Alexander Tran

League of Legends: Predicting a Professional Team's Gold Difference at 15 Minutes Based on Early Performance

Introduction

The research question that will be explored is how various in-game factors influence a team's "GDat15" in a game of professional League of Legends, defined as the difference between a team's and the opposing team's gold totals by fifteen minutes. This question will be answered by constructing a linear regression model that is not only easy to interpret, but retains strong predictive properties.

League of Legends (LoL) is one of the most popular games in the world. The game has fostered a flourishing professional scene, where pros compete in the eSport for cash prizes and glory. As in traditional sports, predicting which team will win in a game is a popular practice. By predicting GDat15, which has historically shown strong correlation with the probability that a team wins, we can predict which team will win a match (Chouhbi, 2020).

The importance of this research lies in the stakes of being a player in the eSports market. As both the industry and the game continue to develop, the potential gains and losses to be had grow as well. A team that consistently performs well will attract a larger audience, garnering attention from sponsors, and thus increasing the value of the team's brand. Losers on the other hand suffer from operating on deficits, and risk being forced out of the market. Knowing this, a team naturally wants to know how to maximize their chances of winning, which is precisely what we will explore.

As a relatively niche topic, there are few studies related to this sort of topic. A search on several libraries yielded few relevant results. I would like to note that despite these studies opting to use various other regression techniques, the linear regression model that we are constructing is still useful because it only uses variables from before fifteen minutes into a game, allowing us to isolate a portion of the game that is arguably the most important.

Methods

In order to construct a linear regression model, several preliminary steps must be taken. The first step was to clean the collected data, split it into training and testing sets, then perform an exploratory data analysis on the training set. This training set was used to construct the model, and the test set was used to validate it. In the EDA, histograms, boxplots, and scatterplots were created for all potential predictors. These plots were then examined to check for signs of non-normality in the data, non-linearity in the relationships between the response and predictor variables, or bias in the data. Potential problems were noted and investigated formally later on.

The next step was to construct an initial model whose properties could be verified, and which could be refined to a more preferable model. Automated selection methods were avoided due to having several limitations, such as ignoring the context of the study, and having potential problems such as assumption violations and bias. It was decided that the initial model would be manually formulated based on preexisting knowledge from literature and experience. All variables that were relevant were included at the start. T-tests were run on coefficients in this model to check for significant linear relationships with the response variable. Results which were not statistically significant suggested that some predictors should be removed.

Certain conditions first had to be met to utilize residual plots in order to verify assumptions. The first condition was that the conditional mean response was a linear combination of the predictors, and was checked by plotting the response against the fitted values from the initial model, and seeing if the points were randomly scattered around the identity function. The second condition was that the conditional means of each predictor were linearly related to every other predictor. We verify this by plotting pairwise relationships and visually check if the relationship appears weakly linear, with no unusual relationship present. If either of these conditions did not hold, a linear regression model would be reconsidered, or the model would be respecified.

Having verified that conditions held, residuals from the model were plotted against fitted values and each predictor. Model assumptions held if there was no discernible pattern and the residuals were uniformly scattered around zero, as residuals are ideally random. Otherwise present patterns may suggest issues with linearity, error correlation, and constant variance. A normal QQ plot was also used to check normality. If the quantiles of the residuals from the model matched those of the standard normal, the points should follow the line with minimal deviation. Severe deviations suggested that normality was violated.

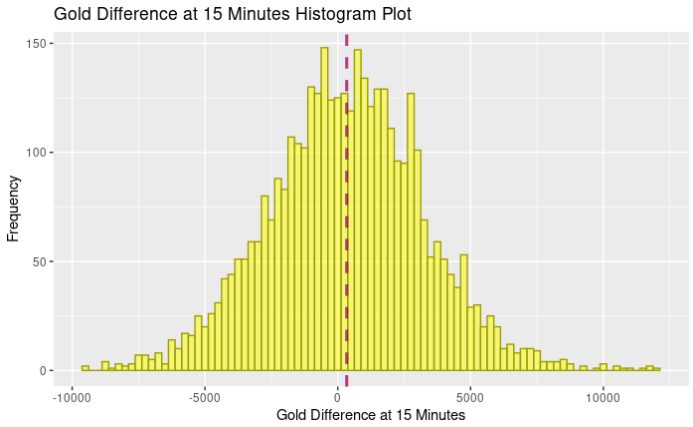
At this point, the model was checked for multicollinearity. Variance inflation factors were calculated for the initial model, alongside its adjusted coefficients of determination, AIC, and BIC. This was also done for other potential models, and their values were compared. Generally, models with low VIF, AIC, and BIC, as well as a higher adjusted coefficient of determination were preferred. Once potential models were narrowed down, ANOVA tables were calculated for each, and the one which explained variation well, while having statistically significant predictors was chosen to move forward with. Assumptions were again verified for this model if not done so previously.

