# Project Report

## STAT-S 670 - Exploratory Data Analysis Spring 2019

**Chrislin Priscilla, Dhivya Swaminathan, Raja Rajeshwari Premkumar, Varun Miranda**
**April 24, 2019**

## 1 PUBG - Overview

PUBG - Player Unknown Battlegrounds is an online multiplayer game developed by PUBG Corporation.

In a PUBG game, up to 100 players parachute onto an island at a place of their choice. They then look for weapons or equipment that can be used to harm or kill others in order to defend themselves. The last person/ team to survive wins the game. During the game, the player should stay inside a blue circle. Failing to do so results in decrease of health points and subsequently death. (2)

## 2 Project Goals

The main goal of this project is to work with the PUBG (Player Unknown's Battle Grounds) Match Deaths and Statistics dataset obtained from https://pubg.op.gg/ and answer the following research questions:

### 2.1 Which weapon is more lethal?

For this dataset, through exploratory data analysis, we intend to answer usage of which weapon yielded a higher kill rate. This is an important question to answer because, based on the type of weapon a player possesses at that point in a game, his probability of making it to the top 10, or subsequently winning changes. This analysis can be used to understand patterns of weapon usage and how it influences the player placement in a game.

### 2.2 Predicting whether a player will be placed in the top 10 roster in the game or not

Using this dataset, we intend to predict if a player would survive and make it to the top 10 roster of game or not. The primary interest behind this is to determine the relationship between the number kills a player makes, the total distance he covers in the game and the probability of the player making it to the top 10 in the game. This analysis can be very vital in determining the big question that most PUBG players have: Can I win the game by just sitting idle, quiet, and staying inside the blue circle? or should I be actively participating in the game?

## 3 Data Description

The datasets (1) "kills" and "aggregate stats" were obtained from Kaggle datasets. The author who published this dataset obtained it from pubg.op.gg, a game tracking website.

The data dictionary of the two datasets - *kill match* and *aggregate match stats* are shown below:

*Dataset 1: Kill Match*

Each entry in the dataset represents a kill in a game.

1. Match id: It is a unique identifier for each match played.
2. Killed by: How was the victim killed. Usually a weapon, or others like hit by a truck, bike.
3. Killer name: Name of the killer.
4. Victim name: Name of the Victim.

5. Killer placement: It has the killer's rank in the corresponding match.
6. Victim placement: It has the victim's rank in the corresponding match.
7. Killer position x, Killer position y, Victim position x, Victim position y: This gives the killer and victim x and y coordinate positions when the kill happened. The coordinates range from 0 to 800000.
8. Time: Time into the game (match) when the kill happened. Essentially the survival time of the victim in the respective match.

*Dataset 2: Aggregate Match Stats*

Contains the player statistics and meta information for the match.

1. Match id: It is a unique identifier for each match played.
2. Match mode: The game can be played in three modes - Solo, Duo and Squad and in two different perspectives- FPP (First Person Perspective) and TPP (Third Person Perspective). For our analysis, we retained only Solo matches.
3. party size: The maximum number of players per team
4. player name: Name of the player
5. player kills: Number of kills by the player
6. player dbno: Number of knockdowns the player has scored; A knockdown does not count as a kill; the opponent is damaged enough to not retaliate.
7. player assists: Number of assists the player has; If a player does damage to somebody but they are eventually killed by another player, the player gets an assist, since he/she assisted in the kill.
8. player dmg: Total hit points that the player has earned on the accounts of all damage done to others
9. player dist ride: Total distance the player has travelled in a vehicle in the match
10. player dist walk: Total distance the player has travelled on foot in the match

The final dataset consisted of 397766 rows. Below are the steps followed to obtain the final dataset:

1. Combine the files kill match and aggregate match stats using the columns Match Id and Killer Name
2. Remove missing values
3. Filter where map is Miramar
4. Consider only solo matches by filtering party size = 1
5. Remove data with missing values

For the research question 2.2, the response variable was derived from the column killer placement by bucketing them into top 10 or not. Other primary features in the dataset include:

1. match_id: This is an identifier for a game
2. killed_by: This represents the weapon that killed the player
3. killer_placement: This represents the killer's ranking in the current game at the current

