Electronic Sports: Winner Prediction

Hao Ding
Department of Statistics
University of Connecticut

December 15, 2023

Abstract

This paper explores factors influencing player success in the popular online multiplayer battle royale game, PlayerUnknown's Battlegrounds (PUBG). Utilizing data from the PUBG Developer API, which includes information from over 65,000 recorded games, we employ correlation analysis, feature importance from tree-based models, and recursive feature elimination to identify key elements impacting player performance. The dataset encompasses various metrics such as assists, boosts, revives, damage dealt, team kills, kill place, distance traveled, weapons acquired, match duration, and win place percentile. Our analysis reveals correlations between different factors and highlights the significance of kills and walk distance in achieving better win placements. The methods section details the approach, including the use of ensemble models like Random Forest and Gradient Boosting, as well as deep learning models for predictive tasks. The discussion emphasizes the need for a large and diverse dataset, periodic model updates due to game dynamics, collaboration with PUBG experts, and the inherent unpredictability of player strategies. While machine learning models offer valuable insights, they should be interpreted cautiously, recognizing the game's dynamic nature and unpredictable elements.

1 Introduction

PUBG (PlayerUnknown's Battlegrounds) represents a pivotal development in the gaming world, particularly within the rapidly growing eSports sector. This game, a brainchild of PUBG Corporation, a subsidiary of the South Korean company Bluehole, has not only garnered immense popularity but also significantly influenced the battle royale genre since its inception in March 2017. Its design, inspired by the Japanese film "Battle Royale", offers a unique blend of survival, exploration, and combat, making it a standout in the gaming community.

The gameplay of PUBG is both engaging and complex. It drops up to 100 players onto an island where they must scavenge for weapons and equipment. The players then battle it out, with the play area gradually shrinking, forcing closer encounters and intensifying the game's pace. This innovative approach has not only attracted millions of players worldwide but also set new precedents for interactive and competitive gaming. As a result, PUBG has been a driving force behind the explosion of the battle royale genre, influencing numerous other games and becoming a staple in competitive gaming events like the PUBG Global Championship.

In recent years, the intersection of data analytics and gaming has opened up new frontiers in understanding player behavior and game strategies. In competitive games like PUBG, where player decision-making and strategy are key, analyzing in-game data can yield insights into effective tactics, player tendencies, and overall game dynamics. This application of analytics is becoming increasingly crucial in the professional eSports arena, where understanding the nuances of player actions can give teams a competitive edge.

Our research focuses on leveraging this intersection of gaming and data analytics. By examining a comprehensive dataset from over 65,000 PUBG games, we aim to uncover the underlying factors that contribute to player success. Specifically, we are interested in identifying which behaviors and strategies lead to more kills or longer survival in the game, ultimately contributing to a player's or team's victory. This analysis not only has implica-

tions for player performance enhancement but also provides deeper insights into the strategic fabric of PUBG. Through this study, we hope to contribute to the growing body of knowledge in eSports analytics, offering both theoretical and practical insights into one of the most popular games in the modern gaming era.

The rest of the paper is organized as follows. The data will be presented in Section 2. The methods are described in Section **??**. The results are reported in Section 4. A discussion concludes in Section 5.

2 Data

Our dataset was sourced from the PUBG Finish Placement Prediction competition on Kaggle. This dataset includes a wide range of variables collected from actual PUBG games, offering a rich foundation for analyzing player behavior and game outcomes.

2.1 Sample Size and Representation

The dataset comprises data from over 65,000 PUBG matches, encompassing a diverse array of player skills and strategies. This extensive collection allows for a detailed examination of various factors influencing player success in PUBG. we can see in the table 1

2.2 Data Processing

The data preprocessing was performed using R, employing a series of steps to ensure the integrity and suitability of the data for subsequent analysis. Initially, missing values within nominal predictors were handled by introducing a new level to represent novel categories, thus accounting for potential new factor levels that were not present during the training phase. For numeric predictors, missing values were imputed with the mean of the respective variables to preserve the dataset's integrity without introducing significant bias.

Furthermore, numeric predictors were standardized, scaling each feature to have a mean

Variable	Description
Assists	Number of enemy players this player damaged that were killed by teammates.
Boosts	Number of boost items used.
Revives	Number of times this player revived teammates.
Heals	Number of healing items used.
Damage Dealt	Total damage dealt. Self inflicted damage is subtracted.
TeamKills	Number of times this player killed a teammate.
Kill Place	Ranking in match of number of enemy players killed.
Kills	Number of enemy players killed.
Ride Distance	Total distance traveled in vehicles measured in meters.
Swim Distance	Total distance traveled by swimming measured in meters.
Walk Distance	Total distance traveled on foot measured in meters.
Weapons Acquired	Number of weapons picked up.
Match Duration	Duration of match in seconds.
Win Place Perc	Percentile winning placement, where 1 corresponds to 1st place, and 0 to last place.

