SN Computer Science (2024) 5:968
https://doi.org/10.1007/s42979-024-03327-w

ORIGINAL RESEARCH

Assessing Player Contributions in League of Legends Matches: An Analytical Approach

Manuel Perez[1] · Cesar O. Diaz[1] · Pau Soler[1] · Aitor Mier[1]

Received: 29 August 2024 / Accepted: 12 September 2024

Abstract

This paper presents a comprehensive study on the contribution of various factors in "League of Legends" (LoL) matches. The research focuses on multiple aspects, including match analysis, data reduction strategies, predictive models, and exploratory data analysis (EDA). We predict match outcomes with significant accuracy using machine learning techniques such as Binary Logistic Regression (BLR). Additionally, we employ Principal Component Analysis (PCA) and Gradient Boosting Regressor (GBR) for dimensionality reduction to simplify the complex interactions among game variables. Our exploratory data analysis identifies key patterns, trends, and relationships within the game data, helping to optimize gameplay strategies. The models demonstrate high accuracy and robustness through rigorous evaluation, and the findings provide valuable insights for players, researchers, and the eSports industry, highlighting the potential of real-time data and machine learning models in enhancing game performance and strategic decision-making.

Keywords eSports analytics · Machine llearning · Game data analysis · Data reduction · Predictive modeling

Introduction

In recent years, the analysis of video game data has become a pivotal area of research, particularly in the realm of eSports. "League of Legends" (LoL), a highly popular multiplayer online battle arena (MOBA) game, presents a rich dataset that offers numerous opportunities for analysis and insights. Understanding the dynamics and outcomes of LoL matches can provide significant advantages for players, coaches, and analysts, driving performance improvements and strategic advancements.

This study aims to explore the contributions of various factors in LoL matches through a comprehensive analysis. Using machine learning techniques, we can accurately predict match outcomes, providing valuable insights into player performance and team dynamics. Additionally, we investigate data reduction strategies to manage the vast amounts of data each match generates, ensuring that the most critical variables are identified and utilized effectively.

Furthermore, we delve into exploratory data analysis (EDA) to uncover patterns and trends within the game data, enhancing our understanding of the underlying mechanics and strategies. This multifaceted approach highlights the potential of real-time data and predictive modelling in eSports and underscores the importance of data-driven decision-making in optimizing gameplay and achieving competitive success.

This paper is structured as follows: we begin with a review of related work, followed by a detailed analysis of match data, data reduction strategies, predictive models, and exploratory data analysis. Finally, we present our findings and discuss their implications for future research and practice in the field of eSports analytics.

Manuel Perez, Cesar O. Diaz, Pau Soler, Aitor Mier have contributed equally to this work.

✉ Cesar O. Diaz
cesar@omashu.gg

Manuel Perez
manuel@omashu.gg

Pau Soler
pau@omashu.gg

Aitor Mier
aitor@omashu.gg

[1] Informatics Department, OMASHU, Carrer de Pascual i Vila, 08028 Barcelona, BCN, Spain

Published online: 16 October 2024

Related Work

The analysis of video game data, particularly from "League of Legends" (LoL), has received considerable attention in recent academic research [1–5]. This section reviews key works that have explored relevant aspects of game dynamics and data modelling in LoL.

Analysis of LoL Matches: Jailson B. S. Junior and Claudio E. C. Campelo (2023) studied predicting match outcomes in LoL using machine learning techniques [6]. They explored various models, including LightGBM, which achieved an accuracy of 81.62% in the intermediate stages of matches. Logistic Regression and Gradient Boosting models were also found effective in the game's early stages. Their research highlights the importance of real-time data in predicting match results and provides insights for both players and the betting industry.

Data Reduction Strategies: Data reduction is essential for handling large volumes of information coming from different sources [3, 7]. Sharma and Srivastava (2020) applied dimensionality reduction techniques [3], such as Principal Component Analysis (PCA), to simplify complex interactions among game variables. Their research demonstrated how PCA can identify the most influential variables in predicting victories and defeats.

Predictive Models and Evaluation: A study by Do, Wang, Yu, McMillian, and McMahan (2021) explored the use of various machine learning models to predict match outcomes in LoL based on pre-match data [4]. They investigated Support Vector Classifiers, k-Nearest Neighbors, Random Forests, Gradient Boosting, and Deep Neural Networks. Their study found that pre-match data, such as champion mastery and player experience, can significantly predict match outcomes, achieving high accuracy rates. This research provides valuable insights into the predictive analytics of eSports and highlights the effectiveness of using machine learning models for game outcome prediction.

Exploratory Data Analysis (EDA): Morales-García, Llanes, Curado, and Arcas-Túnez (2023) conducted an exploratory analysis of professional LoL match data [8]. Their study utilized various EDA techniques to identify patterns, trends, and relationships within the game data. This analysis helps players and researchers optimize gameplay strategies by understanding key statistics, strengths, weaknesses, and meta-changes in LoL matches.

Background

As mentioned in a previous work [1], authors have carried out an Exploratory Data Analysis (EDA) strategy to know more about the data incoming from a particular game, ***League of Legends (LOL)***.[1] Also, the authors added new calculated variables to give a better approximation of what occurs in the match, *goldUsage, midLaneTurretAssits, midLaneTurretKills, baronAssists, midLaneInhibitorAssits* [1] This analysis consists of several aspects, such as correlation matrix analysis and violin graphs of the variables by position, resulting in extensive graphics and a huge data analytical approach. The authors are presenting here a briefly relevant summary:

- Depending on the position, some of the variables analysed have different effects on the game's development: The fact that the variables change depending on the position allows us to define a role for each position.
- Killing minions does not seem to affect the game.
- How a lower value of *goldUsage* contributes to the victory, so it can help to optimise a strategy to spend the gold obtained in the match.

Some of the most relevant variables are shown below to showcase this approximation, which benefited from splitting the same variable by position.

The variable *baronAssist* (Fig. 1) displays significant discrepancies across all positions, with each position exhibiting three distinct density peaks. It is imperative to conduct a detailed case-by-case analysis of these discrepancies:

- First peak, below average: the middle and top positions show a very similar density concentration in the middle and top positions. This indicates that there is usually not much difference between wins and losses. However, this is not the case for the other positions, where being below average is a good indicator to predict defeat in most cases. Even so, when reaching the average, there are more observations of victory.
- Second and third peak, above average: This case indicates a generalised behaviour, where being above average is a strong driver to victory. These graphs conclude that the importance of Bottom and Utility assists is remarkable, although team participation, in general, must not be neglected.

[1] "League of Legends is a team-based strategy game where two teams of five powerful champions face off to destroy the other's base" [2].