time

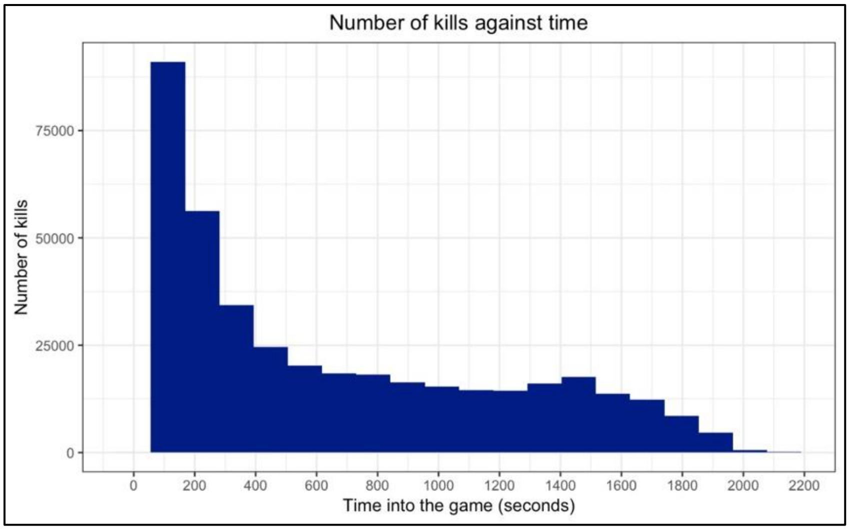
4. killer_position_x and killer_position_y: The coordinates of the killer on the “Miramar” map.
5. time: This represents the duration of the victim/player’s life in the current game
6. victim_position_x and victim_position_y: The coordinates of the victim/player on the “Miramar” map.

## 4 Exploratory Data Analysis

Various graphs were plotted to understand the data and to aid in identifying important factors that help in answering the research questions.

### 4.1 Overall picture of kills in the battleground:

*Graph 1: Capture the number of kills at any time in the game*



Time into the game - Variable used is time. It points to the time elapsed since the start of the match.

Number of kills – Each entry in the dataset accounts for a kill, hence counting number of rows provides the number of kills.

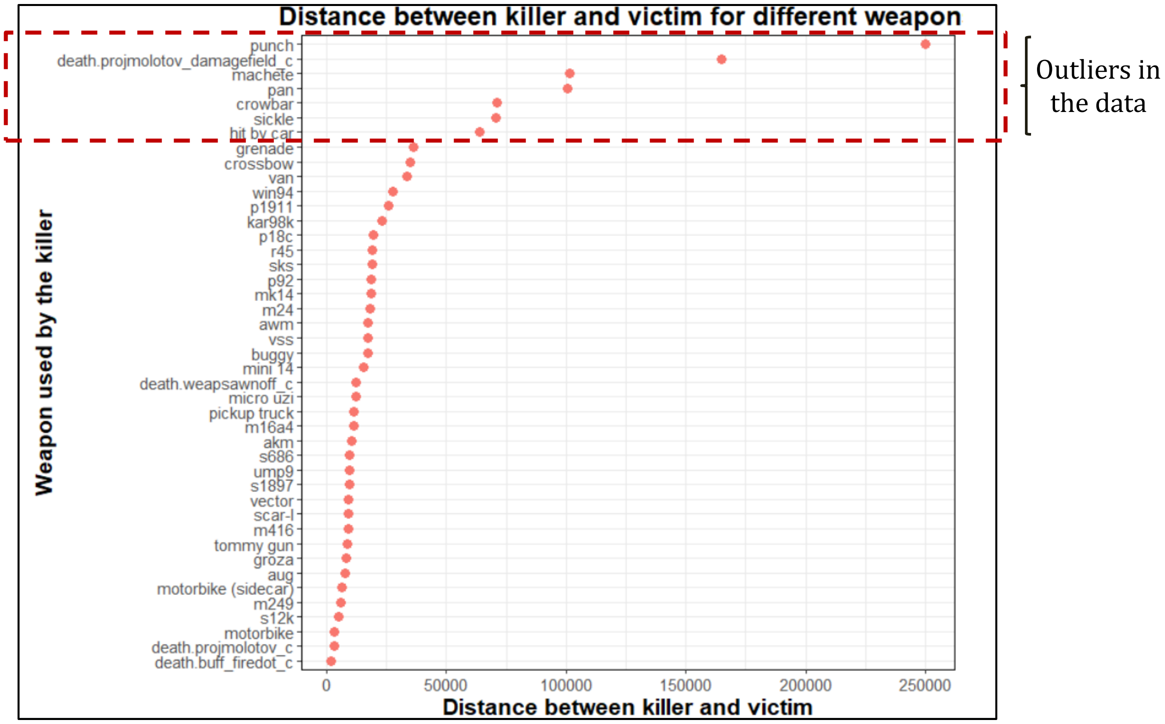
Findings:

- The number of kills is very high in the beginning of the game. This shows that the weaker players get killed off very early in the game. More so because the game begins with players being airdropped on the island in the open, the players who cannot hide instantly or find weapons, get killed off early.
- The count of kills falls steeply towards the middle of the game and remains relatively same throughout the rest of the game.
- The winner of the game is the one who survives the longest, hence it goes without saying that the top 10 players will be having high survival time and are in the last few histograms.

*Graph 2: Average distance at which the weapons were used to kill*

Variables: Killer_position_x, Killer_position_y, Victim_position_x, and Victim_position_y, Killed_by (weapon name)

Average distance between killer and victim: The Euclidean distance between the killer position and the victim position is calculated. The average of this distance is plotted for each weapon or method of kill.



