Chenxi Yang & Yi Tang

Professor Zhang

2 Dec 2018

Final Project: PUBG Final Placement Prediction on Kaggle

1. Introduction

PlayerUnknown's BattleGrounds (PUBG) is a playershooter game that up to 100 players are dropped in each match while the scope of safety zone is constantly decreases. Players can choose to enter the match solo, duo or with a small team of up to 4 people.

In our project, on the basis of over 65,000 games' worth of anonymized player data that is split into training and testing sets, we are supposed to determine the best model for predicting final placement from in-game stats and initial player rating. (2018)

2. Analyzing Preparation

2.1 Factors Involved

According to the information from Kaggle competition page, there are 29 covariates are involved in our projects, which are listed below.

- 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
- 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. This may be misleading, as downing a player and driving away may lead to a large longestKill stat.
- 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. The standard modes are “solo”, “duo”, “squad”, “solo-fpp”, “duo-fpp”, and “squad-fpp”; other modes are from events or custom matches.
- rankPoints - Elo-like ranking of player. This ranking is inconsistent and is being deprecated in the API’s next version, so use with caution. Value of -1 takes place of “None”.
- 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 - The target of prediction. This is a percentile winning placement, where 1 corresponds to 1st place, and 0 corresponds to last place in the match. It is calculated off of maxPlace, not numGroups, so it is possible to have missing chunks in a match.

(Kaggle 2018)

2.2 Data Reading

For testing dataset, the dimension is 1934174 * 28 while for training dataset 4446966 * 29.

head(test.r)

dim(test.r) #1934174 28

head(train.r)

dim(train.r) #4446966 29

2.3 Data Cleansing

2.3.1 Bug Consideration

As we observed that there are 1535 players with 0 ride and walk distance, we considered them as bugs and removed them.

dim(cheater) #1535 29

2.3.2 Other Changes

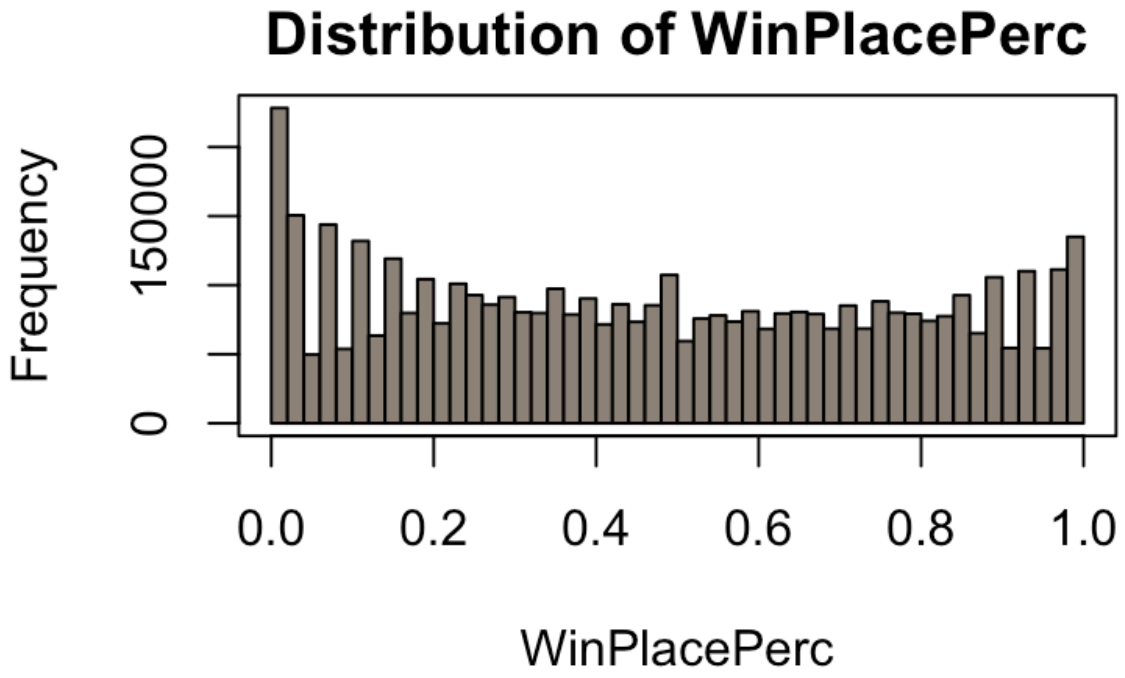
Removed all unused variables (the first three columns), changed -1 into 0 in covariate rankPoints, and checked if NA existed inside and remove them.

3. Data Exploration

3.1 Basic Description Data

We firstly analyzed the Distribution of target variable and got the graph below:

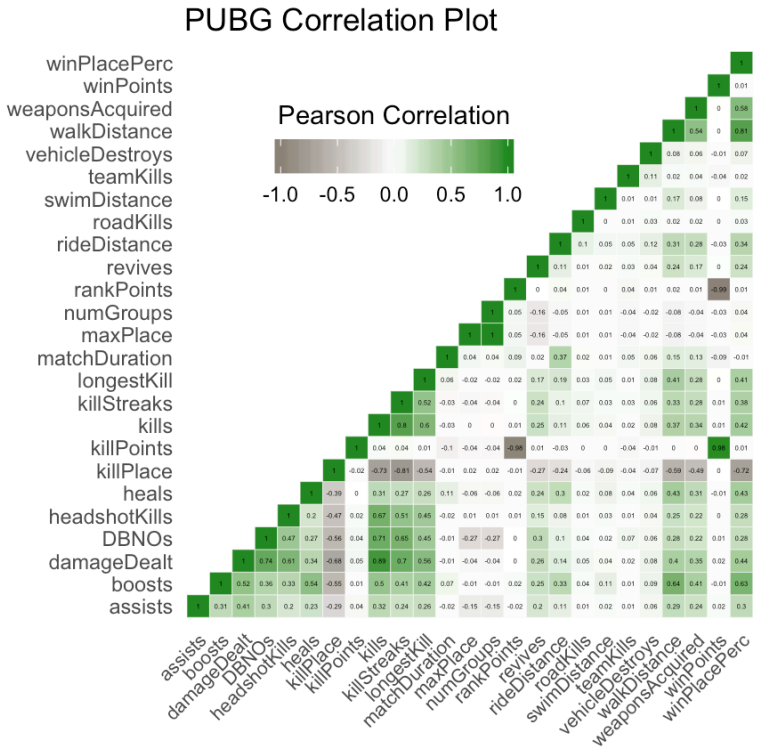
According to this graph, when a player enters a match, he is more likely to be the last place as the



frequency for 0 is the highest.

Secondly, the Pearson correlation between these covariates are posted below:

As the graph indicates, maxPlace and numGroups are totally correlated with coefficient 1 while killPoints and winPoints are highly positively correlated with coefficient 0.96; rankPoints are highly



negatively correlated with winPoints and killPoints with coefficient -0.99 and -0.98 respectively.

3.2 Data Splitting

Training data are splited into 16 types based on matchtype, including crashfpp, crashtpp, duo, duo-fpp, flarefpp, flaretpp, normal-duo, normal-duo-fpp, normal-solo, normal-solo-fpp, normal-squad, normal-squad-fpp, solo, solo-fpp, squad and squad-fpp.

crashfpp	crashtpp	duo	duo-fpp	flarefpp	flaretpp
6285	371	313480	996351	717	2504
normal-duo	normal-duo-fpp	normal-solo	normal-solo-fpp	normal-squad	normal-squad-fpp
199	5489	326	1681	516	17171
solo	solo-fpp	squad	squad-fpp		
181888	536569	626308	1755575		

3.3 Model Selection

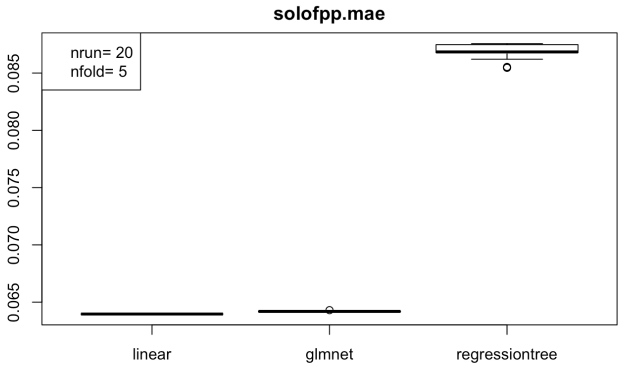
As different types of match may have different best choice of model, so we choose models for each type respectively. In this project, we mainly focus on linear model, penalized linear regression model and regression tree model as classification methods cannot be used in this dataset.

