Predicting the Outcomes of Professional League of Legends Matches Based on Early and Mid-Game Performance Metrics

Background & Introduction

Over the last decade, League of Legends has risen to immense popularity, consistently topping charts as the most played online game globally. This surge in player engagement has significantly expanded the ecosystem of professional League of Legends competitions, now boasting over 11 major leagues and a multitude of secondary leagues worldwide. The evolution of professional play, coupled with the escalating competitive intensity, has enhanced the game's appeal and engagement levels.

As a 5V5 MOBA game, League of Legends assigns 10 players to either the red or blue side, with each player occupying a distinct position and fulfilling specific team functions. In professional matches, a team's victory is intricately linked to its collective performance, encompassing strategy, coordination, and execution. Recognizing this, our focus is on harnessing data that reflects team-level performance in each game to develop a predictive model for victory or defeat.

Central to competitive League of Legends are the early and mid-game phases, which are pivotal in determining a team's likelihood of victory. These stages, marked by strategic movements, resource accumulation, and critical skirmishes, significantly influence the game's momentum. Metrics such as gold difference, kills, deaths, and objectives secured during these phases offer deep insights into a team's performance trajectory. Early game actions set the match's pace and control, impacting resource distribution and map dominance, while mid-game performance either cements these early advantages or highlights a team's capacity for recovery and adaptation. Analyzing these early and mid-game metrics, therefore, provides a comprehensive snapshot of a team's strengths, weaknesses, and strategies, making them reliable indicators for predicting match outcomes in the high-stakes arena of professional League of Legends.

Predictive task & research question

Predicting the Outcomes of Professional League of Legends Matches Based on Early and Mid-Game Performance Metrics

Dataset Description

Our dataset comes from the Internet, a website called *Oracle's Elixir*, which is a website dedicated to collecting data on every professional game. The data we used this time is the game data of the League of Legends professional competition throughout 2022. There are two reasons for using data in 2022: On the one hand, it is to ensure timeliness. League of Legends' game modes and game updates over the past 10 years have made a lot of difference between past game play and recent game play. Therefore, using data from 2022 can ensure that our model has better interpretability when facing new data. On the other hand, all competitions in 2023 have not yet completely ended, which means that if we use the data in 2023, we may miss some data at the end of the year, making our data set potentially biased.

Oracle's Elixir websites link: https://oracleselixir.com/about

About Dataframe, the number of columns has reached 123. Each row in the Dataframe represents the data of a game with the player as the unit. The period includes the position of the player in a certain game, the selected champion, and the data during the game. The data of the game includes the player's damage to the enemy per minute, the gold obtained per minute, and KDA (Kill/Death/Assist rate), etc. In addition, it also includes the player's name, the player's location club, and the league the club is in.

There are a total of 149232 rows in the Dataframe. This data represents the sum of the number of games played by all players in each league in 2022. In addition, the distribution of the dataframe is characterized by a group of 12 rows from top to bottom in order. The first 10 rows in each group are the game data of ten players, while the 11th and 12th rows are the overall data of the two teams.

More details about the columns name in the dataframe:
https://oracleselixir.com/definitions

Data Visualization & Exploration

Because this dataset isn't structured ideally, as both individual and team data are listed as entries, we should speculate the trends among them separately. We have a total of 24900 team results and a total of 124500 player results. However, most of the player data is being aggregated and represented through their team data, so there isn't much need to look into the variation among specific players. Instead, looking into the team data could provide much more holistic insights into each entire matchup.

First and foremost, in each League of Legends match, the two opposing teams are assigned to distinct sides: the red side and the blue side, each with its unique terrain and strategic implications. This difference in terrain is not merely cosmetic but can potentially influence the outcome of the game, tipping the scales towards victory or defeat. Intriguingly, our analysis across various leagues has revealed a subtle yet consistent trend: teams on the blue side tend to have a marginally higher win rate compared to those on the red side. This phenomenon is depicted in Figure 1, which presents a detailed comparison of the win rates for teams on both sides across different leagues. It is noteworthy that only 9 out of the 49 leagues analyzed exhibit a scenario where the red side holds an advantage over the blue. In stark contrast, the majority of leagues demonstrate a preference towards the blue side. When averaged across all leagues, the data points to a distinct pattern - the win rate for the red side stands at 47.5%, which is approximately 5 percent lower than that of their blue counterparts. This statistic not only highlights the subtle asymmetries in gameplay dynamics based on side allocation but also underscores the importance of side selection in strategic planning and its potential impact on the game's final 'result'.

