E-sports Data Analysis

Ben Li

Dec. 2022

1 Introduction

League of Legends is an online video game where two teams of 5 players fight over territory and objectives. A team consists of five roles, including a "mid laner" who controls the central area of the map, a "top laner" who controls the top-left corner of the map, a "jungler" who roams around the map assisting other players as needed, and so on. This game has gained massive popularity in its lifetime, particularly in its competitive scene. Every year, 24 teams from regions around the world compete in the League of Legends World Championship, with peak viewership soaring as high as 5 million in this year's world championship.

Like in traditional sports, data analysis can be employed in League of Legends matches to determine what players and teams can do to get an edge over their opposition, allowing them to win more games. This is precisely what I will be investigating in this report. More specifically, I hope to determine whether one can predict a professional League of Legends mid laner's gold difference at 15 minutes given their CS per minute, damage per minute, kills and deaths at 15 minutes, and first blood participation.

To briefly explain what each variable means, gold difference measures how much more or less gold an individual has compared to their enemy counterpart. Having more gold allows a player to be stronger, leading to an easier victory. CS per minute is a measure of how many neutral non-player enemies were killed per minute, with each "CS" giving a small amount of gold. Damage per minute is an average of damage dealt to enemy players per minute. Kills and deaths at 15 measure how many times a player killed an enemy, or died themselves. A kill will grant a fixed amount of gold, and dying will remove a player from the game for a brief period of time. Finally, first blood participation tracks whether or not a player participated in the first kill of the match. This first kill grants bonus gold, and is a great way to get an early advantage.

Niche as it may sound, there is a good amount of prior research done on this very topic. A study based on data from the 2018 League of Legends World Championship found that the percentage of destroyed towers and the number of destroyed inhibitors (both of which are structures which provide control on the map) are significantly related to victory, each at p-values of $p = 0.006$ and $p = 0.029$ respectively (Novak et al, 2020). Another study investigated the impact on a team's overall success when a core player dies within the game, making them unable to play for a fixed amount of time (Eaton et al, 2017). A third study fitted a logistic regression model to predict the outcome of a match based on factors including destroyed structures for either team, kills for either team, etc... (Maymin, 2021). The study I'm conducting is unique in that its goal is to model an individual player's performance rather than that of an entire team. As such, the variables of interest are those which impact an individual player (e.g. gold difference, damage per minute) as opposed to those which impact a team as a whole (e.g. destroyed towers and inhibitors).

2 Methods

Data was obtained from https://oracleselixir.com/tools/downloads, under "2022 match data". This dataset contained statistics for professional games played around the world in the 2022 season. The data was then cleaned, selecting observations from matches that took place in the four major regions in the League of Legends circuit, being North America, Europe, China, and South Korea. The data was then divided into two smaller subsets of equal size, creating one training dataset to fit the model, and one validation dataset to evaluate the model's performance later.

An informal EDA was then run using the training dataset, where each predictor was plotted against the response variable, and a histogram of the response variable was plotted to check that the assumptions for linear regression would hold. A model was then fit using R, allowing us to formally verify these assumptions, first by checking that the response is a simple function of a linear combination of each predictor, and by checking that all predictors are pairwise linear. Then, residuals were plotted against fitted values and against each predictor to look for uncorrelated errors, constant variance, and linearity. A QQ-plot was created to verify normality of the response. Square root transformations to correct non-constant variance were applied to the predictors "kills at 15" and "deaths at 15". Lastly, VIFs were computed for each predictor to ensure that multicollinearity wouldn't impact the model.

With assumptions verified, I transitioned to improving the model. A t-test was run on the coefficients of the model to test for significant relationships between each predictor and the response. Based on the results of the t-test, partial f-test was then run to determine if we could remove potentially insignificant predictors from the model. With a few different candidates for potential models, goodness measures such as adjusted R^2, AIC, and BIC were calculated for each model to determine which would perform the best. The model with a balance of a high adjusted R^2, and low AIC, BIC was chosen as the final model. Observations with leverage, standardised residuals, Cook's distance, and DFFITS / DFBETAS outside of their respective cutoffs were identified as problematic.

Lastly, the model that was settled on was fitted using the validation dataset. The two models' characteristics such as their coefficient estimates, how they satisfied the assumptions for linear regression, and their goodness measures to make sure that these properties remained mostly consistent.

3 Results

First, here are some numerical summaries of each predictor:

Predictor	Mean	Standard Deviation	Maximum	Minimum
CSPM	8.742	1.117	13.23	4.858
DPM	526.72	204.16	1411.58	48.11
Kills at 15	0.652	0.929	5	0
Deaths at 15	0.483	0.693	4	0
XP difference at 15	15.69	626.71	1681.00	-2056.00
First blood participation	0.515	0.389	1	0

Moving on to building the model, the results of the t-tests on each predictor classified them all as extremely significant, with the highest p-value among them still being as low as 9.02e-8.

