Win Prediction of Players Unknow Battle Ground Game

Arth Vyas
x17170516
MSc in Data Analytics

Ranu Parate
x17161452
MSc in Data Analytics

Samir Khan
x17161461
MSc in Data Analytics

Mitul Choksi
x17154855
MSc in Data Analytics

***Abstract*— The craze of mobile gaming is increasing day by day with the development of new high feature games available for cellphone. Among all the new games, Players Unknow Battle Ground (PUBG) has caught tremendous heap for the public. This project aims to predict win place percentile using the different probabilistic method. Different feature selection methods are carried out for the selection of the most important features. The models like linear regression, support vector regression and the random forest have been applied to test the dependent variable. The results are evident that random forest has outperformed the other two regression model. The results and important features can give insight to the player to strategies their game as well as gives an opportunity for the developer to improve the complexity of the game.**

Keywords—Data mining, Multiple Linear Regression, SVR, Random Forest, Win Prediction

I. Introduction

Mobile gaming is known as one of the favorite hobbies among children, but adults too show equal interest in playing games on mobile or tablet. On average there is a 5 percent increase in active players of mobile games yearly. Adults with age group 25 to 44 are found top players of mobile games, they spend roughly around six hours in a week for playing games [1]. The games like Pokémon GO, Players Unknown Battle Ground (PUBG) and others have attracted the players a lot. Due to the growing popularity of advanced mobile games like PUBG, mobile phones are projected to reach 1.8% of total cellphone shipment to India by end of 2019[1]. The popularity of PUBG has tempted several cellphone manufacturers to make deals with Tencent Games which can help to attract customer and buy more cellphones. Vivo which is one of the leading cellphone company of India has announced the partnership with PUBG MOBILE by Tencent Games as the title sponsor for its upcoming event[2]. The biggest mobile gaming tournaments in the world, PUBG MOBILE Club Open 2019, where players will fight for the winning prize of $2.5M USD. Vivo promises to bring the ultra-smooth gaming experience to attract customer[2].

Tencent Games and PUBG Corporation has developed PUBG MOBILE, it is based on “PLAYERUNKNOWN’S BATTLEGROUND”, available for cellphone, PC and Xbox.

PUBG is one of the latest game which has caught tremendous attention of people in the year 2017. In this game up to 100 players parachute on a remote island where the battle begins, and the player compete for the winner. Players must scavenge their own weapons, vehicles, and supplies. Finally, they must defeat every player in a tactical and visual rich battleground which forces players into a shrinking play zone. PUBG being new in the mobile gaming market and got fantastic welcome by people with the subscriber of 400 million worldwide, in that 50 million copies were sold in the form of PC copies by PUBG corporation as of 19th June 2018.

Various researchers are interested in predicting the winner of the game, the prediction is leveraged by the improved data mining techniques hence, they wish to analyze how accurately a model can predict the winner. Win prediction is in the trend for different games like Dota 2, NFL, soccer, tennis, and war craft with different machine learning algorithms. Win prediction has plethora feature which can significantly affect the dependent variable hence, feature selection becomes vital in win prediction. Since PUBG is new mobile game with a different concept of visualization and tactical survival with the shrinking zone has checked the eye of researcher and they have analyzed the strategies for the player to land in the safe zone,

1. https://economictimes.indiatimes.com/markets/bonds/bharti-airtelinternational-begins-cash-purchase-of-its-usd-1-5-billion-bonds-to-reducedebt/articleshow/66566013.cms
2. https://www.prnewswire.com/news-releases/vivo-announces-to-empower-gamers-conquest-at-pubg-mobile-club-open-2019-by-tencent-games-and-pubg-corporation-300816977.html
3. https://www.kaggle.com/michaelapers/deprecated-pubg-finish-placement-prediction#train.csv

jumping time, and judgment of azimuth and so on. Till date there is no research has been done on win place prediction of PUBG hence, we pose a question *"To what extent PUBG win place can be predicted by using the different probabilistic method?"*

The dataset for this project is taken from Kaggle dated 31st Jan 2019, the data is about PUBG game for various match Id in which different participants have taken part in the form of solo or in the group. The data also contain different features of players which leverage the prediction of win place for a player to the machine learning model. The rest of the paper is as follows: section II consists of a literature review, section III consists of methodology. Section IV Implementation Section V consists of evaluation, Conclusion, and future scope.

