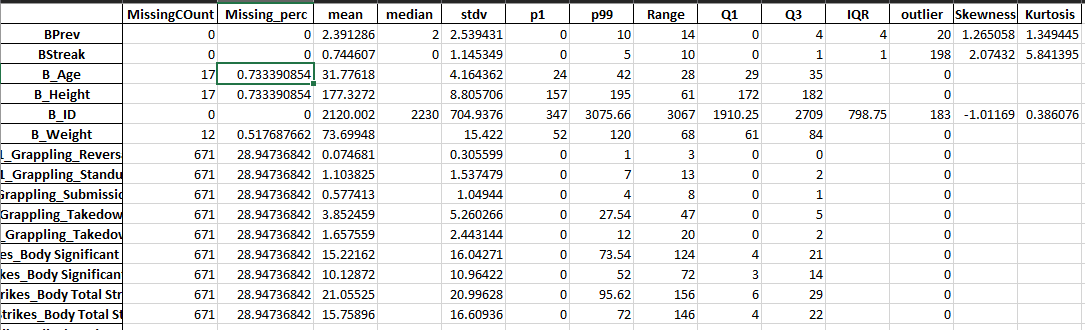
Kuchen, Robert. "Model Based Prediction of Starcraft II Match Outcomes."

El-Fouly, T. H. M., E. F. El-Saadany, and M. M. A. Salama. "Improved grey predictor rolling models for wind power prediction." *IET Generation, Transmission & Distribution* 1.6 (2007): 928-937.

Soman, Saurabh S., et al. "A review of wind power and wind speed forecasting methods with different time horizons." *North American Power Symposium 2010*. IEEE, 2010.

