Predicting The Probability For Winning A League Of Legends Game Based On The First 10 Minutes Of A Match Using Logistic Regression Model

Trong Tuan Hung Dao 1005093337

21/12/2020

Link to the Github repo: https://github.com/hungdaouoft/STA304-Final-Project.git

Abstract

In this report, we construct a multiple logistic regression model on a sample dataset contains statistics on the first 10 minutes of 500 Diamond or higher rank games of League of Legends, by utilizing backwards AIC and BIC elimination. We found out that the most significant factors in the first 10 minutes of the game are dragon control and money farming to wider the financial gap with the opponent team. Winning a computer game may sound silly to the general public since playing games is just an outlet for stress relief, but for a lot of people, it is their main income source. They can make a living because the industry is booming at an extraordinary rate and the demand for entertainment in the current global pandemic is inarguably among the top in the past decades.

Keywords

Logistic Regression, Esports, Gaming, League Of Legends, backward AIC, backward BIC.

Some other useful terminologies:

Minions: NPC that reward golds to an individual who defeat it.

Dragons: Elite monster which reward a buff and money for the whole team when defeated. There can only be one Dragon on the map at any time.

Introduction

The current pandemic is putting a lot of people under stress. In a report released on October 20, 2020, Statistics Canada stated that "Since COVID-19, fewer Canadians report having excellent or very good mental health - 55% (July 2020) down from 68% (2019)". To fight against the unwanted pandemic, citizens are expected to obey social distancing rules and other safety health measures. Because more individuals are staying home, there is a huge increase in the demand for entertainment outlets, one of which is through gaming. According to a recent post from The Verge, based on Youtube's 2020 statistics, more specifically Youtube Gaming head Ryan Wyatt, the author Nick Statt expressed "users of the video-sharing site watched 100 billion hours of gaming content on the platform this past year, double the number of hours watched in 2018". When speaking about gaming, we usually imagine an easy-going atmosphere that comes with it since

the most primal purpose of it is for stress relief, but it is not always the case. Competitive gaming exists and the population of viewers is expecting to grow at a 9% compound annual growth rate between 2019 and 2023, from 454 million in 2019 to 646 million in 2023 (a preview from Business Insider). From the same report, Mariel Soto Reyes suggested that the investments into this industry were up to 4.5 billion dollars in 2018 from just 490 million dollars in the year before. This shows an outstanding interest for the field which leads to many professions being born from it, the most notable one is probably professional players (gamers).

In the world of competitive gaming, the major contributor to its revenue is from advertisements through competitions (called esports) and among the largest scale games is League Of Legends. It is a team-based strategy game published by Riot Games where 5 players of the same team assume the role of *champions* to compete with other 5 players of the opposite team to claim victory at the end (the condition to win is depended on the game mode). In 2019, the game had generated 20 billion dollars in revenue since its first release (Forbes). Because of this, it has been one of the main income sources for both professional (pro) players and streamers alike. A report from the Business Insider from 2015 stated that, in daily basis, players from Team Liquid, one of the top League Of Legends team of North America, spend a minimum of 50 hours weekly, 8 hours daily, practising together. So for them, this is no longer just a game, is it a way to make a living, and the more match that players win, the higher rewards they can obtain.

Given the scale, it is natural that a lot of effort is being put into finding strategies to win a game, and in this report, we are aiming to predict a winning team based on their performance in the first 10 minutes of the games. Since the outcome is either win or loss (binary) we will build a multiple logistic regression model using some of the quantitative variables in our original dataset, the choice of variables will be decided using backwards AIC, BIC elimination methods to pick out which variables will be more useful in predicting the outcome of a game, hence arriving at our final model.

We will briefly describe the original data, outline the construction of our sample data while also performing AIC, BIC backward elimination to choose a fitting model (by removing redundant features) that we can use in the Methodology section. Results from the attained logistic model are provided and interpreted in the Results section. In the Discussion section, we will provide a summary of the report, giving some conclusions that we had drawn based on our results while at the same time, discussing possible limitations and future approaches of this report.

