C7081 Assignment for Harry Buckley

Harry Buckley

25/11/2021

Link to Github repository.

Background

League of legends is a team-based game developed by the company Riot Games for PC. The game is a red verses blue objective based game where player pick characters with differing abilities to attempt to gain an advantage. The objective of the game is to destroy defending towers and kill enemy minions to gain entry to the enemy team base and destroy the "Nexus". Killing minions and objectives, such as towers and other elite monsters, rewards players with gold which is used to buy items which enhance the characters attack, defence, hit points and ability powers. The game has a following of around 115 million competitive players and splits these up into ranks from bronze to challenger (Kou and Gui, 2020), ranks are shown in table 1 below.

Table 1: Number of players in each rank.

Rank	**player percentage(%)**
Iron	7.100
Bronze	22.000
Silver	35.000
Gold	23.000
Platinum	7.900
Diamond	2.500
Master	0.032
Grandmaster	0.040
Challenger	0.017

It can be seen from the table that the top three ranks make up less than 1% of the total playing population of the game, these tend to be sporting professional who compete globally.

The game can be broken up into three segments: Early game, Mid game, and late game, where the prediction of who will in the game can swing from team to team. Early game is defined as the first 20 minutes where, for the first 15 minutes, there is an inability to surrender (forfeit) the game; a unanimous decision to surrender must be achieved between the 15-to-20-minute mark. For this report the data for the first 10 minutes of games is to be used for games in the platinum ranking. With the outcome of the game being successfully predicted after these 10 minutes, clarity can be given to players who are unsure whether forfeiting and starting a new game is the correct decision.

It has been shown that video games and esports share a link to the human desire to gamble (Macey and Hamari, 2018; Fisher and Griffiths, 1995; Johansson and Gotestam, 2004; Wood et al., 2004). Understanding the risks and wanting to swing the odds in favour of the player is well documented (Ore, 2017), and first published in Cardano's Book on Games of Chance in 1564 (Cardano, 2015). Incorporating the use of probability and understanding when to forfeit can save a player money, but more commonly time.

This report will endeavour to successfully predict the outcome of League of Legends game played in the platinum rank dependant on the outcomes of variables in the first 20 minutes of a game, giving the player an understanding of the odds comparable to what Cardano achieved in 1564. A similar study, conducted by de Souza and Cortimiglia (2017), was seen to achieve a 75% accuracy in predicting the outcome of a league of legends game using Logistic Regression and Random Forests; 75% accuracy will be taken as the benchmark for success in the methods implemented for this report.

Objectives

- Find the highest accuracy possible from a range of models to see if we can successfully predict the outcome of the game looking only at variables from the first 10 minutes.

- State the most influential variables to the outcome of the game and theorise as to why this might be.

Methods

Data

The data was taken from the Kaggle dataset search engine and contains 9880 data points. A subset of 30 variable were taken forward from the original data set as some were removed as they were seen to have little relevance on the outcome of the game from a user's perspective. Sub setting occurred when the data was turned into tidy form in Microsoft Excel. The data contains data points recorded for the first 10 minutes of games in the diamond rank of Leagues of Legends for both the blue and red team. The variables and their description can be found in table 2.

Table 2: Table of variables and description.

Variable	Class	Description
b_wins	Factor	Dependent Variable - 1 if Blue team wins, 0 if Red team wins
b_war_pl	Numerical	Number of wards the Blue team placed
b_war_des	Numerical	Number of wards the Blue team destroyed
b_fir_blo	Factor	1 if Blue team achieves the first kill, 0 if Red team achieves first kill
b_k	Numerical	How many kills the blue team achieved
b_d	Numerical	How many deaths the Blue team achieved
b_a	Numerical	How many assists the blue team achieved
b_e_mon	Numerical	Number of dragons and heralds the Blue team killed
b_tow_des	Numerical	Number of towers the Blue team destroyed
b_tot_gp	Numerical	Total amount of team gold earned by Blue team
b_av_lvl	Numerical	Average level of the Blue team
b_tot_xp	Numerical	Total amount of team experience earned by Blue team
b_tot_mons	Numerical	Total number of minions killed
b_cs_min	Numerical	Number of minions, monsters and wards Blue team have destroyed per minute
b_gp_min	Numerical	Amount of gold the Blue team earned per minute
r_war_pl	Numerical	Number of wards the Red team placed
r_war_des	Numerical	Number of wards the Red team destroyed
r_fir_blo	Factor	1 if Red team achieves the first kill, 0 if Blue team achieves first kill
r_k	Numerical	How many kills the Red team achieved
r_d	Numerical	How many deaths the Red team achieved
r_a	Numerical	How many assists the Red team achieved
r_e_mon	Numerical	Number of dragons and heralds the Red team killed
r_tow_des	Numerical	Number of towers the Red team destroyed
r_tot_gp	Numerical	Total amount of team gold earned by Red team
r_av_lvl	Numerical	Average level of the Red team
r_tot_xp	Numerical	Total amount of team experience earned by Red team
r_tot_mons	Numerical	Total number of minions killed
r_cs_min	Numerical	Number of minions, monsters and wards Red team have destroyed per minute
r_gp_min	Numerical	Amount of gold the Red team earned per minute

