Final Project – Who will lead to win?

Jeremy Zeng

11/01/2020

Summary

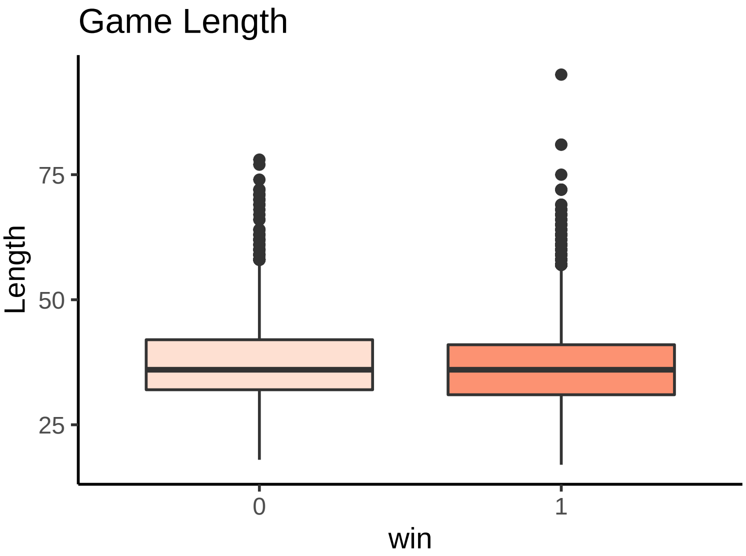
League of Legends is an online 5 vs. 5 competitive PC game, starting with two teams, for each team having 5 different roles, Top, Mid, Jungle, ADC and Support. This project is going to conduct some analyze workes on this game. The goal of this report is going to answer the key question "Is earning more gold will lead team to win a match?". In other words, this report is trying to measure the effects of gold difference between two teams earned by different roles on the winning of the game. Boxplots and binned plots will be used to analyze the association between binary response variable "win" or "loss" and numeric predictor variables. Joint probability table and conditional probability table will be used to analyze the association between between binary response variable and categorical predictor variable. Binned residual plot will be used to evaluate the overall fit of the regression model, and check if the function of numeric predictors is well specified. Chi-squared test will be used to compare the deviance of null model and new model. Confusion matrix and ROC curve will be used to validate the performance of the model through calculation of sensitivity, specificity, accuracy, and AUC curve. VIF will be used to calculate the multicollinearity of the function. The outcome of the study shows that earning more gold will increase the odds of winning. However, support is not an important role to lead the team to win since it is not significant.

Introduction

The response variable in this study is if the blue team win or loss. And the predictor variables, this study is interested in looking into are topAvg, MidAvg, JungleAvg, ADCAvg, SupportAvg, KillDiff, GameLength, Season, Carry and TeamofCarry. After completing data modeling, model validation will be performed. Interesting associations with odds of win will also be highlighted. The study will begin with EDA with the goal of checking the association of predictor variables and response variables, and highlight the preliminary concerns for the response and predictor variables. Next preliminary logistic regression model fitting will be conducted (excluding interactions and transformation) to understand the significance of coefficients for each selected predictor variable. Binned residual plots will be utilized to assess the overall fitted quality and check if the function of predictors is well specified. Next, preliminary model validation will be performed to the preliminary model. Confusion matrix and ROC curve will be utilized to validate the performance of preliminary model through calculating the sensitivity, specificity and accuracy. Through carefully analyzing the results from EDA, preliminary model fitting and validation, interactions will be added to imporve the fit of the model. Logistic regression model fitting and validation will be performed to justify each modification made to the model. In addition, chi-squared test will be implemented to compare the performance of the new model to the original. Lastly, stepwise function will be performed to find the optimal model with the lowest AIC score. To ensure the final model fulfills the assumptions of logistic regression, model validation will be performed again. Last but not least, VIF will be used to check for any multicollinearity for the final model.

Data

The original data has 7620 observations[1], with winner, “blue” and “red” as the response variable. The primary step is to convert the response variable to categorical variable. Since if the winner is blue team, red team would be defeated. This report is going to perform analysis based on blue team, so if the winner is blue team, the response variable will be equal to 1, otherwise, it will be equal to 0. In this data, the number of blue team win is 4146, and the number of blue team lose is 3474. There are 10 predictor varibales: TopAvg, JungleAvg, MidAvg, ADCAvg, SupportAvg, GameLength, KillDiff, Season, Carry, TeamOfCarry. The first 7 are numeric predictors, while other 3 predictor varibales are categorical. RoleAvg means that the gold difference between two teams for a certain role. If positive, blue team has advantage; if negative, red team has advantage. The boxplots of RoleAvg vs win showed that for every role, the distribution of gold difference will be higher for winning a match than that of losing a match. In other words, earning more golds tend to have a higher chance to win the game. The boxplot of KillDiff vs win showed that the distribution of kill difference by team of win is higher than that of lose. The boxplots of GameLength vs win showed that the distribution of win and lose is about the same, which means that the length of a match has little effects on winning a match (see plot below). This report will investigate the effects of GameLength further in the model fitting process.