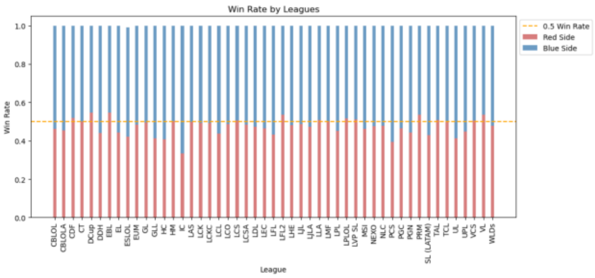

Fig.1

Secondly, our analysis delves into the crucial early and mid-game indicators that are pivotal in securing map resources, such as achieving first blood, destroying the first tower, capturing the first dragon, and being the first to demolish three towers. These elements are not just milestones in the game; they potentially wield significant influence over the final outcome. To illustrate this point, Figure 2 presents a side-by-side bar chart comparison, vividly demonstrating the correlation between securing these early advantages and the subsequent rate of victories and defeats. The data paints a compelling picture: teams that secure first blood, for instance, show a markedly higher probability of winning the match, with a win rate of 61%. This trend is similarly observed in other indicators. Teams that destroy the first tower and those that capture the first dragon have win rates of 70% and 58%, respectively. The statistical significance of these findings cannot be overstated. They unequivocally suggest that gaining an early upper hand in these key areas substantially increases a team's chances of emerging victorious. As a result, incorporating these four factors into our feature selection emerges as a logical and crucial step in our predictive modeling process.

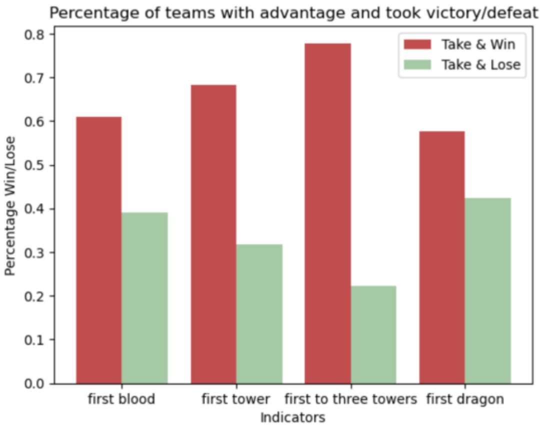

Fig.2

Thus far in our analysis, we have primarily focused on binary variables to understand their influence on the outcomes of League of Legends matches. However, to further enrich our predictive model, it's essential to explore the role of numerical variables derived from early and mid-game stages. Figures 3 and 4 skillfully utilize violin plots to showcase the distribution patterns of two such key numerical variables: the gold difference and the kills difference during the early to mid-game phase, segmented according to the distinction between winning and losing teams. The gold difference, directly sourced from the 'golddiffat15' column, reflects the economic disparity between teams at the 15-minute mark, a critical juncture in the game's progression. Meanwhile, the kills difference is a calculated metric, derived by subtracting the number of deaths at 15 minutes ('deathsat15') from the number of kills at the same timeframe ('killsat15'), thus encapsulating the combat dynamics of the teams during this pivotal period.

Upon a detailed examination of these figures, a pattern emerges: the median values for both the gold and kills differences are slightly higher for the teams that secure victory. This observation extends beyond just the median figures; the larger distribution of these statistics for the winning teams predominantly surpasses that of the losing teams. Such a pronounced divergence in the distribution patterns underscores the critical role these early and mid-game numerical variables play in determining the trajectory of the game. Hence, their inclusion as features in our predictive model is not only strategic but also imperative, enhancing the model's capability to accurately predict the outcome of a match based on these key early and mid-game performance indicators."

