Capstone Project 2: PUBG Finish Placement Predictions

Overview:

PUBG is a game where up to 100 players start in match onto an island empty handed and must explore, scavenge and eliminate other players in a free for all until only one player/team is left standing. Kaggle has given us 65,000 games' worth of anonymized data and we're asked to predict the final placement in a game given the in-game stats and initial player ratings. Kaggle has asked that we evaluate our model on Mean Absolute Error between the predicted win placement and the observed win placement.

Target Audience:

Within the last 2 years, Battle Royale (BR) styled video games have exploded in popularity, giving us new IPs such as Fortnite and Realm Royale and even affecting long standing franchises like Black Ops. This surge in numbers can arguably be traced back to PUBG, one of the first BR games to receive a substantial amount of development and sustain a large, consistent player pool. The premise of the game is simple: You and up to total of 100 players are dropped onto a large island empty handed and you must explore, scavenge and eliminate other players in a free-for-all/team vs team first person shooter styled battle until only one player or team is left standing. This project aims to use 65,000 games' worth of anonymized data to predict final placements in a game using over 20 different in-game stats. There are a couple groups of people that would be interested in the results:

- The game's developers: The data analyzed could be used to help conduct balance changes based off of impact of each feature.
- The players: The data analyzed could be used to help pinpoint weaknesses in a player's performance, giving them an objective avenue of improvement to go through.
- Tournament organizers: High impact features could be used when casting tournament games, being able to select the top 5 or 10 most impactful stats that players should keep an eye on.

Data:

The data I will be using can be found on Kaggle (here). The file list includes:

- A sample submission file in the correct format
- A training dataset (4.45m rows)
- A test dataset (1.93m rows)

The columns found in our data are as follows:

- DBNOs
- Assists
- Boosts
- damageDealt
- headshotKills
- Heals

- Id
- killPlace
- killPoints
- killStreaks
- Kills
- longestKill
- matchDuration
- matchId
- matchType
- rankPoints
- Revives
- rideDistance
- roadKills
- swimDistance
- teamKills
- vehicleDestroys
- walkDistance
- weaponsAcquired
- winPoints
- groupId
- numGroups
- maxPlace
- winPlacePerc - This is our prediction target

Method:

- Data Gathering:
  - Download datasets from kaggle and import into pandas dataframes
- Data Wrangling:
  - Determine if I need to split up the data into game mode (FFA, duos, teams)
  - Find missing values and replace after determining reason for missing values
  - Consider possible outliers (are there any cheaters, etc)
- Exploratory Data Analysis and Inferential Statistics:
  - Make visualizations for each of the features to determine normality
  - Find highly correlated values w/r/t the win place percentage
  - Possible z-tests for correlated values
- Machine Learning and Model Building:
  - Build regressive models using data as a whole vs data split into game modes
  - Possible deep learning model by splitting up data into iterations of different games