Next, EDA of analyzing the association of the response variable and categorical predictors, Carry, TeamOfCarry and Season, will be conducted. Three tables are drawn for each predictor. The first table shows the number of win and the number of lose for each level of categorical predictor. The second table shows the probability of each combination of response variable and categorical predictor with the denominator as the total number of data. The third table shows the conditional probability for each combination of response variable and categorical predictor. Next, we calculate the p-value of Chisq-test to see if the predictor is significant. According to the p-value table below, only Carry and TeamOfCarry variables seem to have association with response variable (more investigation is needed). Since Carry is the role who has the highest gold difference of two teams, which means that for a certain Carry, it could be came from team blue and team red. And TeamOfCarry is the Carry from which team. This report will drop Carry for further studies.

Variable	Carry	TeamOfCarry	Season
p-value	< 2.2e-16	< 2.2e-16	0.3457

Model

After conducting EDA, the next step is to construct the preliminary logistic model and conduct model validation with all major predictor variables for this study (excluding transformation and interaction), which are TopAvg, JungleAvg, MidAvg, ADCAvg, SupportAvg, GameLength, KillDiff, and Season. The summary table of the preliminary model shows that only the coefficients of TopAvg, JungleAvg, MidAvg, ADCAvg, GameLength, KillDiff are significant, the rest of the predictor variables have a p-value above 0.05. In addition, according to the binned residuals versus predicted probabilities plot, all points are within the standard error bounds and the overall plot appears to be random. For RoleAvg, the binned residual plot displays a random pattern with all points except less than 3 within the standard error bounds (transformation is not to be needed). According to the confusion matrix of the preliminary model with 0.5 threshold, the optimal sensitivity and specificity is (0.956, 0.942) and the overall accuracy is 0.95. In addition, the ROC curve shows the optimal 1-specificity and sensitivity is (0.94, 0.96) and the AUC value is 0.987. Although the result looks good, interactions and model selection will be performed for analyzing reason.

Through analysis the summary table of preliminary model, this study is trying to convert the RoleAvg to a categorical predictor, that is if RoleAvg is positive, which means that blue team has advantage, RoleWinningTeam will be equal to 1, otherwise, it will be equal to 0. Another model will be fitted with predictor RoleWinningTeam. The result of the model showed that only the p-value of Support is lower than 0.05, which means it is significant. So this report will still conduct further study on the RoleAvg.

The next step will be investigating potential interaction in logistic regression. Since the main purpose of this study is to find the association of gold and win, interaction of RoleAvg and all other predictor variables will be investigated. Only boxplot for numeric variable SupportAvg vs Win by TeamOfCarry shows a little difference in distribution between TeamOfCarry and SupportAvg. Therefore, the interaction of SupportAvg and TeamOfCarry appears to be a potential limitation. After fitting another model with the interaction of SupportAvg and TeamOfCarry, the anova test showed that the p-value is 0.15, which is greater than 0.05 and even 0.1. Therefore, adding this interaction does not improve the model.

For this report, stepwise selection method will be implemented to find the lowest AIC because BIC generally places a heavier penalty on models with more than 8 variables. Main predictors that will be included to the full models are TopAvg, MidAvg, JungleAvg, SupportAvg, ADCAvg, KillDiff, GameLength, TeamOfCarry and Season. Besides the main predictors, we will include 2 major kinds of interactions, including: 1) RoleAvg and all other variables 2) TeamOfCarry and Season. After performing stepwise selection, the final model ended up with 8 predictors: KillDiff, TopAvg, GameLength, ADCAvg, JungleAvg, MidAvg, ADCAvg:JungleAvg, TopAvg:JungleAvg, . These six major predictors match the finding from EDA. However, interactions of ADCAvg and JungleAvg, and TopAvg and JungleAvg are kept from the model. Through the results of EDA and potential interaction investigations, another potential interactions is identified: SupportAvg and TeamOfCarry. Chi-squared tests are conducted to decide whether to include this interaction in the final model. The result of Chi-Squared test shows including SupportAvg:TeamOfCarry has a p-value 0.48 compare to excluding these interactions. Therefore, this study will not add this interaction into our final model. The binned residual plots of our final model look random, and all points are within the standard error bound. In addition, all VIF values are smaller than 10, which is good (no violation of multicollineaity). According to the confusion matrix with 0.5 threshold and ROC curve of the final model, the optimal sensitivity and specificity is (0.955, 0.944), accuracy is 0.95, and AUC value is 0.988, which are slight better than the preliminary model.