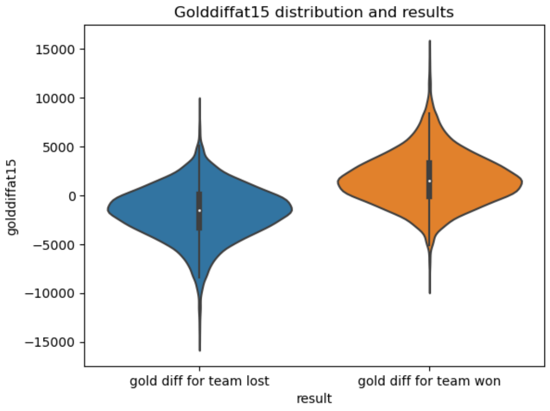

Fig.3

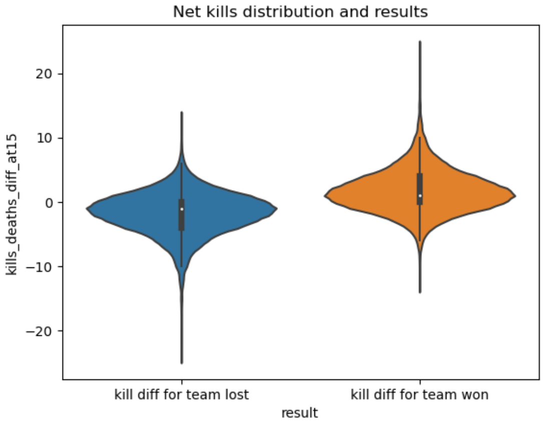

Fig.4

Moreover, to further explore the dynamics of League of Legends matches, we investigated the relationship between the gold difference at 15 minutes and the kills difference at the same stage. This analysis aimed to understand their correlation and covariance and how these two metrics collectively influence the outcomes of matches. To visualize this relationship, Figure 5 presents a scatter plot with the kills-death difference plotted on the x-axis and the gold difference on the y-axis. Each dot in the plot, colored in shades of green and red, represents an individual match outcome, signifying victory and defeat, respectively. This scatter plot not only reveals that a larger kill-death difference often correlates with a higher gold difference but also suggests that the victorious teams are typically characterized by a substantial positive divergence in both gold and kills-death metrics. The observed high covariance between these variables indicates that they tend to mirror each other's trends in our predictive model, without raising concerns of multicollinearity, given the variance present in the dataset. Therefore, it is logical and justifiable to incorporate both gold difference and kills-death difference as predictive features in our model to forecast the outcomes of League of Legends matches.

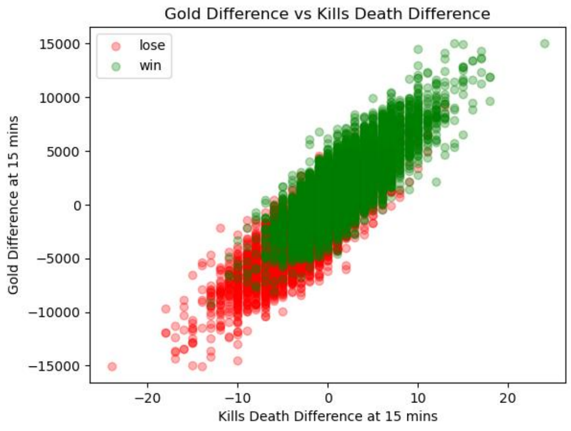

Fig.5

Variables(features) Chosen

The predictive task we explore is: *Predicting the Outcomes of Professional League of Legends Matches Based on Early and Mid-Game Performance Metrics.* Based on our comprehensive exploration of these stages of the game, we have identified a set of variables that are instrumental in shaping the course of a match. These variables, chosen for their significance and impact on the game's dynamics, include: 'side', 'firstblood', 'firsttower', 'firsttothreetowers', 'firstdragon', as well as quantifiable metrics like 'killsat15', 'deathsat15', and 'golddiffat15'. Each of these elements plays a pivotal role in determining the trajectory of a match, making them invaluable in constructing a robust predictive model.

Data Cleaning (Missing Value Assessment)

When we extracted all the data about the team in the dataset, we found that there were some missing values in the data. Table. 1 records the number of times each feature is missing.