Table 1: Description of Variables in the PUBG Dataset

of zero and a standard deviation of one. This normalization is a crucial step, especially when employing algorithms that assume data is centered and scaled, such as support vector machines or k-means clustering.

The combined use of these preprocessing techniques ensures that the analysis is conducted on a dataset that is free from common issues that could undermine the validity of the results. Each step was carefully documented to maintain the reproducibility of the tudy.

Listing 1: R code for data preprocessing

```
# Create the recipe
rec <- recipe(~., data = data) %>%
step_novel(all_nominal_predictors()) %>%
step_unknown(all_nominal_predictors()) %>%
step_impute_mean(all_numeric_predictors()) %>%
step_normalize(all_numeric_predictors())
```

```
# Prepare and bake the recipe
prepped_data <- prep(rec, training = data)
data_transformed <- bake(prepped_data, new_data = NULL)
```

2.3 Statistical Summary

A statistical summary is provided, showcasing the distribution and range of key variables such as damage dealt, distance traveled, and number of kills. This summary offers preliminary insights into general trends and patterns within PUBG gameplay.

Statistic	Damage Dealt	Ride Distance (meters)	Kills
Minimum	0.00	0.00	0.0000
1st Quartile	0.00	0.00	0.0000
Median	84.16	0.00	0.0000
Mean	129.74	590.99	0.9184
3rd Quartile	185.40	0.01	1.0000
Maximum	6229.00	40700.00	58.0000

Table 2: Statistical Summary of Key Variables in PUBG Dataset

In this table 2, three key variables (Damage Dealt, Ride Distance, and Kills) from the PUBG dataset. It is observed that while most players have a lower number of kills (median of 0), some players achieve a significantly high number of kills (maximum of 58). Similarly, although most players travel short distances by vehicles, some cover distances as long as 40,700 meters. These figures reflect the diverse player behaviors in PUBG gameplay.

3 Methods

3.1 Data Preprocessing

Data preprocessing was a pivotal step in preparing the PUBG dataset for subsequent analysis. The process commenced with exploratory data analysis (EDA), which involved employing descriptive statistics and visualizations to unearth underlying trends, detect outliers,

and observe patterns within the data. After EDA, we addressed missing values through appropriate imputation techniques, and feature engineering was undertaken to enhance the predictive capability of the dataset. This included encoding categorical variables and normalizing numerical variables to ensure uniformity.

A crucial part of preprocessing was the division of the dataset into training and testing sets. This split was essential for validating the model's performance and ensuring its ability to generalize to new, unseen data. Typically, a portion of the data (e.g., 70-80 percentage) was allocated to the training set, with the remainder reserved for testing.

Moreover, to refine the feature set and improve model efficiency, a correlation analysis was conducted. This analysis helped identify and eliminate highly correlated variables, thereby reducing dimensionality and potential multicollinearity issues. Variables with high correlation were evaluated, and those offering less informational value were selectively removed. This step was crucial to simplify the model, enhance interpretability, and potentially improve performance by focusing on the most relevant predictors.

In addition to the initial preprocessing steps, the data was further manipulated to facilitate a more straightforward interpretation of the results. Specifically, the continuous variable 'winPlacePerc', representing the percentile rank of a player's finish in a match, was transformed into a categorical factor with two levels: 'defeat' and 'win'. This dichotomy was established based on whether the player's rank was less than the top percentile (1), with the rationale that a victory can be considered as achieving the highest possible rank in a match. The code snippet below illustrates this transformation in R:

Listing 2: Creating a binary outcome variable

```
df3 <- df2 %>%
mutate(win = factor(winPlacePerc < 1, c(TRUE, FALSE),
c("defeat", "win")))
```

This transformation was pivotal in simplifying the analysis, allowing subsequent models to predict a binary outcome of either winning or not winning, which is more aligned with

common objectives in game outcome predictions.