II. Literature Review

A. *Prediction based Researches in Online Gaming Domain (Dota 2,Destiny,PUBG)*

[4] were the first to make use of information from the hero's draft and applied logistic regression and k-Nearest Neighbours (kNN) for the purpose of model building. The resulting test data accuracy was 69.8% for the trained data of 18,000. The issues with this model were that it wasn't able to record antagonistic as well as synergistic relations amongst the heroes in and between the teams. Using customized weights & distance metrics the optimal d-dimension value was chosen using kNN with 2-fold CV. For the dataset of 20,000, the recorded accuracy was 67.43% and for the test dataset of 50,000, the accuracy was 70%. A recommendation system was developed based on a web interface was developed based on the obtained results. The point of concern for this technique was its speed which was very slow (12 Hours for C-validation) whereas this approach seems to be quite simple which could affect the interaction capturing. The results could be improved by selecting more amount of data for the research.

Another interesting study by [5] introduced the in-game hero's roles as the features for the win prediction model. The study is an extension of the previous work conducted by [4] where recommendation engines for heroes were build based on predictive algorithms. In this approach, all the game types such as "All Pick", "Least Played", "All Random" etc., were used along with matches played by the skilled players (match duration of more than 900 secs) were filtered based on matchmaking rating (MMR). Using 220 matches, the previously studied logistic regression was ensembled with genetic algorithm and the resulting accuracy stood out to be 74.1% whereas the logistic regression simply gave 69.42% accuracy. Certain drawbacks that were understood was that there was a lack of information for ROC - AUC and also the reliability of the technique is the point of an issue as a small sample size of the matches were taken into consideration.

A novel attempt to enhance the work done by [5] was made by another research group [6]. It considered 62,000 matches of highly skilled players and the games having a duration of more than 10 mins was taken into consideration. Logistic regression and random forest were used for creating two different win predictors i.e. using post-match data and others using given picks. For the given picks data, the model performance was positively affected as it gained an accuracy of 72.9% for logistic regression. This is because the features selected were similar to that done by [4] and were able to record relationships like countering, matchup, synergy and offset. The point of concern was in the random forest which overfitted the data and gave a test accuracy of 67% after performing tuning. The baseline predictor used for comparison had a very high combined individual win -rate for the heroes. Another point of significance in this study was that for the prediction purpose, it made use of hero countering, but the downside was that synergy metric used to represent the hero in [5] gave higher accuracy that the proposed predictor.

[7] extended the work by including other relevant game data which included information like the gold, kills, deaths assists for each hero per minute along with the draft information that was being used by previous researches. They performed training on random forest and achieved an accuracy of 82.23% at the game duration of 5 mins. Though the accuracy achieved using this method is quite high, the research was based on the fact that it requires the use of real-time data which limits the use of this approach.

The authors [8] had proposed using a mixed-rank based dataset which consists of 270 matches of professional games and 1663 public matches with extremely high skills where score is in excess of 6000. Mixed-rank based approach was used as data for professional match was not enough to train model and only could be used to test data. The data has been split in two parts – pre-match features and in-game features. The main purpose of using mixed-data was to use it as a proxy data for prediction against professional data. Two popular algorithms Logistic regression and Random forest for Dota-2 win prediction has been used for prediction. Three features were derived using WEKA framework for feature selection. In logistic regression, parameter ridge in log-likelihood and parameter number of trees in random forest were changed for parameter tuning. An accuracy of 76.17% and 75.22% was achieved for Random forest and logistic regression respectively. The authors concluded that using mixed-rank data for prediction reduces accuracy compared to using only professional match data.