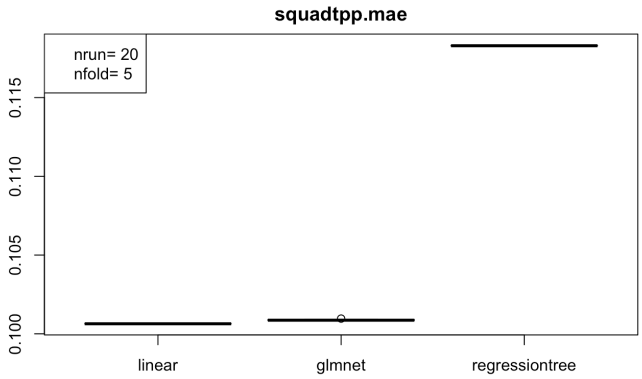
For the purpose of choosing the most suitable seeds and cv folds, we split data into 4 cases: nrow(data)$\leq$500, 500<nrow(data)$\leq$10000, 10000<nrow(data)$\leq$100000 and nrow(data)>100000. The number of seeds are 100, 100, 20 and 20 while that of cv folds are 5, 10, 10 and 5 respectively. All not available terms are converted into 0.

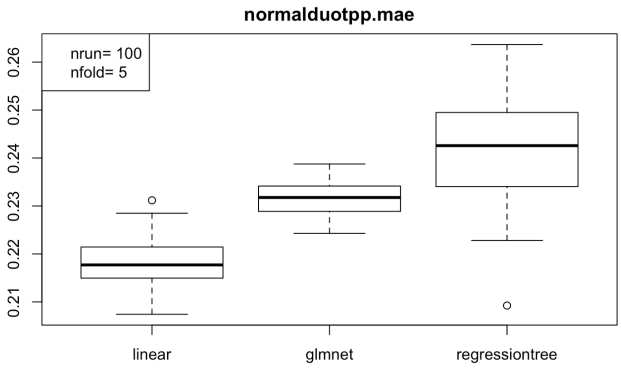
The criterion for best fitting model is finding the least mean absolute error. For each type, firstly using the model given by the training data to predict values on the testing data in each seed. Then, predicted values and observed values are used to calculate mean absolute values. At the end, after several seeds (set above), the one with least mean absolute error are considered as the best fitting model. Furthermore, for linear and penalized linear regression model, the average model is the final one chosen while specially in penalized linear regression model model, the most frequent value of α

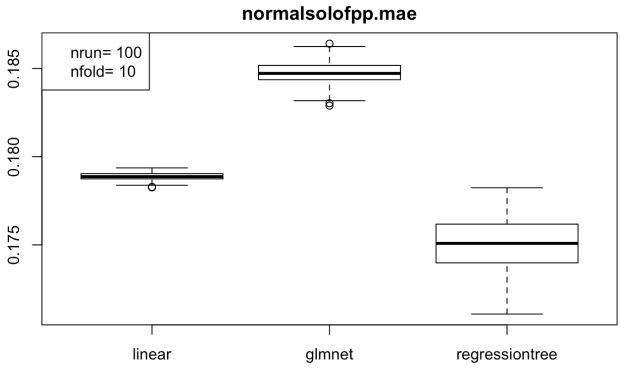
are chosen in the final model; for tree model, choosing the one with best seed.

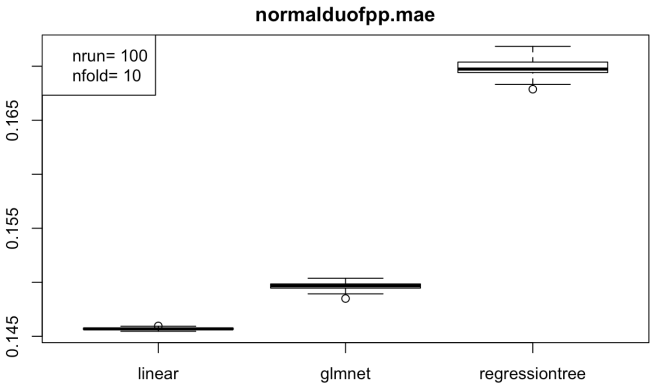
After these steps, we got 16 box plots to compare the best performed model for each type of game.



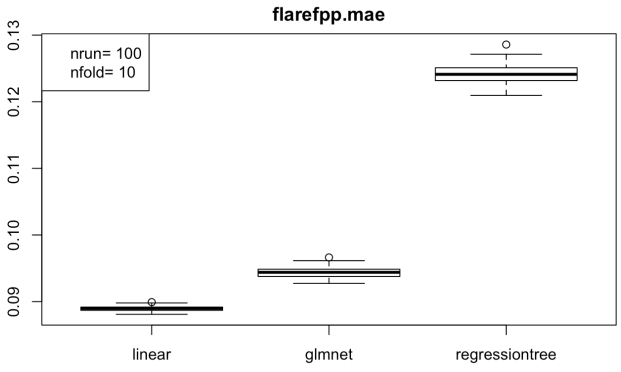


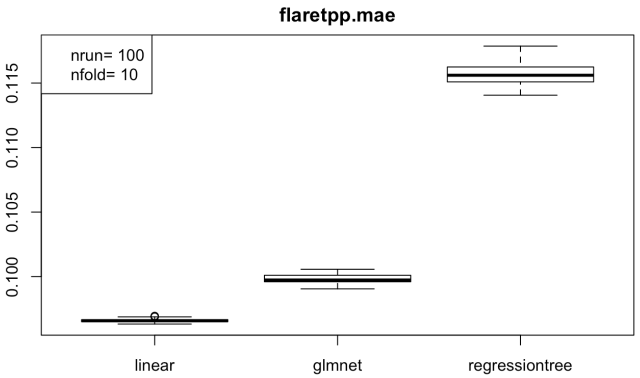


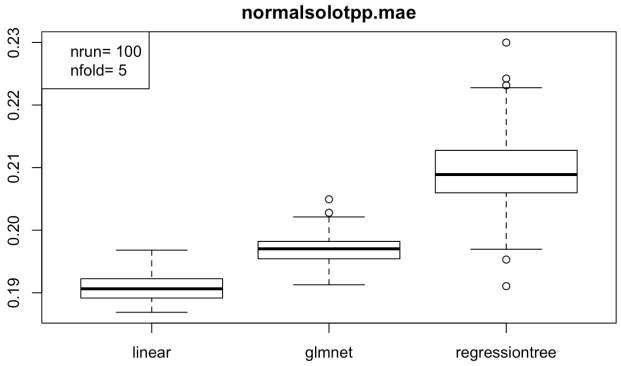


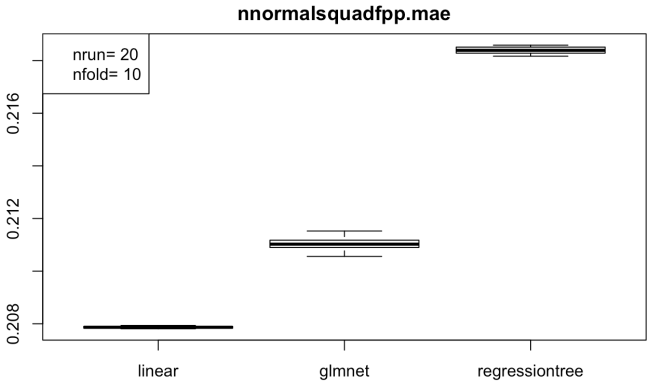


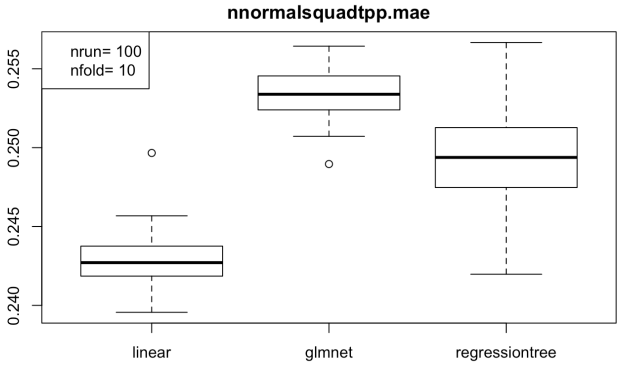
As a result, the linear model performed best in most type of game with least mean absolute error