**APPEN****DIX**

1. **Univariates: First few var****iables**

****

1. **EDA ILLUSTRATIONS:**

Illustration 1: Fights Vs Win Type

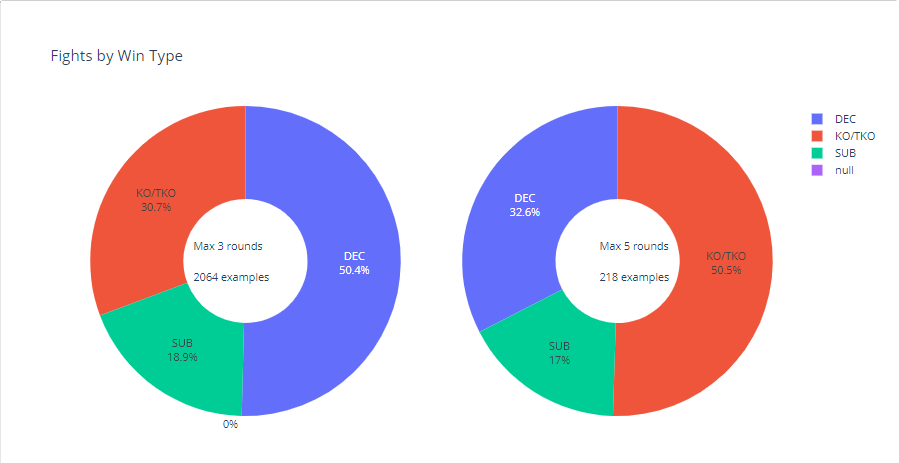


Illustration 2 : Weight Category vs Decisions

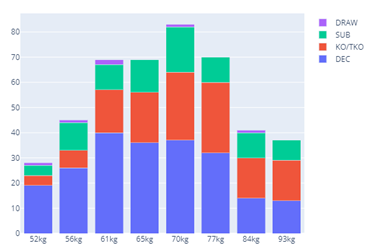


Illustration 3 : Weight Category Vs The Outcome

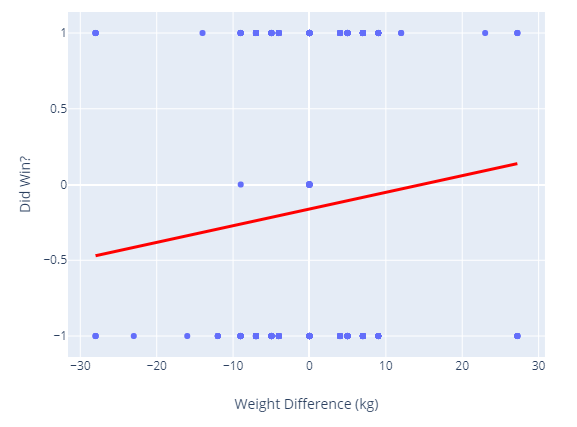
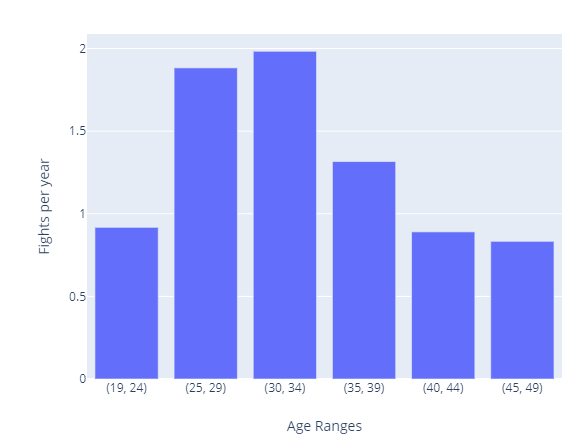
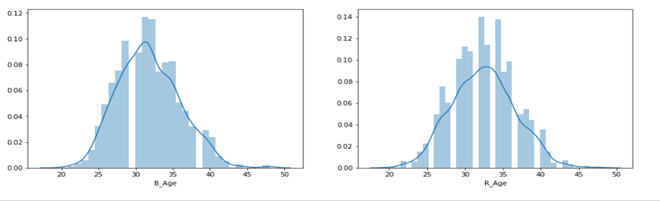


Illustration 4: Age distribution analysis





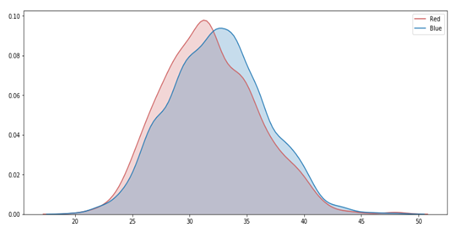
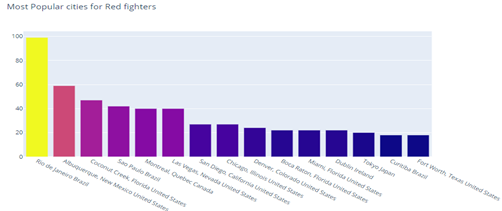


Illustration 5: The hometown of the fighters



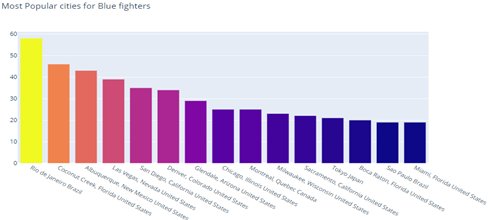
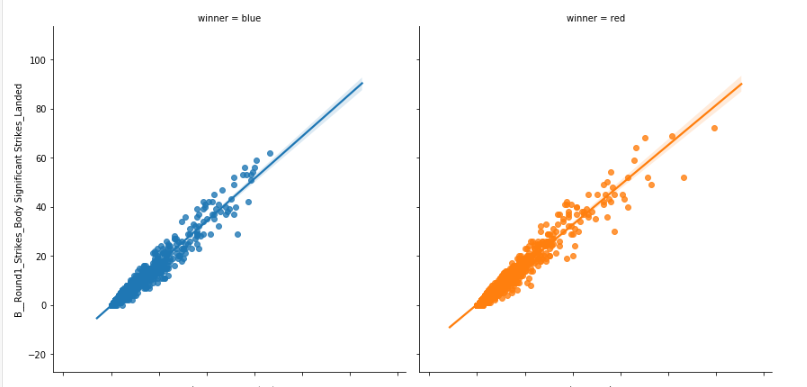


Illustration 7 : Relationship between the number of aggressive strokes in Round 1 vs the outcome