[9] made a research based on the game Destiny which is a First-Person Shooting game. In this research prediction of victory based on classification is studied along with the gamer’s performance metrics effect on the match outcome. Unique features from the game were used for study purpose. The uniqueness of this research was that it formed two model groups i.e the first combined model covers all game modes of Destiny and each individual game mode is predicted by is another group of the model. Random forest and Gradient Boosting is used as the method of classification and the performance accuracy for this research stands out to be between 63-99%. The top 5 performance metrics varied as per the different game modes but SPL and KDA remained constant. Even though this research resulted in a greater accuracy range by using post-game data, this limitation could be set aside if in-game data could be brought into use for further research.

For the selected topic of PUBG, one recent study by [10] has been done based on mathematical modeling to analyze major tactical features of the game. The driving force behind this study is to assist gamers in certain ways. First is to decide the location where the players could opt to jump as well as timing the jump properly. Considering the hit rate, certain small alternatives are analyzed for improvement based on aspects such as azimuth, the distance between two teams, etc. Another problem that was considered was for evaluating the points and performance of individual players based on KDA to determine the gaming contest winner.

B. *Prediction based Researches in Sports Domain*

The authors in [11] have used 10 years data of Turkish super league to predict outcome of football game based on win-loss-draw and point/no point scenario using Naïve Bayes, Decision tree and Ensemble models – Random Forest and Gradient Boosting trees. The missing values accounting to 11% of the dataset were removed as no significant relation could be found to perform imputation. Using 10-fold cross-validation, different classifiers were evaluated based on accuracy, sensitivity and specificity. In win-loss-draw scenario, Random forest performed the best amongst all the classifiers with accuracy of 74.60 %. In point/no point scenario, minority class oversampling was performed and an accuracy of 86.3 and 86.4 was achieved for Random Forest and Gradient boosted tree. The main drawback of this study was significant amount of missing values and exclusion of it from the model.

The authors in [12] have used Genetic Programming to predict the outcome of English Premier league in win, loss, draw scenario. The main advantage of Genetic Programming (GP) is its ability to generate high-quality functions as and when required. A set of 25 features from each game was selected and given as input to GP system. Bayesian network along with decision tree and ANN have been used for prediction. Bayesian network provides an accuracy of 52.21% whereas the best accuracy of 68.8% was achieved using ANN. The best overall accuracy of 76% was achieved by GP system consisting of 43 different GP-generated functions. The drawback of this system is that only 25 features were selected for analysis purpose.

The authors in [13] have used machine learning techniques to train classification models to predict outcome of the game. They evaluated decision trees, random forest, Naïve Bayes, multilayer perceptron (neural network) and rule learners for the model and have used adjusted efficiency statistics that were calculated relative to all the team for their attributes rather than raw average statistics. They claimed that modelling the difference between teams' attributes could not improve the overall accuracy and arrive at a conclusion that model selection is not as important as the feature selection in their study. They concluded that there is an upper glass ceiling of 74-75% accuracy for their models.

The authors in [14] have proposed a neural network-based prediction model to predict the outcome of final stage of 2006 Football world cup. A multi-layer perceptron with back propagation with 8 inputs, 11 hidden nodes and 1 output (8-11-1) has been used. Instead of doing feature selection, the authors have hand-selected feature set such as indirect free kick goals, direct free kick goals, goals scored, corner kicks, shots, shots on target, fouls suffered and possession percentage using their domain knowledge. This approach correctly predicts 10 out of 16 games in the final stage of world cup which includes win-loss-draw scenarios with accuracy of 62.5%. If the draw scenario is taken out, the model accuracy achieved is 76.9%.