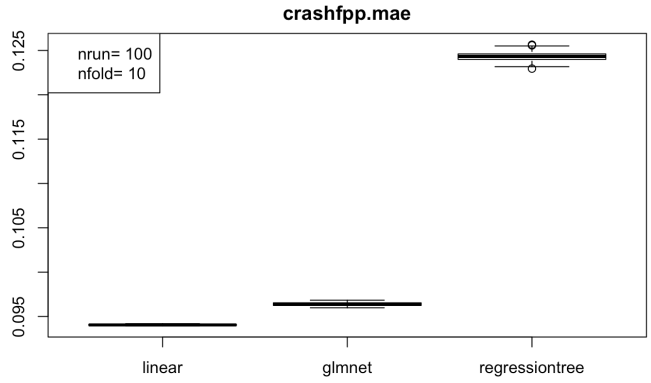

flarefpp.mae
nrun= 100
nfold= 10
0.09
0.10
0.11
0.12
0.13
linear
glmnet
regressiontree

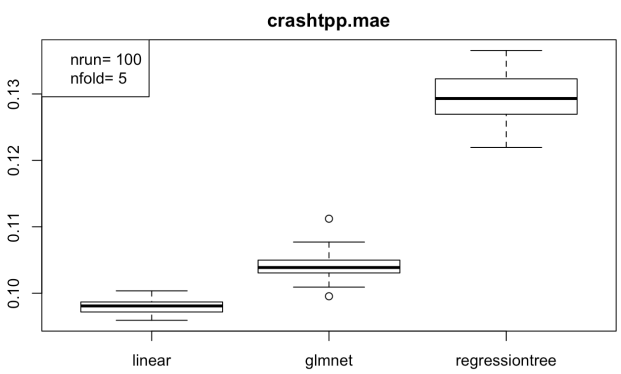

flaretpp.mae
nrun= 100
nfold= 10
0.100
0.105
0.110
0.115
linear
glmnet
regressiontree

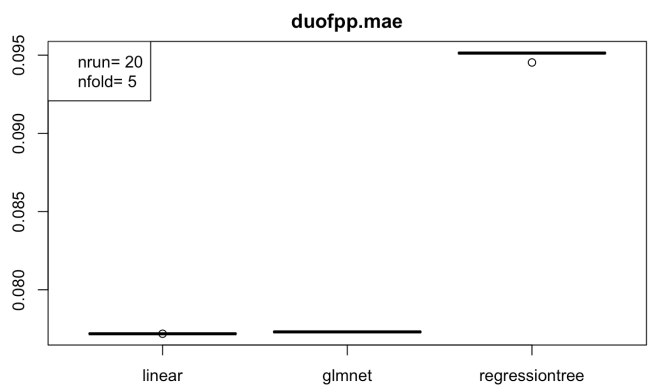

normalsolotpp.mae
nrun= 100
nfold= 5
0.19
0.20
0.21
0.22
0.23
linear
glmnet
regressiontree

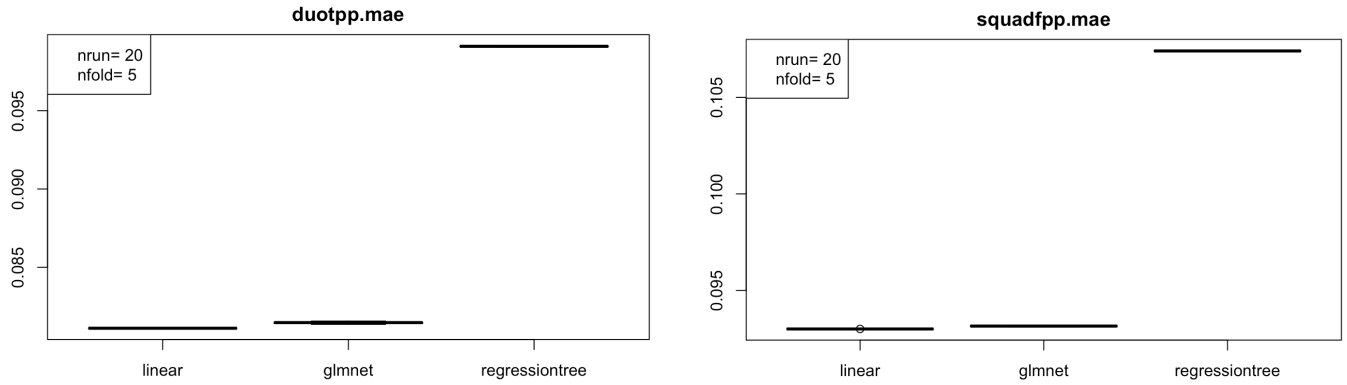

nnormalsquadfpp.mae
nrun= 20
nfold= 10
0.208
0.212
0.216
linear
glmnet
regressiontree


nnormalsquadtpp.mae
nrun= 100
nfold= 10
0.240
0.245
0.250
0.255
linear
glmnet
regressiontree


crashfpp.mae
nrun= 100
nfold= 10
0.095
0.105
0.115
0.125
linear
glmnet
regressiontree


crashtpp.mae
nrun= 100
nfold= 5
0.10
0.11
0.12
0.13
linear
glmnet
regressiontree


duofpp.mae
nrun= 20
nfold= 5
0.080
0.085
0.090
0.095
linear
glmnet
regressiontree