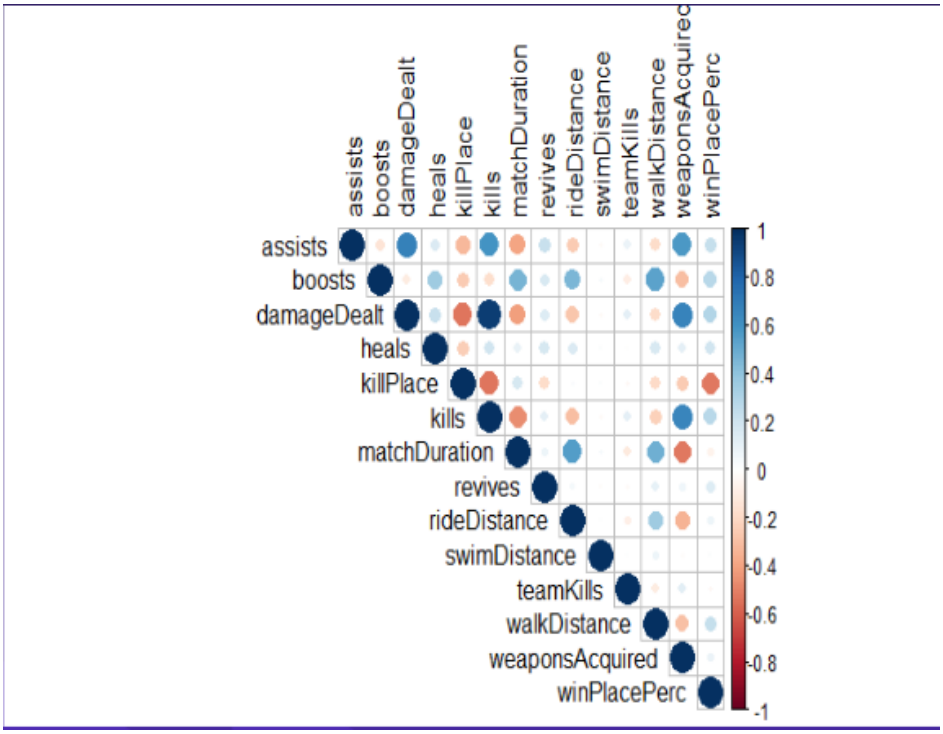

Figure 1: Correlation matrix of PUBG gameplay variables.

Figure 1 presents the correlation matrix of selected gameplay variables from the PUBG dataset. Each cell in the matrix provides the correlation coefficient between two variables, ranging from -1 to 1. A positive correlation (closer to 1) suggests a direct relationship, where an increase in one variable tends to be associated with an increase in the other. Conversely, a negative correlation (closer to -1) indicates an inverse relationship. Notably, the variables such as 'kills' and 'winPlacePerc' show a strong positive correlation, implying that higher kills are strongly associated with better placement in a match.

3.2 Model Selection

Considering the project's aim to predict player success in PUBG, both regression and classification models were deemed suitable. Initial models included decision trees to establish a baseline performance. To capture more complex relationships within the data, advanced

models such as Least Absolute Shrinkage and Selection Operator(Lasso Regression), Specified Lasso, and Support Vector Machines (SVM) were also employed. These models were chosen for their ability to handle large datasets and provide insights into feature importance.

