League of Legends 2022: Secret Winning Tips from Top-Tier Players

Group Name: One&Only

Github Link:https://github.com/zoexiao0516/lol2022

Bei Zhang	Cindy Luo	Mengzhe Cai	Yajie Xiao
bz2428@nyu.edu	kl3108@nyu.edu	mc9441@nyu.edu	yx1750@nyu.edu

20 Dec 2022

1 Introduction

League of Legends (LoL), developed by Riot Games, is the most widely played multiplayer online battle arena (MOBA) video game in the world. In 2020, LoL had more than 100 million monthly active players (Active Player, 2022). Due to its entertaining and fast-paced gameplay, LoL is also one of the most popular eSports in the industry, with an international competitive scene composed of 12 regional leagues. Each year, more than 100 top-tier teams in professional leagues compete against each other in worldwide tournaments to win an enormous prize pool reaching several million dollars (Gough, 2022). Therefore, it is crucial for professional teams to identify the determining winning factors and develop appropriate gaming strategies to maximize the probability of victory.

LoL's gaming mechanics are deliberately designed to have a rigorous evaluation system while preserving unpredictability by inducing complex features across different game stages. In the game, two teams (red or blue) of five players battle in player-versus-player combat to occupy or defend their team's part of the map. Each player controls a character, known as a "champion," with unique abilities and different play styles. During the match, the champions collect experience points, upgrade skill levels, earn gold, and purchase items from the store to become more powerful. The goal is to destroy the other team's base, also called "Neux." Each base contains a set of resources, including a series of turrets on three lanes, waves of minions that constantly spawn from the inhibitors, monsters within the Jungle, and heralds and drakes. Experience points and gold could be earned from these resources, as well as from killing other champions of the opponent team.

With the hope of learning from LoL's top-tier players about their keys to winning, this study aims to identify critical winning factors in professional matches of the game, using the League of Legends eSports match data in 2022. This report consists of three parts. In part one, we tested whether the team's pre-game factor, being on the red/blue team, is independent of winning the match. In the second part, we looked at the contributing factors to predict the team's total gold, one of the commonly recognized winning factors. In the last part, given the gaming statistics, we developed machine-learning models to identify a winning/losing team.

2 Data describtion

2.1 Original dataset

The original dataset was aggregated and released by Tim Sevenhuysen of OraclesElixir.com[1]. As the dataset is updated daily, the dataset used in this report was obtained on Dec 1, 2022. It included player- and team-level 2022 match data from all leagues, of matches between Jan 1, 2022, and Dec 1, 2022. There were 148140 rows and 123 columns containing information from 12332 games. Each game had ten rows of data: five rows of player-level data from the red team, five rows of player-level data from the blue team, and one row of team-level data from each team. A detailed data dictionary with data type and data description labeled can be found on the GitHub page.

2.2 Data Preprocessing

For the purpose of our analysis, we decided to focus on the team-level data, which are the rows marked as 'team' under the "position" column. After checking missing values for all columns, we discovered that more than 1/4 of the variables are missing for all the teams in the following leagues: LDL, LPL, and WCS (partially missing). By checking, the variable "Data Completeness" in the original data set was an indicator of this missing feature. Therefore, we only selected rows labeled "complete" under the variable "Data Completeness".

We then selected 95 of the 123 variables in the original dataset, excluding some unnecessary text data for the analysis. Again, we checked missing values for the new dataset obtained. Three additional variables, "dragons (type unknown)", "monsterkillsownjungle", "monsterkillsenemyjungle", were removed due to large missing values. 92 rows were dropped due to missing entries in 'elemental drakes' and "firstmidtower". Since there were two rows for one match in the dataset, all rows removed were double-checked that they belonged to the same sets of matches. The final dataset has 21114 rows and 93 columns, containing team-level information from 10557 matches. Each row represents a team's performance in a match, and each column represents the specific performance under each game metric. The correlation matrix of all the numeric features is below.

