MINING INSIGHTS FROM PUB-G DATASET FOR WINNING

Sri Varsha Chellapilla(srchell) Vanita Lalwani(vlalwan) Naba Kishor Sahoo(nabsahoo)

***Abstract*—Electronic sports have become the preferred mode of entertainment for players, who now support a global media outlet just like bystanders. Esports analysis has progressed to meet the need for data-driven critique and is now centered on digital competitor evaluation, methodology, and expectations. Previously, game data from a wide range of players, from casual (non-professionals) to proficient, has been used in studies. In contrast to hobbyists and less talented players, however, proficient players had persevered. Given the scarcity of expert data, a significant question is whether the given match dataset can be used to create data-driven models that predict winners in completed matches and provide real-time in-game analytics for fans and broadcasters to see the acquired data has been mined for predicting the winner.**

Keywords— Battlegrounds, player kills, player damage, player survival, guns, weapons, survivor, team, team placement.

Introduction:

In this report, we will show our approach in exploring and predicting final leader board placements in Player Unknown's Battlegrounds matches. We first give background information on the video game and context for the problem. We do exploratory data analysis. Afterwards, we use K-means for creating more insightful features that better predict the target variable. We also discuss the interesting discoveries we made when solving this problem. Finally, we discuss future steps to improve our models.

Dataset:

The dataset has been acquired from Kaggle which has the following two csv files 'aggregate_match_stats' and 'kill_match_stats'. The aggregate_match_stats csv had the information regarding the match in which the player played and match statistics like player mode, the match duration, the total no. of players in the team etc. The kill_match_stats csv had player information on the weapons used, the weapon used to achieve the maximum damage, the player position etc,

We will discuss in the EDA section how the dataset has been used to find the winning player inferences.

Methods:

1. Data Preprocessing:

After the process of collecting data of size 2GB, the data had to be preprocessed to remove the noise and outliers. This is the most time-consuming technique because there are so many possibilities that your data is still noisy. In our model we did the following:

- As the data obtained was with over 67M records in aggregate csv and over 65M records in deaths csv. We down sampled the data for training our model.
- We cleaned the dataset by eliminating the records with missing values, NaN and dropped the columns that were not required.
- We labelled players ending the game within top 3 as winners and others as losers.
- While down sampling we have identified the imbalance in our class labels hence we down sampled by maintaining a 1:1 ratio. As a match can have only 3 players as top position and the rest 97 were considered as losers the imbalance occurred. So we chose random data from each csv to handle this imbalance.
- The new data set obtained after data preprocessing has been saved as '*_mini' and this has been used throughout the clustering.
- The 'match mode' was a string categorical datatype with two categories like 'FPP and TPP shooter' which was label encoded so that we could use the clustering on it.

2. *Exploratory Data Analysis (EDA):*

The EDA technique was used to explore the cleaned data to find the correlation between the aggregate and kill dataset that we have. Then the explored data has been plotted using various charts to visualize the data.

In the EDA we have focused on the following findings:

Finding 1: **Finding the famous guns used by the players in the winning positions (top 3):**

To explain how a weapon can be a winning factor for a player to be able to survive in the game till the end we have a histogram graph plotted (figure 1). A weapon which can help a player achieve maximum number of kills throughout the duration of the match is considered as one of the best weapons to have.

- From the deaths csv we have obtained the winners from which we have explored to find the guns used by the players in top 3 ranks in the match.
- From the plot, we can infer that M416 has the highest kill count which was used by most players in the top ranks.
- The guns have been color coded as below:
- Red: Assault weapons
- Yellow: Snipers
- Green: Submachine gun
- Black: Shot guns

This information could be useful for a player to understand which weapon needs to picked while playing the game.