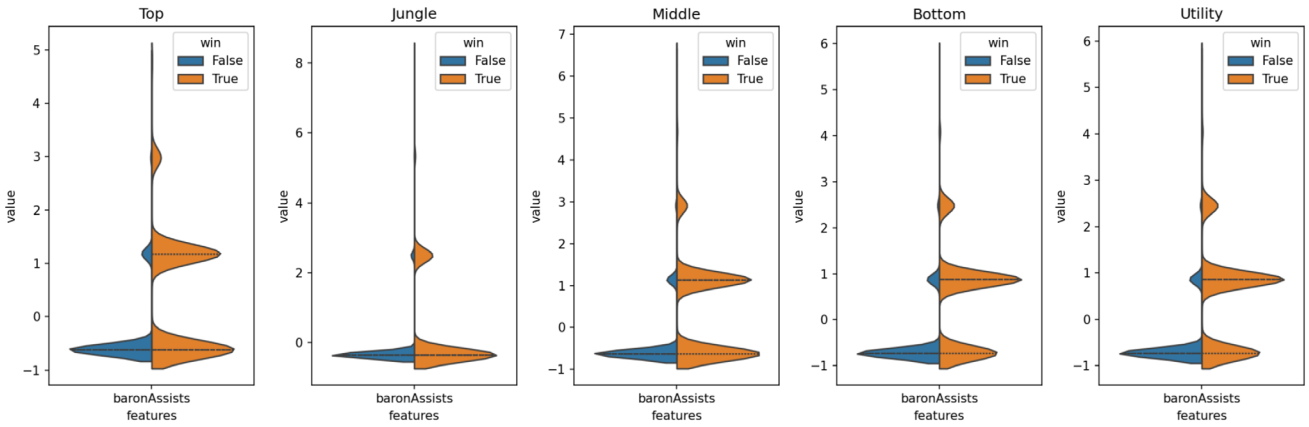


Fig. 1 Violin plot of the variable *baronAssists*

Regarding Kills (Fig. 2), for Top, Bottom, Middle, and Utility, there is a fairly symmetrical distribution around the average, although there is a slightly higher concentration on "win". For Jungle, being below the average is not enough, as it induces a higher degree and probability of defeat, which changes around being above the average.

In the development of the article, models such as GBR (Gradient Boosting Regressor, Jerome H. Friedman [9], in 1999), BLR (Binary Logistic Regression, Joseph Berkson [10] in 1944) and PCA (Principal Components Analysis, Karl Pearson [11] in 1901, Harold Hotelling [8] 1930) were used.

Reduction Strategies

The match data to be analyzed is obtained from the RIOT Developer API [12]. A match can be downloaded in two formats: a summary of several metrics that happened in the match and the timeline representation, which contains events with the timestamp that occurred in the match, together with some additional information. To sum it up, the two data forms available are:

- **Postmatch**: summary of the match when finished. The data is organized on a player basis, with the exact same variables for every 10 players in the match.
- **Bytime**: description of the match on an event basis, structured by minute. The data is not organized by the player, and not every minute has the same number of events, depending on what happened in the match.

In this data exploration stage, some data pruning was quickly discarded due to the possibility of missing some key insights. Despite ruling that option out, it was quickly noticed that some data treatment was needed to make it more manageable. On average, the combination of Bytime and Postmatch representations and the player rank variables weight about 694.19 kB for just one match. The approximation was taken to change the original representations downloaded from RIOT into a representation that (1) was more accessible to extract info from and (2)

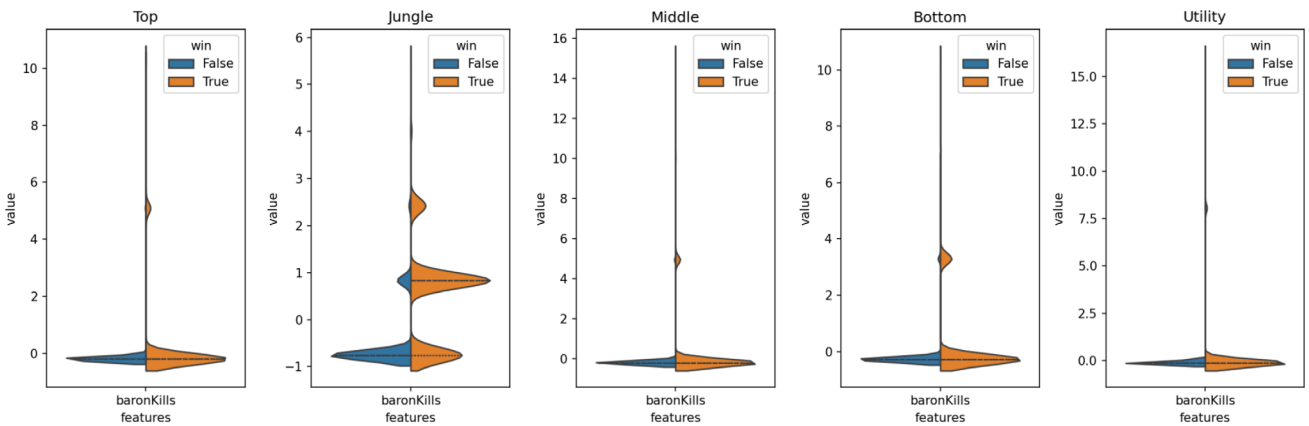


Fig. 2 Violin plot of variable *baronKills*

was smaller. Before describing the process, the following table shows the average results reduction and the reduced proportion (see Table 1):

The table results show that our new and innovative approach to reorganizing the data allows us to store all data more optimally, enhancing data management and drastically reducing costs.

Data Reduction

The two reduction strategies implemented for the distinct data types exhibit slight variations while sharing a common underlying principle. Our discussion will commence with the Postmatch representation.

Postmatch

This section will explain the algorithm using the example shown in Fig. 3 as an input and Fig. 4 as its output. The input comprises data not from a real match, shortened and simplified to showcase the algorithm and its results.

For the sake of the example, the number of players has been reduced to 4, two per team. As you can see, "m1" gets repeated for every player in the list. The metrics usually are between 6 and 18 characters long, and with approximately 300 metrics per player, the size can get substantially bigger when the number of metrics grows.

Our approximation is to de-nest the list into the metrics, not to repeat the metric for every player, instead just to repeat the player metric, like it can be seen in the figure 4:

Rewriting the longest metric as near to the base (the more outer level of the JSON) as possible reduces the number of times it is repeated, making the data much lighter. On the other hand, this process also accomplishes a summary function, deleting data that is the same across the whole team (m2), and it's just written once per team instead of once per player.

Bytime

The Bytime data, as opposed to Postmatch, holds the most significant representation and offers substantial room for improvement through data indexing and reorganization.

Table 1 Table showing size before and after the reduction of the total, the Bytime and the Postmatch

	Original	Reduced	Improvement
Download	694.19 kB	328.84 kB	2.11
Bytime	619.81 kB	111.03 kB	5.58
Postmatch	73.22 kB	35.23 kB	2.08

Before delving into the optimization process, it is imperative to comprehensively understand the data and its structural layout conveyed in the Figs. 5 and 6.