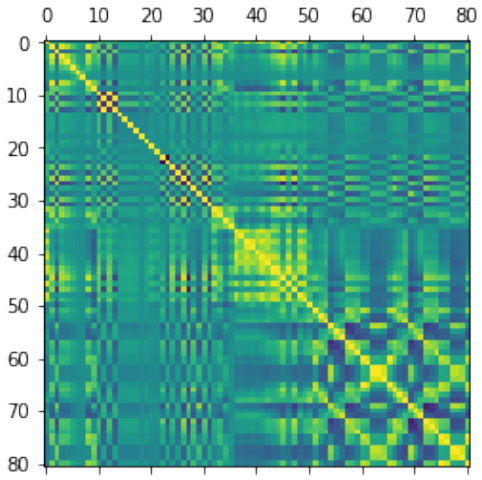

Figure 1: the correlation matrix of all the numeric features

[1]https://oracleselixir.com/tools/downloads

3 Inference question

3.1 Question

In a League of Legends game, two teams pick their sides as blue and red before all players select their champions. There is a slight difference between the two sides. For example, during the phase of picking and banning champions, the blue team can pick a top-tier meta champion first; correspondingly, the red team can pick the champions to counter the champions selected by the blue team. Moreover, due to the layout of the game map, the players on the blue side have easier access to the dragons, while red players usually get more heralds and barons.

In the first part, we aimed to explore whether side selection significantly influences the game result.

3.2 Approach

The features used in this question are side ("red" or "blue"), result ("0" =lose,"1" =win), and patch (the version of the game).

The null hypothesis, H_0, is: side is independent of the result; the alternative hypothesis, H_1, is: side is not independent of the result.

Since the blue team wins in one game means that the red team loses, we randomly picked one record for each game. Then we did a Chi-square test to check the null hypothesis. In addition, we analyzed each patch and saw the trends for this year. Because of the large sample size of our data, statistical power should not be an issue in our analysis.

3.3 Analysis

First, we did the global analysis for all data this year. There are 10557 games in total, and the win rate for the blue side is 52.28%. It seems that the blue team has an edge this year.

Side Result	0	1	All
Blue	2561	2710	5271
Red	2809	2477	5286
All	5370	5187	10557

Table 1: the cross-tabulation for all match data

After randomly picking one record for each game, we got a simple cross-tabulation of sides and results (Table 1). The p-value for this Chi-square test is 3.10e-06, which is smaller than 0.05. So, we have significant evidence to reject H_0. That's to say; side selection does influence the game result.

Then we repeated the steps for each patch. The number of games, the win rate of the blue side, and the p-values for the Chi-square test are shown in Table 2.

patch	Number	BlueWinRate	p-value
12.01	697	49.641320	0.925711
12.02	719	50.486787	0.873517
12.03	870	50.689655	0.734817
12.04	897	54.515050	0.006831
12.05	1099	47.224750	0.076466
12.06	187	56.684492	0.106289
12.07	81	53.086420	0.671599
12.08	181	50.276243	1.000000
12.09	352	55.113636	0.066350
12.10	611	55.482815	0.008722
12.11	768	53.645833	0.051082

patch	Number	BlueWinRate	p-value
12.12	1166	51.715266	0.267029
12.13	738	53.523035	0.072524
12.14	494	56.680162	0.003940
12.15	395	49.113924	0.819737
12.16	285	57.894737	0.018227
12.17	24	50.000000	1.000000
12.18	519	49.710983	0.965323
12.19	196	59.693878	0.010892
12.20	183	52.459016	0.601859
12.21	95	58.947368	0.127099

Table 2: Win rate in blue side and p-value of the Chi-square for each patch

Table 2 shows that in most patches, the blue-team-win rate is higher than 50%. The red team has a higher win rate in only 4 patches out of 21. Moreover, 5 patches significantly favor the blue side, respectively 12.04, 12.10, 12.14, 12.16, and 12.19.

Picture 2 gives a more intuitive insight into the side advantage across patches for this year. The color of the bar indicates the side with a higher win rate, and the color intensity represents the p-value—the deeper the color is, the more significant evidence we have to reject H_0. For example, from patch 12.02 to 12.04, the blue side always has an edge, so designers adjusted some champion properties and settings, which led to a red-favor patch—12.05. When serious inequity arises, we can see some adjustments made to reduce the unfairness. So, there are no consecutive patches that are significantly blue-favor or red-favor. However, in general, side selection does have a significant influence on game results, and the blue side has a higher chance of winning.