Outliers in the data

Observation:

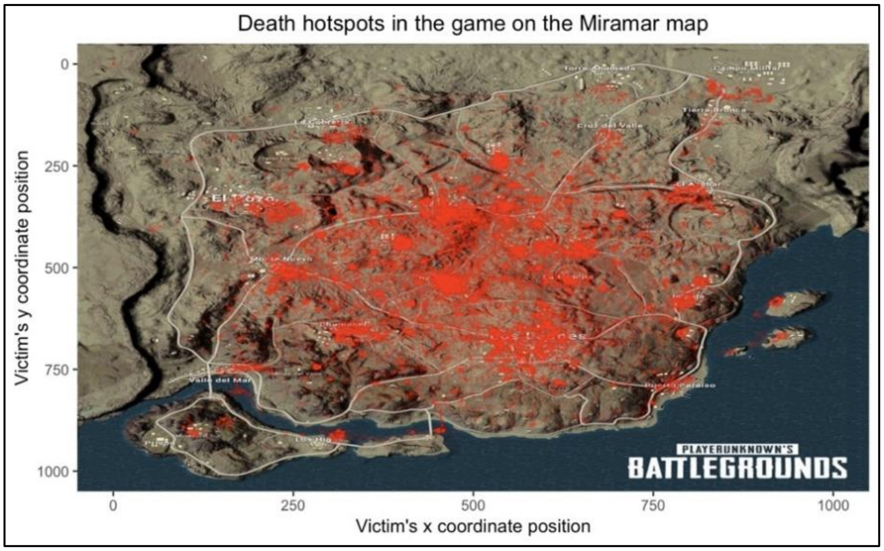
The distance between the killer and victim is very high for some weapons marked by the dotted red block. This is unexpected because the weapons like crowbar, sickle, pan, and machete can be used or killing by punch or car can happen only when the killer and victim are close to each other. Clearly these are outliers.

Findings:

- On taking a closer look at the data, we found that the outliers were being caused by the data which had Victim_position_x and Victim_position_y at 0 or when the Killer_position_x, Killer_position_y were at 0. There is no data with both killer and victim at 0.
- There are about 9526 such rows.
- On removal of these entries, the outliers were removed and distance for different killing methods looked reasonably good.

*Graph 3: Victim Deaths on the Miramar Map*

The Miramar Map has been provided by PUBG which is taken from Kaggle [4]. Here, the victim_position_x and victim_position_y variables has been normalized to fit the scale of the map and the victim hotspots has been plotted by regulating the alpha value. So that all the points on the map can be represented, and at the same time the map isn't overcrowded with points.

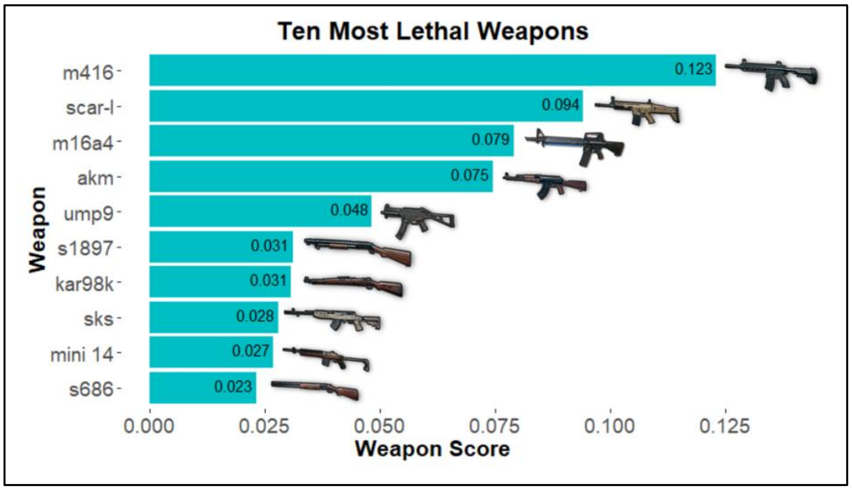

Death hotspots in the game on the Miramar map
Victim's y coordinate position
0
250
500
750
1000
0
250
500
750
1000
Victim's x coordinate position
El Pozo
PLAYERUNKNOWN'S
BATTLEGROUNDS

Findings:

- Victim deaths are mostly observed in the cities or in mountain ranges.
- Deaths in cities are mostly due to the fact that a lot of players take shelter there and target potential victims out in the open.
- Deaths in mountain ranges are because killers can utilize the altitude advantage to target victims in the game.