The research put forward by [15] suggested a system that could be used for winner prediction in any general sporting event in any particular league. The unique and systematic approach to process the data (algorithm implementation) was done in a modular way so as to enable independencies in terms of modification for each module. All the modules are linked to a central module which also allows linking of a new module to the system. Data that was used for the purpose of the research was taken for all the teams in any sporting league either by scraping the data from the internet using a crawler or from the databases available. Based on manual feature selection, the researchers didn't include certain starting games for every team as they lacked data for the features representing a particular match. The major drawback of this research was that there was no base model for the purpose of comparing and evaluating the predictions done by the proposed system, they made use of referent classifiers which was not so reliable as well. Results suggested that a subset of classifiers that are used from most of the WEKA classifiers have better precision compared to the greedy logic-based referent classifier.

III. Methodology

A goal-driven methodology is an important part to carry out any data mining project because it gives the proper direction to the project, the various researcher has incorporated different methodology to successfully achieve their targeted results and evaluate in a systematic manner. For implementing this project successfully, the knowledge data discovery (KDD) methodology has been selected. KDD will guide properly to implement the project and this proposed structure will be followed throughout the project to successfully predict win place of players playing PUBG using different probabilistic methods.

The selection of methodology differs from researcher to researcher, based upon the need of the project must be selected. The type of data and final approach of researcher for the project plays a vital role in selection. Knowledge data discovery (KDD) helps to extract knowledge from targeted data.

KDD follows basic steps to achieve knowledge shown in the below figure.

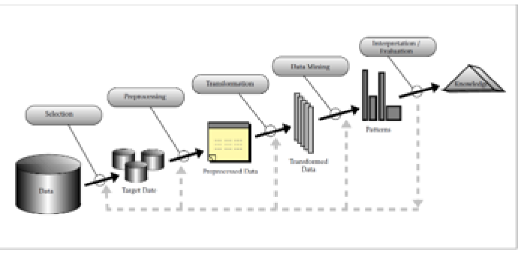

Figure 1: KDD Process [3]

The below-mentioned steps will give define approach from the extraction of data to knowledge discovery. Each step process used for this project will be shown in the next section.

- Domain
- Data selection
- Data preprocessing
- Data transformation
- Data mining
- Interpretation or evaluation
- Knowledge

IV. Implementation

Below steps mentioned are involved in the KDD process for this project.

A. *Data Selection and Extraction*

The PUBG data to predict win place percentage was taken from Kaggle[3]. The dataset contains various fields which leverage the prediction of win place percent prediction. Some fields present in datasets are player Id, match Id, assists, longest kill done by the player, walk distance, winning position of a player, match type in which player is involved, group Id if players are playing in groups, and so on. There are almost 26 features in the dataset which will be explored, and appropriate feature selection will be done with the help of feature selection techniques.

B. *Attributes*

Id: Unique Id of player, Match Id: Unique Id, groupId: Unique Id for group, DBNOs: Number (Nos) of enemy players knocked, assists: Nos of enemy knocked by teammates, boosts: Nos of boost item used, damageDealt: Total damage dealt, headshotKills: Nos of player killed by headshot, heals: Nos of healing item used, killPlace: ranking of player in the match killed, killPoints: external ranking of player, killStreaks: Max Nos of player killed in short duration, kills: Nos of enemy killed, longestkill: distance between player and enemy got killed, matchDuration: Duration of match in seconds, mactType: "Solo","duo","squad","solo-fpp","duo-fpp","squad-fpp" other modes are custom matches, rankpoints: ranking, revives: Nos of time this player revived teammates, rideDistance: Distance travelled by player on vehicles in meters, roadKills: Nos of kills while in a vehicle, swimDistance: Distance travelled by swimming measured in meters, teamKills: Nos of times player killed a teammate, vehicleDestroys: Nos of vehicles destroyed, walkDistance: Distance traveled on foot measured in meters, weaponsAcquired: Nos of weapons picked up, winPoints: win based external ranking, 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 the last place in the match. It is calculated off maxPlace, not numGroups, so it is possible to have missing chunks in a match.