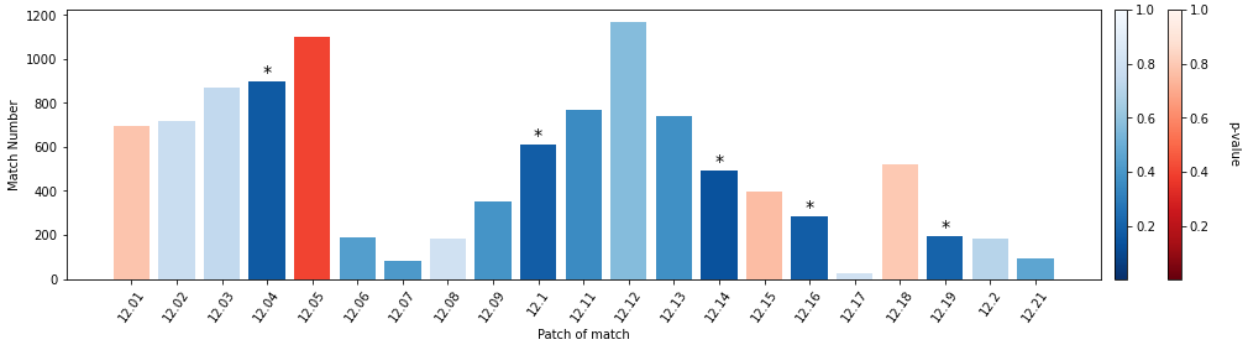

Figure 2: Change trend of all patches in 2022

3.4 Summarized Results

In a balanced system, the side of the team should be independent of the game result. However, we could tell from the percentages that blue teams have a higher winning rate than red teams. Taking a closer look at the proportions across different patches, we discovered adjustments made by the developers over time.

4 Prediction question

4.1 Question

Total gold is one of the commonly recognized winning factors. The more gold a champion earns, the more advanced items they can buy from the store to strengthen themselves. So, we defined the research question for prediction as follows: what factors contribute significantly to each team's total gold?

4.2 Approach

To answer this question, we selected a continuous numeric variable called "totalgold" as our target variable. As for predictors, since there was a significant difference in scale between our continuous features, which might impact the prediction performance, we first performed standardization on the data of continuous features. Then, we combined the data of binary features with the standardized data of continuous features. Both the binary and continuous features would be predictors in our prediction models. Also, we split the dataset into training and testing data by 4/1. We ran regression models for the dimension-reduced and originally high-dimension datasets to select the most important features in predicting total gold and compared their performance on the test dataset.

4.3 Analysis

4.3.1 Regression for dimension-reduced dataset

Given that our dataset had a large number of features and many of them are correlated, we conducted PCA to reduce the dimensionality of the continuous data. As for choosing the optimal number of principal components, we eyeballed the scree plot(Picture 3) and found that, at point 4th principal component, the proportion of variance explained by each subsequent principal component dropped significantly. So, we decided to choose 4 principal components, which explained around 53.15% of the variance in the data.

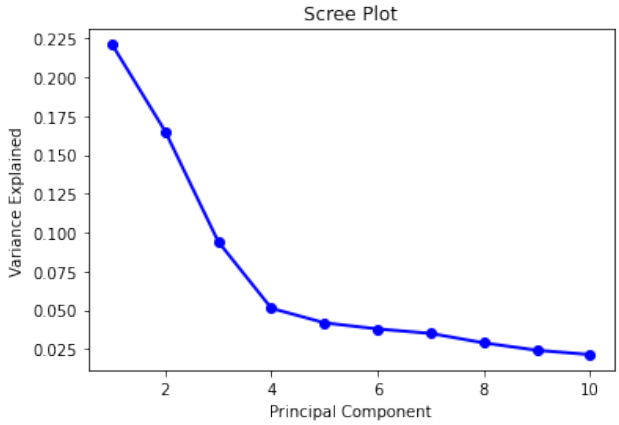

Figure 3: Scree Plot for PCA model

To learn more about the composition of the principal components, we analyzed the PCA loadings. Component 1 is related to overall resources gathered by the team, like towers and dragons; Component 2 can be explained mainly by inter-game killing data; Component 3 contains variables relevant to game length; Component 4 is about the opponent's performance during the game.