Predictor	Coefficient Estimate	p-value
CSPM	86.06	5.25e-8
DPM	0.41	9.02e-8
Square root of Kills at 15	362.4	Less than 2e-16
Square root of Deaths at 15	-203.8	2.85-e10
XP difference at 15	0.573	Less than 2e-16
First blood participation	2.675	7.65e-10

In spite of this, I still tried to run partial f-tests on the predictors "dpm" and "cspm" as they had the relatively highest p-values among the predictors. Assumptions for the reduced model were verified and no violations were found so I continued with the partial f-test, and unsurprisingly the null hypothesis was rejected at a p-value of 1.02e-14, so I proceeded without removing any predictors from the model.

Moving on to computing goodness measures and model validation, I chose to keep the reduced model around just to see how its AIC, BIC, and adjusted R^2 would compare to the full model. It had both a slightly lower adjusted R^2 value, and slightly higher AIC and BIC values, which gave reassurance that the full model was indeed the one to stick with.

Measurement	Full Model	Reduced Model
Adjusted R^2	0.5591	0.5309
AIC	15585.5	15646.2
BIC	15624.7	15675.73

Lastly, I fit the aforementioned model using the validation dataset. The same issue with non-constant variance in kills and deaths at 15 appeared in this dataset, so the same square root transformation was applied to these variables. Below is a table comparing the key features when fitting the model with the testing and validation datasets:

Characteristic	Model fit with training set	Model fit with testing set
Violated assumptions	None, aside from constant variance in kills and deaths at 15	None, aside from constant variance in kills and deaths at 15
Largest VIF value	1.312	1.368
Intercept	-1.152e-3	-1.008e-3
CSPM	86.06	64.04
DPM	0.41	0.34
Sqrt Kills at 15	362.4	332.8
Sqrt Deaths at 15	-203.8	-180.6
XP difference at 15	0.573	0.596
First blood	267.5	183.9

I felt that the two were in an acceptable range from each other, and thus I chose to consider the model validated. Hence, I settled on the following model as the final model:

$$\begin{aligned}\text{goldDiffAt15} = &-1.152 \times 10^{-3} + (86.06)\,\text{cspm} + (0.41)\,\text{dpm} + (362.4)\,\text{sqrtKillsAt15} \\ &- (203.8)\,\text{sqrtDeathsAt15} + (0.573)\,\text{xpDiffAt15} + (267.5)\,\text{firstBlood}\end{aligned}$$

4 Discussion

To conclude the research question, I found that there does indeed exist a strong linear relationship between a professional mid laner's gold difference at 15 minutes and their cs per minute, damage per minute, kills and deaths at 15 minutes, experience difference at 15 minutes, and first blood participation. For instance, using the model that was fit, we can expect a mid laner's gold difference to increase by around 86.06 for every cs they get per minute. With these in mind, players can get a better idea of what they should be focusing on to improve.

As for limitations of this model, I noticed that the adjusted R^2 was pretty low. Looking into some of the problematic observations, I found that many of them included a player who played a game where they obtained a very high gold difference while not standing out in terms of their stats, with even some cases of players having very negative stats but still accruing a gold lead. I believe the reason for this is that I simply didn't include data about all the things within the game that influences a player's gold for the sake of not having an overly complex model. The inability to account for these other variables will affect the model's ability to properly describe the relationship and account for all the variance. These issues could be corrected by considering more of the variables that affect a player's gold difference and adding them to the model.

5 References

Eaton, J. A., Sangster, M.-D. D., Renaud, M., Mendonca, D. J., & Gray, W. D. (2017). Carrying the Team: The Importance of One Player's Survival for Team Success in League of Legends. Proceedings of the Human Factors and Ergonomics Society, 61(1), 272–276. https://doi.org/10.1177/1541931213601550

LoL Worlds 2022 - Viewership and Detailed Stats. Escharts. (2022). Retrieved December 14, 2022, from https://escharts.com/tournaments/lol/2022-world-championship

Maymin, P. Z. (2021). Smart kills and worthless deaths: eSports analytics for League of Legends. Journal of Quantitative Analysis in Sports, 17(1), 11–27. https://doi.org/10.1515/jqas-2019-0096

Novak, A. R., Bennett, K. J. ., Pluss, M. A., & Fransen, J. (2020). Performance analysis in esports: modelling performance at the 2018 League of Legends World Championship. International Journal of Sports Science & Coaching, 15(5-6), 809–817. https://doi.org/10.1177/1747954120932853