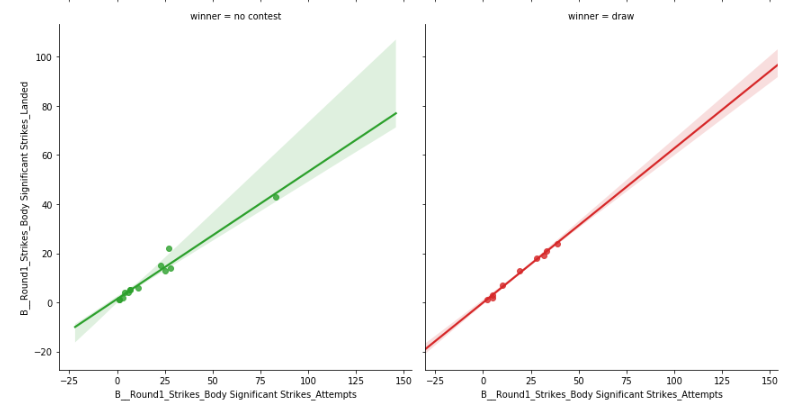
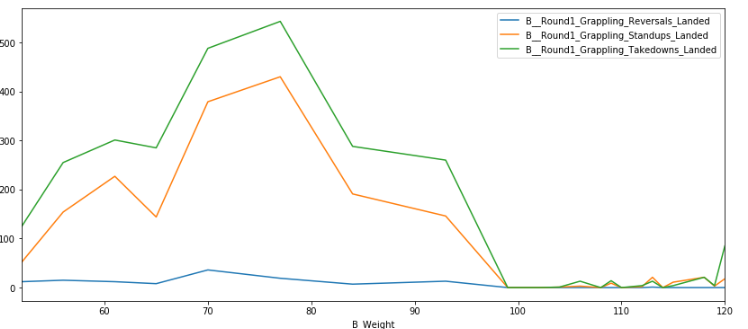
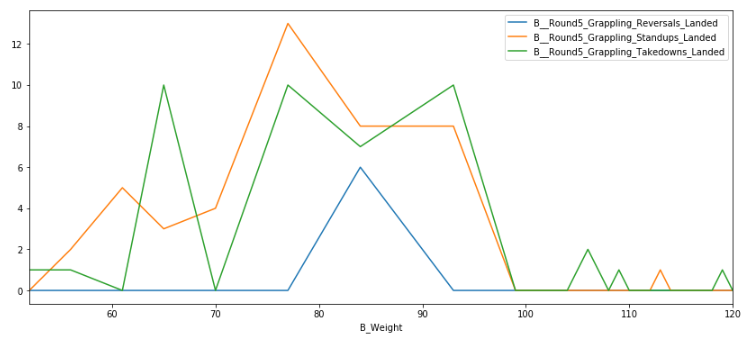
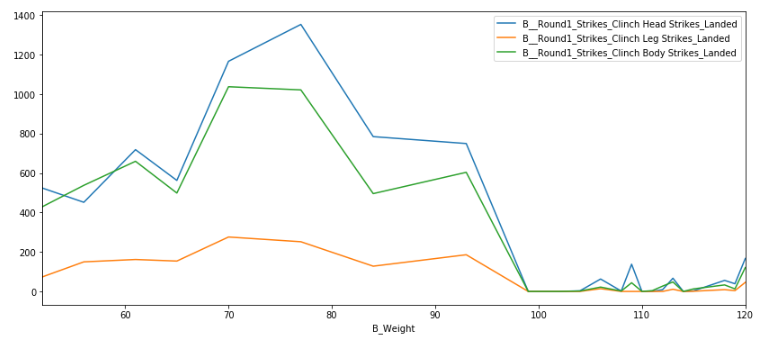
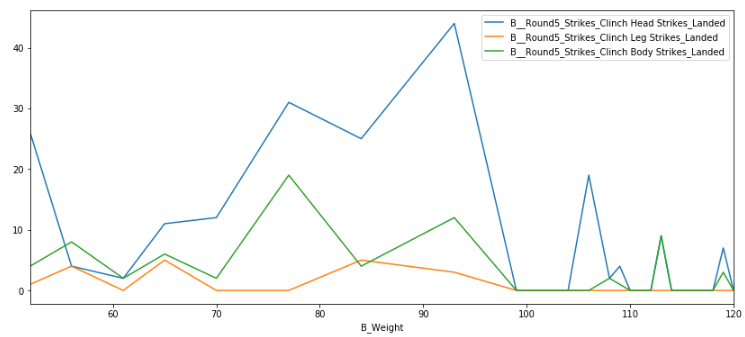


Illustration 8 : Comparing techniques across Rounds ( Round 5 vs Round 1) for each of the fighters