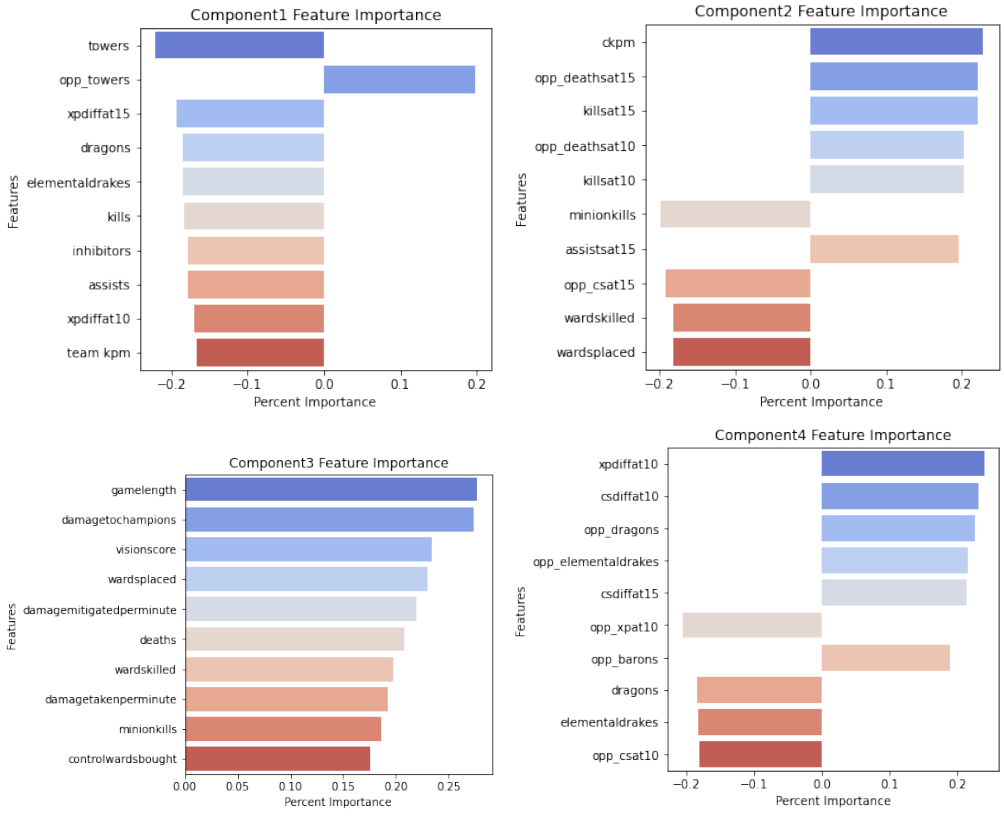

Figure 4: Feature Importance for four Components

Then, we reduced our continuous features to 4 dimensions. Plotting the first three principal components in an interactive 3D scatterplot, these data points are clustered together, indicating no apparent clusters among them.

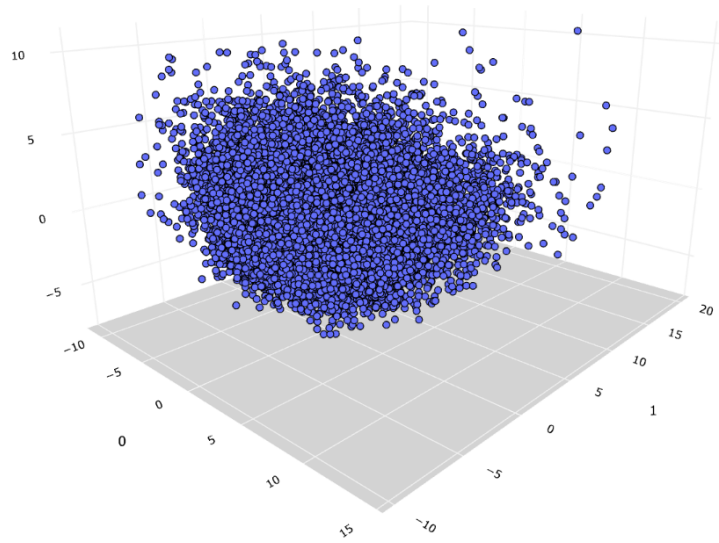

Figure 5: 3D plot for the first three components

Linear Regression

To better explore how these features predict "totalgold", we used dimension-reduced continuous features and binary features to build a "totalgold" prediction model. The Linear regression model achieves an R2 score of 0.908, and the RMSE of the model is 10780. From the models' coefficients, Comp3 has the highest coefficient, indicating game's overall data is important in predicting "totalgold". More specifically, "gamelength" and "damagetochampions" are essential variables in the prediction.