Problematic observations were then identified for the model. Leverage points, outliers, and influential points were detected through measures including the leverage statistic, standardized residual, Cook's distance, DFFITS, and DFBETAS, and their potential impact on the regression model was noted.

Finally, the final model was validated on the test data set. The same model was constructed from the test set, and it was verified whether the regression coefficients were similar, as well as whether they were statistically significant. Furthermore, all model assumptions were verified to see if they held. If these were all true, then the model was validated. Reasons why the model would not have been able to be validated include an abundance of influential observations, data set size, transformations being too specific to the training set, or simply randomness in the splitting of the original data.

Results

The data that was used was gathered by Fernando Rubio Garcia at https://www.kaggle.com/fernandorubiogarcia/2020-league-of-legends-competitive-games, who collected data directly from the API made publicly available by Riot Games, the developer of LoL. The data set contains 5149 observations from professional games in 2020. After cleaning, there remain 17 variables that pertain to this study, each supposedly related to a team's GDat15.



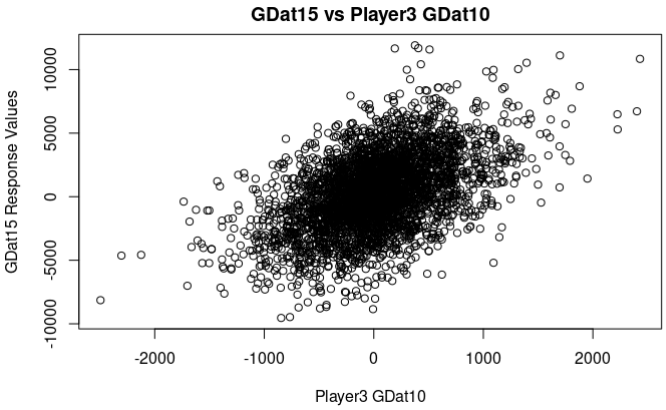
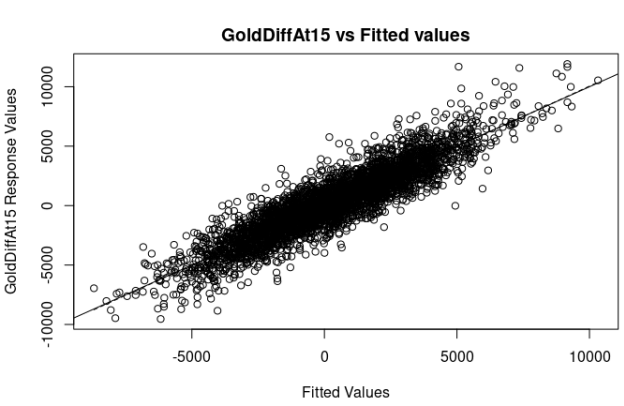


Figure 1: Graphical Summaries of GDat15 and its Relationship with Player3's GDat10

Performing the EDA shows no unusual results. The distributions of each variable appear to resemble normal distributions. Though, as can be seen from *Figure 1*, the mean and median (shown by the dotted lines) of GDat15 suggest some non-normality.

The initial model consists of indicator variables FirstBlood, FirstDragon, FirstHerald, FirstTower, as well as the GDat10, CSDat10, and CSDat15 of each player on the team, with player 1's and player 2's being aggregated into single variables to account for their intertwined nature.

Running T-tests on the model showed that all predictors other than FirstDragon were statistically significant on the 0.001 level. FirstDragon was not removed at this stage for further investigation.