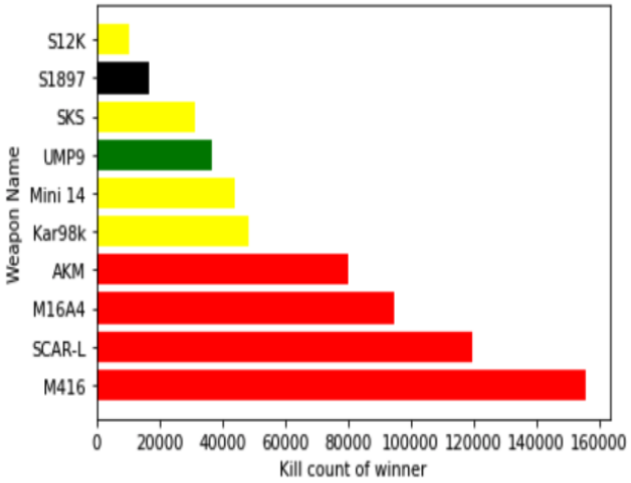


Figure 1

Finding 2: **Explore the data to find the top 3 guns used in every 5 min time interval:**

A typical match in the game lasts for about 35 mins, which has been inferred from the dataset we have. To identify which weapon should be used by a player to be able to survive which is the ultimate goal to be a winner, we now find the which weapons help the player achieve the greatest number of kills which in turn adds to player stats and increase player level. Here we have found the max kills by a gun in 5 mins time intervals.

- To find the top 3 guns used in time intervals, the features that were considered are Time Range, MostKillsWeapon, MostKillsCount from the death csv data.
- In the table below it is seen that the Most kills were obtained by the M16A4 gun in the first 5 mins interval followed by AKM and M416.
- The Table 1 gives us insight on the most kills achieved by the gun. This finding corroborates our previous finding.

Finding 3: **Finding the top 10 weapon combinations used through every match:**

A player cannot always survive till the end of the match with just one weapon. The best way to win is always use multiple weapons as and when required. This finding aims to find such weapon combinations which were used by players in the winning positions. This can help future player understand which weapons to have to win the game.

- To explore various weapon combinations used by players throughout the match which obtained maximum kills.
- This data gives insights on which weapon combinations work best in case of survival.
- In the table the weapon combination that works is M416-Kar98k-M416.
- We can correlate this with our above Findings which match with the top 3 guns.
- Also, we can infer that a player with M416 also has better survival chances.
- In table 2 we can find the info on finding 3.

Time_Range	MostKillsWeapon	MostKillsCount	SecondMostKill
5.0	M16A4	1442.0	
10.0	M416	1057.0	
15.0	M416	1484.0	
20.0	M416	1892.0	
25.0	M416	3498.0	
30.0	M416	7798.0	
35.0	M416	232.0	

Table 1

Count	Weapon_Combinations
1581	M416
1128	SCAR-L
705	M16A4
531	AKM
316	M416 Kar98k M416
307	SCAR-L M416
298	AKM M416
261	M16A4 M416
229	M16A4 SCAR-L
204	SCAR-L Kar98k SCAR-L

Table 2

Apart from the above 3 finding a heatmap has been generated to identify the correlation between different aspects. We achieved this correlation matrix using the aggregation dataset.

From the heatmap (Figure 3) produced we could infer that for team placement is highly correlated to player survival time, and next major factor is player_assist, player_dist_ride, player_dmg and player kills.

This would help in future to gain more insights on how factors like player_dist_rise (which means a player using a vehicle to travel in the game) will help to survive the game and be in top 3 positions.