The data's main body is the frame's key, which contains a list. Every element is a minute of the match, and per every minute, we have two pieces of information:


```
{
    "info": {
        "i1": 1,
        "i2": 2,
        "participants": [
        {
            "m1": 11,
            "m2": 101,
            "id": 1,
            "teamId": 100
        },
        {
            "m1": 12,
            "m2": 101,
            "id": 2,
            "teamId": 100
        },
        {
            "m1": 13,
            "m2": 202,
            "id": 3,
            "teamId": 200
        },
        {
            "m1": 14,
            "m2": 202,
            "id": 4,
            "teamId": 200
        }
        ]
    }
}
```


Fig. 3 Postmatch style data obtained from RIOT API


```
{
    "info": {
        "i1": 1,
        "i2": 2,
        "players": {
            "m1": {
                "1": 11,
                "2": 12,
                "3": 13,
                "4": 14
            }
        },
        "teams": {
            "m2": {
                "100": 101,
                "200": 102
            }
        }
    }
}
```


Fig. 4 Reduced result of the data shown in Fig. 3

```
{
  "info": {
    "i1": "c",
    "frameInterval": 60000,
    "frames": [
      {
        "participantFrames": {
          "1": {
            "pf1": 1,
            "pf2": 2
          },
          "2": {
            "pf1": 1,
            "pf2": 2
          },
          "3": {
            "pf1": 1,
            "pf2": 2
          },
          "4": {
            "pf1": 1,
            "pf2": 2
          },
          "timestamp": 0
        },
```


Fig. 5 Bytime style data containing ParticipantFrames

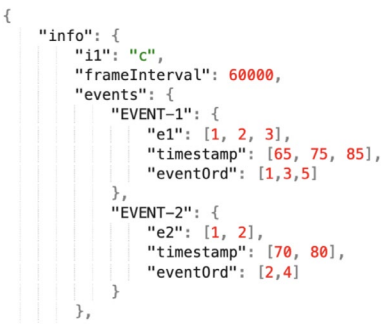
```
        "events": [
          {
            "e1": 1,
            "timestamp": 1000,
            "type": "EVENT-1"
          },
          {
            "e1": 2,
            "timestamp": 1200,
            "type": "EVENT-1"
          },
          {
            "e1": 3,
            "timestamp": 1400,
            "type": "EVENT-1"
          },
          {
            "e1": 1,
            "timestamp": 900,
            "type": "EVENT-2"
          },
          {
            "e1": 2,
            "timestamp": 1100,
            "type": "EVENT-2"
          }
        ]
      }
    ]
  }
}
```


Fig. 6 Bytime style data containing all the events in the match

- **Events**: holds all the events that happened in that minute. This is known in the *timestamp* variable, which will be a multiple of 60000, the value of *frameInterval*. Every event has different attributes inside, but it always has the key *timestamp* and *type*
- **ParticipantFrames**: Holds diverse player data and statistics. It happens every 60,000 ms.

```
        "participantFrames": {
          "pf1": {
            "1": [11],
            "2": [12],
            "3": [13],
            "4": [14]
          },
          "pf2":{
            "1": [21],
            "2": [22],
            "3": [23],
            "4": [24]
          }
        }
      }
    }
}
```


Fig. 7 Bytime ParticipantFrames reduced



```
{
    "info": {
        "i1": "c",
        "frameInterval": 60000,
        "events": {
            "EVENT-1": {
                "e1": [1, 2, 3],
                "timestamp": [65, 75, 85],
                "eventOrd": [1,3,5]
            },
            "EVENT-2": {
                "e2": [1, 2],
                "timestamp": [70, 80],
                "eventOrd": [2,4]
            }
        },
```


Fig. 8 Bytime events reduced

The Figs. 5 and 6 are just one minute long, so as the match gets longer, the more data will hold the Bytime, contrasting with the Postmatch data representation which is approximately the same size regardless of the match duration.

The strategy to reduce the Bytime data is, in general, the same as the Postmatch: the longest keys are going to be put as outside the JSON as possible, but there will be too more factors added. The first one is how the events are restructured, which will be grouped not by the minute but by event type. The second is the structure of the data inside the events. The result is shown in the following figures 7 and 8:

Before proceeding, it is necessary to analyse the occurrences illustrated in Fig. 8. In the minute treated, we merged the events by *type*. The contents are merged under the key name, and the list contains all the data of all the events. This means that the length of the list is the number of events that happened in the minute described. To exemplify this, notice that the amount of *EVENT-2* in the first figure was 2, and the length of the list of event two is two, coinciding the events of the metric, being the index of the list of the event we are referring to (e.g., the second element in the list in the e1 metric is the exact value of the

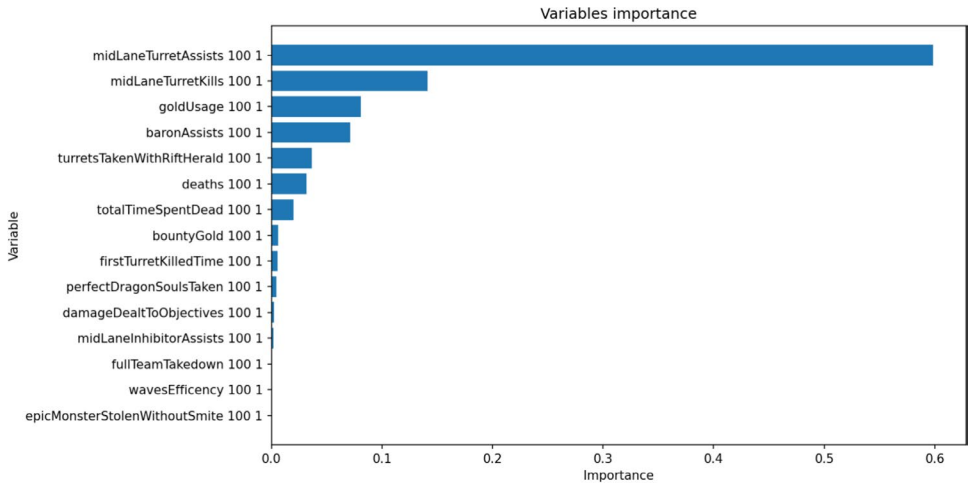


Fig. 9 Results of the GBR model for Top variables

Table 2 Results of the three principal components of the PCA

	0	1	2
midLaneTurretAssists	2.9416E–05	−0.00027834	−0.00017283
midLaneTurretKills	3.5176E–05	−0.00013776	1.1691E–05
baronAssists	2.5179E–05	−0.00010004	– 4.8729E–05
goldUsage	– 1.417E–05	0.0001	– 1.4732E–05
deaths	– 5.0476E–05	0.0016401	0.00177223
turretsTakenWithRiftHerald	7.6133E–5	−0.00285938	−0.00039176
totalTimeSpentDead	– 1.1656E–05	0.04128914	0.07195481
firstTurretKilledTime	0.01378454	−0.95987577	0.27938542
bountyGold	0.00274434	0.27706882	0.95745547
perfectDragonSoulsTaken	4.2377E–06	– 5.9891E–05	– 5.2482E–05
damageDealtToObjectives	0.99990122	0.01247311	−0.00647847
midLaneInhibitorAssists	7.2333E–06	– 8.6835E–05	– 4.2384E–05
fullTeamTakedown	2.8456E–05	5.7964E–05	0.00014106
wavesEfficency	– 1.7007E–06	2.8571E–05	7.3781E–06
epicMonsterStolenWithoutSmite	5.1715E–07	3.0503E–06	9.6146E–07

second *EVENT-1*. Lastly, to be able to reconstruct the order, we have the *eventOrd* list, where the number on the index is exactly the number of the event in the whole match. So, by reconstructing the events sorted by event order, you obtain all events of the match sorted.

The same principles apply to the *participantFrame* (7). The list holds the metric's value for a given player in a given minute. The length of the list is the length of the match, and the value of a given index is the value of that minute. The participant frames are kept aggregated by minutes because they appear without timestamps, just the same exact values for each minute.

