Title: Player Unknown Battle Grounds Analytics

|  |  |  |
| --- | --- | --- |
| First Name | Last Name | Email-ID |
| Charmi | Kalani | [ckalani@hawk.iit.edu](mailto:ckalani@hawk.iit.edu) |
| Anwesha | Kakoty | akakoty@hawk.iit.edu |

* **Introduction:**

Analysis can be performed on anything in this world, as long as we can acquire the required data. We ourselves perform analysis on a lot of day to day things without realizing that we are performing an analysis. For our final project, we have decided to work on one of the most popular games worldwide, PlayerUnknown's BattleGrounds (PUBG). This multiplayer online game has enjoyed massive popularity. With over 50 million copies sold, it's the fifth best-selling game of all time, and has millions of active monthly players. Factors like, killstreaks, kill time, boosts used, walking distance, swimming distance etc. can be used to analyze the overall model and predict information like the probability of a player winning the match in the end. Using the already provided data, we can work on them and predict the future outcomes. The goal and motivation of this project is to work on something which is trending these days and relates to the young generation and also to make the players and the makers of the game aware of the techniques that can be utilized to gain victory in this game, so that the creators can provide the best, most realistic and challenging gaming experience for the players. Being avid players of PUBG ourselves, we believe that a lot of people are into entertainment sources like gaming which could be a great way to divert and refresh our minds from the mundane activities of our daily lives along with connecting with people and players across the world, hence we have taken up this interesting topic which relates to an enjoyable part of our lives to work on and provide appropriate results on.

In a PUBG game, up to 100 players start in each match (matchId). Players can be on teams (groupId) which get ranked at the end of the game (winPlacePerc) based on how many other teams are still alive when they are eliminated. In the game, players can pick up different munitions, revive downed-but-not-out (knocked) teammates, drive vehicles, swim, run, shoot, and experience all the consequences – such as falling too far or running themselves over and eliminating themselves. After we consider all of these factors, how can we figure out what's the best strategy to win in PUBG? Should a player just sit in one spot and hide their way into victory, or do they need to be the top shot? Let's let the data do the talking!

* **Dataset:**

The dataset “pubgdata.csv” as attached in the same folder has been obtained from Kaggle. The original dataset has more than 10,00,000+ tuples, however we would be using a part of them, around 10000 tuples of data to train and create the model. The data has following attributes:

1. DBNOs - Number of enemy players knocked.
2. assists - Number of enemy players this player damaged that were killed by teammates.
3. boosts - Number of boost items used.
4. damageDealt - Total damage dealt. Note: Self-inflicted damage is subtracted.
5. headshotKills - Number of enemy players killed with headshots.
6. heals - Number of healing items used.
7. killPlace - Ranking in match of number of enemy players killed.
8. killPoints - Kills-based external ranking of player.
9. killStreaks - Max number of enemy players killed in a short amount of time.
10. kills - Number of enemy players killed.
11. 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 longest kill stat.
12. matchType – The type of match a player played which is either FPP or TPP, squad, duo or solo mode. A categorical variable.
13. matchId - Integer ID to identify match. There are no matches that are in both the training and testing set.
14. revives - Number of times this player revived teammates.
15. rideDistance - Total distance traveled in vehicles measured in meters.
16. roadKills - Number of kills while in a vehicle.
17. swimDistance - Total distance traveled by swimming measured in meters.
18. teamKills - Number of times this player killed a teammate.
19. vehicleDestroys - Number of vehicles destroyed.
20. walkDistance - Total distance traveled on foot measured in meters.
21. weaponsAcquired - Number of weapons picked up.
22. winPoints - Win-based external ranking of player.
23. groupId - Integer 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.
24. numGroups - Number of groups we have data for in the match.
25. maxPlace - Worst placement we have data for in the match. This may not match with numGroups, as sometimes the data skips over placements.
26. 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.

Although the dataset has a lot of attributes, not all of them will be useful for prediction of the winning place. We will obtain the significant variables by using feature selection and then build the model.

* **Research Problems:**
* The primary research goal of this project is to **predict the probability of a player ending up in the first place in the match** by analyzing the match statistics of the player.
* The second research goal is to analyze the data in order to discover patterns which make a player better in the game, which includes factors like increasing the survival rate, gaining more killstreaks and ending up with a higher score. We could analyze strategies, like whether travelling longer distances and having more number of kills, ambushing strongholds actually decide a player’s probability of winning, or is staying put and stealth play like hiding away and sniping from a concealed location a better strategy to win than the former. This could be done by applying the fundamental concepts learned in the ITMD 527 course/class like performing regression analysis, building and evaluating different models with the help of R programming language.
* We can also predict how much the winning chances, survival rates and gaining a higher score depends on which mode a player plays the match, whether on First-Person Perspective (FPP) or Third-Person Perspective, as a solo, duo or squad of four and vice versa, like predicting the match mode using given statistics. Here, the match mode is categorical data and hence we will use classification techniques like logistic regression to predict which mode the player played the match on.
* The hypothesis testing can be done on claims like, players who play in duo mode have higher killPlace ranking than players who play in squad mode, or players who play in duo mode have significantly more number of kills than players who play in solo or squad mode, or players who play in Third person Perspective survive longer in comparison to players who play in First-person perspective.
* **Potential Solutions:**
* For this project, the main goal is to predict the chances or probability of a player ending up in the first place and what are the factors (variables) that influence the probability values.
* We will build multiple linear regression models where:
  + **Y-variable:**  winPlacePerc – This is 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. The values range from between 0 to 1.
  + **X-variables:** DBNOs, assists, boosts, damageDealt, headshotKills, heals, killPlace, killPoints, killStreaks, kills, longestKill, matchId, revives, rideDistance, roadKills, swimDistance, teamKills, vehicleDestroys, walkDistance, matchType, weaponsAcquired, winPoints, maxPlace