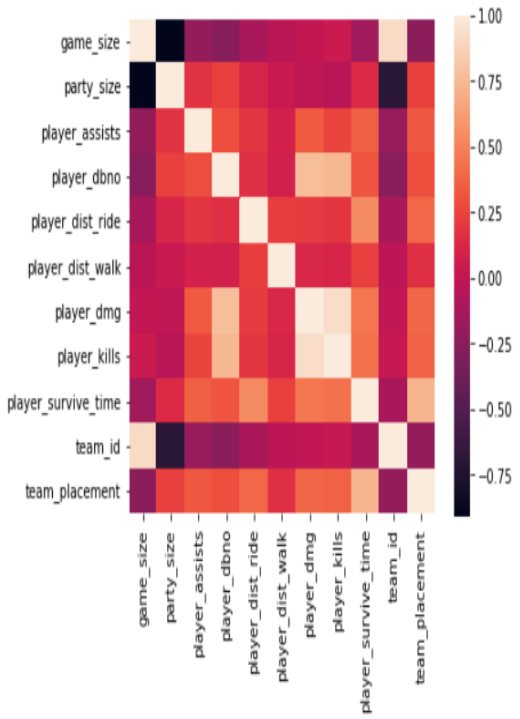
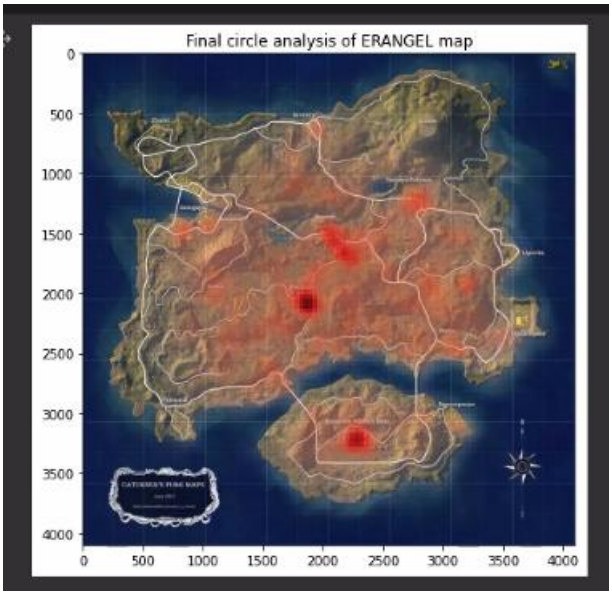


Figure 3

Finding 4: **Final circle Analysis for ERANGLE map:**

In this game a player can chose to play in different match which could yield different results. A map circle usually reduces in size as the number of players decrease and the match begins to approach the end.

In the below figure we see the heatmap for the ERANGEL map.



- In the above map the darker spots represent where the last circle usually is present in map. This helps the player to move in that direction so that he is not killed by the other factors.

3. *Clustering*:

The process of making a group of abstract objects into classes of similar objects is known as clustering. This is an important part to identify the winning cluster. To achieve this kmeans clustering has been used.

K-means clustering:

It is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K.

- The K-means has been run on the aggregate data The optimal value for k is determined as 5.
- All the records of players were extracted belonging to the cluster which had mean centroids of team placement close to 1 which belongs to all the winners.
- The extracted dataset is merged with the dataset of deaths by having the primary key as player_name and match_id.

We have determined the optimal value for k as 5 from the elbow curve (Figure 4).

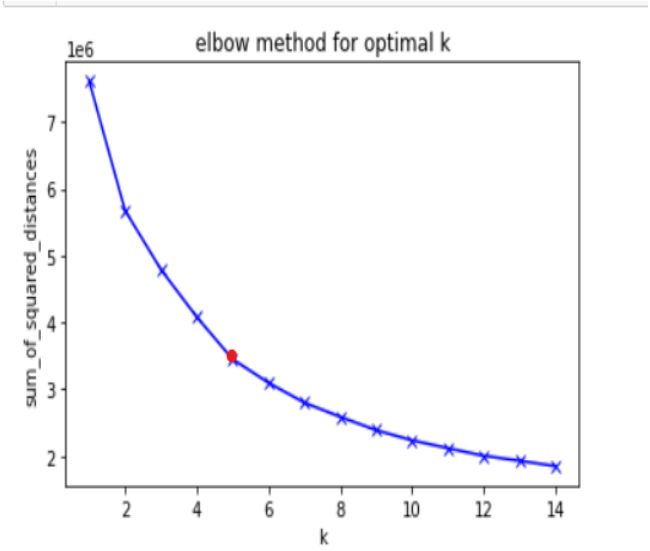


Figure 4

From the clusters we have taken cluster which have more than 90 % winners, see the below table.