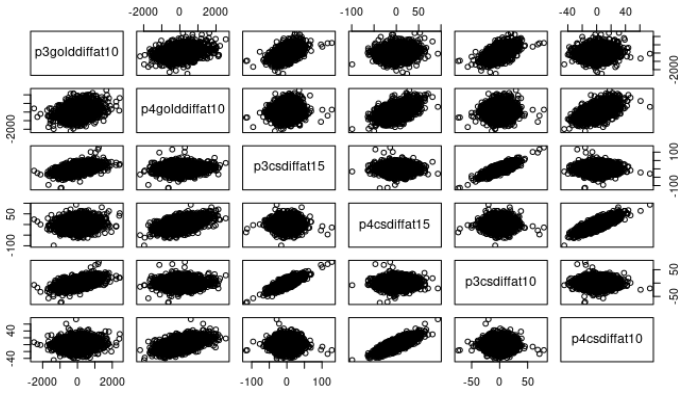


Figure 2: GDat15 Plotted against Fitted Values, Pairwise Scatterplots between Predictors

From *Figure 2* it can be seen that the points scatter randomly around the identity function, and most pairwise plots show weak linear relationships. There is evidence of strong linear relationships between CSDat10 and CSDat15 for a given player, so we considered removing one of these two due to multicollinearity issues. At this point we considered a model without CSDat10 or without CSDat15, and one without FirstDragon and either of the CSDatX predictors.

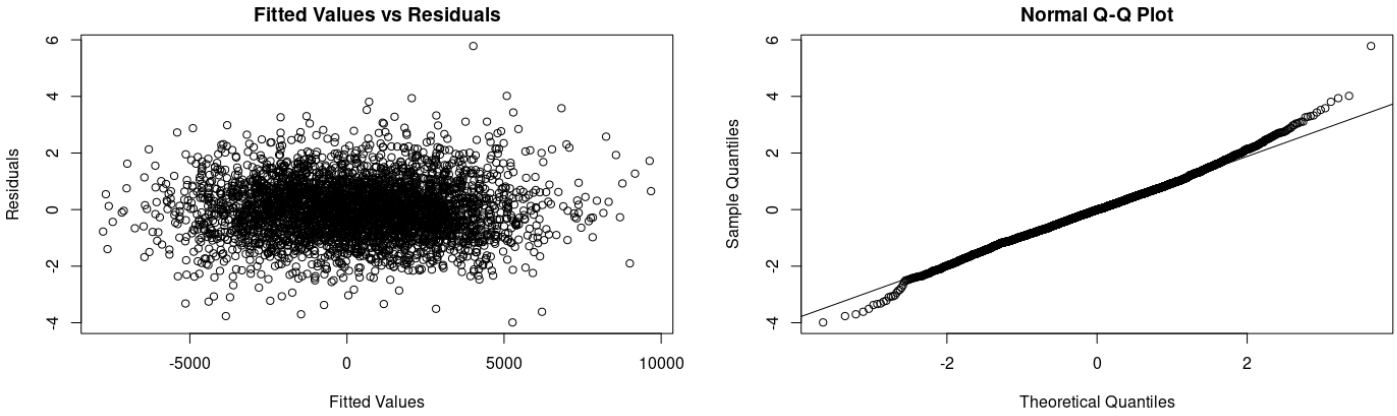


Figure 3: Residual Plots - Residuals vs Fitted Values Scatterplot, Normal Q-Q Plot

Residual plots were created. *Figure 3* demonstrates that there is no discernible pattern between the residuals and the fitted values, with similar results for each predictor. The normal Q-Q plot suggests minor normality violations. A transformation is not necessary, allowing us to maintain interpretability.

	Predictors	Adjusted R^2	AIC	BIC
Without CSDat15	12	0.7505629	66490.49	66577.95
Without CSDat10	12	0.7952655	65736.86	65824.32
Without FirstDragon, Without CSDat10	11	0.7953070	65735.09	65735.09

Table 1: Summary of Multicollinearity Measures for Potential Models

In *Table 1* it can be seen that the model containing CSDat15 performs better as opposed to CSDat10, and the model without FirstDragon performs similarly, but with one less predictor. Calculating VIF values for this model yielded acceptable values, so this model was selected and assumptions were then reassessed.