Dimensionality Reduction

Given such a vast array and depth of variables, one of our priorities when building the model is to optimize the number of variables. To achieve this, we apply GBR (*Gradient Boosting Regressor*) and PCA (*Principal Component Analysis*) methods. Without delving into too many technical details, these non-black box models allow us to extract the relative importance of variables concerning a common target variable.

In Fig. 9 is shown GBR model, where a significant portion of importance is attributed to a single variable, *midLaneTurretAssist*, with nearly 60%, followed by *midLaneTurretKill*. This sheds light on the significance of the Top position in dismantling turrets over the mid-lane[2] and the substantial impact of including these variables in the model. Table 2 shows the application of Principal Component Analysis (PCA) on these same variables.

Table 2 provides information about the PCA process applied to our data. The cumulative explained

[2] These kinds of turrets include the nexus turrets.

variance in the first principal component amounts to a non-negligible 99.6%. Analyzing the weight of the variables in this component, it can be seen that there is a variable that weights 0.99990122, the variable *damageDealtToObjectives*. This suggests that this variable may be fundamental to the underlying structure of the data, in addition to reaffirming the importance of dismantling buildings such as turrets. Now, tests without this variable are performed to check the effect on the data set and the model.

Watching Fig. 10, it can be appreciated that the first component has a relative importance of 61%, and the second is already close to a little more than 95%, very significant changes from the previous case. Now, examine some graphs like Fig. 11 that can help illustrate this point:

Examining the coefficients in Table 3 of the linear combinations that constitute each Principal Component:

Now, by excluding the variable *damageDealtToObjectives*, a significant change is observed in the structure of the principal components and in the GBM model, which indicates that this variable has a substantial impact on the model and the interpretation of the results.

Model Description, Implementation, and Evaluation

Notation

To be able to estimate separately the contributions of the 10 players in a LoL game using the metrics selected in Section "Dimensionality Reduction", we need to choose an appropriate model. The proposed case is a BLR, as it is a non-black-box model, and we can easily handle

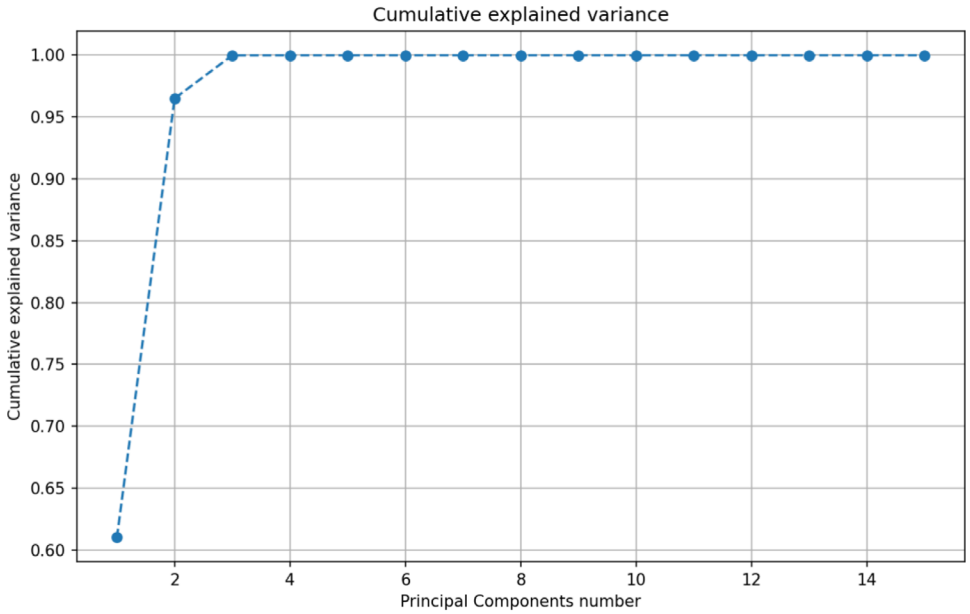


Fig. 10 Results of PCA after excluding *damageDealtToObjectives*

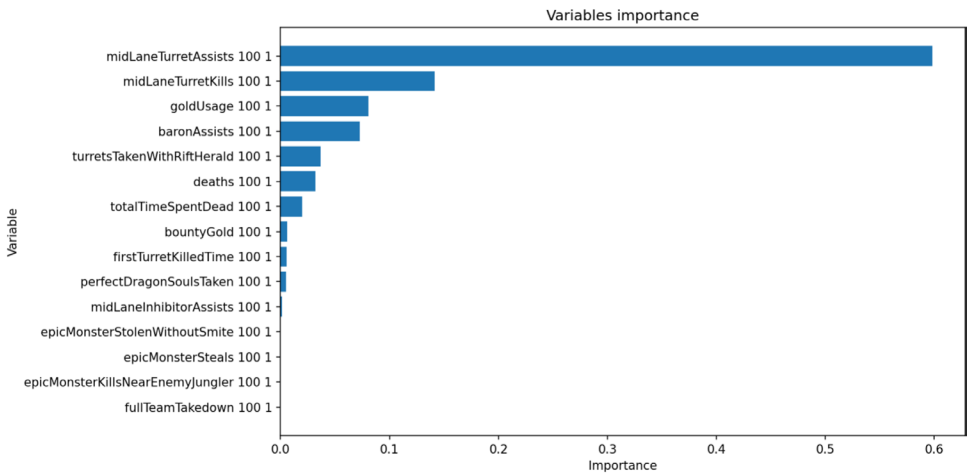


Fig. 11 Results of GBR after excluding *damageDealtToObjectives*

Table 3 Results of the three principal components of the PCA

	0	1	2
midLaneTurretAssists	−0.00038133	−0.00010345	−0.00144008
midLaneTurretKills	−0.00027953	0.00010991	−0.00038848
baronAssists	−0.00019936	2.2728E–05	– 9.1425E–05
goldUsage	0.00015504	– 4.9453E–05	0.00043569
turretsTakenWithRiftHerald	−0.00301084	−0.00043508	−0.00454749
deaths	0.00169944	0.00174031	0.02567846
totalTimeSpentDead	0.03645881	0.07456874	0.99620306
firstTurretKilledTime	−0.97580453	0.2177588	0.01944609
bountyGold	0.21555514	0.97314814	−0.08073499
perfectDragonSoulsTaken	– 7.2645E–05	– 4.3908E–05	– 4.2249E–05
midLaneInhibitorAssists	−0.00011116	– 2.7232E–05	−0.00026004
epicMonsterStolenWithoutSmite	6.5918E–07	2.9138E–06	3.387E–05
epicMonsterSteals	8.8696E–07	2.7778E–06	3.6772E–05
epicMonsterKillsNearEnemyJungler	– 1.0908E–05	1.1667E–05	3.8272E–05
fullTeamTakedown	– 7.078E–05	0.00023508	0.00077978

Table 4 Details of the evaluation of the model

Type	Accuracy	Precision	Recall	F1-Score
Test	0.99898477157	1	0.99775280899	0.9988751406
All sample	0.999898477	1	0.999782229965	0.9998911031

the coefficients and weights that the model gives to each variable. To this end, it seeks to estimate the following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n = \beta_0 + \sum_{i=1}^{n} \beta_i X_i, \qquad (1)$$

for certain coefficients β_i. By performing the process with the n variables extracted per position from the RBM models, we obtain a model with an Accuracy of 99.989%.