Methodology

Data

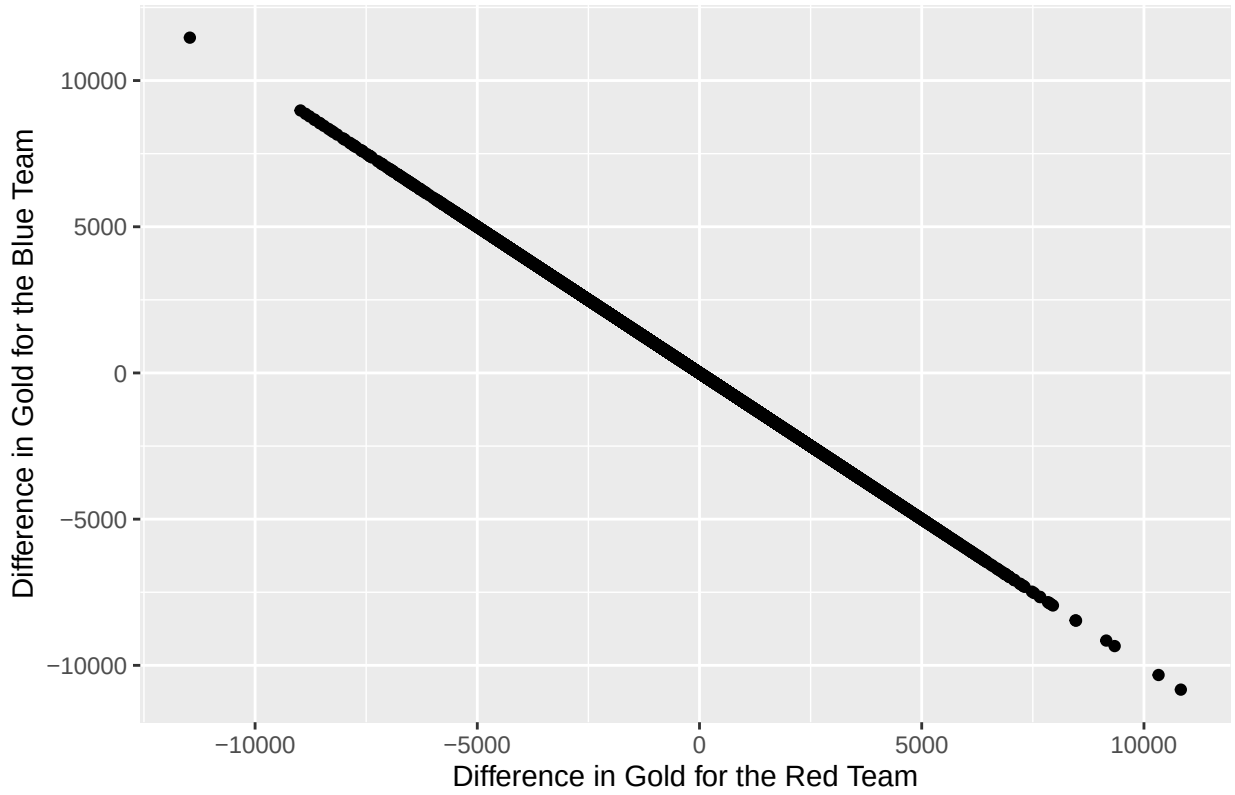
League of Legends has a competitive ranking system, in which the upper-tier (serious players) would most likely be in the rank Diamond and above. We will only be looking at matches that fall in this category for our report. The dataset that we obtained from Kaggle contains detailed information on the first 10 minutes of approximately 10,000 ranked games. According to the user that produced this dataset, he/she went on a live database online (na.op.gg) and randomly picked out a lot of users/players that are Diamond rank or higher. For each of the player profile that the publisher picked out, he/she tried to get the latest 10 ranked games' ids and called the Riot API for each game to get the statistics. Lastly, he/she did some data cleaning (mostly removing all of the observations that are passed the 10 minutes mark) to produce this dataset. By doing this, it is very simple to obtain the dataset since all of the statistics is in the game itself and by grouping the variables by teams, we do not have to deal with information from each specific player, hence making it easier to perform further analysis. But one drawback of this data is the potential of it being outdated since there may exist some games that happened in the previous version of the game.