Variable	Coefficient
Comp3	2776.02
Comp1	-1769.1
firstbaron	1439.4
Comp2	-1073.2
firstdragon	-987.99
firsttower	197.25
firsttothreetowers	159.44
firstmidtower	-123.41
firstherald	123.19
firstblood	58.43
Comp4	-5.56

Table 3: Top 10 importance variables for linear regression model

Lasso Regression

We punished the model by using Lasso and found the best Lasso parameter is alpha = 300 (the best value to achieve feature selection), with R2 = 0.905 and RMSE = 3471. Under this alpha, only component features were kept in the model, indicating Comp1, Comp2 and Comp3 are the three most essential features for "totalgold" prediction. The R2 performance and coefficients of the Lasso Regression are similar to those of the Linear Regression model, but Lasso Regression has a smaller RMSE. Comp3 still has the highest coefficient, showing its importance in predicting "totalgold".

$$y = -1857.20Comp1 - 1033.12Comp2 + 2716.84Comp3 + intercept$$

4.3.2 Lasso without PCA

Then, we implemented regularized regression with LASSO on standardized, non-reduce data to find the most important determining factors of the contribution to total gold without principal component analysis.

We found an appropriate Lasso parameter is alpha = 300, with a test R2 = 0.985 and an RMSE = 1395.36. After several hyperparameters tuning using GridSearch, we established that we would need to set a high alpha level to achieve feature selection. The rationale was that our target variable is total gold, which has a high variance across matches and a relatively wide range of values as compared to the standardized predictors. In addition, the high R2 score suggested that our model could capture the variance in the total gold values very well and that penalizing the coefficients would not result in much loss in R squared, so our focus here should be finding the fewer parameters to capture this variance as defined in the research question, that is, dimension reduction and feature selection rather than finding the best fit model.

The alpha level (300) ensures a high R squared value while implementing a feature selection resulting in the 11 most significant factors. The 11 most significant factors are shown below, with their corresponding coefficients (sorted in descending order):

Variable	Coefficient
gamelength	5128.926063
kills	2621.127989
minionkills	2001.172417
towers	1870.846789
monsterkills	1188.599623
barons	786.0405668
damagetochampions	592.6211873
visionscore	210.9220621
elders	84.36344984
assists	63.86484866
turretplates	27.63784349

Table 4: Top 11 importance variables for lasso model

We can observe that game length is the dominating factor for a team's total gold. The longer the game lasts, the more gold a team cumulates. Moreover, different types of kills, including monster kills and minion kills, are also important actions in the game to gain more gold. One thing that is particular to LoL is that taking over the tower is also a method to collect gold. Elemental drakes and heralds are not presented as important factors here, suggesting that they might not bring much gold earning to the team.

As a next step, we plotted the correlation heatmap between the 11 most significant factors to evaluate the correlation among these significant factors better:

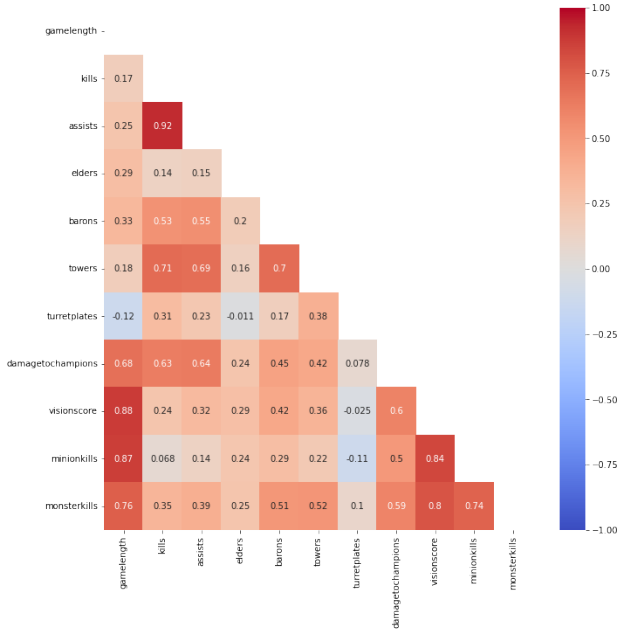

Figure 6: the correlation heatmap between the 11 most significant factors

We can observe that game length significantly correlates with vision score and minion kills, and vision score is also highly correlated with minion kills.