The data was read into RStudio (2021) using the readxl package due to the tidy data being created in Microsoft Excel. The data frame was initially put into a pairs plot to visualise correlation; however, the number of variables and data points made this challenging. An approach was taken to look at half the variables at a time against the dependent variable of Blue team wins. This saved computing power and made the pairwise plot more readable. Within this plot, the outcome of the game was highlighted using colour for the Blue team wins variable. This was done for both the blue and red team variables.

Boxplots were made to view independent variables individually against the dependent variable. This led into the analysis of the data where the summary of the data frame was printed.

Correlation was checked to test for collinearity, as there was evidence of this it was decided that some variables should be removed. In a study by Dormann et al. (2012) it was shown that collinearity and cause severe problems when a model is trained on data with non-independent data, as a test for validity a model was run without collinear variables removed, a total of 8 variables were removed.

Binary Logistical Regression

It was decided that due to the dependent variable being a win or lose outcome that a binomial logistic regression would be appropriate for the data set. To correctly use the logistical regression, it was necessary to use the logistic function, this prevented values for the probability from being negative or greater than one, this also meant that the method of maximum likelihood was used due to its statistical properties (James et al.,2013).

The data was then split into test and train data using an 80:20 split of the data. This put 7902 data points into the train data and 1977 in the test.

A histogram with the fitted values was produced for visual clarification of the probability and frequency of blue team winning. The predictions were then looked at with they type being response to view the probability that blue team will win throughout the original data frame. The coefficients were viewed and variables with significant coefficients were selected (figure 1).

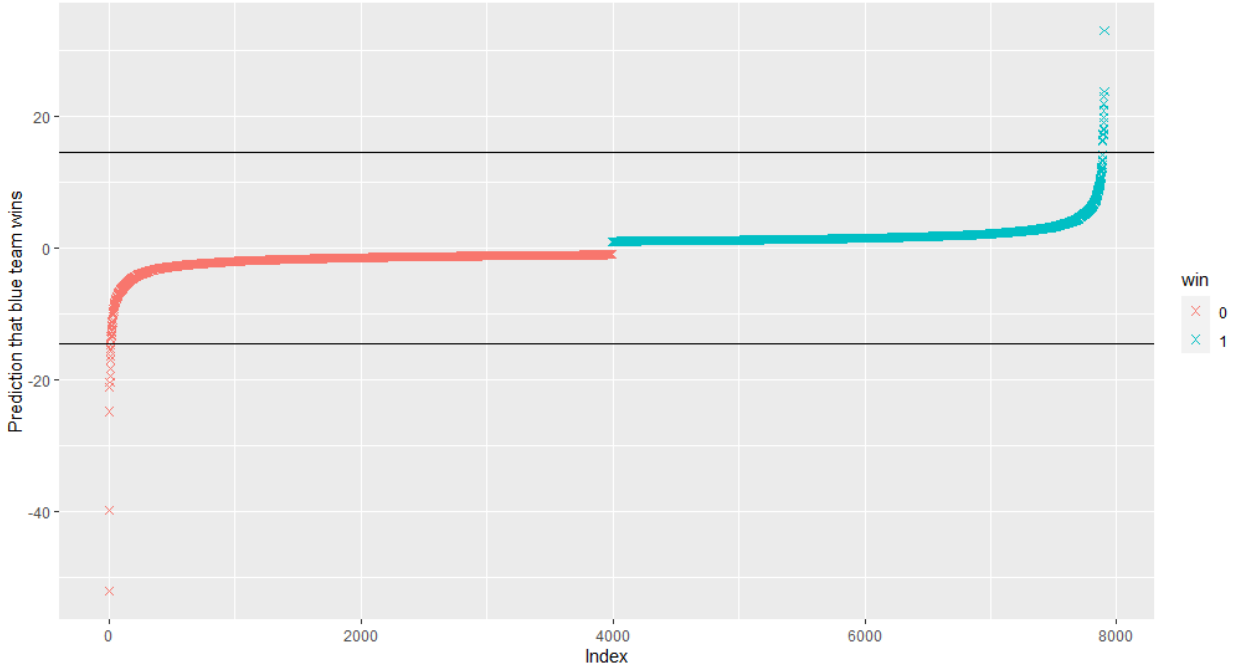


Figure 1: Histogram of fitted values

The model showed that there were six significant predictor variables table 2. As two of the variables with the lowest p-values were total gold earned by each team it can be noticed that this might have the largest influence on the outcome of the game.

