Early-Game Advantage: Predicting League of Legends Match Outcomes Using Early-Game Statistics

Authors:

Mouad Zemzoumi, Saad Elmilouay, Salaheddine Boumazough

University of Helsinki

Introduction to Data Science

October 17, 2025

Abstract

This project explores the use of early-game statistics from Riot Games' League of Legends (LoL) API to predict match outcomes using machine learning. By focusing on the first ten minutes of gameplay, the study aims to identify the early-game performance indicators most predictive of victory. Data was collected through Riot's API, preprocessed to extract key metrics such as kills, deaths, assists (KDA), gold, and objective control, and then modeled using logistic regression. The results demonstrate that the gold difference, KDA ratio, and early objectives, such as dragons, play a significant role in determining the probability of winning. The final model achieves strong classification performance and provides players with actionable insights through a simple web interface.

1. Introduction & Motivation

League of Legends (LoL) is a team-based strategy game where two teams of five compete to destroy the opposing team's base. The outcome of a match is heavily influenced by early-game performance, where small advantages in kills, gold, or objectives can snowball into decisive victories. This project seeks to quantify those relationships by using statistical modeling to predict match results based on the first ten minutes of play. The motivation stems from the desire to assist competitive and casual players in understanding how early-game decisions correlate with success, thereby improving their strategy.

2. Data Collection & Preprocessing

Data was collected using Riot's official API endpoint, which provides timeline data for each match. The dataset includes early-game metrics such as kills, deaths, assists, team gold, dragon kills, and void grub captures. Preprocessing involved cleaning missing or inconsistent values, computing derived metrics such as KDA ratio $(Kills + Assists)/(Deaths + 1)$, and calculating the gold difference between teams. Categorical features such as objective control were encoded as binary variables. The processed data was stored in CSV format for reproducibility and model training.

3. Modeling Approach & Evaluation

The modeling process followed a structured machine learning workflow that combined data cleaning, feature engineering, and evaluation through iterative experimentation. The primary goal was to build a simple yet interpretable model capable of predicting a team's probability of winning a League of Legends match based only on early-game performance data extracted from the Riot API.

The first step was to construct a clean feature set from the preprocessed dataset. Each match entry represented one team's statistics within the first ten minutes, including kills, deaths, assists, total gold, experience gained, and objective control such as early dragons. Derived metrics such as gold difference and KDA ratio were computed directly from these values to summarise combat efficiency and resource control. Outliers caused by early surrenders or incomplete matches were filtered out to avoid introducing bias.

For model selection, a logistic regression classifier was implemented using `scikit-learn`. This choice was guided by its interpretability, computational efficiency, and suitability for predicting binary outcomes. During preprocessing, continuous features were normalized to a common scale, while categorical ones (like objective control) were encoded as binary indicators. The dataset was divided into training and testing sets with an 80/20 ratio, ensuring both contained a similar distribution of wins and losses.

Training and evaluation were performed iteratively. The model was first trained on the cleaned dataset, and its performance was tested using accuracy, precision, recall, and F1-score to ensure balanced predictive ability. Additionally, a confusion matrix and classification report were generated to visualise how well the model distinguished between winning and losing teams. Cross-validation was also applied to confirm that results were not dependent on a specific data split. The evaluation results showed that the model consistently performed well across different folds, demonstrating its stability and reliability.

In practice, the model's workflow from the notebook involved loading the cleaned CSV, splitting the dataset, fitting the logistic regression model, and visualizing predictions. The simplicity of this approach made it ideal for practical use cases, allowing efficient training on a standard laptop while maintaining clarity in how each variable contributed to the outcome.

4. Results & Discussion

The model produced promising and interpretable results. On the test set, it achieved around 70% accuracy, which is a strong result given the natural complexity and variability of online matches. This indicates that early-game data alone can provide a reliable estimation of a team's chance of victory. Precision and recall values were balanced, showing that the model could effectively identify both winning and losing teams without excessive bias toward one outcome.

Detailed inspection of the feature contributions revealed that **gold difference** and **KDA ratio** were the strongest predictors of a win. Teams that accumulated more gold and maintained higher kill participation early on were significantly more likely to convert that advantage into victory. Early objectives such as **dragon control** were also key indicators: teams securing these objectives displayed a clear statistical edge, confirming the competitive value of early map dominance.