Features	# Entries Missing
firstblood	2
firstdragon	3638
killsat15	3638
deathsat15	3638
golddiffat15	3638
firsttower	3638
firsttothreetowers	3638
side	0

Table.1

In our examination of the dataset, we noticed a significant pattern in the missing values, particularly in early and mid-game data such as 'firsttower', 'firsttothreetowers', and 'firstdragon', with the counts of their missing values being remarkably similar. A closer inspection of the rows where these gaps occur revealed a consistent trend: these omissions predominantly exist within four specific leagues - 'LPL', 'LDL', 'WLDs', and 'DCup'. This pattern suggests that the missing data likely stems from league-specific inconsistencies in data recording during certain stages of the game. Such discrepancies point to the conclusion that the missing values in our team data are not just random occurrences but are Missing at Random (MAR), where the absence of data is influenced by identifiable league characteristics

Through our discussion, we reached a consensus that there are two ways to handle the missing values:

The first way is to fill these anomalies by imputation. Imputing would allow our data to be complete and consistent without sacrificing existing data. However, this approach also comes with its risks. The major downsides of this method would be the lowered accuracy in prediction and authenticity of the data. Firstly, to impute numbers into our dataset, we have to follow existing patterns and constraints. The data within the columns of missing data are all binary, and every two rows in our dataframe are binded as pairs and regarded as a single game, which means the value between two adjacent rows should be mutually exclusive. For example, in a pair of adjacent rows, when the first team's value in 'firsttower' is 1, the second team's data should be 0, indicating that the first team destroyed the first defense tower in the game. As such, during the imputation process, we likely need to randomly choose a team within each matchup to be assigned a positive and the other to have a negative, resulting in an occurrence probability of 50% for each. The imputed results would likely violate the underlying mechanisms that are unobserved, and likely strongly correlated with these binary data. Such as some teams with better strategies and tactics will have a greater than 50% probability of obtaining these values as 1, and some doesn’t. Hence, the imputation process will insert a lot of noise into our dataset, and then affect the accuracy of the prediction model.

The second solution for dealing with missing values is to drop all rows in the dataset that contain NaN. The benefit of this method is better prediction accuracy, as there wouldn’t be random entries within our model. But the risk of this solution is the large proportion of missing values relative to the whole team dataset. 3638 out of 24900 rows, close to 15%, contain missing values for the essential features, which could remove potential correlations contributing to the prediction. Although the total number of rows retained in the end is large enough to build a prediction model, there may occur local biases. When predicting the results of a game for certain leagues that have a significant number of rows dropped, the accuracy of our prediction for those leagues could be reduced. The four leagues that generated the missingness, 'LPL', 'LDL', 'WLDs', and 'DCup', would become less representative and interpretable.

However, even with its risks, dealing with missing values using deletion is a much better option in this scenario, since it provides more interpretable and reasonable statistics, resulting in an, overall, more comprehensive solution. Imputing would leave too much noise within our dataset and couldn’t be addressed through the model fitting process, and would in turn lower the quality of our predictions. Another reason why deletion is more applicable is because of the missing values occurring all together in a given selection of observations, meaning that columns such as 'firstdragon', 'firsttower', and 'firsttothreetowers' contain NaN values for the same games. This implies that deleting entire rows wouldn’t affect other metrics and features of our model as much, and less sacrifices are made to other potential correlations. Thus, in the pursuit of prediction accuracy, we utilized deletion for our missing values.

In addition, we have also considered grouping rows with missing values, which account for about 15% of the overall team's dataset, into our test set, so as to make full use of the dataset. But, from the above analysis, we can see that most of the rows with missing values come from these four leagues, 'LPL', 'LDL', 'WLDs', and 'DCup', which belong to Missing at Random (MAR). In other words, their missing values do not occur randomly. This violates our random principle in splitting the test set and the training set, so this solution is considered unfeasible. To explain from a practical perspective, if we use rows with missing values as our test set, since most of the rows with missing values come from those four leagues, this means that our test set and training set do not come from The same collective. That is, we have Sample Selection Bias. The prediction model we build is based on the training set. When the training set and test set do not come from the same collective, our prediction model will not match the test set.