## 4.2 Research Question 1: What weapons should one use or not use while playing a PUBG game?

*Graph 4: Top 10 lethal weapons*



**Weapon Score**:

Step1: We find the average number of kills per match using the below formula for each weapon

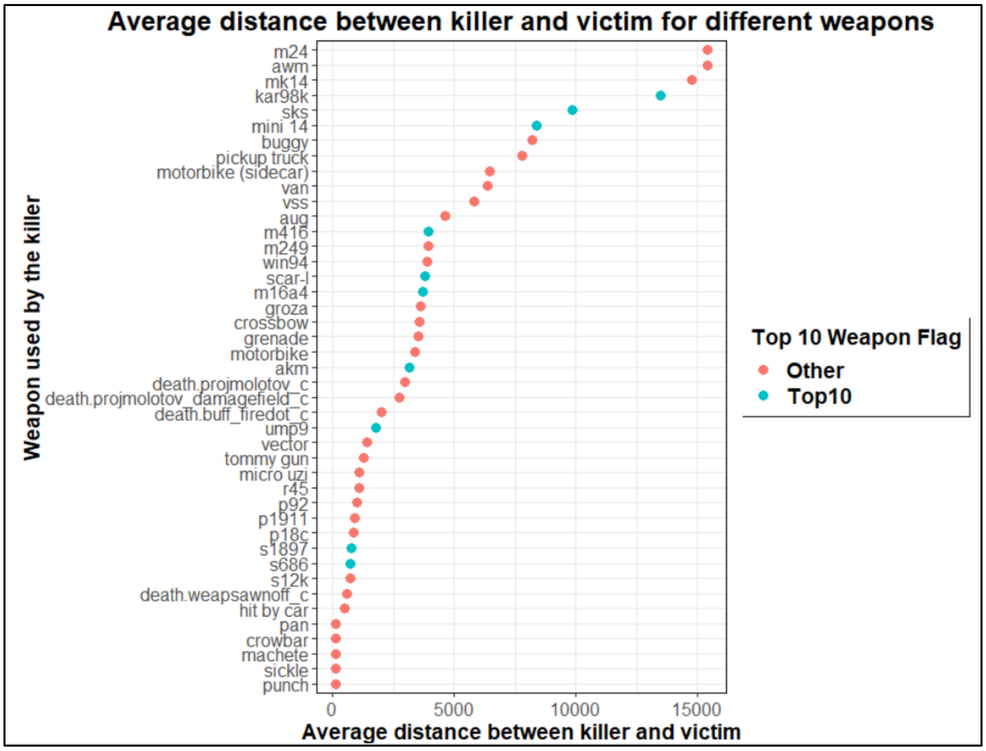
$$\frac{\text{Total \# kills by that weapon}}{\text{Total \# matches where the weapon is used}}$$

Step2: We normalize the average kills by dividing each value by the sum of average kills across the 43 weapons.

Findings:

- The top 10 has only guns. Other methods such as vehicles and tools do not figure in the chart for obvious reasons- guns are better.
- M416 is the top ranked weapon. This is not because it can be used at far away distances as we can see in the below graph5. It could be because it is most efficient in killing instantly and most easily available when compared to the other weapons.

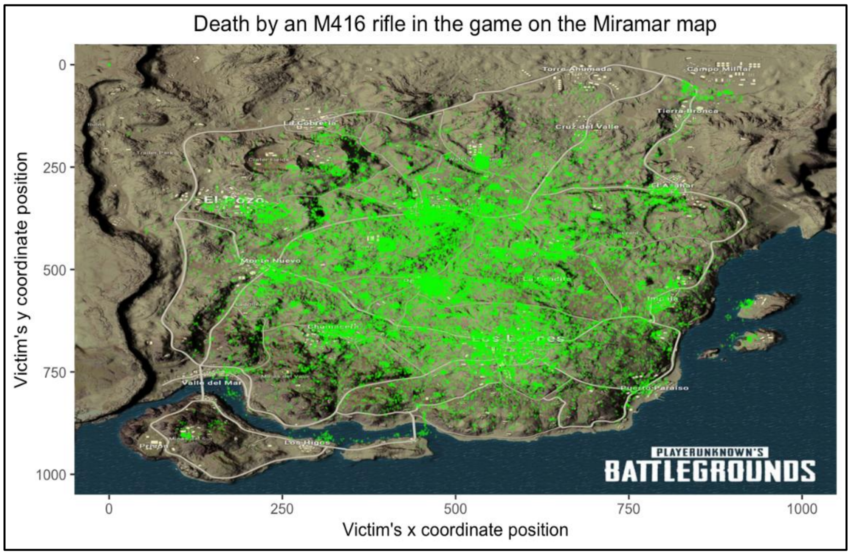
*Graph 5: Average distance of kill for weapons*



Findings:

- The blue points on the graph mark the top 10 weapons that were ranked based on the “killer_score” metric which meant that those weapons have a higher kills per match than the weapons denoted by red points (not a top 10 weapon)
- The top 10 weapon with the highest average distance between the killer and victim is the Karabiner 98 bolt-action rifle weapon which is a gun designed to lock on a target approximately at 13000 units distance away from the killer on an average
- The M416 machine gun has the best weapon score and is used for mostly short distance aim and kills a target at nearly 4000 units distance away from the killer on an average