So our population is all of the League Of Legends game, while the frame population is every game that is covered by the sampling frame, which means that it is all the Diamond or higher rank games. The sampled population is the 9879 Diamond or higher rank games that are in our dataset.

There are a total of 40 variables in the original dataset, the ID of the match, the result of the game, which is also what we are trying to predict is listed under the variable *blueWins*, 1 being the blue team won that

game, 0 otherwise. The rest 38 features are different statistics of each team at the 10 minutes mark of the game, we will only be focused in the features of the blue team since most of the feature between the 2 teams are inversely correlated (an example would be the number of gold difference in each team, if the blue team has a positive 10 gold differences then the red team will have a negative 10 gold difference, just like in Figure 1 below).

Figure 1. Scatterplot between the difference in gold for both the red team



Model

Going over roughly 10,000 games would take too long so we will randomly pick out 500 games from the original dataset to create our sample data. This sample data is then used to construct a *rough* logistic model (using Rstudio) with all of the predictors present, this is **not** going to be our final model because there may exist redundant features that contribute little to nothing in our prediction for the outcome of a game. The only aspect that we will be looking at is the significant level of each variable in predicting our response variable *blueWins*, this is illustrated as the p-value of each variable in the last column below. The reason for this being that if the p-value is large, then that feature will have very little to none effect on our prediction of the result (since larger p-value suggest that the data is consistence with the null hypothesis that the slope for the feature is 0).

Based on the *rough* logistic model, there are only 2 variables that are significant in predicting the result of a game:

blueDragons: The number of dragons that the blue team capture (in the first 10 minutes).

blueGoldDiff: The amount of gold that the blue team has more than the red team (in the first 10 minutes).

So the suggested model based on the above information is roughly:

$$log(\frac{p}{1-p}) = slope_1 * blueDragons + slope_2 * blueGoldDiff$$

where p is the probability that the blue team will win that particular match.

To confirm this model, we carried out backward AIC and BIC elimination. Both of these processes agree with the model above so we will pick it to be our final model, that will be used to predict whether the blue team will win or not based on how they performs in the first 10 minutes of the game.

Result

Based on the model arrived in the Model section above, we can construct a multiple logistic regression model as follows:

$$log(\frac{p}{1-p}) = 1.342 * blueDragons + 0.0005879 * blueGoldDiff$$

where:

blueDragons: The number of dragons that the blue team capture (in the first 10 minutes).

blueGoldDiff: The amount of gold that the blue team has more than the red team (in the first 10 minutes).

p: the probability that the blue team will win that particular match.

In the most simplest form, our final model suggest that both an increase in the amount of dragon captures and the positively bigger gaps in terms of money in the first 10 minutes will, although it is small, boost the blue team winning chance by an appropriate amount. To be really specific, an addition dragon captured will result in a multiplicative increase in the blue team winning chance by a factor of

$$\frac{e^{1.342} + e^{1.342*(blueDragons+1)+0.0005879*blueGoldDiff}}{1 + e^{1.342*(blueDragons+1)+0.0005879*blueGoldDiff}}$$

Similarly, an additional gold gap between the two team will result in a multiplicative increase in the blue team winning chance by a factor of

$$\frac{e^{0.0005879} + e^{1.342*blueDragons+0.0005879*(blueGoldDiff+1)}}{1 + e^{1.342*blueDragons+0.0005879*(blueGoldDiff+1)}}$$

This is based off our multiple logistic regression model which accounted for the number of dragons the blue team capture and the amount of gold that the blue team holds more compared to the red team in the first 10 minutes of the game.

Discussion

Summary

After we obtaining the dataset from Kaggle, we decided to clean the data by removing all of the variables that surround the red team since there a lot of those variables are inversely correlated with the variables for the blue team (an example for this is from Figure 1 above). We then randomly selected 500 Diamond-ranked or higher games to construct a *rough* multiple logistic model to get an idea about which features can be significant for our prediction based on their respective p-values. Next, we confirmed this choice by applying backward AIC and BIC elimination, starting with a *full* model with all of the predictors present

and removing each unnecessary variable each step at a time. Since both of these processes arrived at the same choice of variables, we constructed our multiple logistic model using them to predict the probability that the blue team will win at the end of each game.