4.4 Summarized Result

The predictor "game length" is the variable with the highest coefficient value in both Lasso regression without PCA and principal component analysis (component 3). In addition, "damage to champions", "vision score", and "minion kills" are selected by PCA (component 3), "Assist", "kills", and "towers" are selected by PCA (component 1), as they are also selected by Lasso regression, indicating they are the significant predictors. In Lasso regression, the binary predictors are non-significant in predicting the total gold, while in regression of principal components predictors combined with binary variables, the binary variable is also not as important as if we set the alpha value to be high. All these regression models showed "game length" is the most crucial variable in predicting "total gold"; it makes sense because the longer the game played, the higher the killing number and damage to champions, etc., contributing to higher "total gold", which is the determinant factor of "total gold".

5 Classification question

5.1 Question

Given the competitive nature of LoL, win or lose is always the most essential part of the game. In this section, we try to predict the result of a game based on gaming statistics. Since there are only two possible results-win or lose, it is a 0-1 classification problem.

5.2 Approach

The features used in this question are: result ("0" =lose, "1" =win) as the dependent variable, and all other features except text features as the independent variable. There are 88 predictor variables in total in this question.

Notice that some variables are data of the opponent team, like "opp_dragons", "opp_glodat10", which means two records from the same game strongly correlate with each other. So, like the inference question, we also randomly picked one record for each game. After that, data was split into a training set (80% of all the data) and a testing set (20% of all the data).

5.3 Analysis

5.3.1 Logistic Regression Model

Using 88 gaming metrics as predictors, we ran a logistic regression model with built-in cross-validation on the non-standardized training data. The default cross-validation generator used by the function is Stratified K-Folds, and the default value was 5-fold. This cross-validation process (which happened while training the model using the training set) allowed us to perform regularization on the logistic model and find the inverse of the regularization parameter value that map to the best score of the model.

The logistic model predicted the game outcome with an accuracy of 97.59%, and its AUROC value was 0.9965. The ROC curve and the confusion matrix are shown below.

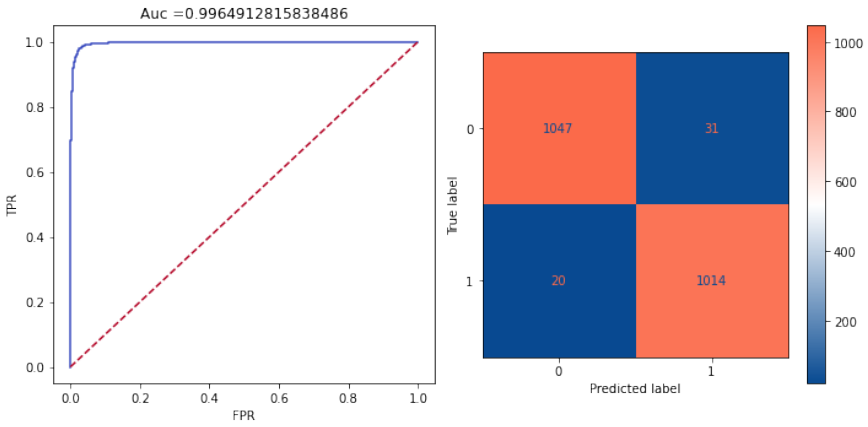

Figure 7: ROC curve and confusion matrix for logistic regression model

Table included the top 10 most important factors related to the game result, sorted based on the absolute values of the weights. By looking at the weights, we could tell that "earned gold" and "earned gold per minute" played a vital role in the game outcome: the more gold earned per game and per minute, the higher the chance of winning the game. In addition, "deaths" and "assists" were significant predictors. "Deaths" was negatively correlated with winning, suggesting that having more deaths on the team reduced the probability of winning. On the other hand, more assists could lead to a higher chance of winning. The result also emphasized the importance of "vision" in LoL, with a higher vision score correlated with an increased winning rate.

Predictor	Weight
earned gpm	0.005462
minionkills	-0.005420
dpm	0.004086
damagetakenperminute	-0.002261
deaths	-0.001703
visionscore	0.001331
earnedgold	0.001177
wardsplaced	0.001002
opp_towers	-0.000969
assists	0.000964

Table 5: Top 10 importance variables for logistic regression model

One interesting finding to note down here is that while average damage to champions per minute ("dpm") was positively correlated to winning, damage taken per minute was negatively related to winning. This difference in the directions of two similar gaming metrics indicated that damage made to champions was a more effective type of damage in winning the game.

