AUBG - ANALYSTUNKNOWN'S BATTLEGROUNDS 2.0

Report on Methodology and Analysis

Submitted by

TEAM – **INNOVATION GEEKS**

IIM CALCUTTA

Table of Contents

[OVERVIEW AND PROBLEM STATEMENT 2](#_Toc25092422)

[ANALYSIS 3](#_Toc25092423)

[1. EXPLORATORY DATA ANALYSIS 3](#_Toc25092424)

[1.1. CLOUD OF PRINCIPLE COMPONENTS: 3](#_Toc25092425)

[1.2. NORMALISATION OF VARIABLES: 3](#_Toc25092426)

[1.3. COMPARISION AMONG CLASSES: 4](#_Toc25092427)

[2. FEATURE GENERATION 4](#_Toc25092428)

[3. PREDICTIVE ANALYTICS 5](#_Toc25092429)

[RESULT/OUTCOME 6](#_Toc25092430)

# OVERVIEW AND PROBLEM STATEMENT

**UFC®**

The study is pertaining to “the Ultimate Fighting Champion” that features the best athletes skilled in the various disciplines of all martial arts, including karate, jiu-jitsu, boxing, kickboxing, grappling, wrestling, sumo and other combat sports.

PROBLEM STATEMENT: To predict the target variable 'Winner' which tells the outcome of a match.

# ANALYSIS

We begin with exploratory data analysis to analyse various characteristics of the data. During the exploratory analysis, following was observed:

## EXPLORATORY DATA ANALYSIS

### CLOUD OF PRINCIPLE COMPONENTS:

Following is a depiction of the **cloud of principle components (97% variance is covered within these two principle components)**. It is evident from the cloud pattern for first two principle components. It can also be observed that data clouds for **wins for Red is denser than Blue (1st row mid) and draws( 2nd row)** however the pattern of cloud for both red and blue is **similar** hence it would be **difficult to classify**. In another way, model based on this **dataset is not expected to result in a very good accuracy**.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |



### NORMALISATION OF VARIABLES:

It is observed that most of the variables are skewed as example average body attack and landed are depicted below. It is also observed difference of these variables are is following a normal distribution.

|  |  |  |
| --- | --- | --- |
|  |  |  |

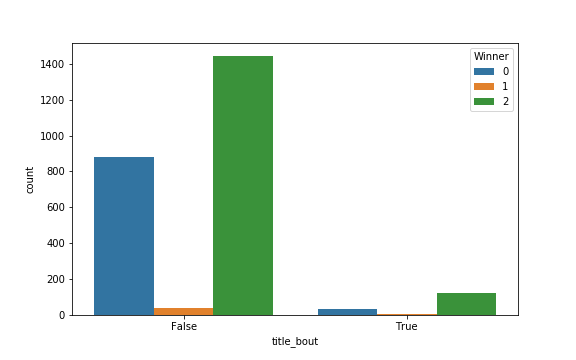
### COMPARISION AMONG CLASSES:

It is observed that variables show similar behaviour for different classes. Box plots for various variables are indicated for reference.