Why ‘side’ Need to be Considered

The permutation test conducted on the 'side' variable in League of Legends revealed a pivotal insight: a p-value close to 0. This statistically significant outcome indicates that the side a team plays on (either red or blue) has a profound impact on the game's result, a deviation unlikely to be attributed to random chance. In our permutation test, we randomized the distribution of wins between the blue and red sides multiple times, creating an empirical distribution of win rate differences. This process was aimed at simulating what the win rate disparity would look like if the side had no real impact on the game's outcome. (Fig.6)

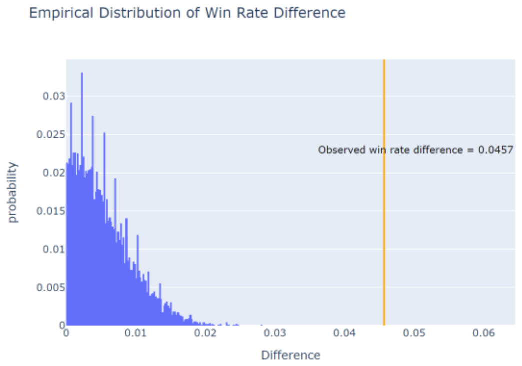

Fig.6

The histogram from this permutation test, shown in Figure 6, starkly demonstrates that the observed win rate difference is an outlier in this empirical distribution, indicating that such a win rate difference is highly unlikely to occur by chance. This finding strongly implies that the choice of side, and by extension, the terrain differences associated with each side, significantly influences the dynamics and ultimately the outcomes of the matches. This insight validates the inclusion of the 'side' variable in our predictive model, acknowledging the strategic nuances introduced by the asymmetrical design of the game’s map, which can sway the balance of the game in favor of one team over the other.

Data Transformation & Standardization

From the dataframe, we observe that there is a correlation between 'killsat15' and deathsat15. For example, in a game, one side's 'killsat15' and 'deathsat15' data are 5 and 3, and the other side's data are the opposite, that is, 3 and 5. This is because the number of kills of one team in a game is the number of kills of the other team. In order to reflect the difference between two adjacent rows, that is, the two sides in a game, it would be a good choice to calculate the difference between these two values, and obtain the net kills per team. Doing so would also align with other features that we employ, using the difference in statistics rather than the absolute value. For example, the first four features, 'firstblood', 'firsttower', 'firsttothreetowers', 'firstdragon', are all binary data, 0 or 1, to indicate whether the resources in these games have been obtained before the other team and gained some advantage. Another example is the column 'golddiffat15', which also represents the difference between the two sides of the game, that is, the difference in gold obtained by both teams in a game at the 15 minutes mark. Therefore, we should also calculate the difference between the number of kills and deaths to reflect the difference between the two teams, similar in principle to 'golddiffat15', which reflects the vectorially through positive and negative.

Furthermore, using the kill difference improves the balance of the dataset. This ensures that new features do not cause one aspect of the data to be overemphasized, biasing the model towards a particular outcome. On top of that, it also reduces the feature dimensions. By combining two related features into one, you reduce the number of features the model needs to process, helping to simplify the model and potentially improve its performance.
Lastly, it addresses the issue of multicollinearity. Since killsat15 and deathsat15 are logically related, meaning they are the exact complement, merging these two features can reduce multicollinearity in the data, which can be beneficial for some models.

Additionally, Standardizing the parameters 'golddiffat15' and 'kills_deaths_diff_at15' is necessary and offers several key benefits. First, it ensures that these features are on the same scale, allowing algorithms to process them without bias towards features with larger scales. Second, for algorithms that use gradient descent, standardization can accelerate model convergence since all features are on a comparable scale, allowing the gradient descent to work more efficiently. Third, it contributes to improving the overall predictive performance of the model, as standardized data reduces the potential imbalance in the impact of different features. In summary, standardization ensures fair and effective learning of the model across different features.

Model Selection

1. Linear Regression (Baseline):
We chose Linear Regression for its transparency in understanding the influence of both binary and continuous variables like 'firstblood', 'firsttower', and 'golddiffat15' on match outcomes. Its straightforward nature facilitates an easy interpretation of how each feature affects the game. However, the drawback of Linear Regression in our context is its limitation in capturing complex, non-linear interactions within the data. Considering the mixed feature types in our dataset of 21,253 rows, this model might not fully leverage the dataset's complexity, potentially oversimplifying the intricate relationships.