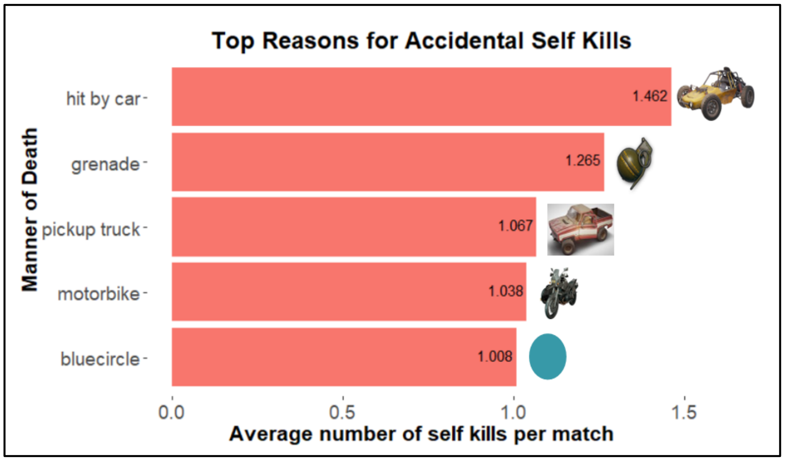
*Graph 6: Victim Deaths on the Miramar Map through M416 rifle (Most lethal weapon)*



Findings:

- Here, the kill pattern is similar to all victim deaths, this could be explained by the fact that M416 has the best weapon score when Miramar map is concerned

*Graph 7: Top 5 propellers of accidental self-kills*

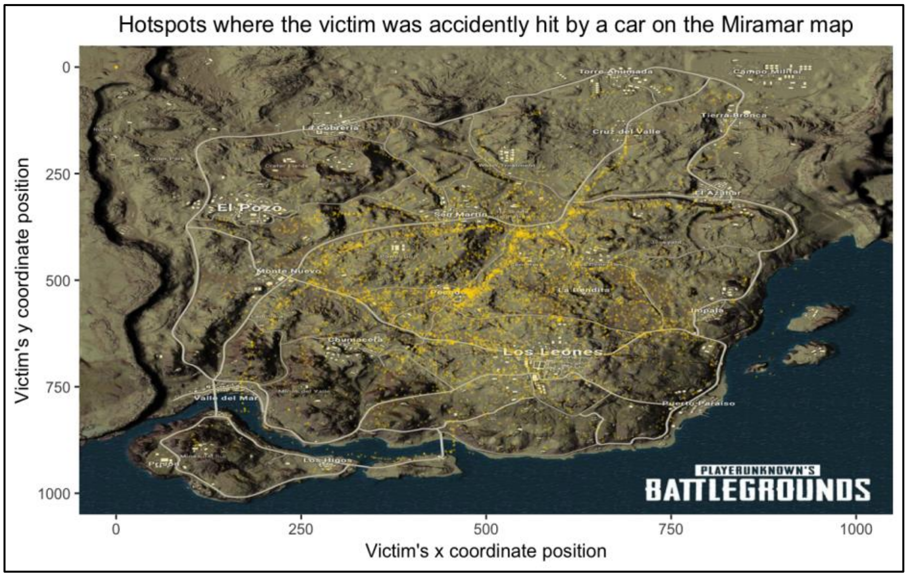


Self-kills – Suicides in the match where killer accidently kills himself/herself

Findings:

- The player usually suffers an accidental death while driving a car as when there is a collision between the car and a building or any object, the driver topples over, and the crash will be fatal. It also explains why pickup truck and motorbike are also in the top five. Although it is not clear why a player should prefer using a car over the other vehicles.
- The top of the list is closely followed by grenade – could be due to accidental self-kill or a suicide.
- We can also see the blue circle in the top 5 which is not surprising since many players are unable to stay in the blue circle and die.

*Graph 8: Victim Deaths who were accidently hit by the car on the Miramar Map*



Findings:

- Here, the map indirectly shows the routes that were frequently taken as a lot of accidental deaths by car happen in the southwest portion of the main island.
- The two mini islands on the map do not have many yellow points as there is no road route from the main land to those islands. People usually travel there by boat or by swimming towards them.
- The northwest portion of the map is either the road that is less frequent or they have less accidental car deaths as players are aware of the blue circle concept (where the battle space keeps getting confined to a smaller area as the game progresses) and tries to stay away from the roads that pass through "La Cobreria" and "Torre Ahumada" for instance.

## 5 Model Building and Evaluation

For predicting whether a player makes it to the top 10 rosters in a game or not, the final feature set had to be finalized. A correlation analysis was done to identify and remove columns with high mutual correlation.