C. *Exploratory Data Analysis*

The exploration of data is the most important part of a project before we start building our model. This exploration gives us fair enough idea about data that what is the nature of data is it linear, does it have missing value, outliers are present or not if yes then is it serious and so on these, all part were checked in the analysis. In this dataset, there were no missing values, but many features present in the dataset required scaling, for example, walkDistance and weaponsAquired all those features were scaled for the better performance of the model.

D. *Removing Inappropriate Attributes*

After the exploratory analysis the cleaning and transformation is performed, then the appropriate features are selected for the analysis. Some unwanted features like matchId, matchDuration, and matchType which are less likely to have any impact on dependent variable were removed for the ease of analysis.

E. *Feature Selection*

In this dataset there are almost 26 features so, to know the feature importance various methods were implemented by which precise and accurate features can be selected for the predictive model. Firstly, a correlation matrix was tested against all the features and to visualize this corrplot package was used which show the correlation matrix between variables. From the figure 2, the dependent variable winPlacePerc is highly correlated with walkDisatnce and boosts from the plot with the high color intensity. There is some negative correlation present in the correlation plot. Correlation matrix has given two most important feature to be selected for the predictive model but the features cannot be selected based on the correlation matrix only.

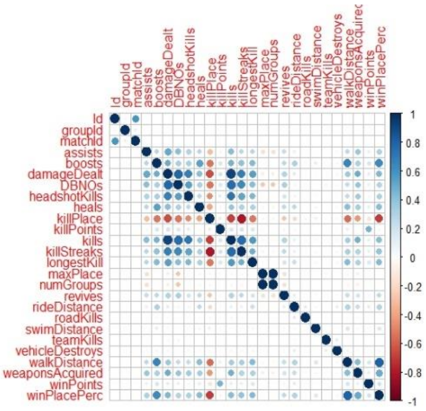

Figure 2: Correlation Plot

To recheck the correlation between dependent and independent variable the other two methods were used they are Bouruta and rank feature. The results were strange as compared to the above correlation matrix Bouruta model has given only one important feature whereas, the rank feature has given top four important feature which is a most important feature for the dependent variable. They are walkDistance, killPlace, killStreaks, and boosts shown in figure 3 which is selected as an independent variable for the predictive model.

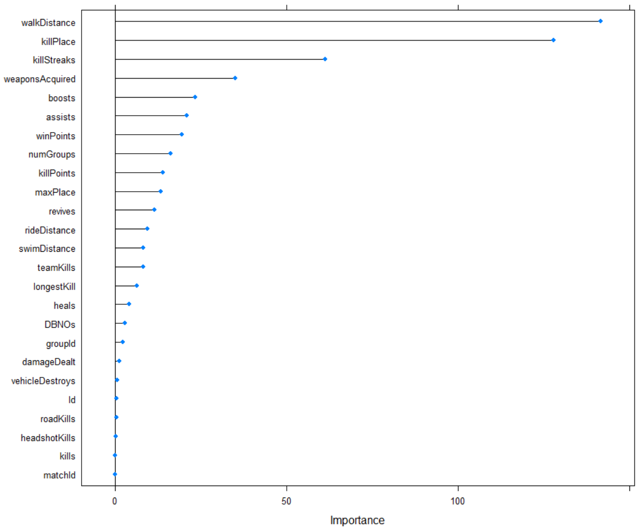

Figure 3: Feature Importance

Recursive Feature Elimination is shown below, the figure shows the accuracy by considering all the feature. Here it can be said that if the model is built by considering 4 features other than 20 features accuracy will be same. So, after analyzing importance of feature and RFE all the models are built with 4 most important feature which are walkDistance, killPlace, killStreaks, and boosts. Apart from this feature scaling was also done to avoid unwanted outcome in results.