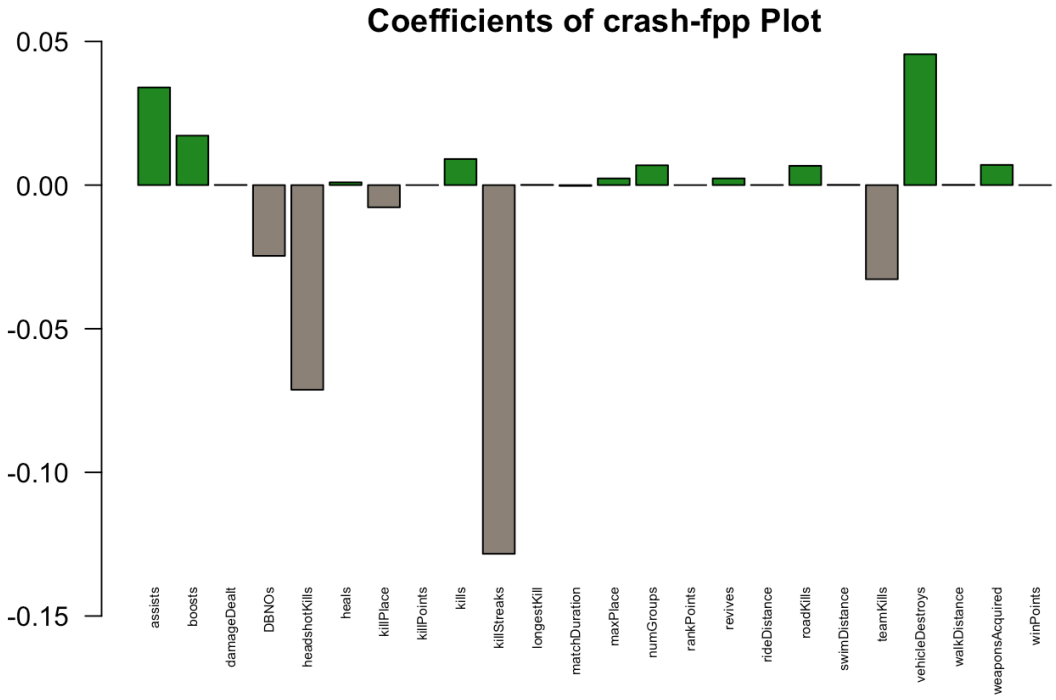


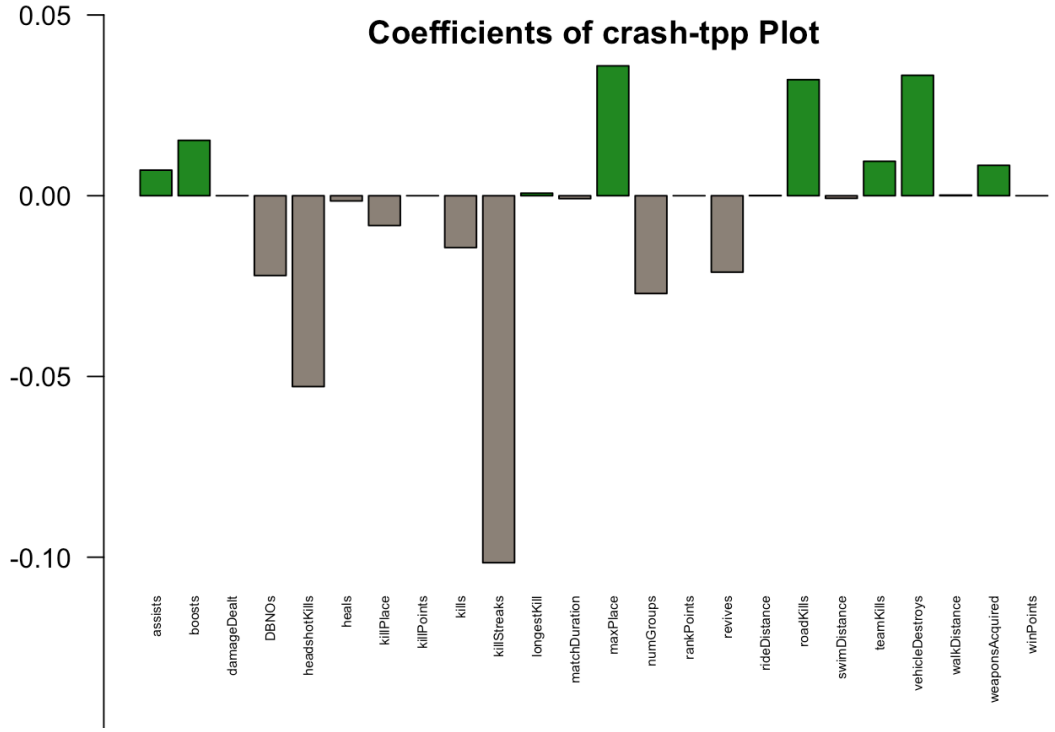
while there is one special case that the regression tree model is the best in normal-solo-fpp.

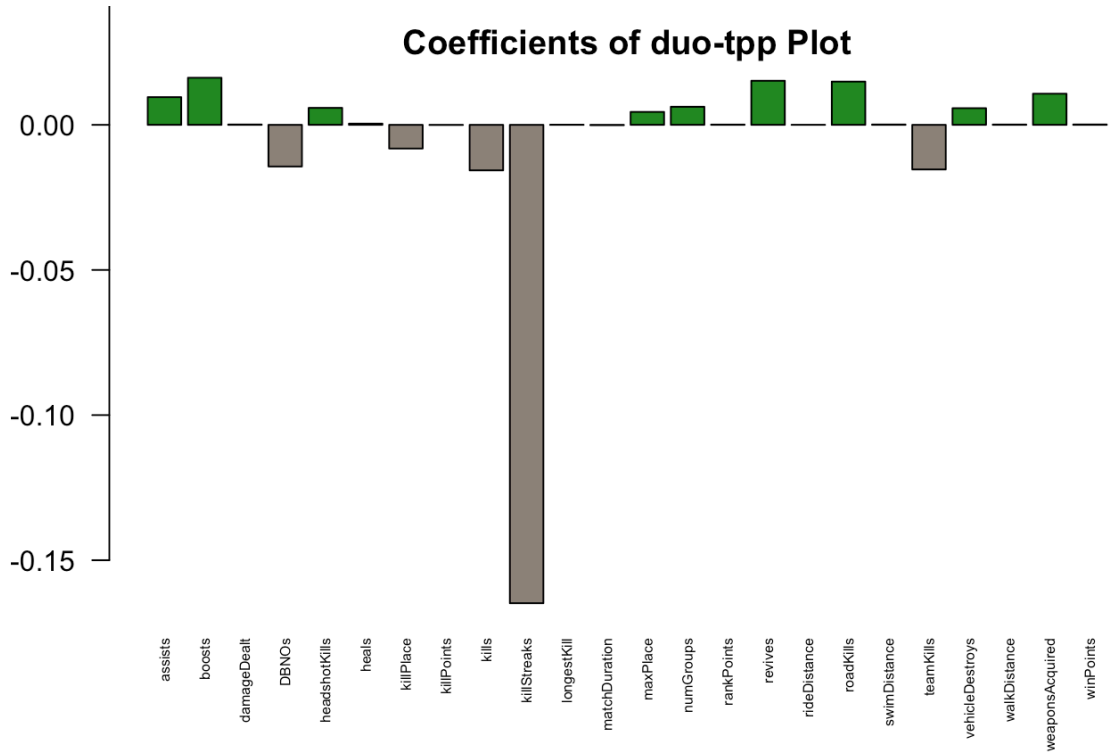
3.4 Data Prediction and Analyzation

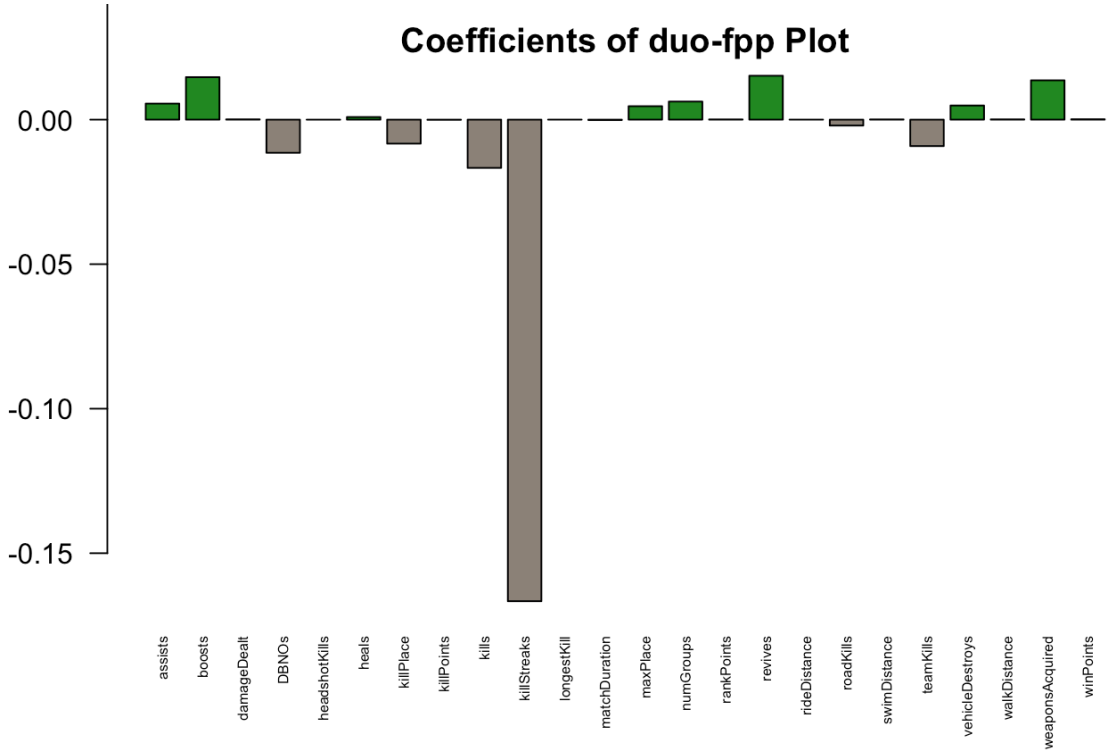
Testing data are used to predict the final placement with the best model chosen.

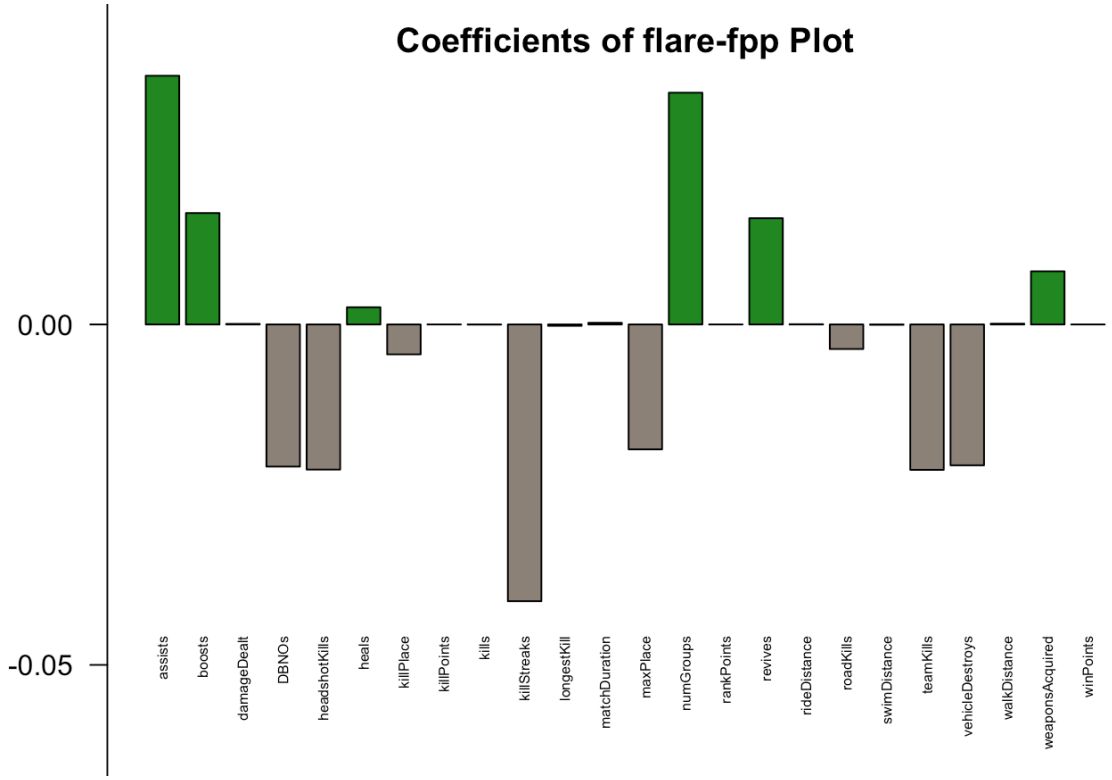
Using the best model for each type of game, we made 16 bar-plots to represent coefficients for each terms determining the final placement.