2. Logistic Regression:
Opting for Logistic Regression was motivated by its suitability for binary outcome predictions, aligning well with our win/loss prediction goal. This model's ability to provide probabilities for outcomes is particularly advantageous for understanding the nuances of match results. However, its limitations mirror those of Linear Regression, particularly in handling non-linear relationships between mixed feature types. In our sizable dataset, Logistic Regression might not adequately capture all the underlying patterns and interactions, leading to a potential oversimplification of the game dynamics.

3. Neural Networks:
Our choice of Neural Networks stems from their proficiency in modeling complex and non-linear relationships, a necessity given the diverse nature of our binary and continuous features. This model's flexibility and adaptability make it well-suited to our dataset. The challenge, however, lies in its

'black box' nature, which could impede our understanding of how specific features influence match outcomes. Moreover, the risk of overfitting is a significant consideration in our context, given the extensive size of our dataset and the potential complexity of the neural network model.

4. Random Forest:

We selected Random Forest for its ability to effectively capture non-linear relationships, crucial for analyzing our dataset comprising a mix of binary and continuous variables. Its insight into feature importance is invaluable for dissecting the impact of each game element. The model's complexity and computational demands, however, pose challenges, especially in a large dataset like ours. This complexity may lead to overfitting and require substantial computational effort for model training and tuning.

Results (Model Accuracy):

We split the data into a test set and a training set to build a predictive model (3:7). Below is the prediction accuracy corresponding to each model:

Linear Regression (Baseline)	77.96424%
Logistic Regression	77.71329%
Neural Networks	77.74466%
Random Forest	76.27038%

Conclusion

The analysis of professional League of Legends matches using various predictive models has yielded insightful results. Our models were applied to a dataset comprising 21,253 rows and 11 columns, with features including binary data like 'side', 'firstblood', 'firsttower', 'firsttothreetowers', 'firstdragon', and transformed continuous variables such as 'golddiffat15' and 'kills_deaths_diff_at15'. The goal was to predict match outcomes based on early and mid-game performance metrics.

Our findings show that the Linear Regression model, serving as our baseline, achieved the highest prediction accuracy of 77.96%. This is particularly noteworthy given the model's simplicity and interpretability. Its success suggests that, despite the complex dynamics of League of Legends matches, a significant portion of the outcome can be explained through linear relationships between the selected features and the match results.

The Logistic Regression and Neural Networks models followed closely, with accuracies of 77.71% and 77.74% respectively. The comparable performance of Logistic Regression indicates its effectiveness in handling binary outcome predictions, aligning well with the win/loss nature of our target variable. The Neural Networks' slightly higher accuracy underscores its capability to model complex relationships and interactions among features, though its 'black box' nature poses challenges in interpretability.

The Random Forest model, while slightly less accurate at 76.27%, provided valuable insights into feature importance and the non-linear relationships within the game. Its slightly lower performance might be attributed to the model's complexity and the potential overfitting to our specific dataset.

Overall, the close performance metrics across all models suggest that early and mid-game dynamics in League of Legends have a significant, yet somewhat predictable, impact on match outcomes. The slight variations in model accuracies could be due to the nature of the features and the inherent randomness in game outcomes. These findings highlight the importance of considering a variety of modeling approaches when dealing with complex datasets in esports analytics. Each model offers unique strengths and can contribute to a more nuanced understanding of game dynamics, ultimately enhancing strategies for players and teams in professional League of Legends competitions.

Discussion and Exploration

Analyzing the prediction results across our four models, we noted that their performance metrics are strikingly similar. The majority of our chosen features are binary, supplemented by some data that have been standardized to represent differences. Generally, linear regression's accuracy in handling high-dimensional data might theoretically lag behind more complex models due to its relative inflexibility. Consequently, we believe that incorporating deeper features could significantly enhance our prediction model.

Beyond early and mid-game performance metrics, it's essential to consider the variability unique to each team. Specifically, we aim to integrate a parameter that encapsulates a team's distinctive characteristics, such as their gameplay style, strategic capabilities, and overall team strength. A team's strength is a composite of various elements, including the players' in-game prowess, the coach's tactical acumen, and the support infrastructure provided by their organization. 'Teamname', as a feature, effectively encapsulates these multifaceted aspects, offering a more comprehensive representation of each team's potential and capabilities. By integrating 'teamname' into our model, we anticipate a more nuanced understanding of the game dynamics, potentially enhancing the predictive accuracy and depth of our analysis.