Visual analyses from the `Visualizations.ipynb` notebook supported these findings. Plots of average gold difference over time showed that winning teams typically established a noticeable lead by the 8–10 minute mark. Distribution plots for kills and assists highlighted distinct separation between winning and losing outcomes. Correlation heatmaps further confirmed that gold difference and KDA ratio had the highest correlation with victory, followed by objective control metrics. These results align with in-game logic: teams with better resource management and map control tend to dictate the pace of play and minimize risks.

The confusion matrix revealed that most misclassifications occurred in matches with very small early-game leads. This suggests that while early-game performance is important, it does not capture mid- or late-game events such as team fights, scaling champions, or strategic comebacks. Nevertheless, the model's consistency across multiple tests proves that early-game indicators are strong, measurable reflections of competitive success.

Overall, these results highlight that League of Legends outcomes are not purely random or dependent on individual skill, but often shaped by quantifiable early advantages. Even modest improvements in early gold efficiency or objective timing can meaningfully shift win probabilities, which makes this analysis valuable not only for competitive teams but also for casual players looking to understand how their decisions affect match progression.

5. Communication & Ethical Considerations

Following the successful development and validation of the predictive model, the project was deployed as a live web application accessible at https://lol-project-gamma.vercel.app/. The web interface was developed using `JavaScript` and the `Express.js` framework to provide a lightweight, real-time prediction service. The design goal was to make the model usable and understandable for players without technical expertise, transforming data analysis into a practical tool for early game winning rate prediction

The backend server, built with Express.js, handles HTTP requests and communicates directly with the trained logistic regression model. When a user submits early-game statistics, which include Total kills, Total Deaths, Total assists, and Total Gold for each team, as well as specifying which team has captured the dragon, it validates the input, processes it into the appropriate format, and passes it to the model. The model then returns a win probability. On the frontend, this probability is visualised as a clear percentage value accompanied by contextual feedback (for example, ”Looking very promising!”) and the calculated inputs that were fed into the model, namely each team's KDA, Gold Difference, Gold per Kill and Dragon advantage. The system's fast response time and modern interface ensure smooth user interaction across both desktop and mobile devices.

Deployment on `Vercel` simplified hosting, scalability, and continuous integration. It allowed the team to update the interface and backend logic rapidly, ensuring that improvements to the predictive model could be published instantly. The API architecture also enables potential future extensions, such as automated data retrieval from Riot's servers or the inclusion of live match-tracking features. Overall, this web deployment transformed the research prototype into an accessible and practical application for players, analysts, and e-sports enthusiasts.

Ethically, the project maintains full compliance with responsible data use principles. All data was collected through Riot Games' public API, which anonymizes personal player information and provides access only to match-level statistics. The analysis was conducted at the team level, not the individual level, thereby avoiding profiling or ranking specific players. The dataset contained no usernames, account IDs, or region identifiers, ensuring total privacy protection.

Potential ethical concerns were still acknowledged. Predictive systems in gaming can unintentionally influence player behavior or expectations if results are misunderstood as guarantees rather than probabilities. To address this, the model output emphasizes that predictions reflect statistical tendencies, not certainties. Furthermore, the model's reliability may vary with game patches or changing metas, which can alter the relative importance of early objectives or champion balance. For this reason, periodic model retraining and transparent documentation of its limitations were recommended.

The broader implication of this project lies in its educational and analytical value. It demonstrates how open data and machine learning can be used responsibly to deepen understanding of complex competitive environments. By quantifying how early decisions shape outcomes, the model promotes strategic thinking while preserving fairness and respect for player autonomy. Ultimately, this balance between technological insight and ethical awareness defines the strength of the project and its potential for future expansion.

6. Conclusion & Future Work

This project demonstrates that early-game performance metrics in League of Legends can reliably predict match outcomes using interpretable machine learning techniques. By leveraging Riot API data and logistic regression modeling, we established that gold difference, KDA ratio, and objective control are strong indicators of victory. Future work could include integrating champion-specific features, expanding to deep learning architectures for improved accuracy, and deploying real-time prediction through API streaming. Ultimately, this system empowers players to make more informed strategic decisions during gameplay.