Table3: Significant variables

Coefficient Estimates	P-Values
0.4622157	-0.9716654
0.2132727	-0.0018572
0.6130473	-0.0153836
0.1639536	-0.0178232
0.0000031	0.2324458
0.2124811	-0.1969427
0.0000000	0.0004892
0.8637778	0.0350526
0.0000897	0.0002520
0.0406133	-0.0400201
0.6143698	-0.0007293
0.1090655	-0.0940674
0.8448133	-0.0059220
0.0494095	0.0245986
0.0000022	-0.2341810
0.0915447	0.2899970
0.0000000	-0.0004726
0.7643765	0.0612556
0.0003408	-0.0002308
0.0972084	0.0322561

Validation of the binomial logistical model was completed where the variance of the residuals was plotted to check that 95% of the data points lay within +/- 2 standard error of the mean.

The standard residual plots were also viewed, as seen in figure 2, and it was found that the data was normally distributed with few outliers.

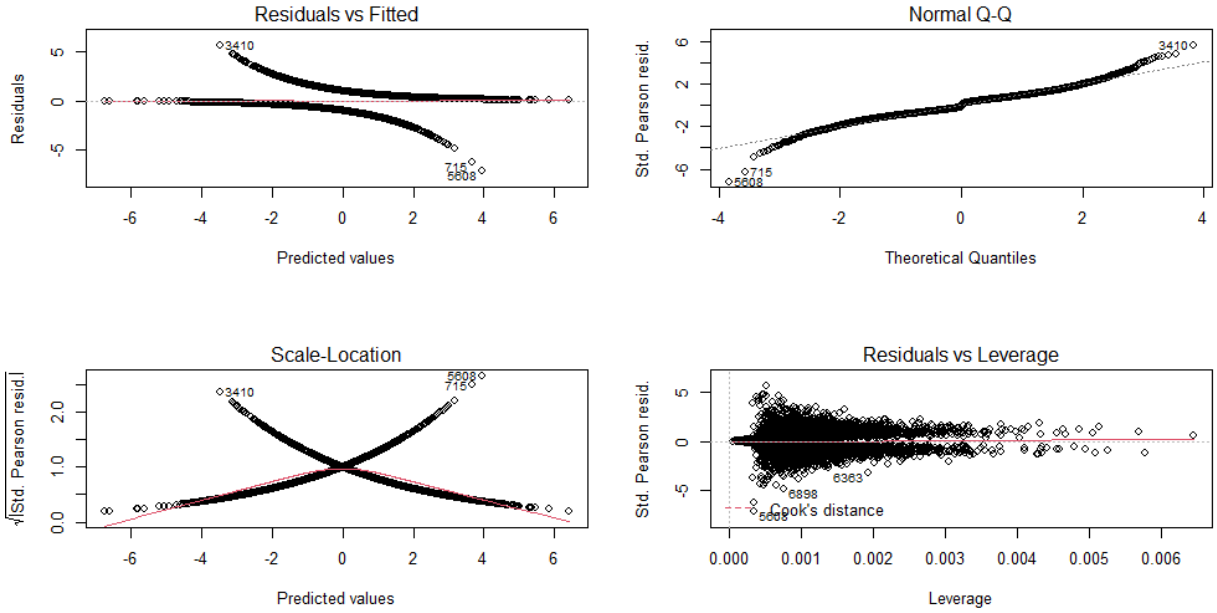
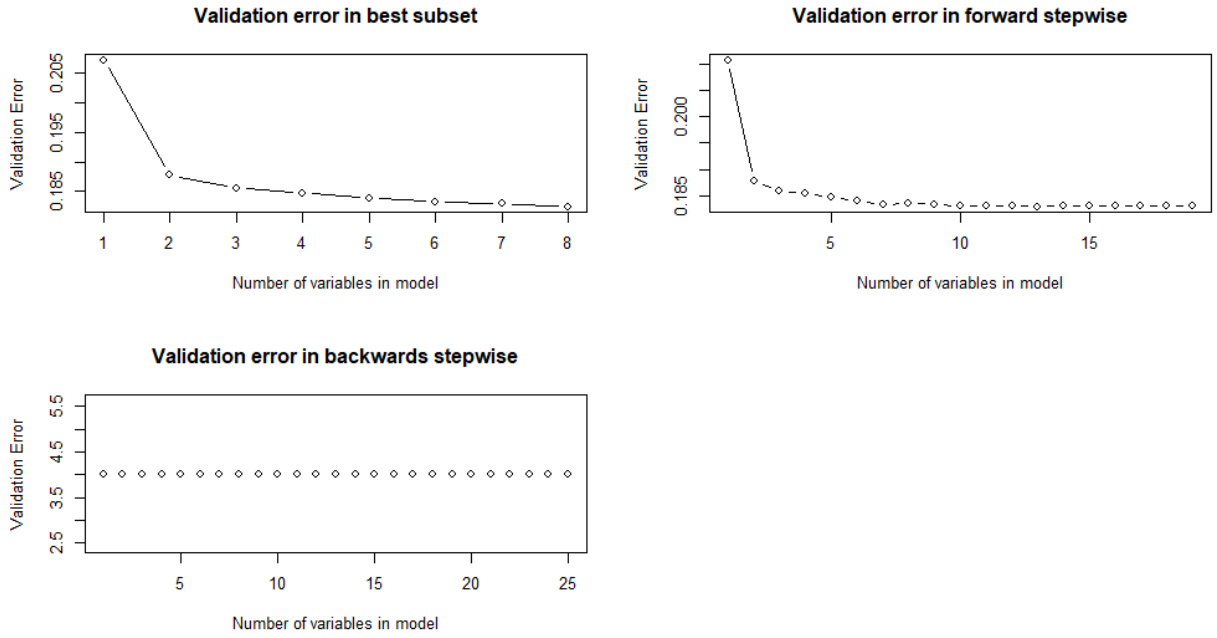