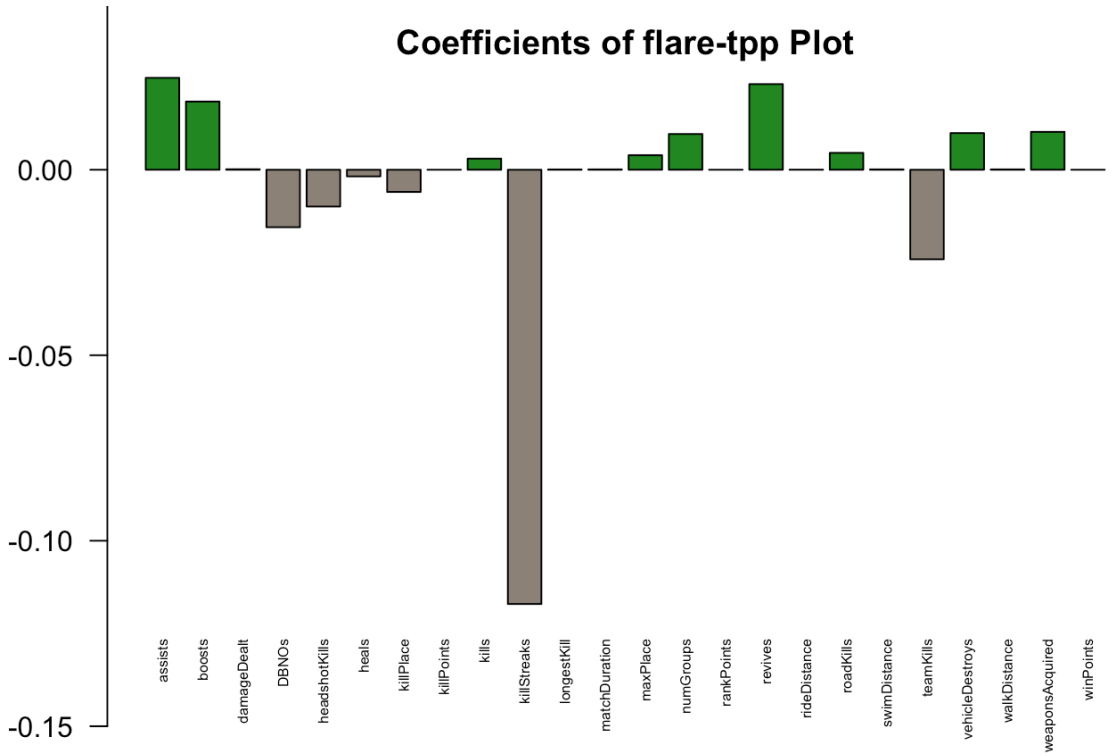


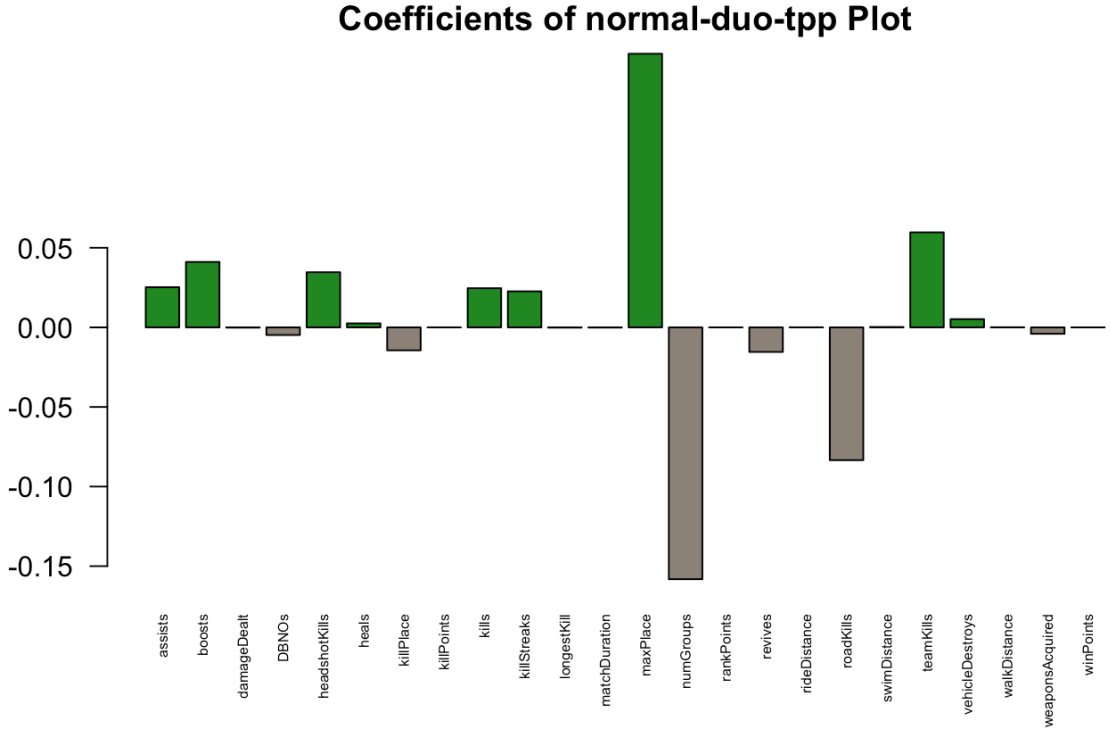

Coefficients of crash-tpp Plot
0.05
0.00
-0.05
-0.10
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints

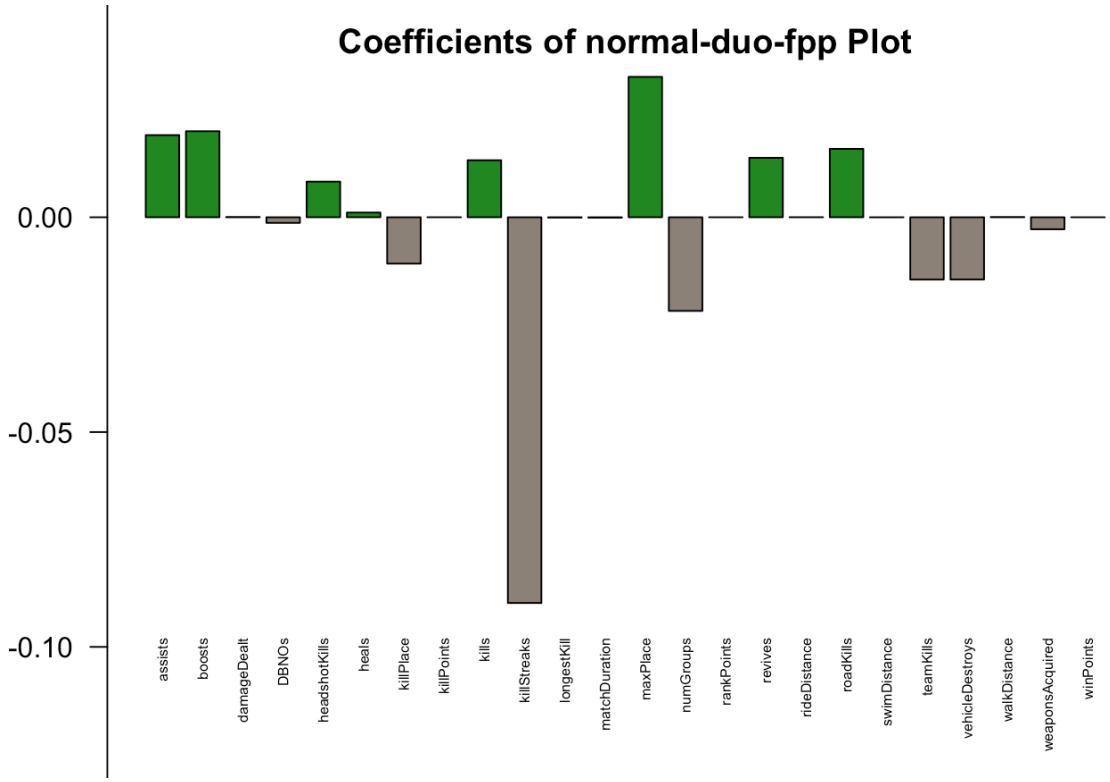

Coefficients of duo-tpp Plot
0.00
-0.05
-0.10
-0.15
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints

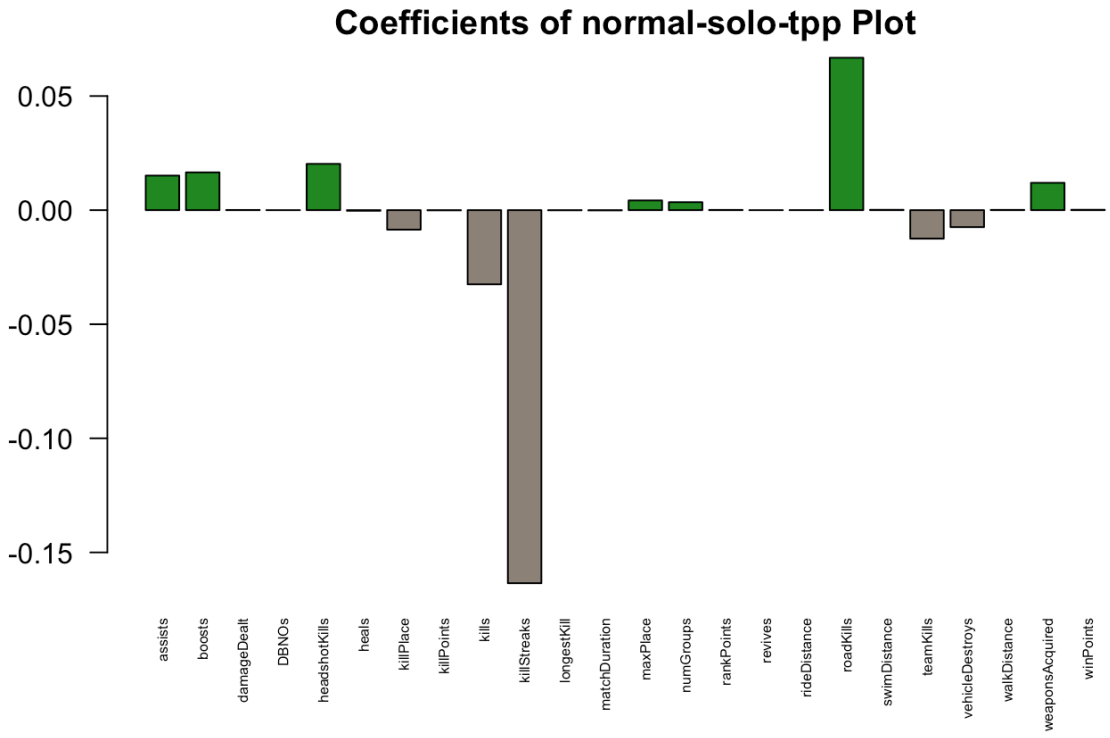

Coefficients of duo-fpp Plot
0.00
-0.05
-0.10
-0.15
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints

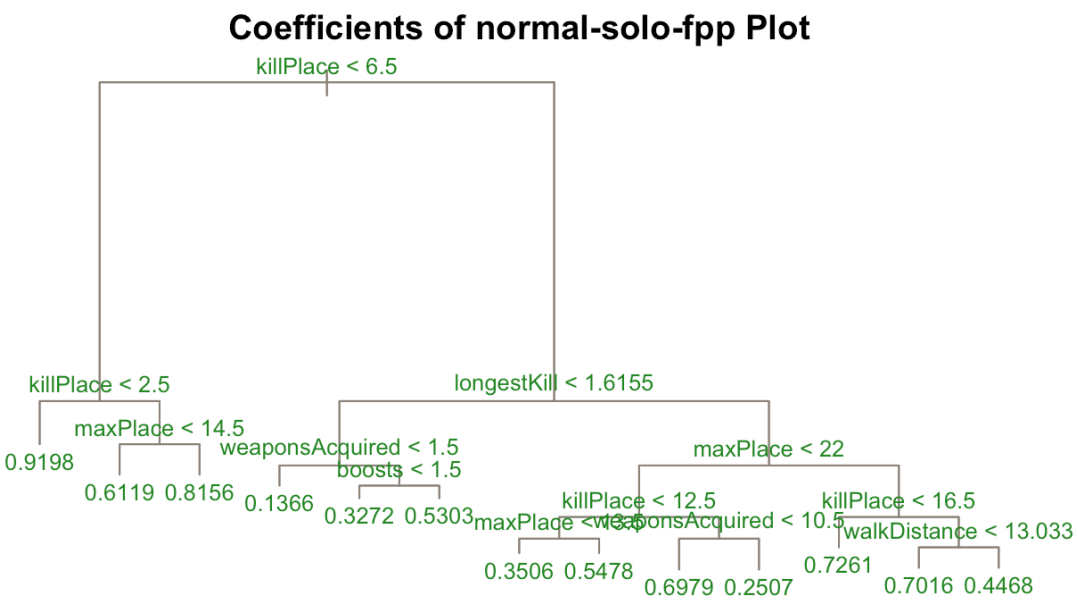

Coefficients of flare-fpp Plot
0.00
-0.05
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints

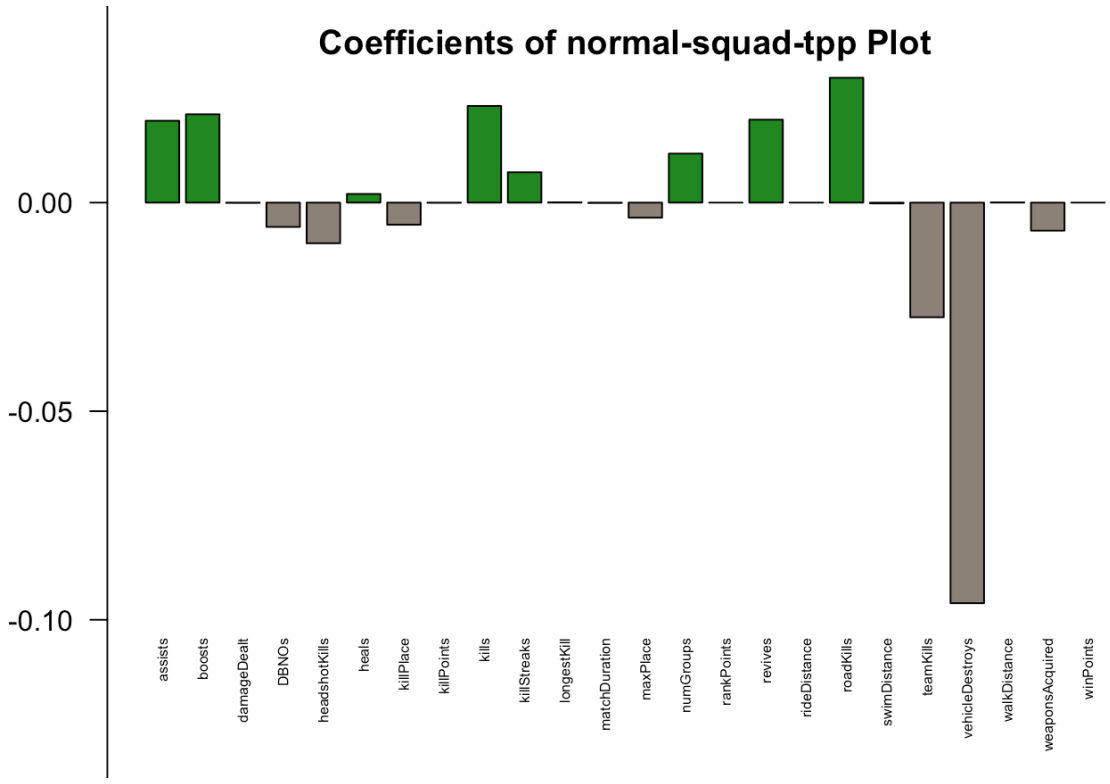

Coefficients of flare-tpp Plot
0.00
-0.05
-0.10
-0.15
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints

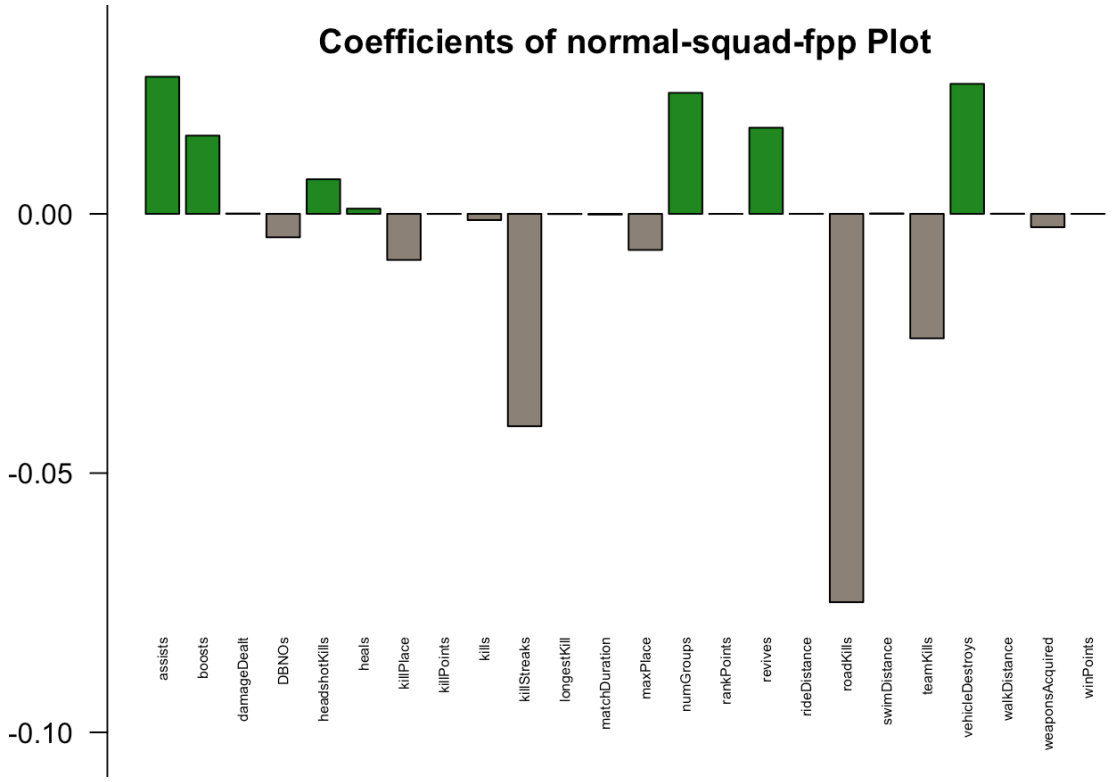

Coefficients of normal-duo-tpp Plot
0.05
0.00
-0.05
-0.10
-0.15
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints

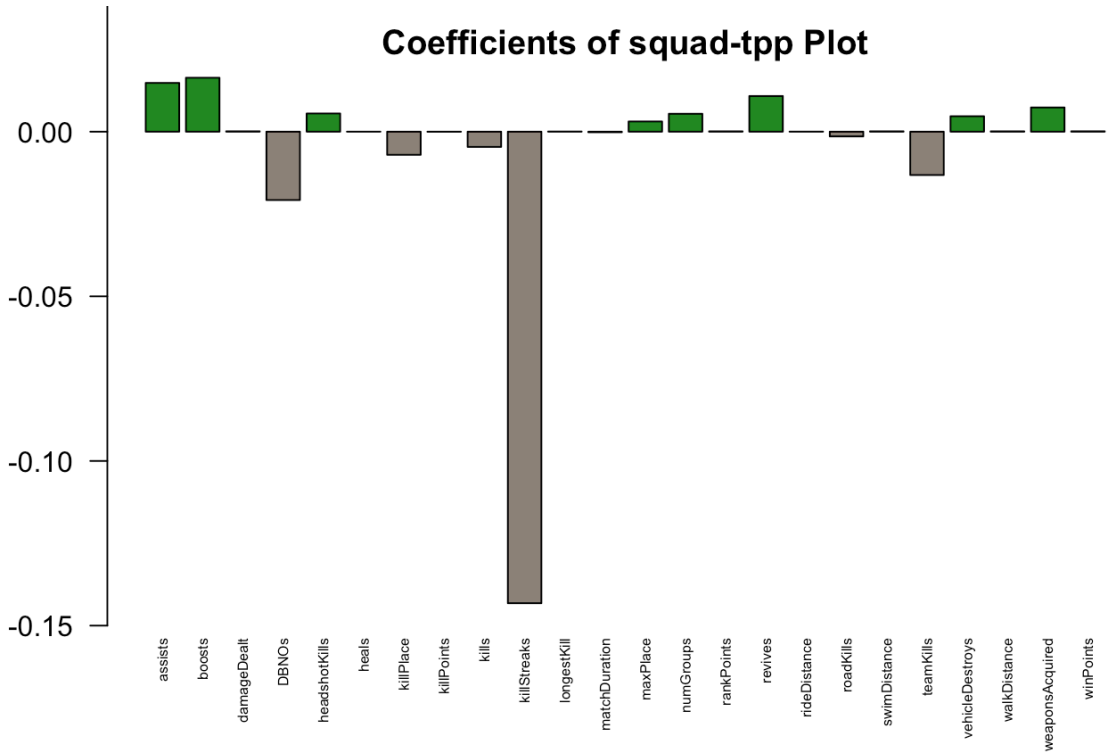

Coefficients of normal-duo-fpp Plot
0.00
-0.05
-0.10
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints

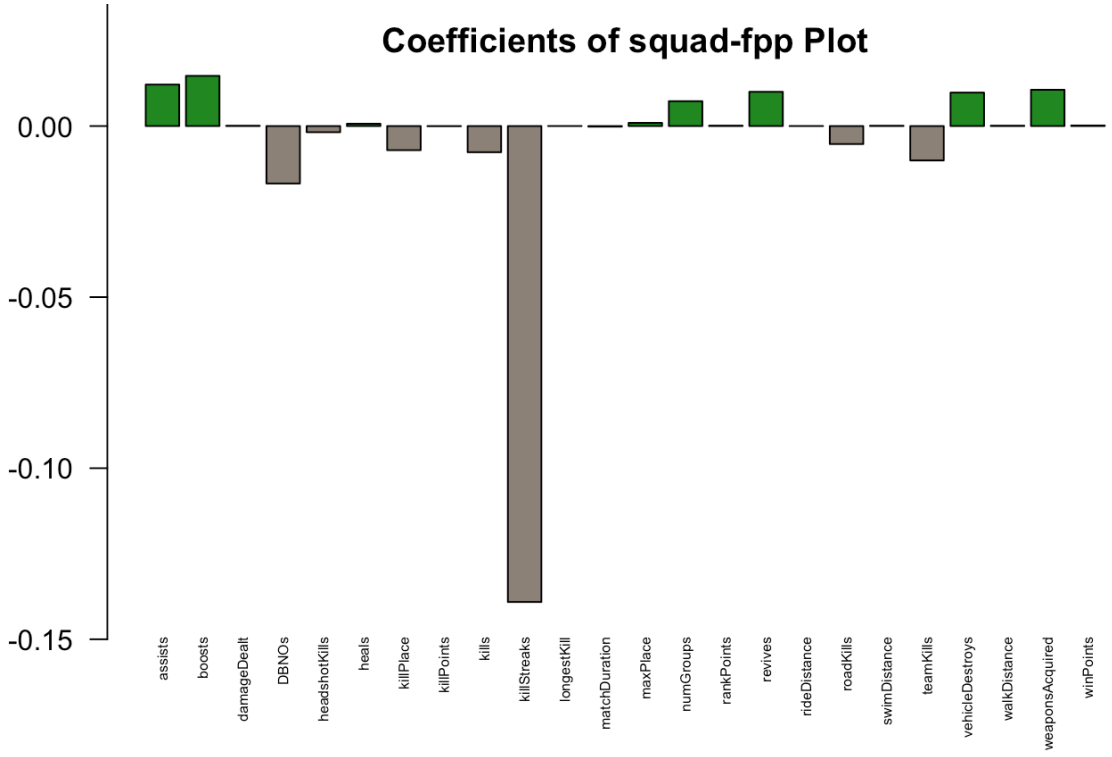

Coefficients of normal-solo-tpp Plot
0.05
0.00
-0.05
-0.10
-0.15
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints

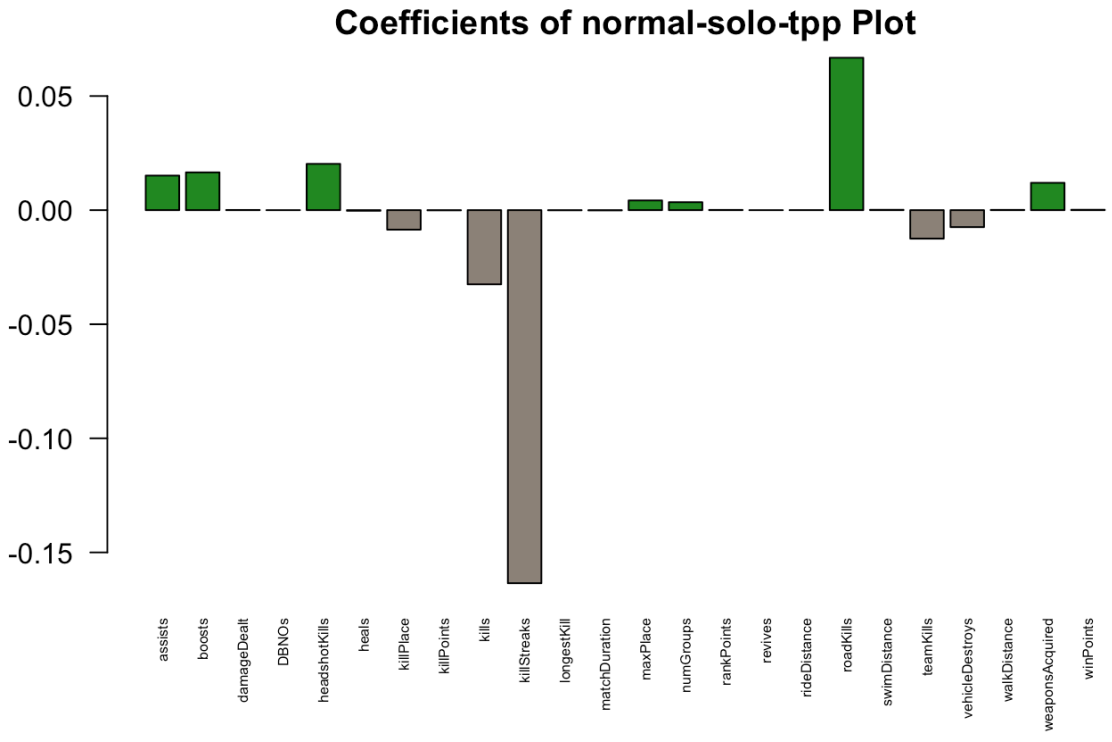

Coefficients of normal-solo-fpp Plot
killPlace < 6.5
killPlace < 2.5
longestKill < 1.6155
maxPlace < 14.5
weaponsAcquired < 1.5
maxPlace < 22
boosts < 1.5
killPlace < 12.5
killPlace < 16.5
maxPlace < 13.5
weaponsAcquired < 10.5
walkDistance < 13.033
0.9198
0.6119
0.8156
0.1366
0.3272
0.5303
0.3506
0.5478
0.6979
0.2507
0.7261
0.7016
0.4468

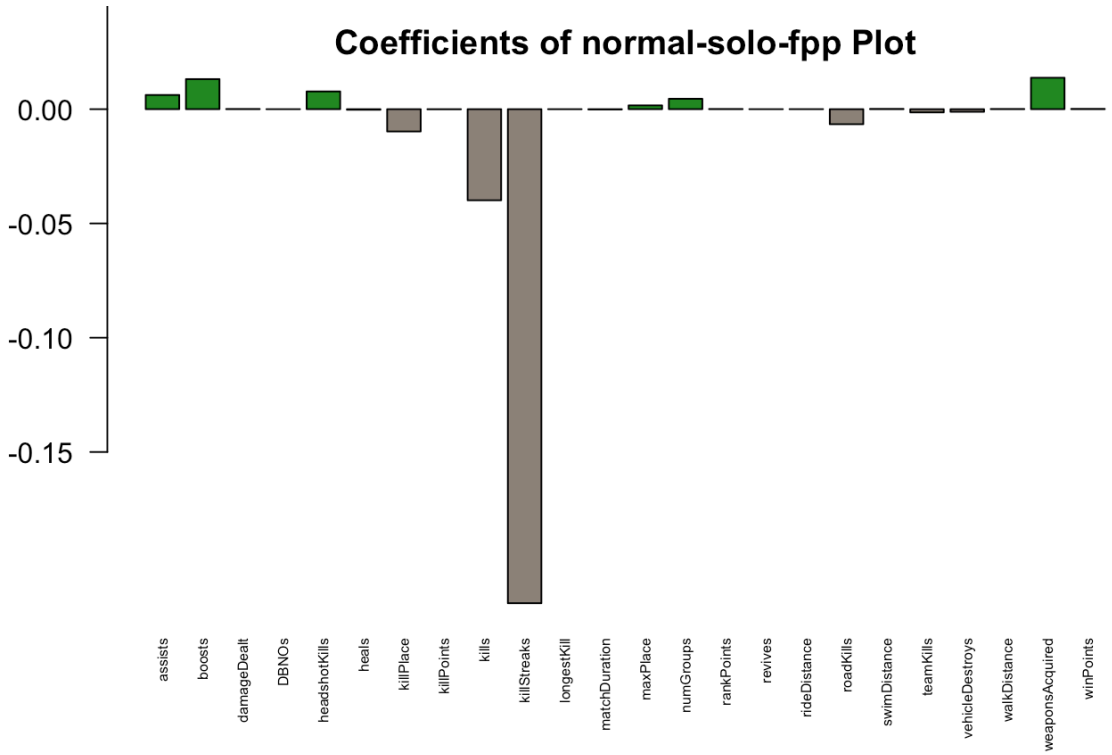

Coefficients of normal-squad-tpp Plot
0.00
-0.05
-0.10
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints


Coefficients of normal-squad-fpp Plot
0.00
-0.05
-0.10
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints


Coefficients of squad-tpp Plot
0.00
-0.05
-0.10
-0.15
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints


Coefficients of squad-fpp Plot
0.00
-0.05
-0.10
-0.15
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints


Coefficients of normal-solo-tpp Plot
0.05
0.00
-0.05
-0.10
-0.15
assists
boosts
damageDealt
DBNOs
headshotKills
heals
killPlace
killPoints
kills
killStreaks
longestKill
matchDuration
maxPlace
numGroups
rankPoints
revives
rideDistance
roadKills
swimDistance
teamKills
vehicleDestroys
walkDistance
weaponsAcquired
winPoints



The result can be concluded in the table below:

Type of Game	Best Strategy	Worst Strategy
crash-fpp	vehicleDestroys	killStreaks
crash-tpp	maxPlace	killStreaks
duo-tpp	boosts/revives/roadKills	killStreaks
duo-fpp	revives	killStreaks
flare-fpp	assists	killStreaks
flare-tpp	assists/revives	killStreaks
normal-duo-tpp	maxPlace	numGroups
normal-duo-fpp	maxPlace	killStreaks
normal-solo-tpp	killStreaks	teamKills
normal-solo-fpp	killPlace	weaponsAcquired
normal-squad-tpp	roadKills	vehicleDestroys

normal-squad-fpp	assists/vehicleDestroys	roadKills
normal-solo-tpp	roadKills	killStreaks
normal-solo-fpp	boosts/weaponsAcquired	killStreaks
squad-tpp	boosts	killStreaks
squad-fpp	boosts	killStreaks

We can observe that the best strategy for winning the game varies in different types of game. Similarly, the worse strategy shows the same outcome. To be specific, boosts, maxPlace, assists and roadKills are the common best strategies for being the winner while killStreaks has negative effect on the final placement of 11 types of game.

4. Conclusion

According to the result above, both best strategies and worst strategies can be hints for the final placement. If a player or a team wants to be the winner, he or they should achieve the best strategy listed and avoid the worst one at the same time based on what kinds of game he or they attend.

Regarding to model selection process, the linear model is the best one in most types except for normal-solo-fpp. As classification model cannot be used in this project, so there are small number of models we learned cannot be used. We assume that if more models apply, there may be better performed one than the linear model. Besides, for small number of covariates, the linear model performed better than penalized linear regression model. The last point is that a tree model surprisedly performed best, which may lead to overfitting phenomenon.

Work Sited

Kaggle, 2018, https://www.kaggle.com/c/pubg-finish-placement-prediction/data. Accessed 28 Nov 2018.