Evaluation

We show details of the evaluation of the model. The main objective is to show that the method described below does not incur any overlapping. For this, 10% of the sample is extracted, and with the remaining 90%, 90% is used for training and 10% for the test. The results are shown in Table 4:

It can be observed that the accuracy in the test sample and the whole sample is really elevated, being 99.899% in the first case and 99.9899% in the complete sample. With the precision values, it can be said that the wrong predictions occur when it's predicted a victory of the *100 team* is predicted when it really loses the match.

A series of tests were conducted to assess our methodology's robustness. Initially, we instituted a cross-validation process involving the manipulation of sample proportions within the test set. Subsequently, we proceeded to analyze the outcomes yielded by this approach. The results are shown in table 5.

Table 5 Details of the cross-validation process

Train Size	Test Size	Accuracy
0.9	0.1	0.998990
0.8	0.2	0.997980
0.7	0.3	0.999663
0.6	0.4	0.998737
0.5	0.5	0.998384
0.4	0.6	0.998485
0.3	0.7	0.998701
0.2	0.8	0.997980
0.1	0.9	0.997531

Based on the results provided, we can make several observations and draw some conclusions regarding the accuracy of the model, and it's behaviour about overfitting:

- The model's accuracy is consistently high across all test set sizes, with values above 99.7%. This suggests that the model performs well overall.
- There is slight variability in the accuracy values as we change the test set size. However, these variations are minimal (differences in the third or fourth decimal place), indicating that the model maintains high performance regardless of test set size.

- It can be observed that the accuracy is slightly higher in intermediate configurations, such as in the case of 70% training and 30% test, where the accuracy reaches 0.999663.

Consistency of accuracy suggests that the model is avoiding overfitting. An overfitted model generally shows a significant drop in accuracy when the size of the test set is increased. In this case, consistently high accuracy indicates that the model fits the training data well without losing generality. In addition, the high accuracy in different settings also suggests that the model is robust and reliable in making accurate predictions in various data-splitting scenarios.

Secondly, we will proceed with the hold-out procedure as shown in Table 6, which is to evaluate the model with the 10% we extracted before. Accuracy can be extracted from the tested model, which can be compared with the Accuracy by applying the model to the reserved data. We do this in our case. Having a dataset with a total of 9900 rows (items), we have to choose 990 random rows. We perform this process and obtain the following data:

It can be seen how the values are very similar in predicting new data that have not been used in the training and testing of the model, which confirms the robustness and generality of the developed model. This behaviour indicates that the model is not overfitted to the training data and can adequately handle new samples, providing reliable predictions.

Furthermore, analysing evaluation metrics such as accuracy, recall, and F1 score further supports the model's effectiveness. These positive results suggest that the model can be useful for practical applications, offering consistent performance across different datasets.

It is important to note that although the current results are promising, it is always advisable to continue validating the model in different scenarios and with other datasets to ensure its robustness and adaptability to varied situations. This ensures that the model is accurate and reliable in controlled conditions and in real environments where variables may change.

Analysis of the Contributions

This section will analyse the suitability of various distributions to our calculated contributions. The initial step involves conducting different tests to determine the best-fitting distribution for our numbers. Subsequently, we will examine the contributions of a specific position, namely the *Top* position of the *blue team*, and evaluate the match contributions.

As can be seen in Fig. 12, there is no precise distribution. Note that there are very heavy queues on the left, with a high concentration on the opposing side. However, the highest concentrations are around 0.15 and 0.25. It can be anticipated that the analysis will be complicated without using various distributions.

Tests to Find the Best Fitting Distribution Function

To perform tests of normality, it will be used the Shapiro–Wilk [13] and the Kolmogorov-Smirnov test [14, 15].

- Shapiro-Wilk p-value: 2.6309875465813093e−05
- Kolmogorov-Smirnov p-value: 3.368574598744161e−31

The p-values for all scenarios are significantly low, indicating a rejection of the null hypothesis that our data follows a normal distribution, as they are below the 0.05 threshold. Conducting a comparison to another distribution, such as the exponential (Fig. 11), can be beneficial. In this regard, the Akaike Information Criterion [16] can be employed for comparison purposes (see Fig. 13) (see Table 7):

- Normal distribution AIC: 101.7651113362715
- Exponential distribution AIC: −1.9075708098152386

Upon utilizing the AIC, the Exponential distribution appears better suited to our data due to its lower value. However, upon conducting a test, the p-value is determined to be 6.242223266773024e−17, indicating statistical insignificance as shown in Table 8. Therefore, affirming that any distributions are appropriate for the data is inconclusive. Let us conduct an analysis using additional and more

Table 6 Results of the hold-out procedure

Type	Accuracy	Precision	Recall	F1-Score
Test	0.99898477157	1	0.99775280899	0.9988751406
All sample	0.999898477	1	0.999782229965	0.9998911031
Hold-out	0.9878787879	1	0.9744136461	0.98704103672