Figure 2: Plots of the standardised residuals

Stepwise Regression

Stepwise regression was chosen for exploratory analysis as the backwards selection of variables would allow for the variables with the least effect will be removed. Stepwise regression does have its issues; it has been reported that the forward method of stepwise regression can produce suppressor effects within variables. It also can cause issues as it underestimates the relationships between variables when it removes them from the model (McElreath, 2018). Results from the stepwise model should be validated carefully and both forward and backwards stepwise were used to determine the difference in accuracy. This method was chosen due to its wide use in determining accuracy in models.

Validation error depending on model size was viewed here for the types of regression used, this is shown in figure 3.

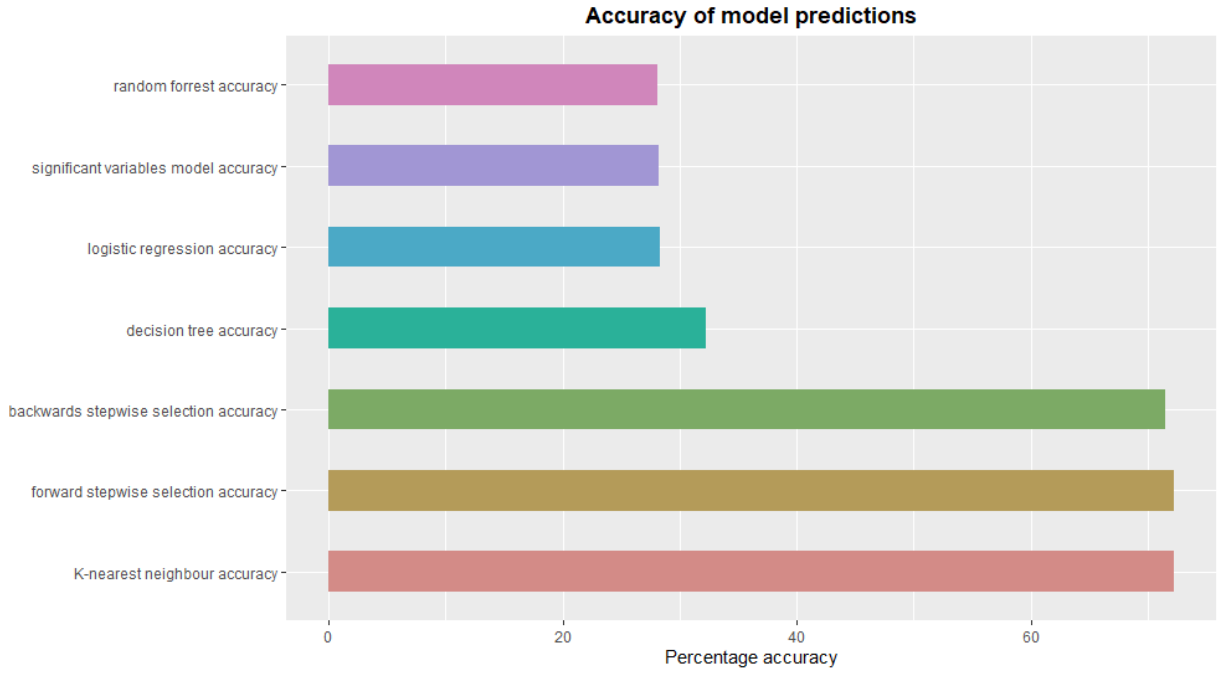


Tree and Random Forest

Both tree and random forest methods were implemented in this study to determine their accuracy on the outcome. As the dependant variable was binary, a classification regression method was used. In this it was necessary to work out a value for m, as this is the number of predictors used at each split. For classification problems, like the one we are studying, it is considered correct to take m = square root of p when p = the total number of predictors. In this study m = 4. As random forest was a method used to attain a 75% accuracy in the study conducted by de Souza and Cortimiglia (2017) it was used to assess its outcome using this dataset.

Results

After running the binary logistic regression it is seen that there are 10 variables that show significance in the result of the blue team winning or losing. The most significant variables, shown by the lowest p-value, are the total gold for both blue and red team. Accuracy of the model was attained by showing test data to the model and validating the outcome with the original data set. K – nearest neighbour was seen to produce the highest accuracy with 72.25% closely followed by forward stepwise regression with 72.23%, as shown in figure 4.



Interestingly the decision tree and random forest methods used did not achieve a high accuracy, with the study of de Souza and Cortimiglia (2017) producing their results using this method it would be worthwhile comparing the methods and datasets used. It should be noted that the data used in this study only records within the first 10 minutes of the game. Games can last for over 60 minutes and this is where the 2017 study could have different results.