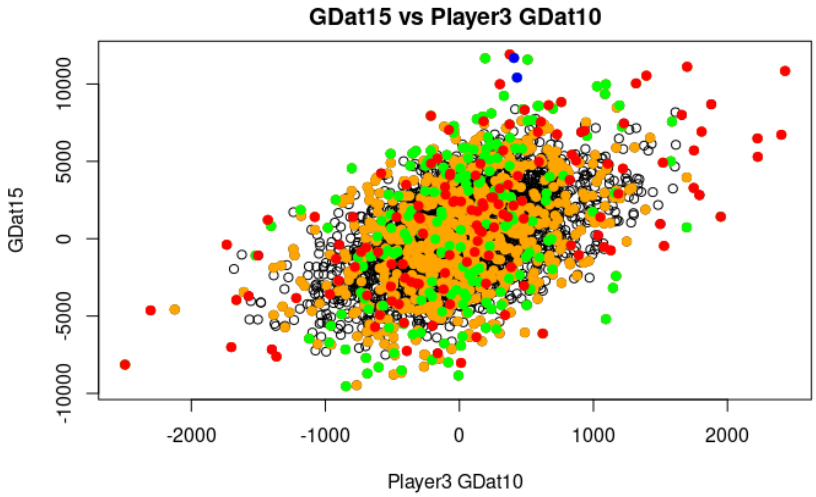


Figure 4: GDat15 Plotted Against Player3 GDat10, Problematic Observations Highlighted

Identifying problematic observations in *Figure 4*, it can be seen that there are quite a few of them, making up over a quarter of observations in the training data set.

Finally, validating this model on the test data set shows that while most coefficients are not too far off, some lie outside of their respective standard errors.

Discussion

The final model, given by *Table 1* in the appendix gives us key information: for a given non-indicator predictor, the coefficient represents change in the response based on a 1-unit change in GDat15 keeping all other variables held constant. This implies that player 3's GDat10 and player 1 and 2's CSDat15 are most impactful in their groups. This means that teams should look to invest in player's 1, 2, and 3 in the early stages of the game to maximize their chances of winning later on.

One limitation to discuss is the inability to fully validate the model. This likely stems from the large number of problematic observations identified previously, which pull the regression line towards them. These observations could not be removed due to ethical reasons, and had to be kept in the data.

Word Count: 1499

References

Chouhbi, K. (2020, February 2). *What is it like to be a data scientist with a passion for gaming ...* Retrieved October 21, 2021, from https://towardsdatascience.com/what-is-like-to-be-a-data-scientist-with-a-passion-for-gaming-43c067ad6415.

Lin, L. (2016). League of Legends Match Outcome Prediction.

Quintana, D. (2019). *RPubs - Predicting Wins in League of Legends*. Retrieved October 21, 2021, from https://rpubs.com/diegolas/LogisticLoL

Appendix

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.039e+03  7.945e+01 -13.082  < 2e-16 ***
firstblood       3.099e+02  8.678e+01   3.571 0.000369 ***
firstherald      5.234e+02  9.011e+01   5.808 7.99e-09 ***
firsttower       1.264e+03  9.614e+01  13.142  < 2e-16 ***
p12golddiffat10  9.546e-01  6.427e-02  14.852  < 2e-16 ***
p3golddiffat10   1.041e+00  9.585e-02  10.863  < 2e-16 ***
p4golddiffat10   1.036e+00  9.248e-02  11.202  < 2e-16 ***
p5golddiffat10   8.117e-01  8.481e-02   9.571  < 2e-16 ***
p12csdiffat15    2.307e+01  2.082e+00  11.080  < 2e-16 ***
p3csdiffat15     1.906e+01  2.276e+00   8.375  < 2e-16 ***
p4csdiffat15     1.784e+01  2.375e+00   7.514 1.08e-13 ***
p5csdiffat15     2.180e+01  2.248e+00   9.699  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1317 on 1261 degrees of freedom
Multiple R-squared:  0.7769,    Adjusted R-squared:  0.7749
F-statistic: 399.2 on 11 and 1261 DF,  p-value: < 2.2e-16
```

Figure 1: Summary of the Final Model