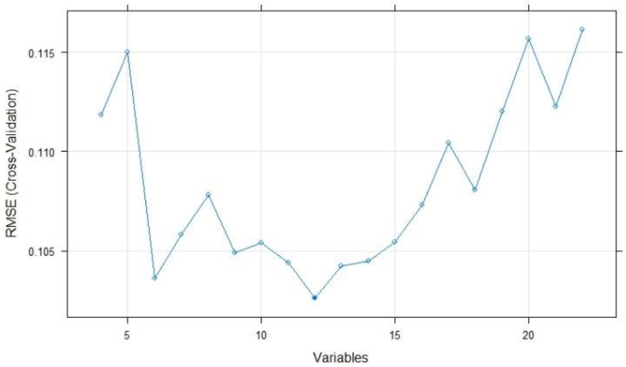

Figure 4: Feature Selection

F. *Outliers*

Outliers can affect the model as well as the outcome of model hence, the outlier is checked for the dataset by using boxplot. Below figure 5 shows some of the outlier present in the dataset.

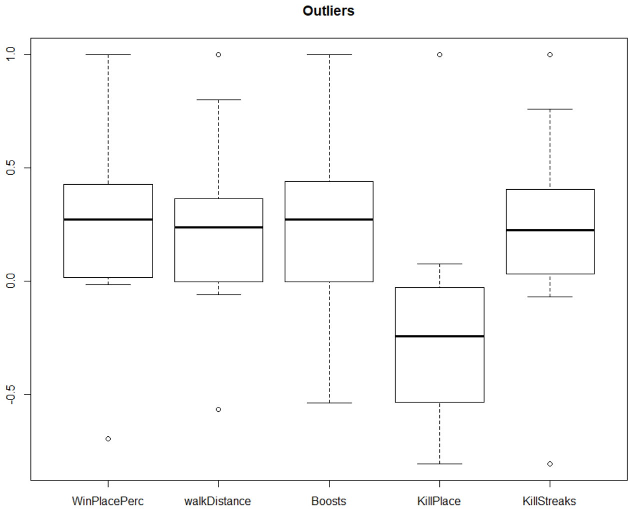

Figure 5: Outliers

G. *Models*

The following models are used in this project.

1. Linear Regression

In regression one of the most simple and accurate models for prediction is linear regression. To confirm the important features achieved from feature selection, initially in the lm regression all the features were considered for prediction of the dependent variable. The less important features were neglected one by one and finally, the four features walkDistance, killPlace, killStreaks, and boosts outstand among all the features. To improve model RMSE value was rerun with only highly correlated features mentioned above. RMSE value of test and train is compared to check the overfitting of the model. Finally, the plot was obtained for the actual and predicted value. To decrease the error rate of RMSE and to improve adjusted R square K-fold cross-validation is used with 10-fold in the cv method and 2-fold in repeated method.

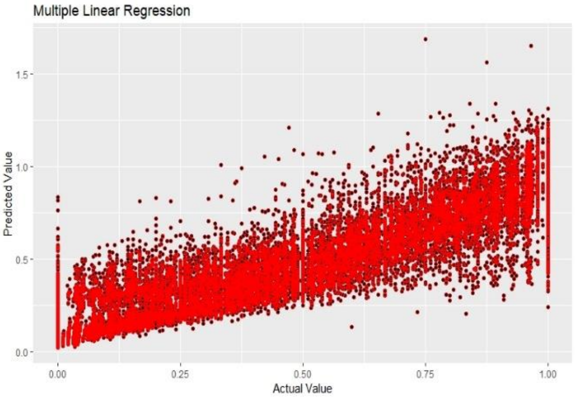

Figure 6: Linear Regression

2. Support Vector Regression

Researchers have shown faith in Support Vector Regression (SVR) for a better accurate prediction for regression. After the linear regression model now the most significant features have been selected and SVR model is applied. Normally, SVR model works on Support Vector Machine (SVM) principle for regression a small change in regressor formula is required which is a selection of eps-regression in type and in kernel linear. These change in regressor will allow the model to work for regression type data. Here, the dependent variable is winPlacePerc whereas walkDistance, kilPlace, killStreaks, and boosts are considered as an independent variable. For the better optimization of result, k-fold cross validation is used with 10-fold in cv method and 2-fold in repeatedcv method. The plot for predicted and actual value is shown below in the figure.