The results of the backwards stepwise model show us that the variable total gold for both teams has the lowest p value, this is consistent throughout all the models; for example, in the decision tree method it was the only variable that was chosen to make the tree. Taking this into account, it can be said with confidence that total gold earned by a team in the first 10 minutes of a game highly influences the outcome of the game. Gold allows players to buy items which improves damage as well as other major character abilities giving the player with the most items purchased a distinct advantage over others.

It should be taken into consideration that a lot can happen in the rest of the game, and the outcome can be influenced by other factors: such as players, deliberately or unintentionally, leaving the game; players making mistakes; or players having picked a wrong match up between characters.

Conclusion

From the range of models used it was seen that the k-nearest neighbour model produced the highest accuracy regarding outcome of the game. Stepwise regression is a method that is used commonly when reporting accuracy of outcomes which is used by current governments to inform policy, but there are issues regarding its use. As a 75% accuracy was stated as the benchmark in reporting the accuracy it can be said that this study did not hit the benchmark. Were this study to be improved a Bayesian method might be used to predict the outcome. It would also be useful to compare the results from other ranks of league of legends to see if the variables that influence the outcome are similar

References

Kou, Y. and Gui, X., 2020. Emotion Regulation in eSports Gaming: A Qualitative Study of League of Legends. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), pp.1-25.

Souza, R.T.D., and Cortimiglia, M.N. 2017. Aplicacao de algoritmos classificadores para previsao de vitoria em uma partida de League of Legends.

Ore, O., 2017. Cardano: The gambling scholar (Vol. 5063). Princeton University Press.

Cardano, G., 2015. The book on games of Chance: the 16th-century treatise on probability. Courier Dover Publications.

Johansson, A. and Gotestam, K.G., 2004. Problems with computer games without monetary reward: similarity to pathological gambling. Psychological reports, 95(2), pp.641-650.

Wood, R.T., Gupta, R., Derevensky, J.L. and Griffiths, M., 2004. Video game playing and gambling in adolescents: Common risk factors. Journal of Child & Adolescent Substance Abuse, 14(1), pp.77-100.

RStudio Team (2021). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Marquez, J.R.G., Gruber, B., Lafourcade, B., Leitao, P.J. and Munkemuller, T. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), pp.27-46.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

McElreath, R., 2018. Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.