|  |  |
| --- | --- |
|  |  |
|  |  |

## FEATURE GENERATION

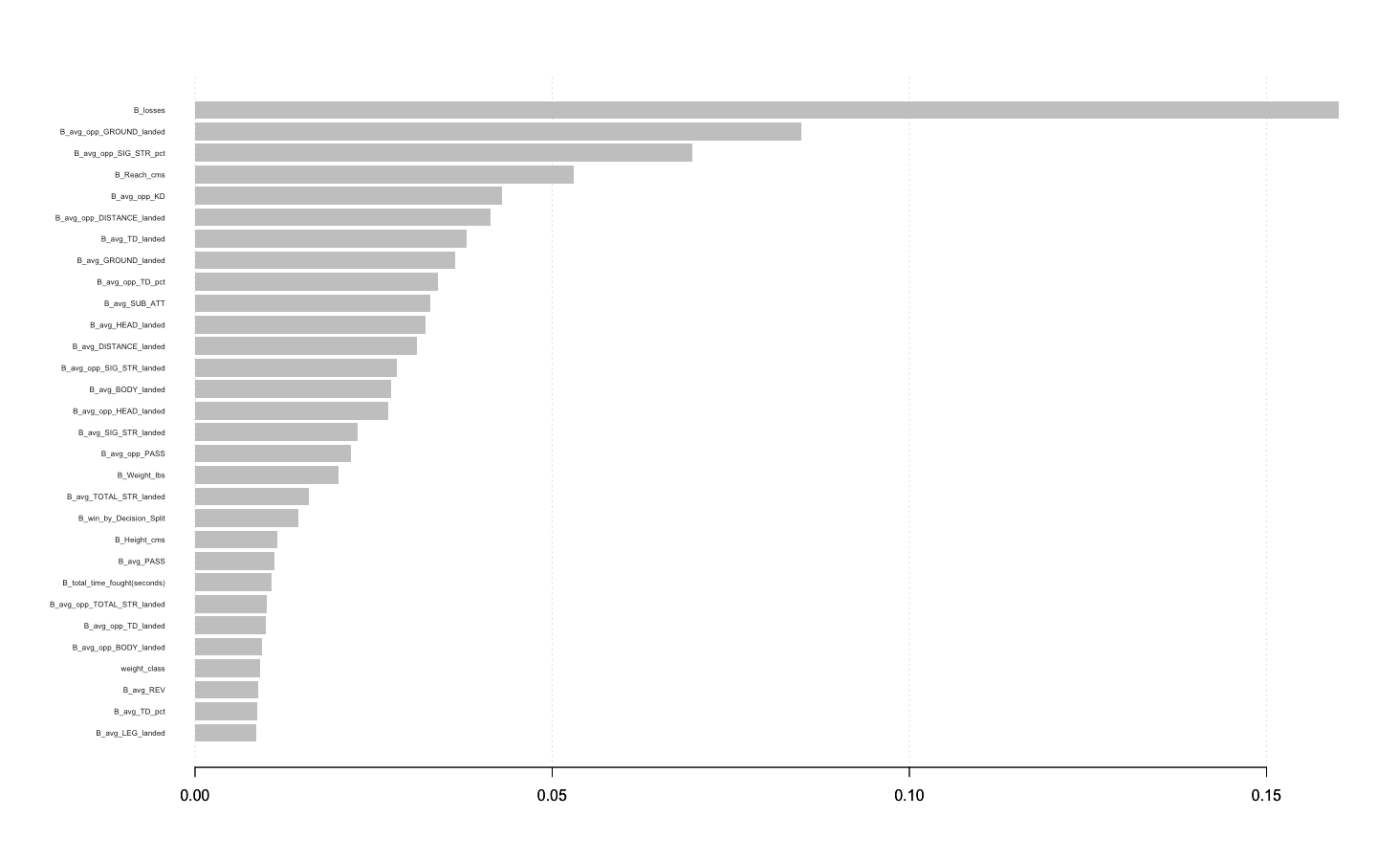
* 1. Features like Names of the Players, Referee Name, Date, title bout and Locations are **removed** before training the model as per our EDA. Their addition to the model was **reducing its accuracy**.
  2. As we can see in this bar plot, the proportion of winner as per title bout are **not significantly different** and hence is dropped from our consideration



* 1. We took the **difference** of all the **respective attributes of Blue team and Red team to reduce the features (comparative)** and still retain the same information to the model being trained.
  2. One more technique was used with respect to **proportions of landings to the attacks** which indicated the **efficiency of the attacks** made by the players at red or blue corner. However, significant improvement was not observed in on the test data prediction (**cross validation discussed later**). Consequently, in this method attacks and landed columns were removed.
  3. All the 18 types of 'ATTACKED' variables have been **removed from the data as they were highly correlated with the respective** 'LANDED' providing similar information also taking the attacks being landed into model makes more sense.

## PREDICTIVE ANALYTICS

* 1. Since it was a multiclass prediction, we used **tree based models like Bagging, Random Forest, AdaBoost, Xgboost** among which performance of Xgboost was the **best with 10 fold cross validation**.
  2. As per the importance matrix (the variables whose splits in the tree produces maximum improvement in accuracy) generated by these tree models, the **top 45 important variables are later used to re-train the model to improve the accuracy even further.**



# RESULT/OUTCOME

Test error was estimated to be **~ 64%.** **Prediction of various categories on train data** is presented below:

|  |  |  |  |
| --- | --- | --- | --- |
| y/y\_pred | 0 | 1 | 2 |
| 0 | 553 | 0 | 363 |
| 1 | 4 | 21 | 18 |
| 2 | 33 | 0 | 1530 |

1. Blue , 1- Draw, 2 - Red