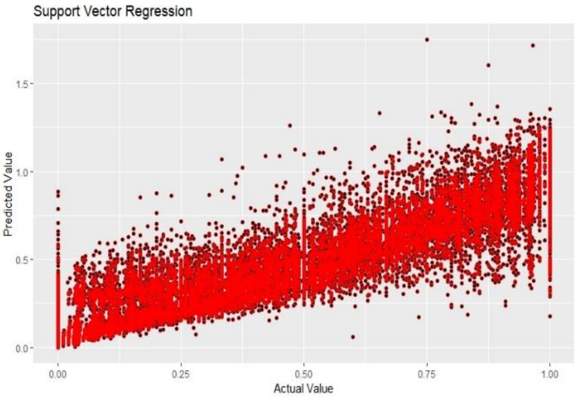

Figure 7: Support Vector Regression

3. Random Forest

Random forest is also used to check if the model is able to perform better and give less error. As discussed, the same method is followed for the important feature selection and checked the actual value against predicted value. The model is checked based on adjusted R square, Root Mean Squared Error (RMSE), and model accuracy apart from it for the tuning of parameter k-fold cross validation is used where the mtry is tweaked in both method cv and repeatedcv with the 100 trees and checked the RMSE error rate. Finally, the plot has been shown below for the actual against predicted value.

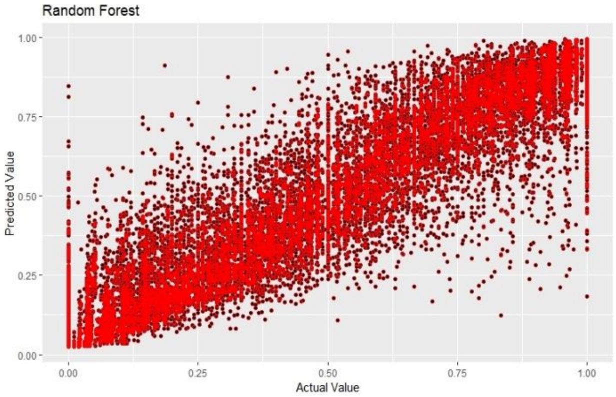

Figure 8: Random Forest

V. RESULT

KDD methodology is applied in the project and all the models were tested based on RMSE value and how accurate is the model apart from it adjusted R square, gives us fair enough information about the model and its performance. On the other hand, MAE value is also tested for the evaluation, both the RMSE and MAE metrics work on the term errors. In RMSE weight is assigned based on absolute error value where larger the value larger the weight is assigned on the other hand MAE assigned the same weight to all the errors [2]. The observation from all the models are given below:

RMSE for Linear Regression (LR) = 8.56

RMSE for SVR = 0.06

RMSE for Random Forest (RF) = 13.17

The accuracy of models is as follows 59.74 %, 74.60%, 76.08% for SVR, Linear regression, Random forest respectively. Here the implementation of k-fold cross validation is done using 10 as the number of folds and RMSE value is taken as the average of the values for the results. As it is evident from the results that least RMSE is obtained by SVR, but the accuracy of the model is not acceptable with the value of 59.74%. on the other hand, Random forest gives the larger RMSE error with more accurate model accuracy with 76.08% hence, we accept Random Forest over the other two models.