Conclusions

Our final model is:

$$log(\frac{p}{1-p}) = 1.342 * blueDragons + 0.0005879 * blueGoldDiff$$

where:

blueDragons: The number of dragons that the blue team capture (in the first 10 minutes).

blueGoldDiff: The amount of gold that the blue team has more than the red team (in the first 10 minutes).

p: the probability that the blue team will win that particular match.

By using backward AIC and BIC elimination, we reached our final model by removing all of the insignificant features. This means that in the first 10 minutes of a game, the two teams should focus on controlling the Dragons while also doing their best to dominate their opponent in term of money since these 2 variables will play significant roles in increasing their winning chance at the later stage of the game.

The sky-rocketing growth of the Esports industry plays a major role in putting unnoticed careers related to games on the spotlight nowadays. Combined with the current global pandemic, the demand for entertainment services also rise greatly, this is evidenced by the rise in the number of viewing time on top video platforms such as Youtube and Twitch. Streamers and gamers are now having a mean of making a living doing what they love (and what they are good at), and the amount of money/viewers that they received can be affected heavily by their performance in the game. Therefore, being aware of the factors that can help you increase the winning chance, even by a little, is an essential skill to have if they want to pursue this career path in the future.

Weakness and Next Steps

There exists many weakness in this analysis, one of which is the existence of unaccounted variables. Since the data is grouped by team, we do not have access to each member's performance in the game, this results in our prediction to be unreliable. For example, the gold difference between the two teams can be the result of one outstanding player defeating all of the minions while the other four members cannot compete as well. It is worth to mention that 10 minutes in a professional game is too short and uneventful since all players are just focusing on making money to upgrade themselves. The game is much more in-depth, money and buffs are just minor factors in this team-oriented game. Another drawback is that both the raw data used and the sample data is obtained through the process of random sampling, which can be very time consuming and costly.

For improvement, we might want to gather more specific data by allowing each player's performance to be recorded. This can help us to perform a more meaningful and precise analysis because gaining accesses to each player can open up many possibilities and interpretations. For instance, instead of grouping the data by team, we now can assemble the observations based on the role each individual fulfil. This will allow us to assess on which front is the team winning and on which they are struggling, these then can be used to make a better prediction.

References

Government of Canada, S. (2020, October 20). Impacts on Mental Health. Retrieved December 18, 2020, from https://www150.statcan.gc.ca/n1/pub/11-631-x/2020004/s3-eng.htm

Statt, N. (2020, December 08). YouTube Gaming had its best year ever with more than 100 billion hours watched. Retrieved December 18, 2020, from https://www.theverge.com/2020/12/8/22163728/youtube-viewers-100-billion-hours-gaming-videos-2020

Reyes, M. (2019, December 18). Esports Ecosystem Report 2020: The key industry players and trends growing the esports market which is on track to surpass $1.5B by 2023. Retrieved December 18, 2020, from https://www.businessinsider.com/esports-ecosystem-market-report

Fanboi, M. (2020, April). League of Legends Diamond Ranked Games (10 min). Retrieved December 18, 2020, from https://www.kaggle.com/bobbyscience. doi:https://www.kaggle.com/bobbyscience/league-of-legends-diamond-ranked-games-10-min

Perez, M. (2019, October 15). Is There Life After 'League Of Legends'? Riot Bets Big On Its First New Game In 10 Years. Retrieved December 18, 2020, from https://www.forbes.com/sites/mattperez/2019/10/15/is-there-life-after-league-of-legends-riot-bets-big-on-its-first-new-game-in-10-years/?sh=7453c5812edc

Jacobs, H. (2015, May 11). Here's the insane training schedule of a 20-something professional gamer. Retrieved December 18, 2020, from https://www.businessinsider.com/pro-gamers-explain-the-insane-training-regimen-they-use-to-stay-on-top-2015-5

RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Thomas Lumley based on Fortran code by Alan Miller (2020). leaps: Regression Subset Selection. R package version 3.1. https://CRAN.R-project.org/package=leaps

Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2020). skimr: Compact and Flexible Summaries of Data. R package version 2.1.2. https://CRAN.R-project.org/package=skimr