DNSC 6279 Data Mining

Final Report - PUBG Finish Placement Prediction Analysis

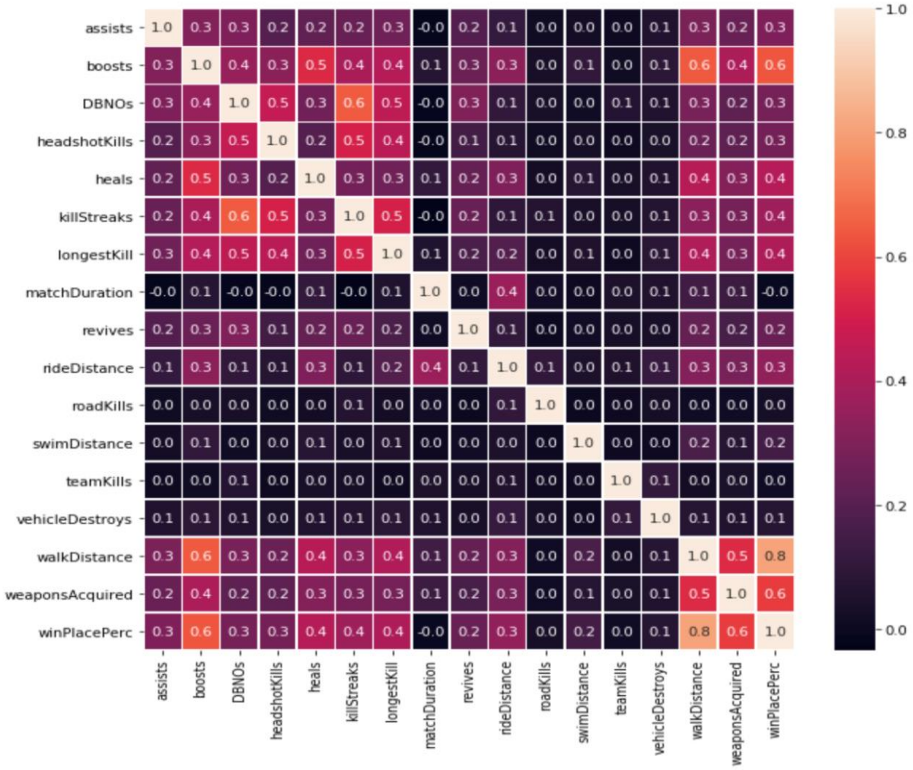
Xinrong Chen, Qianying Diao, Jingyi(Abby) Liu, Qiang Wang

1. Dataset Introduction

Player Unknown's Battle Grounds (PUBG) has enjoyed massive popularity. In the dataset, we are given over 65,000 games' worth of anonymized player data, split into training and testing sets, and asked to predict final placement from final in-game stats and initial player ratings. A detailed description of variables can be found from the following link:
https://www.kaggle.com/c/pubg-finish-placement-prediction/overview

2. EDA

(1) Plot the histogram, get familiar with the property of each variable
(2) Match type matters: The game has 4 modes: solo, double and square. The relationship of independent variables and win percent differs in different match types.
(3) Using heat map, we identify 4 the most correlated variables as 'totaldistance', 'boots', 'weaponAcquired' and 'damageDealt'



3. Data preparation

(1) Dimension Reduction: The original dataset has 27 independent variables which should be reduced. We identify 3 similar variables 'rideDistance', 'walkdistance' and 'swimdistance'. We merged them into one variable 'totaldistance'
(2) Removing outliers: Remove the records showing the player kill others without moving as it is unreasonable.
(3) Binning the levels of the categorical variable: For variable ”matchType”, some records contain unwanted information such as ‘solo-fpp‘ since there are only three types of match, convert records to only 3 levels: ‘solo’, ‘duo’ and ‘squd’.
(4) Multicollinearity: Multicollinearity is not allowed for generalized Linear Model., We generate a heap map. The coefficients of rankpoints, kill points and winpoints, kills and damagedealt, kills and killplace, number of groups and maxplaces are high, so we delete the kill points, kills and rankpoints, maxplaces as well as damagedealt.

4. Modeling

First, we identify whether the dependent variable should be treated as categorical or numerical by comparing model performances of the linear model and logistic model. Generalized linear model turned out to be better so we further our analysis expecting a numeric output in GBM and MLP model.
Overall, we used 4 different models to do the prediction. Generalized Linear Model, Logistic Regression, Multilayer perceptron Model and Gradient boosting machine. Gradient boosting machine is the best performing model with an MAE(kaggle score criterion) of 0.05671.
GBM model tuning: run the origin model, check the parameter of top 10 models and modify the parameters, repeat this process until the MSE narrow down to a steady level.

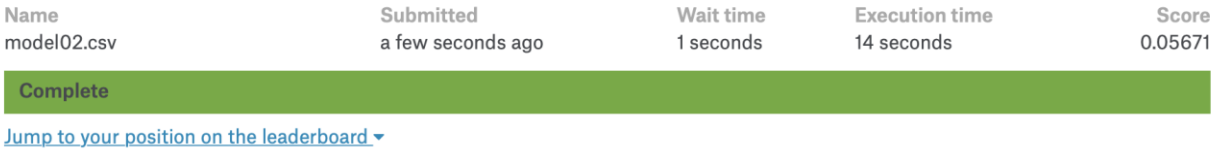
5. Model Comparison (Kaggle scores):

The score of the kaggle is MAE of the models.

(1) Generalized Linear Model

(2) Logistic Regression

(3) Gradient boosting machine



Name	Submitted	Wait time	Execution time	Score
model02.csv	a few seconds ago	1 seconds	14 seconds	0.05671

Complete

Jump to your position on the leaderboard

(4) Multilayer perceptron Model

Name	Submitted	Wait time	Execution time	Score
model03.csv	a few seconds ago	0 seconds	16 seconds	0.08020

Complete

6. Recommendation

The best model GBM shows the 5 best variables are totalDistance, killPlace(the rank of player kills), weaponAcquired, Boosts(number of boost item used) and heals.
Thus, our suggestion for PUBG player are:
(1) being active on moving;
(2) practicing shooting skills and use it in the battle;
(3) sufficient weapons are important to win/ properly and frequently change weapon based on current situation;
(4) sufficient boost and heal items.