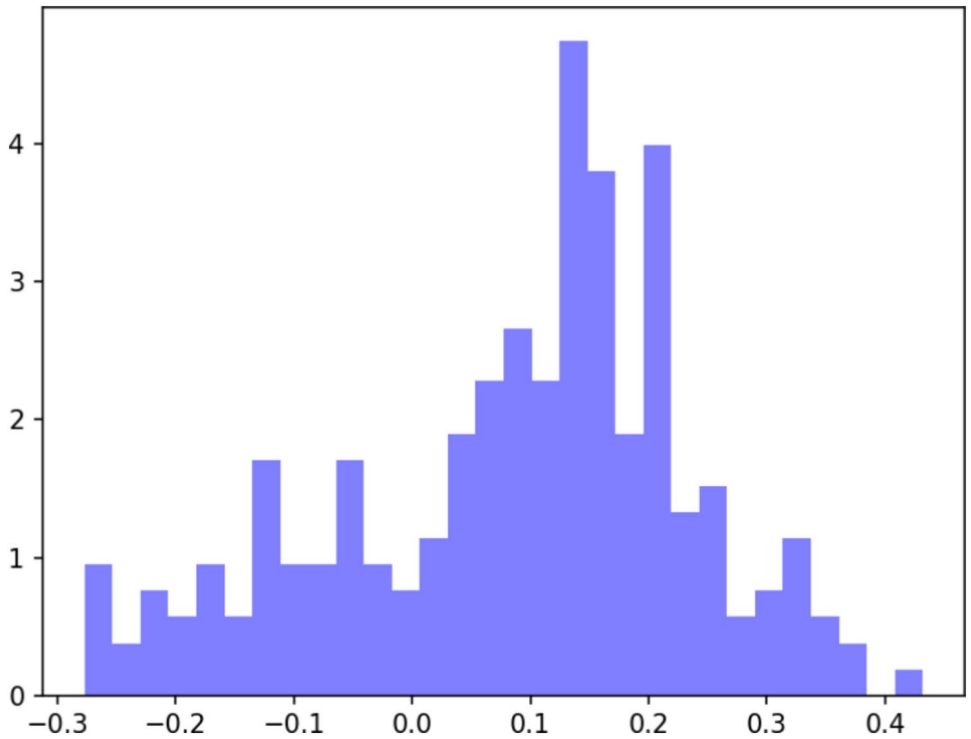


Fig. 12 Plot of the contributions of the Top

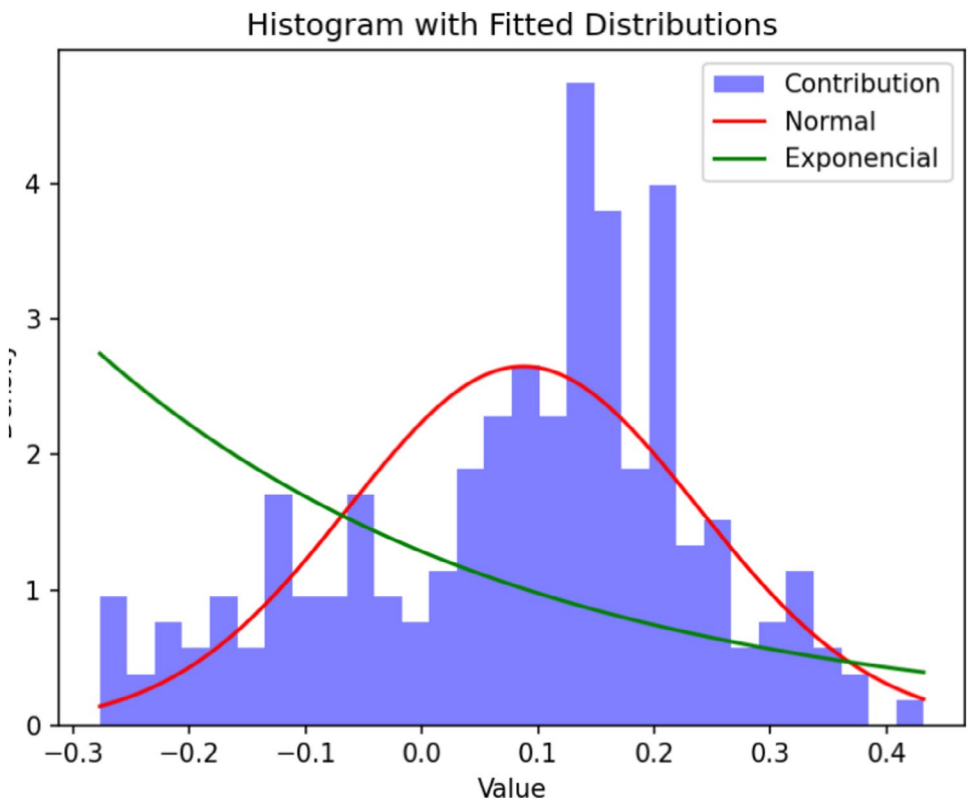


Fig. 13 Plot of the fitting of the Normal and Exponential distributions

intricate distributions, like Log-normal [17], Weibull [18] or Pareto [19]:

Based on the analysis of the Pareto distribution (see Fig. 14), it's evident that the observed p-value is statistically significant, leading to the rejection of the hypothesis that our

Table 7 Results of p-value and AIC

Distribution	p-Value	AIC
Log-normal	0.004375075293021211	−205.0999781232991
Pareto	6.242223683997611e−17	–
Weibull	0.17881847840595044	−219.5301214521686

data adheres to this distribution. Conversely, the hypothesis testing results differ for the Log-normal and the Weibull distributions. Expressly, while the null hypothesis is rejected for the Log-normal distribution, it is not rejected for the Weibull distribution. Visual representations of these findings are available for further examination.

Figure 15 shows that the Weibull and Log-normal distributions only plot positive data values. This is due to the nature of the distribution functions, as they use logarithmic functions. It can solved by adding the absolute value of the minimum of these values as shown in Fig. 16:

With this small translation, the functions fit the complete data. The updated values of the statistics are shown in Table 8:

Differences were not observed when comparing the values. Finally, we will explore the application of various tests to intricate distributions, including Laplace [20] and Cauchy [21].

The p-values obtained are lower than 0.05, suggesting sufficient evidence to reject the null hypothesis that the data follow these distributions. According to this test, the data do not fit the Laplace and Cauchy distributions well. The Akaike information criterion compares statistical models where the AIC value is minimized. In this case, AIC values indicate the model's goodness of fit to the data, whereas lower values indicate a better fit. Here, the Laplace distribution model has a lower AIC than the Cauchy distribution model, suggesting that it fits the data better according to this criterion, as shown in Table 9.

Let us examine both distributions that correspond to our data, as shown in Fig. 17:

The statistical figures are approaching 0.05, even though they remain lower. Further tests should be conducted, as it appears to be a more appropriate fit for this particular type of distribution:

Chi-square [22] test for Laplace distribution:

- Chi-square statistic: 0.2930479719201573
- p-value: 0.9999995017734875

Chi-square test for Cauchy distribution:

- Chi-square statistic: 0.36206120194844665
- p-value: 0.9999986061880377

The p-values obtained from the chi-square test for the Laplace and Cauchy distributions are both close to 1. This suggests insufficient evidence to reject the null hypothesis,