We will perform feature selection method like backward and forward elimination using p-value or AIC as metric, stepwise regression, subset regression with hold out evaluation as the dataset is large. We are mainly going to focus on values like R2 and Adjusted R2 (more preferably Adjusted R2) i.e. goodness of fit (F test), individual parameter test, residual Analysis and P values of independent X variables in the model, and then predict the accuracy of the model.

* We would also predict what match mode a player played the match in using the below x-variables.

For categorical variables like matchType we would like to apply some supervised learning techniques for classification like Naïve Bayesian classification and Logistic regression.

* + **Y-variable:** matchType. The type of match a player played which is either FPP or TPP, squad, duo or solo mode.
  + **X-variables:** DBNOs, assists, boosts, damageDealt, headshotKills, heals, killPlace, killPoints, killStreaks, kills, longestKill, matchId, revives, rideDistance, roadKills, swimDistance, teamKills, vehicleDestroys, walkDistance, weaponsAcquired, winPoints, maxPlace, winPlacePerc
* We can create boxplots and graphs of boosts, weapons acquired, heals, headshots to see if performing actions like collecting supplies, healing teammates, increasing shooting accuracy etc have any effect on the chances of winning.
* Since there are a lot of variables, multicollinearity problem is sure to exist. We can check the correlations of the x-variables and eliminate the variables which have high correlations with other variables.
* **Hypothesis testing:**

**Claim 1:** Players who play in duo mode have significantly more number of kills than players who play in solo or squad mode

**H0:** There is no significant difference between the number of kills between players who play in solo and squad mode.

**H1:** The number of kills when played in solo mode is significantly higher than when played in squad mode.

**Claim 2:** Players who play in Third person Perspective survive longer in comparison to players who play in First-person perspective.

**H0:** There is no difference between the survival times of Third person perspective and First Person perspective

**H1:** Survival time when played in Third Person perspective mode is significantly higher than when played in First person perspective.

우리는 또한 모델이 qualified 되었는지를 알기위해 goodness of fit, individual parameter test, residual analysis를 실행하고 adjusted R2와 RMSE를 비교할 예정입니다.

가설검정의 경우에는,

Claim Example 1:동북아시아가 동남아시아보다 대체적으로 기대수명이 높다.

H0:동북아시아와 동남아시아는 평균 기대 수명이 같다.

H1:동북아시아는 동남아시아보다 평균 기대 수명이 높다.

Claim Example 2 : 선진국이 개발도상국보다 대체적으로 기대수명이 높다.

H0:선진국과 개발도상국은 평균 기대 수명이 같다.

H1:선진국이 개발도상국보다 평균 기대 수명이 높다.

Like above, 원하는 가설에 대한 검정을 먼저 한 후, 그렇다면 A가 B보다 기대수명이 높은 이유는 무엇인지에 대해서 연구를 이어나갈 수 있습니다.

또한, 두 집단 비교함에 있어서, 필요하다면 ANOVA test를 이용할 수 있습니다.

* **Evaluations:**
* We opt for hold-out evaluation for splitting the data set as per the instructions in the class because we have enough data for the splitting. We go for hold-out evaluation when the sample or data set is large and n-folds when the sample data set is small. Here, the entire data set has been split into 2 parts, hence 80 % of the original data (8000 cases of player data) will be considered as training data set (to build the model) and remaining 20 % (2000 cases of player data) as the testing data set (for evaluations).
* We will use Confidence interval and RMSE i.e. Root Mean Square Error as evaluation metrics to provide a not only the best but accurate model for prediction and solve the regression problems for the evaluation of the output.
* **Expected Outcomes:**

Some of the outcomes we expect upon the completion of the project are:

1. To find out the best fitted model for the given real-world data set of the game.
2. To find out the relationships between the dependent y-variable and independent x-variables, mainly winning place to survival time, number of kills and distance travelled and derive appropriate and accurate strategies to land the winning place in the game.
3. To find out if factors like boosts and weapons acquired determine the probability of winning, hence if a player should also concentrate on collecting and finding supplies and looting other players and locations.
4. To find out if the match type or the number of teammates a person plays with increases their individual scores and ranking at the end of the match.
5. To analyze player statistics like their shooting accuracy, headshots, kill streaks and damage dealt to predict their patterns of playing, if they like to go for more open action and close combat or if they prefer stealth operations. **For example, if walking distance is a significant variable which explains the winning place predictions y-variable, and we observe that the walking distance of a player is less than average and shooting accuracy is high (number of headshots), we can deduce that the player must have played a stealth game (hiding at one place and killing from a distance rather than travelling more and ambushing). If we observe that the win place probability of that player is high, we can derive a conclusion that playing a stealth game with high accuracy shooting is probably a more effective way of ending up in a higher-ranking place.**