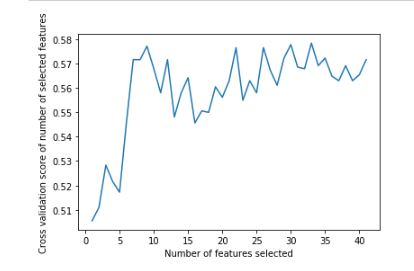
Round 5

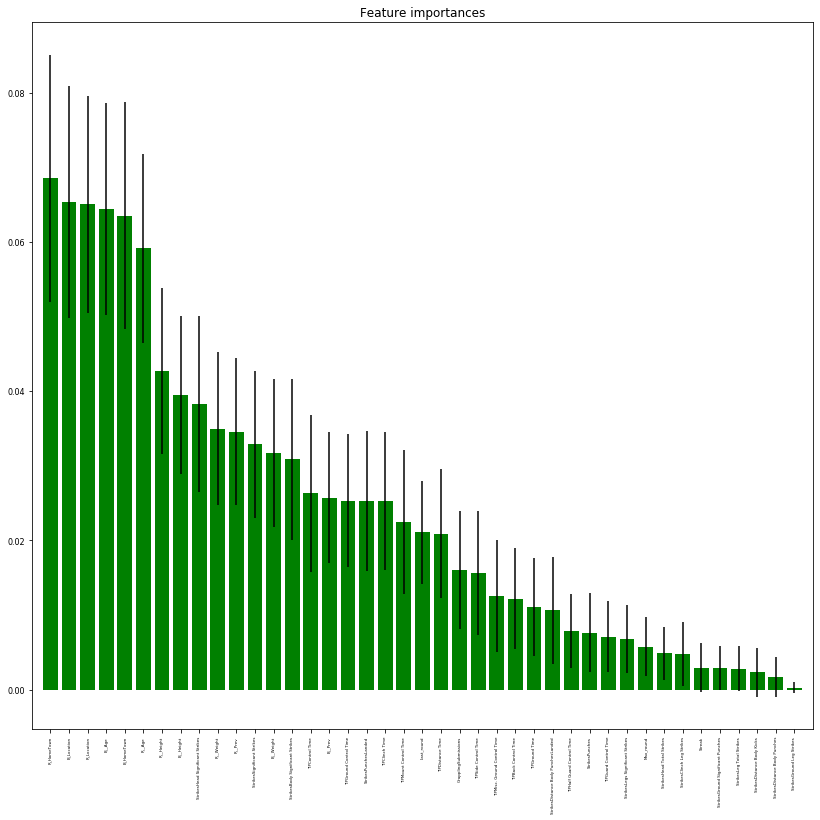
Round 5

Round 1

Round 1

1. **Fe****ature Selection Illustrations:**

****



1. **Predictive Models Evaluations :**

Illustration 9 : Logistic Regression

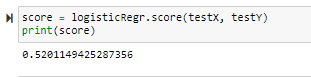


Illustration 10 : Random forest (Base Model)

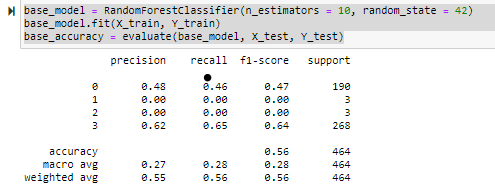


Illustration 11 : Random forest (Random Grid Serach)

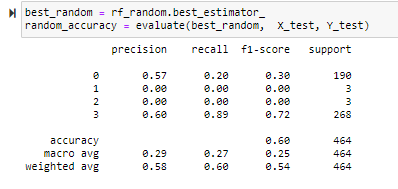


Illustration 12 : Random forest (Tune Grid)

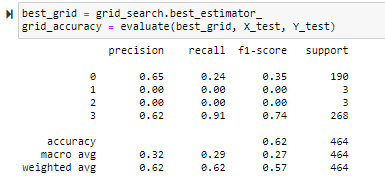


Illustration 13 : GBM (Base Model)

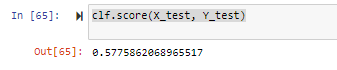


Illustration 14 : GBM (Tune Grid)

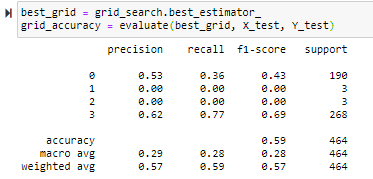


Illustration 15 : Neural Networks (Tune Grid)

