PUBG Top 10% Placement Analysis

Chance Robinson, Allison Roderick and William Arnost

Master of Science in Data Science, Southern Methodist University, USA

1 Introduction

[Intro]

2 Data Description

The source data is available on Kaggle.com under the competition PUBG Finish Placement Prediction. The files used in our analysis were transformed to fit the requirements of a binomial logistic regression classifier. The data has also been pre-split into training and test files for consistency when comparing our different model types. Additionally, as the percentage-based nature of the top 10% of players is inherently unbalanced, we've also down sampled the higher frequency data to match that over the lower frequency outcome of the top 10% of players.

`pubg_solo_game_types.csv`

- Filtered for solo only game types

`pubg_solo_game_types_test_full.csv`

- Pre-split for test data

`pubg_solo_game_types_train_full.csv`

- Pre-split for train data without downsampling for the unbalanced response variable

`pubg_solo_game_types_train_downsampled.csv`

- Pre-split for train data with downsampling for the unbalanced response variable

2.1 Data Dictionary

Column Name	Type	Description
DBNOs		Number of enemy players knocked.
assists		Number of enemy players this player damaged that were killed by teammates.
boosts		Number of boost items used.
damageDealt		Total damage dealt. Note: Self inflicted damage is subtracted.
headshotKills		Number of enemy players killed with headshots.
heals		Number of healing items used.
Id		Player's Id

Column Name	Type	Description
killPlace		Ranking in match of number of enemy players killed.
killPoints		Kills-based external ranking of player. (Think of this as an Elo ranking where only kills matter.) If there is a value other than -1 in rankPoints, then any 0 in killPoints should be treated as a “None”.
killStreaks		Max number of enemy players killed in a short amount of time.
kills		Number of enemy players killed.
longestKill		Longest distance between player and player killed at time of death.
matchDuration		Duration of match in seconds.
matchId		ID to identify match. There are no matches that are in both the training and testing set.
matchType		String identifying the game mode that the data comes from.
rankPoints		Elo-like ranking of player.
revives		Number of times this player revived teammates.
rideDistance		Total distance traveled in vehicles measured in meters.
roadKills		Number of kills while in a vehicle.
swimDistance		Total distance traveled by swimming measured in meters.
teamKills		Number of times this player killed a teammate.
vehicleDestroys		Number of vehicles destroyed.
walkDistance		Total distance traveled on foot measured in meters.
weaponsAcquired		Number of weapons picked up.
winPoints		Win-based external ranking of player. (Think of this as an Elo ranking where only winning matters.) If there is a value other than -1 in rankPoints, then any 0 in winPoints should be treated as a “None”.
groupId		ID to identify a group within a match. If the same group of players plays in different matches, they will have a different groupId each time.
numGroups		Number of groups we have data for in the match.
maxPlace		Worst placement we have data for in the match. This may not match with numGroups, as sometimes the data skips over placements.
winPlacePerc		This is a percentile winning placement, where 1 corresponds to 1st place, and 0 corresponds to last place in the match. (to be removed from our binomial classfier so as not to influence our predictive results)
top.10		The target of prediction. This is a percentile winning placement, where 1 corresponds to a top 10% placement a 0 in the lower 90%.

2.2 Exploratory Data Analysis

3 Objective I Analysis

3.1 Question of Interest

3.2 Model Selection

[Logistic Regression, Ridge, Lasso and Elastic Net]

3.3 Comparing Competing Models

3.4 Model Interpretation

3.5 Conclusion

4 Objective II Analysis

4.1 Question of Interest

4.2 Model Selection

[Logistic Regression with interactions, LDA and Random Forest]

4.3 Comparing Competing Models

4.4 Model Interpretation

4.5 Conclusion

5 Appendix

5.1 Exploratory Data Analysis

5.2 Code