5.3.2 Random Forest Model

Considering the number of dependent variables in this question, we tried the Random Forest model and used random search to do hyperparameter tuning for the number of estimators, the max features, the max depth of the tree, the minimal leaf number, and whether to use bootstrap. Standardized data was used for the model to compare the importance of the features. After doing hyperparameter tuning with random search, the accuracy of the random forest model reaches 98.15%, and the area under the ROC curve is 0.9991, quite close to 1. The ROC curve and the confusion matrix are shown as follows.

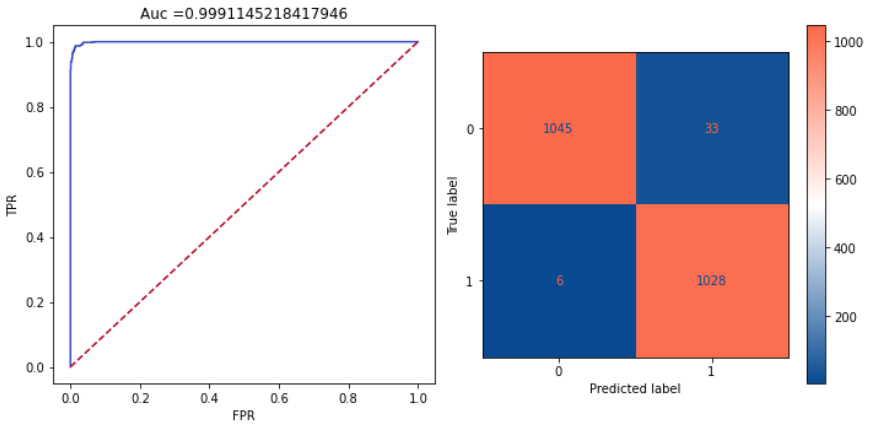

Figure 8: ROC curve and confusion matrix for random forest model

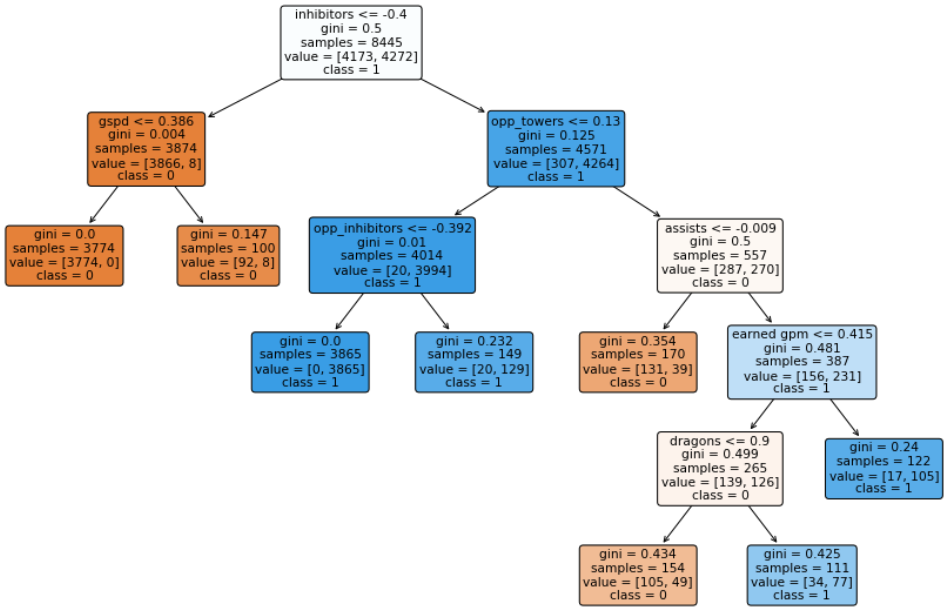

Figure 9: Partial of random forest model

Table 6 lists the top 10 critical variables in the random forest model. “towers”, “opp_towers”, “inhibitors”, “opp_inhibitors” occupy the top 4. It is quite reasonable since destroying the opponent’s towers and inhibitors are necessary conditions to win. Besides those variables, “earned_gpm” and “totalgold” are also important for winning, which means gold plays a significant role in the LoL game. “Death” (which can also be seen as kills of the opponent team) and “assists” are also on the list. Therefore, we may infer that killing with teammates’ assistants may be better than a solo kill. Regarding the resources taken by the team, “dragons” and “opp_elders” seem more important than other types of resources.