$$log(\pi_i/1-\pi_i) \ = \sum \beta * x_i; \ Bernoulli(pi_i).$$

Above is the equation of final model. pi/(1-pi) is the odds of win for observation i, and x_i is the vector containing the corresponding values for predictor variables. The AIC value of final model is 2103.

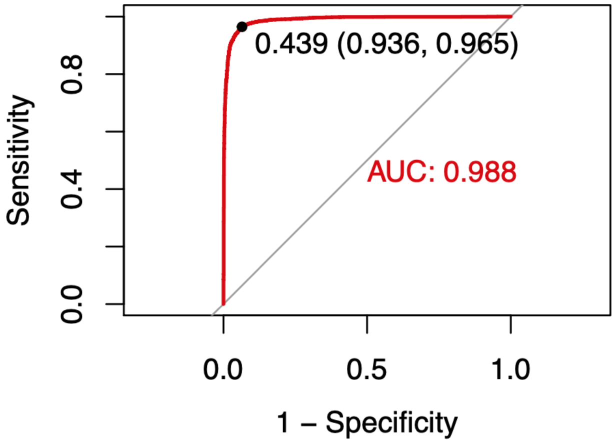
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.6929	0.4554	1.522	0.1281

	Estimate	Std. Error	z value	Pr(>|z|)
KillDiff	0.4944	0.01476	33.49	7.884e-246
TopAvg	6.093e-05	0.0001779	0.3426	0.7319
GameLength	0.06007	0.01088	5.523	3.328e-08
ADCAvg	0.0009759	0.000173	5.641	1.695e-08
JungleAvg	0.0004559	0.0002183	2.088	0.03676
MidAvg	0.0001914	0.0001005	1.904	0.05693
ADCAvg:JungleAvg	4.2e-07	8.46e-08	4.965	6.867e-07
TopAvg:JungleAvg	-2.488e-07	9.166e-08	-2.714	0.006647

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	10504 on 7619 degrees of freedom
Residual deviance:	2086 on 7611 degrees of freedom

The table shows the summary of final model including the interactions of JungleAvg with ADCAvg and TopAvg.



Conclusion

According to the summary table of final model, predictors that are significant at the 95% level are KillDiff, GameLength, ADCAvg, JungleAvg, ADCAvg:JungleAvg, TopAvg:JungleAvg with p-values lower than 0.05. The p-value of coefficient for MidAvg is 0.057, which is slightly higher than 0.05, but still significant at 90% level. As TopAvg increase by 1 unit, which means that blue team Top earn 1 more gold than red team Top, the odds of blue team win will increase by approximately 0.0061% with all other variables constant. Since the p-value of treat is 0.7319, the impact of TopAvg on the odds of win is not significant. As ADCAvg increase by 1 unit, which means that blue team ADC earn 1 more gold than red team ADC, the odds of blue team win will increase by approximately 0.0976% with all other variables constant. As MidAvg increase by 1 unit, which means that blue team Mid earn 1 more gold than red team Mid, the odds of blue team win will increase by approximately 0.0191% with all other variables constant. As JungleAvg increase by 1 unit, which means that blue team Jungle earn 1 more gold than red team Jungle, the odds of blue team win will increase by approximately 0.0456% with all other variables constant. The interaction between Jungle and ADC will have positive impact on the odds of win, while the interaction of Jungle and Top have negative impact on the odds of win. As KillDiff increase by 1 unit, which means that blue team kill 1 more red team player, the odds of blue team win will increase by approximately 63.95% with all other variables constant. As GameLength increase by 1 unit, which means that the game last for more than 1 minute, the odds of blue team win will increase by approximately 6.19% with all other variables constant.

To sum up, earning more gold will help the team to win a match. Mid and ADC players play essential roles to help the team to win a match. If Jungle help the bottom lane more, which is the lane of ADC and Support, the chance of win the match will be higher.

This study have some limitations, first, the data used for this analysis may not accurate. Second, there might be some overfitting problems. And the predictor gold earn by role itself may not very important to predict the win of a match. The damage did to champion by roles could be more intuitively suitable for predicting the winning.

Data Source

[1] https://www.kaggle.com/chuckephron/leagueoflegends