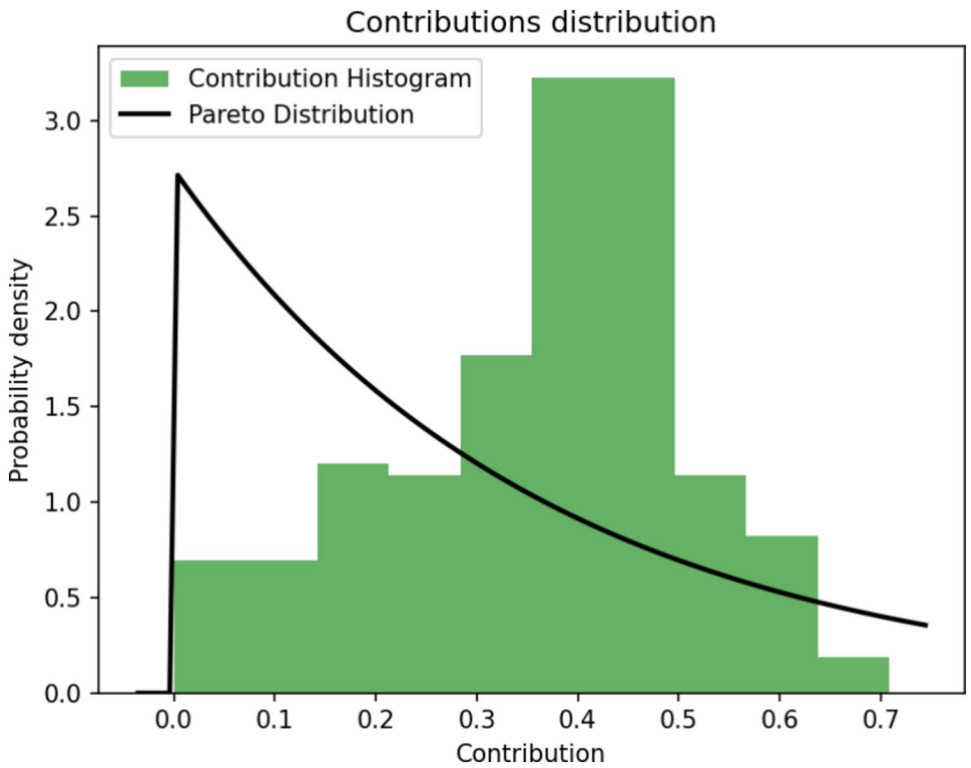


Fig. 14 Plot of the fitting of the Pareto distribution

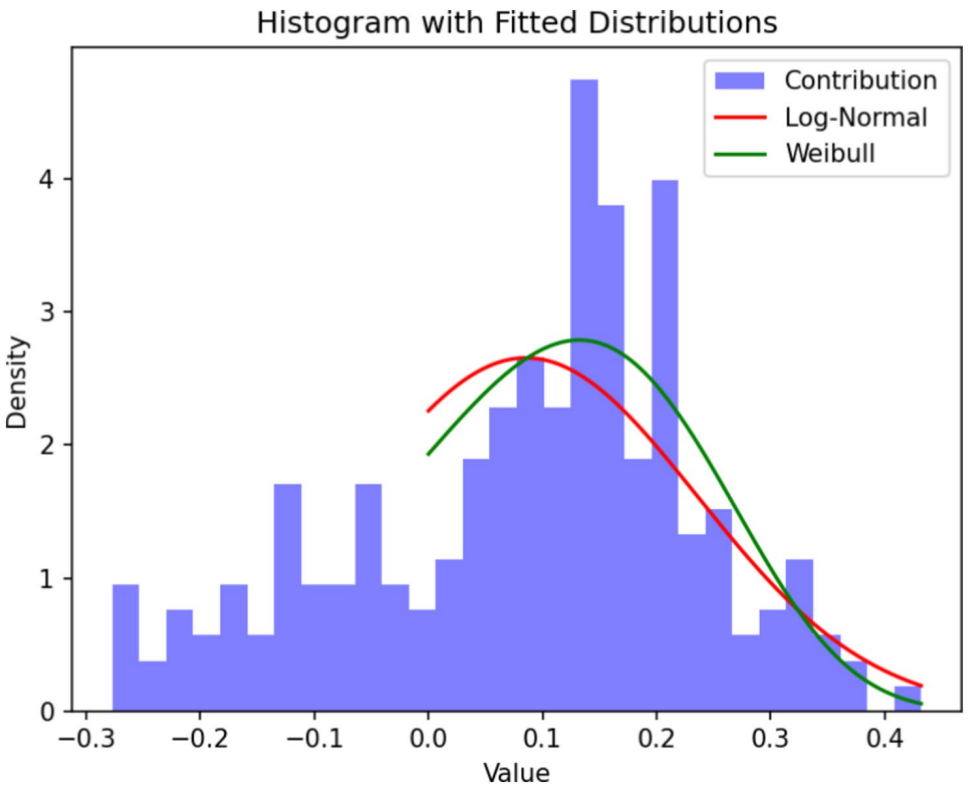


Fig. 15 Plot of the fitting of the Weibull and Log-normal distributions

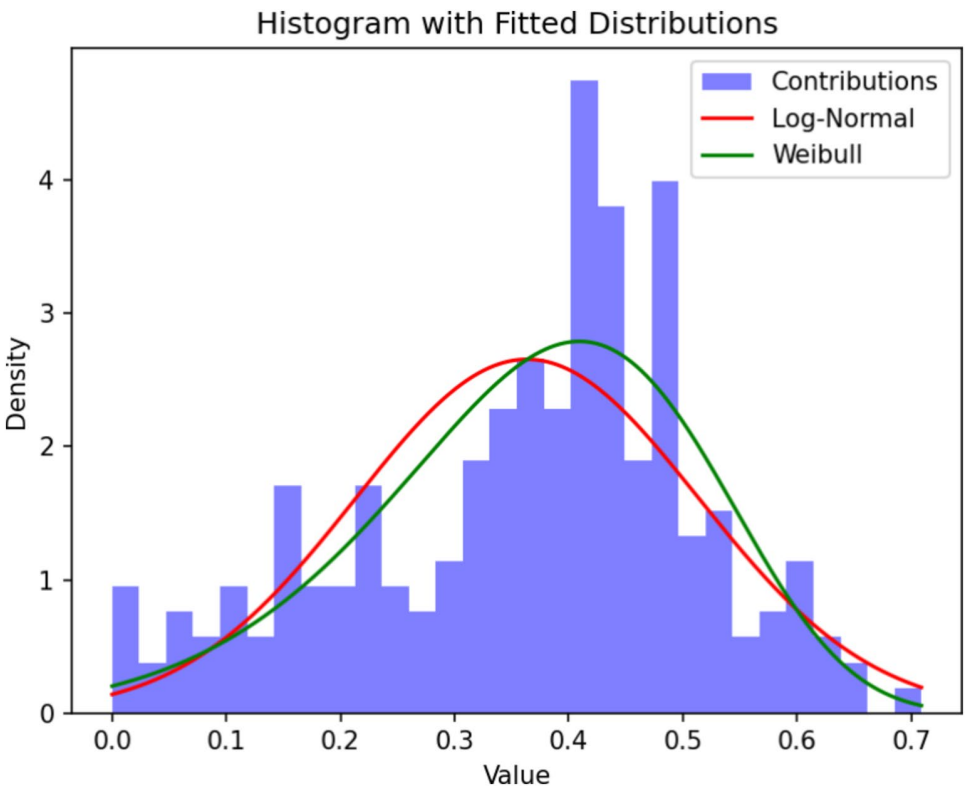


Fig. 16 Plot of the fitting of the Weibull and Log-normal distributions after translation

indicating that the data may be derived from the respective distributions. Nonetheless, it is essential to note that the chi-square statistic value is relatively low, hinting at potential minor disparities between the observed and expected frequencies. It is worth highlighting that these disparities do not hold statistical significance.

Table 8 Statistics values of the functions updated

Distribution	p-Value	AIC
Log-normal	0.005223934193212648	−205.217878391981
Pareto	6.242206228431947e−17	–
Weibull	0.17882331929709394	−219.53012145240803

Table 9 Results of the hold-out procedure

Distribution	Kolmogorov-Smirnoff test p-Value	AIC
Laplace	0.01770109305635071	−196.12725019166515
Cauchy	0.027057473535869958	−144.48472193470096

In summary, none of the distributions examined fit our data perfectly. Still, the Weibull, Laplace, and Cauchy distributions appear to be acceptable based on the analysis and the p-values of the chi-square test, with the Laplace distribution being more appropriate than Cauchy.

Conclusions

This study provides a comprehensive framework for analyzing League of Legends match data using advanced data reduction and predictive modelling techniques. The key findings highlight the critical role of specific in-game metrics in determining match outcomes and demonstrate the effectiveness of PCA and GBR in managing large datasets.

The high accuracy and robustness of the predictive models suggest that these methodologies can be valuable tools for players, coaches, and analysts in the eSports community. By identifying and focusing on influential variables, stakeholders can develop more targeted strategies to improve performance and achieve better results.

Overall, this study contributes significantly to the field of eSports analytics, offering practical insights and a solid foundation for future investigations into the complex dynamics of competitive gaming.

Future Work

To enhance the analysis of League of Legends games, our next goal is to study the game in discrete time slots to examine the game's progress thoroughly. Using a finite set of variables, we can analyze how different metrics change over time and identify various game phases. Additionally, we plan to study the changes in the probabilities of winning or losing throughout the game. This will help us gain a deeper understanding of the game's dynamics.