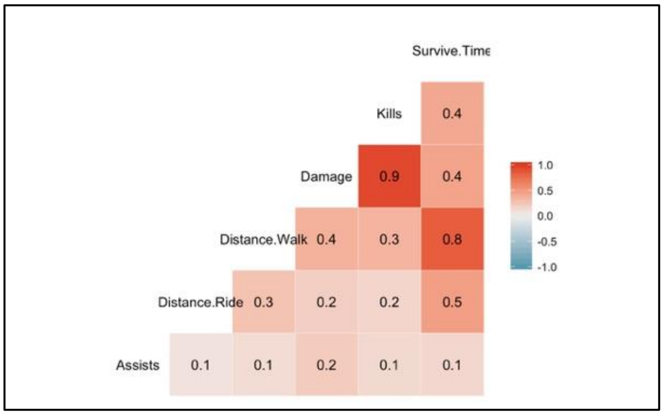


Figure 1: Correlation Matrix

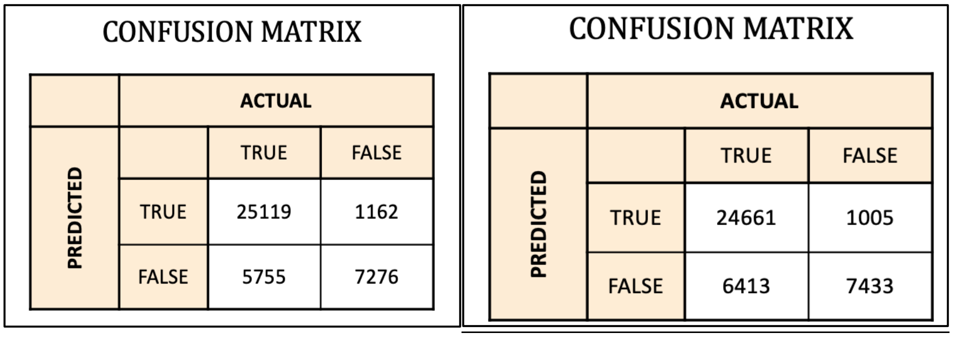
Removing Multicollinearity: Based on the correlation plot as shown in the figure 1, we can see that player dmg (amount of damage done by player) and player kills (number of kills player has achieved). Therefore, player dmg feature was dropped. Also, a high correlation was observed between player_dist_walk and player sunrvive time and therefore the feature of player survive time was dropped. Survival time was removed as it is also a very strong predictor of killer placement.

Sampling and Data Split: Now with the remaining features as the final feature set, the data was down sampled in such a way that a balanced dataset was obtained for classes top 10 or not. Post this, a train-test split in ratio 80-20 was done and a logit model was built. *The match ids in train and test set are exclusive.*

Models: Generalized Linear Models (Logistic Regression) and Random forest models were built on training data and used to predict test data. Advantage of both the models for our data set is that they do not require the independent variables to be normally distributed.

Evaluation:

- The test set was predicted using glm and an F1-Score of 0.68 was recorded.
- By using a random forest model an F1- Score of 0.80 was reported; Also, the feature importance graph was plotted as shown in Graph 9.



GLM

CONFUSION MATRIX

| | | ACTUAL | |
|---|---|---|---|
| | | TRUE | FALSE |
| PREDICTED | TRUE | 25119 | 1162 |
| | FALSE | 5755 | 7276 |

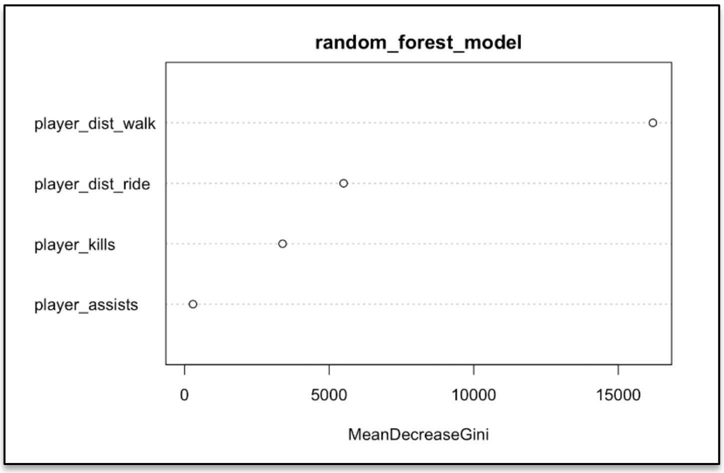
Random Forest

CONFUSION MATRIX

| | | ACTUAL | |
|---|---|---|---|
| | | TRUE | FALSE |
| PREDICTED | TRUE | 24661 | 1005 |
| | FALSE | 6413 | 7433 |

Figure 2: Confusion Matrix

*Graph 9: Feature importance plot from Random Forest*

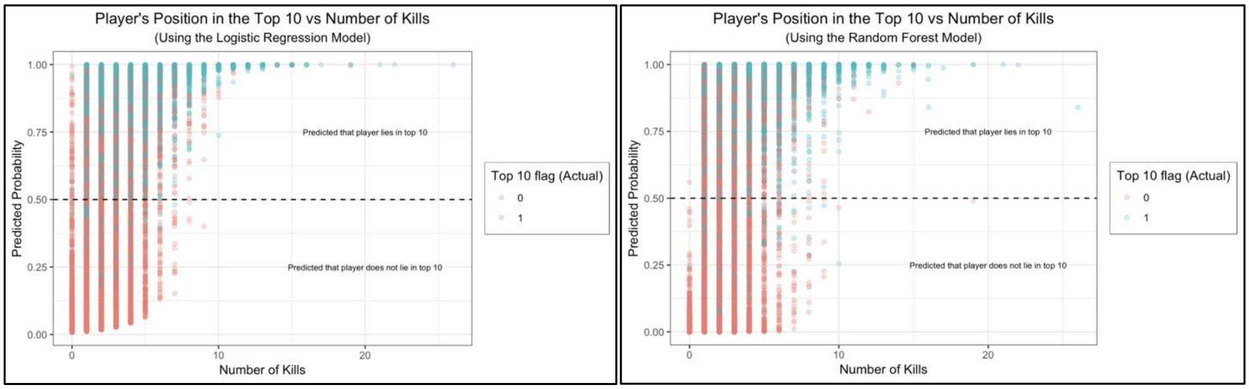


Findings:

- While the predictions from GLM has more false positives, the predictions from Random forest has more true negatives.
- Distance walked and rode by a player are the most important features as per Random Forest.

Predicted Probabilities: A graph between the predicted probabilities of players making to Top 10 roster and the features were plotted to understand the relationships: player kills, player dist walk. player dist ride, and player assists.

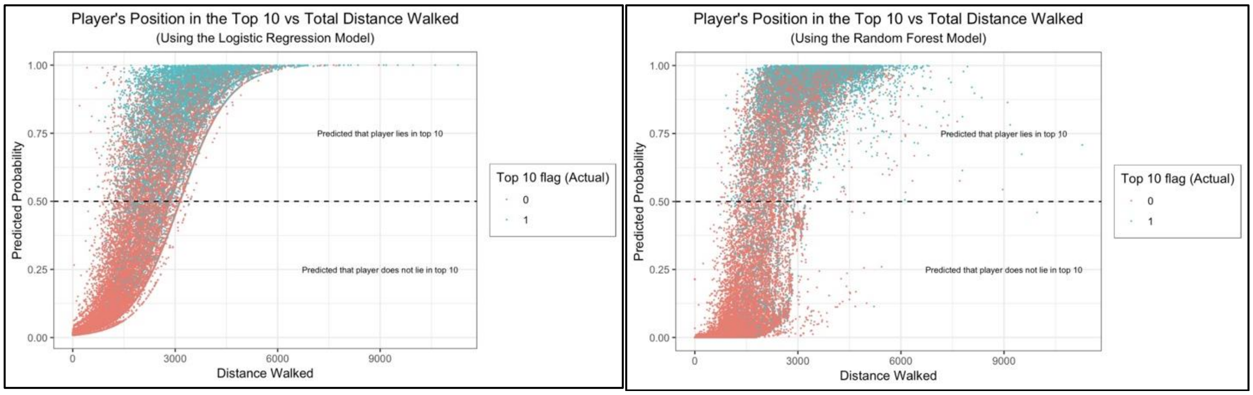
*Graph 10: Plot b/w player kills and probability of making it to Top 10 from Logit and Random Forest models*.


GLM

Random Forest

From the above graphs, it can be concluded that while not exactly a linear relationship, it was noted from the given data that most of the players that make it to the top 10 rosters in the game, generally make more kills.

*Graph 11: Plot b/w dist_walk and probability of making it to Top 10 from Logit and Random Forest models*.

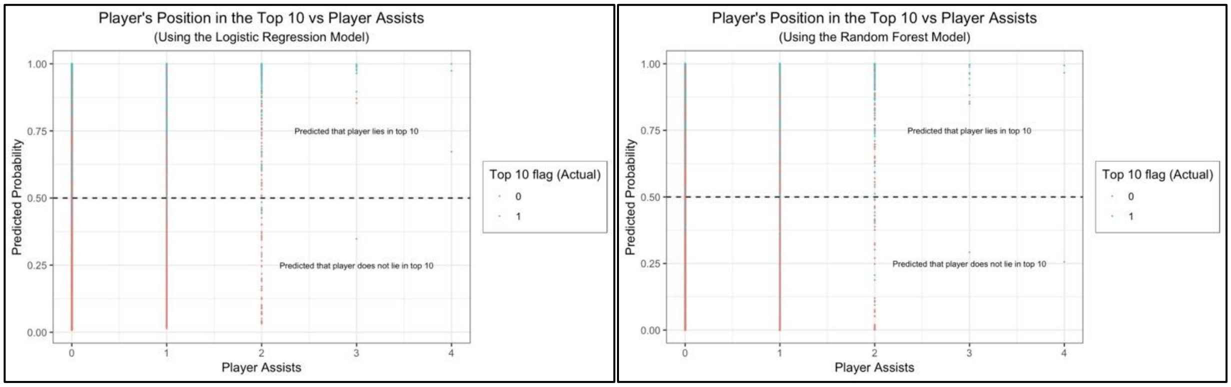

GLM

Random Forest

There is a nearly *linear relation* between player distance walked by players and the probability of a player making it to the top 10 rosters. From this, we can understand that when a player is constantly moving around, there is more chance of him making it to the top 10 rosters. This is understandable as players have to take an aim to kill a person and with a person constantly moving around, the probability of being killed drastically reduces thus increasing one's probability of making it to the top 10 rosters in the game. It is also why the player is able to stay in the blue circle and not get killed. Logit model does a better job at differentiating between the top 10 vs. others than Random Forest model.