CONCLUSION AND FUTURE SCOPE

The PUBG data has been explored in terms of win place prediction. Through explorative analysis of data has been done and models were evaluated by RMSE value as well as the accuracy of models were also kept in mind for choosing the best model for prediction.

Initially, Linear regression model, the simplest method for the regression prediction was carried out. The results form linear regression was promising. The data was further to evaluate with the SVR model since, it is known for it is better performance for both the class classification as well as regression and used by many researchers for prediction. The results were good in terms of RMSE value and accuracy of model. Finally, the Random Forest was applied to dataset and the results were impressive in terms of accuracy than SVR and LR model but RMSE value was quite higher than other two model. Apart from all these model K-fold cross validation was applied to optimise the results of all the model and RMSE for each model were taken as the average of 10-fold RMSE value obtained from k-fold cross validation. The prediction of win place for PUBG can be of great help for players as well as the developer of PUBG. The players can plan different strategies to win the game and developer can build more complex game to restrict the players so they can enjoy the challenge. In future this analysis could be repeated to check does same strategies by players are used or have got changed by which we can get more interesting feature for win prediction of game.

REFERENCES

[1] License!Global (2019), "*Mobile Gaming is No Longer Child's Play: The accessibility of mobile gaming makes it broadly appealing across all age demographics"*, 22(1), p18, 1 p.; UBM LLC Language: English, Database: Academic OneFile

[2] C. Tianfeng and D. Roland (2014), *Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature*. Geoscientific model development, 7(3): pp.1247–1250.

[3] U., Fayyad, G., Piatetsky-Shapiro, and P., Smyth (1996). *The kdd process for extracting useful knowledge from volumes of data.* Communications of the ACM, 39(11):27–34.

[4] Conley, K., Perry, D.: How does he saw me? A recommendation engine for picking heroes in Dota 2. Technical report (2013)

[5]] Kalyanaraman, K.: To win or not to win? A prediction model to determine the outcome of a DotA2 match. Technical report, University of California San Diego (2014)

[6] Kinkade, N., Jolla, L., Lim, K.: Dota 2 win prediction. Technical report, University of California, San Diego (2015)

[7] Johansson, F., Wikstr¨om, J., Johansson, F.: Result prediction by mining replays in Dota 2. Ph.D. thesis, Blekinge Institute of Technology (2015)

[8] V. Hodge, S. Devlin, N. Sephton, F. Block, A. Drachen, and P. Cowling, "Win Prediction in Esports: Mixed-Rank Match Prediction in Multi-player Online Battle Arena Games," no. 2015, 2017.

[9] Norouzzadeh, Yaser & Spronck, Pieter & Sifa, Rafet & Drachen, Anders. (2017). Predicting victory in a Hybrid Online Competitive Game: The Case of Destiny.

[10] Ding, Y., 2018. Research on operational model of PUBG. In MATEC Web of Conferences (Vol. 173, p. 03062). EDP Sciences.

[11] E. Eryarsoy, "Predicting the Outcome of a Football Game : A Comparative Analysis of Single and Ensemble Analytics Methods," vol. 6, pp. 1107–1115, 2019.

[12] T. Cui, J. Li, J. R. Woodward, and A. J. Parkes, "An ensemble based Genetic Programming system to predict English football premier league games," Proc. 2013 IEEE Conf. Evol. Adapt. Intell. Syst. EAIS 2013 - 2013 IEEE Symp. Ser. Comput. Intell. SSCI 2013, pp. 138–143, 2013.

[13] A. Zimmermann, S. Moorthy, and Z. Shi, "Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned," 2013.

[14] K. Y. Huang and W. L. Chang, "A neural network method for prediction of 2006 World Cup Football Game," Proc. Int. Jt. Conf. Neural Networks, pp. 1–8, 2010.

[15] Zdravevski, E. and Kulakov, A., 2009, September. System for Prediction of the Winner in a Sports Game. In International Conference on ICT Innovations (pp. 55-63). Springer, Berlin, Heidelberg.