Also, future research should explore applying these techniques to other eSports titles and consider additional variables that may influence match outcomes. Furthermore,

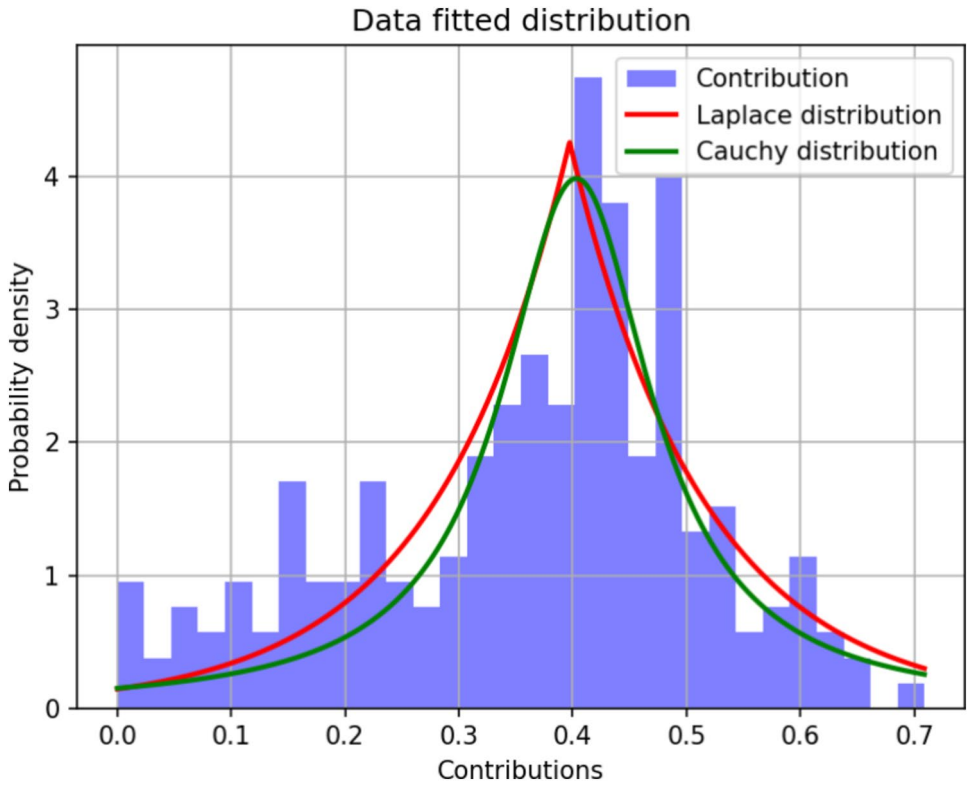


Fig. 17 Plot of the fitting of the Cauchy and Laplace distributions

enhancing data collection methods to include real-time metrics could provide even deeper insights and refine the predictive models. This applies to multiple applications, for instance, well-being and healthcare [23] or smartcites [24], among others.

Acknowledgements We would like to express our sincere appreciation to StartUB from the University of Barcelona and individuals whose contributions and support have greatly enhanced the quality and rigour of this research.

Author Contributions These authors contributed equally to this work.

Funding This research was funded by CDTI (Centro para el Desarrollo Tecnológico y la Innovación) from Spain, under the award number: SAV-20221082

Data Availability This research used public data from RIOT Games servers.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

Research Involving Human and /or Animals Not applicable.

Informed Consent Not applicable.

References

1. Diaz CO, Soler P, Perez M, Mier A. OMASHU: La ciencia detrás del éxito; Big Data e IA en los esports. In: SISTEMAS, ACIS. 2024;170:59–77. https://doi.org/10.29236/sistemas.n170a7.
2. Games R. What is League Of Legends? RIOT. 2024. https://www.leagueoflegends.com/en-us/how-to-play/.
3. Sharma S, Srivastava P. Dimensionality reduction in machine learning: applications and future directions. OSF. 2020. https://osf.io/gpnkb/. Accessed 30 Jun 2024
4. Bahrololloomi F, Klonowski F, Sauer S, Horst RRD. E-sports player performance metrics for predicting the outcome of league of legends matches considering player roles. SN Comput Sci. 2022;4:238. https://doi.org/10.1007/s42979-022-01660-6.
5. Morales-García J, Llanes A, Curado M, Arcas-Túnez F. An exploratory data analysis for league of legends professional match data. In: Workshop Proceedings of the 19th International Conference on Intelligent Environments (IE2023), 2023; pp. 73–81, IOS Press.
6. Junior JBS, Campelo CEC. League of Legends: Real-Time Result Prediction. 2023. arXiv https://arxiv.org/abs/2309.02449.
7. Díaz SD, Díaz CO, Cortés T DF. Electronic system for protection of people victims of domestic violence in areas of interior and exterior. In: AETA 2019-recent advances in electrical engineering and related sciences: theory and application, 2021; pp. 30–41, Springer.
8. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24:498–520.
9. Friedman J. Greedy function approximation: A gradient boosting machine. Ann Stat. 2000. https://doi.org/10.1214/aos/1013203451.
10. Berkson J. Application of the logistic function to bio-assay. J Am Stat Assoc. 1944;39:357–65.
11. Pearson K. Liii. on lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci. 1901;2(11):559–72.
12. Games R. RIOT Developer API Page. RIOT. 2024. https://arxiv.org/abs/2309.02449.
13. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples)†. Biometrika. 1965;52(3–4):591–611. https://doi.org/10.1093/biomet/52.3-4.591.
14. Kolmogorov-Smirnov A, Kolmogorov AN, Kolmogorov M. Sulla determinazione empírica di uma legge di distribuzione. 1933. https://api.semanticscholar.org/CorpusID:222427298.
15. Smirnov NV. Table for estimating the goodness of fit of empirical distributions. Ann Math Stat. 1948;19:279–81.
16. Akaike H. Information theory and an extension of the maximum likelihood principle. 1973. https://api.semanticscholar.org/CorpusID:64903870.
17. Nelson PR. Continuous univariate distributions volume 1. J Qual Technol. 1996;28(2):263–4. https://doi.org/10.1080/00224065.1996.11979674.
18. Cox HL. Fatigue testing and analysis of results. Edited by W. Weibull. Pergamon, Oxford, 1961. 305 pp. 84s. J R Aeronaut Soc. 1961;65(612):844–5. https://doi.org/10.1017/S0368393100076057.
19. Pareto V. Cours D'économie Politique, vol. 1. Paris: Librairie Droz; 1964.
20. Laplace P-S. Mémoire sur la probabilité des causes par les événements. Mémoires de l'Académie royale des sciences de Paris (Savants étrangers). 1774; 4: 27–65.
21. Stigler SM. Statistics on the table: the history of statistical concepts and methods. Boston: Harvard University Press; 2002.
22. Pearson KX. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond Edinb Dublin Philos Mag J Sci. 1900;50(302):157–75. https://doi.org/10.1080/14786440009463897.
23. Bolivar H, Rios S, Garcia K, Castillo S, Díaz CO. Fuzzy logic model for the evaluation of cognitive training through videogames. In: Advances in Computing: 13th Colombian Conference, CCC 2018, Cartagena, Colombia, September 26–28, 2018, Proceedings 13, 2018; pp. 402–417, Springer.
24. Diaz Riveros CA, Beltran Rodriguez KA, Diaz CO, Baena Vasquez AJ. Mobility in smart cities: Spatiality of the travel time indicator according to uses and modes of transportation. In: International Conference on Applied Informatics, 2021; pp. 433–448, Springer.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com