Predictor	Weight
towers	0.389519
inhibitors	0.323722
opp_towers	0.160344
opp_inhibitors	0.098530
earned gpm	0.019952
deaths	0.002486
dragons	0.001269
opp_elementaldrakes	0.001167
assists	0.001114
team kpm	0.000385

Table 6: Top 10 importance variables for random forest model

5.4 Summarized Result

Our analysis in part three confirmed our expectation in part two that gold is a vital gaming metric impacting the game result. Additionally, even though the number of team kills is a crucial factor in total gold prediction, it does not weigh the same importance in winning the game. In contrast, team total deaths became more important in determining the game result. In other words, it might not matter how many kills a team made throughout the game, but to win the game, one must have fewer deaths. Moreover, by ranking the importance of variables in both models, heralds and barons are not as crucial as dragons and elders in winning the game, which could be tied back to our first conclusion about the advantage of selecting the blue side. The reason is that dragons and elders are near the blue side of the jungle, whereas heralds and barons are near the red part.

6 Conclusion

In this report, we used three common data science approaches, inference, prediction, and classification, to answer questions related to the LoL matches.

For the first question about pre-game side selection, by looking at the total winning and losing matches of the red and blue teams, we found that the overall game design was not balanced for the side of the team. Since side selections are not always random in tournaments, selecting the advantageous side might be the first step to winning the whole game. Future questions could be asked to investigate the reason underlying such an advantage in side selection by looking at specific champion selections and bans on each side across patches. One should also notice the self-selection bias in this dataset. If the team won from being on the blue side in its previous matches, it is more likely to choose the blue side again in future games. Thus, we could not interpret the dependence as a causal relationship. An ideal dataset for this analysis would be collected by randomizing side selection for all the matches. Using this dataset, time series analysis focusing on the red/blue side selection of specific teams could also be another interesting direction to explore.

For the second question, based on the results of lasso regressions on both dimension-reduced data after PCA and the original standardized data, the length of the game, damage to champions, towers, vision score, different types of kills, and assists are the contributing factors in predicting the total gold of the team. One limitation of our PCA analysis is that the standardization performed before conducting PCA might amplify variables with low columns. In addition, the variable loadings on the components are not distinct enough for us to clearly interpret each component. Even though the components are not correlated, the collinearity between variables still exists when analyzing the loadings of each component, making it hard to perform efficient feature selection based on the loadings.

For the last question, the logistic regression and random forest model results show that assists, deaths, earned gold, and total gold are critical winning factors. One limitation of this analysis is that we used outcome features to classify the outcome. This issue became more prominent in our random forest model since the

number of towers and inhibitors are part of the embedded rules in evaluating victory. Thus, they would be and should be important winning factors. More than 2/3 of the variables are information collected after the game ends. If we only use the inter-game variables (e.g., gold at 10 min), the accuracy of the logistic regression model would be 74.86%.

All of our models yield great fits of the test data and high accuracy. One of the main reasons behind this is that this project aimed to capture a relatively simple gaming mechanism using a comprehensive dataset. Unlike models that describe real-life events, our model focused on exploring the mechanism in the game of League of Legends, which composes of fewer uncertainties and has an implicit evaluation system designed by the developers. Given that our dataset covered nearly all gaming metrics, it would not be surprising to have great model performance if all variables were added to the model. Aware of this, this report aimed to identify the most important factors rather than develop a best-fit model.

Our results offer insights into pre-game side selection, ways to maximize total gold in the game, and determinant factors of winning. By looking at how professional eSport teams obtained their victory, casual players could learn about strategic decision-making while playing League of Legends. Remember, the next time you are in a ranked LoL game, pick the blue side, die less, assist more, and take more dragons rather than heralds!

Extra Credict

Using a two-sided Welch t-test to compare the number of kills from matches within the LCK leagues to those within the LCS league, we did not detect any significant difference between the two means (p-value: 0.2).

Reference

[1] Active Player. (2022, Dec 20). League of Legends Live Player Count and Statistics. https://activeplayer.io/league-of-legends/

[2] Gough, C. (2022, Nov 15). League of Legends World Championships prize pool from 2012 to 2022. https://www.statista.com/statistics/749024/league-of-legends-championships-prize-pool/