	1	0	Class1percentage	Class0percentage
1	72537	5963	0.924038	0.075962
2	257006	24278	0.913689	0.086311
3	12	24	0.333333	0.666667
4	27501	87378	0.239391	0.760609
0	23680	263092	0.082574	0.917426

Inferences from Clustering:

- The pie chart shows the top 10 guns that were used by the winning players and the top 3 guns are M416, SCAR-L and M16A4 which are the same as our Findings from EDA.
- Hence our clustering was able to identify the top winning guns.
- From the generated cluster we extracted other metadata as discussed in the next slide.
- The 75th percentile of party size was found to be 4, hence having a team size of 4 would make us win the game.
- The player playing solo without any teammates had a higher number of kills.
- Average survival time was 1744 sec which

is around 30 mins.

- The 75th percentile for damage afflicted was at 446 with mean at 326 and with a standard deviation of 327, which means in a team of 4 which is majority in our cluster, the entire team afflicted around 326*4=1304 damage which is reasonable for a winning team.

The pie -chart from the clustering is as below:

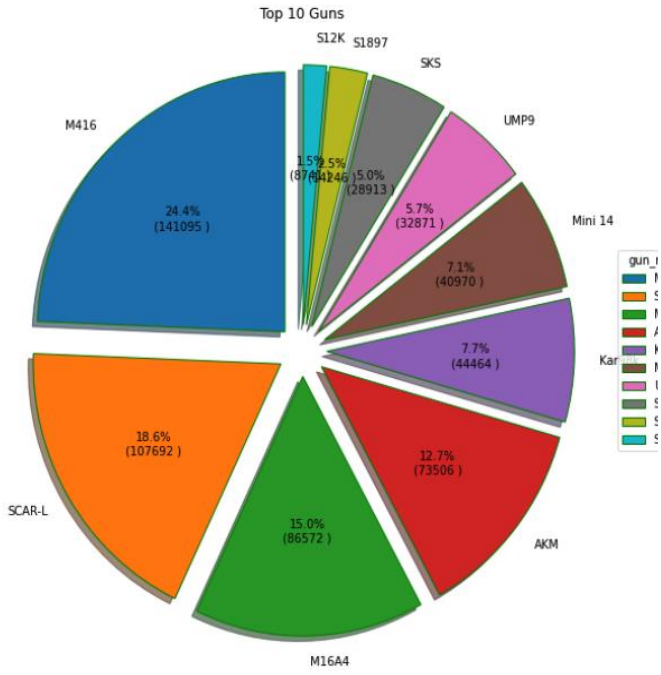


Figure 4

Results:

From our EDA and Clustering on the data we have found that a weapon plays an important role in player survival.

All the findings have corroborated the finding that the best weapon in the game that a player could have is 'M416'.

Conclusion:

We were successfully able to utilize K Means to cluster a subset of the player data which corroborated with our EDA and we were able to extract meaningful insights about the cluster of players by utilizing Kmeans describe feature.

Discussions:

The team approached this PUBG dataset with a focus on exploratory data analysis and mining insights. Since we had domain knowledge regarding the topic, we could find interesting trends and create better features that represented the data. Given more time, we would ensemble high performing models.

In the future all the maps in the PUBG game can be explored to find specific information on how a players performance affects based on the map.

The major problem that was faced during this project was to down sample the data and identify the required data.

References/ Citations:

We contribute a lot of our inspiration from the public kernels for this competition. Here are some of the kernels that were especially helpful in our learning process:

- https://www.kaggle.com/carlolepelaars/pubg-data-exploration-rf-funny-gifs
- https://www.kaggle.com/deffro/eda-is-fun
- https://www.kaggle.com/rejasupotaro/effective-feature-engineering
- https://www.kaggle.com/mlisovyi/relativerank-of-predictions/notebook
- https://www.kaggle.com/jsaguiar/lightgbm-with-simple-features
- https://www.kaggle.com/code/zyabxwcd/rf-classifier-tells-if-player-in-top-10-5-or-3
- https://ieeexplore.ieee.org/document/1017616