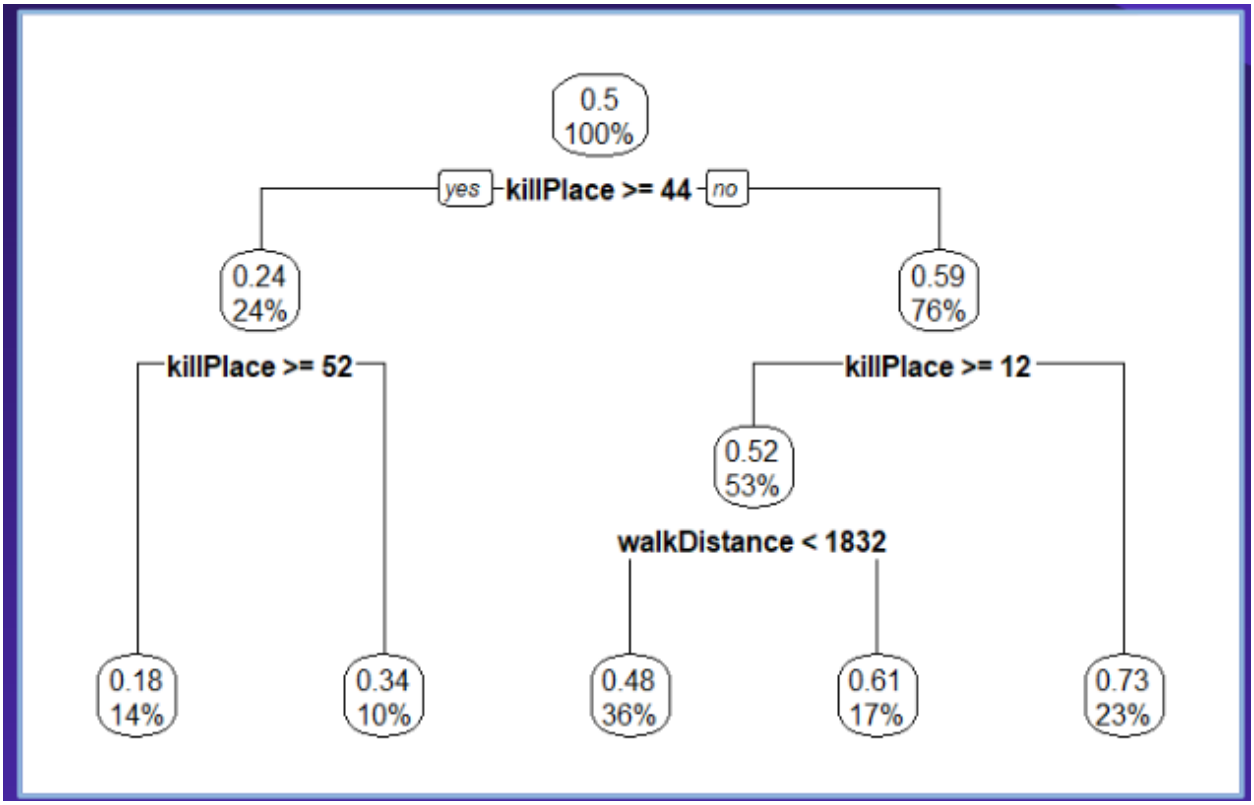

Figure 2: Decision tree model based on PUBG gameplay variables.

In our model training phase, we utilized a decision tree method, leveraging variables identified as significant from the correlation analysis. As illustrated in Figure 2, the decision tree segregates players into groups based on 'killPlace' and 'walkDistance'. The tree structure indicates that 'killPlace' is the most significant variable, initially splitting the players at a 'killPlace' threshold of 44. Further splits are made based on additional thresholds of 'killPlace' and 'walkDistance', highlighting their impact on the prediction of winning. Notably, 'winPlacePerc' was not used as a predictor in this model. Instead, 'win', a binary outcome derived from 'winPlacePerc', was used to train the model, aligning with the objective to predict the likelihood of winning a match.

3.3 Model Training and Evaluation

Models were trained using a split of the dataset into training and testing subsets, ensuring a robust validation process. Cross-validation techniques were applied to assess model performance and prevent overfitting. Model parameters were fine-tuned through a grid search approach, seeking the optimal combination for the highest predictive accuracy. The final model evaluation was based on metrics appropriate to the type of model — Mean Squared Error (MSE) for regression and accuracy, precision, and recall for classification models.

3.4 Interpretation of Results

The final step involved interpreting the model outcomes to draw meaningful conclusions about player behaviors and success factors in PUBG. The analysis of feature importance provided insights into which variables most significantly impacted player performance, guiding strategies for player improvement and game engagement.

4 Results

5 Discussion

We can choose several models, and compared with the accuracy of those models, and we will choose the highest accuracy model as our final model to predict whether the players or teams can win. So, we need to consider more detail for different modeling methods that will be lots of work. Second point is ensure that we have a sufficiently large and diverse dataset training a robust model. Since this is a video game, PUBG is a dynamic game with constant updates and changes. The model might need periodic updates to adapt to the evolving game environment.

If it is possible, we should work closely with PUBG experts to gain insights into the specific dynamics of the game that may not be immediately evident from the data.

Also we should emphasize a important factor, while machine learning models can provide valuable insights, predicting the outcome of a PUBG involves various unpredictable factors, including player strategies, teamwork, and unexpected events during matches. Always interpret the model's predictions cautiously and use them as a supplementary tool rather than a definitive source for predicting winners.

Aggarwal et al. (2021) Liu et al. (2020) Ghazali et al. (2021)

save those cites for late.

References

Harsh Aggarwal, Siddharth Gautam, Sandeep Malik, Arushi Khosla, Abhishek Punj, and Bismaad Bhasin. Rank prediction in pubg: A multiplayer online battle royale game. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 2*, pages 859–867. Springer, 2021.

NF Ghazali, N Sanat, and MA As' ari. Esports analytics on playerunknown's battlegrounds player placement prediction using machine learning. *International Journal of Human and Technology Interaction (IJHaTI)*, 5(1):17–28, 2021.

Jiaxin Liu, Jiaxin Huang, Ruyun Chen, Tianchang Liu, and Liang Zhou. A two-stage real-time prediction method for multiplayer shooting e-sports. 2020.