There is no discernible relationship between distance ridden by car/bike by a player and the probability of making it to the top 10 rosters. A possible explanation for this behavior might be that since this data is of the last quarter of 2017 (the year the game was released), more and more new players were joining the game and were still figuring out how to drive a car or as to how a car could be effectively employed to survive longer.

*Graph 12: Plot b/w player assists and probability of making it to Top 10 from Logit and Random Forest models*.

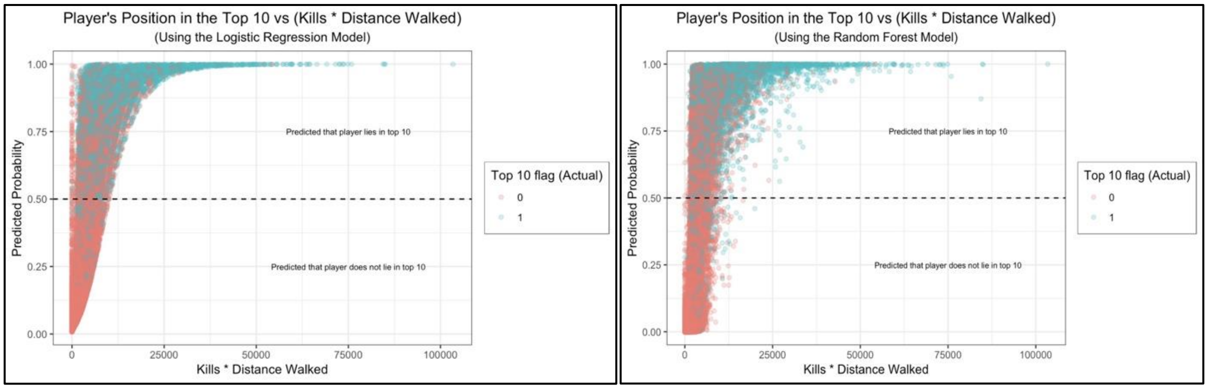


GLM

Random Forest

Clearly, it can be seen that player assist is not a good predictor of a player being in Top 10 or not. It was also observed from Random Forest feature importance plot (Graph 9) that player assist is the least important feature.

*Graph 13: Plot b/w interaction of dist_walk and player_kills and probability of making it to Top 10 from Logit and Random Forest models*.

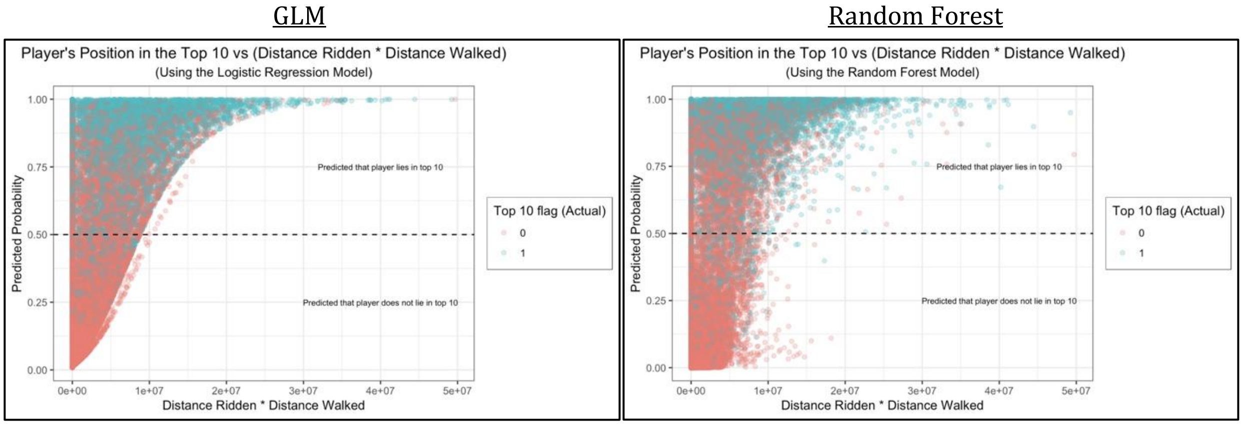


GLM

Random Forest

On adding the interaction between number of kills and distance walked by a player, the transition between predicted probabilities become smoother and makes the differences more discernible.

*Graph 14: Plot b/w interaction of dist_walk and dist_ride and probability of making it to Top 10 from Logit and Random Forest models*.



## 6 References

[1] https://www.kaggle.com/skihikingkevin/pubg-match-deaths

[2] https://en.wikipedia.org/wiki/PlayerUnknown%27s Battlegrounds

[3] https://www.ssc.wisc.edu/sscc/pubs/RFR/RFR RegressionGLM.html

[4] https://www.kaggle.com/skihikingkevin/final-circle-heatmap