Therefore, we introduced the 'teamname' feature into our dataset and retrained our four predictive models. Below is a comparative analysis of the accuracy achieved by each model following the inclusion of 'teamname':

Linear Regression (Baseline)	77.96424%
Linear Regression (with teamname)	78.34065%
Logistic Regression	77.71329%
Logistic Regression (with teamname)	78.07402%
Neural Networks	77.74466%
Neural Networks (with teamname)	70.64363%
Random Forest	76.27038%
Random Forest (with teamname)	81.27038%

The new outcome with 'teamname' added:
The addition of the 'teamname' feature to our predictive models resulted in notable variations in performance, underscoring the complexity and significance of team-specific factors in professional League of Legends matches. The Linear Regression model's accuracy saw a modest increase from 77.96% to 78.34%, suggesting that incorporating team identities adds valuable insights into the prediction process. Similarly, the accuracy of the Logistic Regression model also improved, rising from 77.71% to 78.07%, indicating that the team's unique characteristics have a discernible impact on match outcomes. However, the Neural Networks model displayed a surprising decrease in accuracy, dropping from 77.74% to 70.64%. This decline might be attributed to the model's sensitivity to the added complexity of 'teamname', potentially leading to overfitting or challenges in capturing the intricate relationships associated with team dynamics. On the other hand, the Random

Forest model exhibited the most significant improvement, with its accuracy soaring from 76.27% to 81.27%. This dramatic increase suggests that Random Forest is particularly effective at leveraging the nuanced information captured by the 'teamname' feature, likely due to its strength in handling non-linear relationships and extracting the most pertinent features for predicting match outcomes. These varied results highlight the multifaceted nature of esports analytics, where the integration of specific team characteristics can both clarify and complicate the predictive modeling process, depending on the model's structure and approach.

Past Studies

The burgeoning field of esports analytics has seen researchers delve into the predictive analysis of match outcomes in League of Legends, with a notable focus on the early game phase. Prevailing studies have predominantly concentrated on the first 10 minutes, employing various machine learning techniques to unravel the relationship between early game events—like first blood and first tower—and winning probabilities. These investigations, while utilizing datasets similar to ours, primarily leverage logistic regression, neural networks, and decision trees. However, their insights, primarily predictive in nature, often do not extend into concrete strategic recommendations. This limitation stems from the complex and unpredictable nature of the game, where myriad variables and human decisions intertwine.

Our research aims to build upon and extend these foundational studies. By broadening the analytical timeframe beyond the initial 10 minutes to a more comprehensive early to mid-game phase, our study not only seeks to enhance the accuracy of predicting match outcomes but also to explore deeper into the strategic elements of the game such as which team gets the third tower. This approach is intended to provide more actionable insights into effective early game tactics, thereby addressing a critical gap left by previous research. In essence, while past studies have significantly contributed to understanding the predictive aspects of early game dynamics, our work aspires to link more comprehensive insights, and re-define early games as 15 min with practical, strategic implications in League of Legends gameplay.

One of the past study’s link: https://www.kaggle.com/code/xiyuewang/lol-how-to-win

Ethics & Privacy

We have to collect data from League of Legends tournament matches. We checked out the league of legends’ public statistics website to find the data from all players in the world, covering all the matches in all regions throughout the year 2022. The data collected is comprehensive since all players who played in the tournament has agreed and gave consent for their data being collected. Besides, data collected are also considered accurate, because the data such as kills, deaths, and assists can also be verified through recorded replay videos available. Moreover, the data collected only contain information relevant to the tournament, so no privacy violations occur. The potential biases that the data might have are that players in different regions may have different playstyles and preferences for picking different heroes (characters in game), which may lead to inconsistencies in analysis. However, since the data collected have categorized them into different regions of the tournament, we can analyze them accordingly.

One of the potential privacy issue may occur in our research is that the personal information may be revealed through the data. For example, player names and performance data could be linked to personal social media profiles or other public information, which could compromise their privacy. We will make sure that any personal information, like as player user name, is either anonymized or removed before using the data for analysis.

Riot Games Privacy Policy Link:
https://developer.riotgames.com/policies/general