League of Legends: An Application of Classification Algorithms to Verify the Prediction Importance of Main In-Game Variables

Alexandre C. S. Cruz
Informatics Center - CI
Federal University of Paraíba - UFPB
João Pessoa, Brazil
santacruzalexandre@gmail.com

Thaís G. do Rêgo
Informatics Center - CI
Federal University of Paraíba - UFPB
João Pessoa, Brazil
gaudenciothais@gmail.com

Telmo de M. Filho
Informatics Center - CI
Federal University of Paraíba - UFPB
João Pessoa, Brazil
tmfilho@gmail.com

Yuri Malheiros
Informatics Center - CI
Federal University of Paraíba - UFPB
João Pessoa, Brazil
yuri@ci.ufpb.br

***Abstract*—League of Legends is currently one of the most popular video games with a competitive scene involving regional and international tournaments. In this paper, we predict the result of competitive matches using information about games played between the years 2016 and 2020 in the most well-known regional and international leagues, such as the Brazilian League of Legends Championship and League of Legends Pro League. We used several different approaches to train our models based on different variable categories. First, we used economic variables registered in the first 10 and 15 minutes of each match. Then we considered variables that change only from the beginning to the end of the game and do not suffer interference before and after the game. The accuracy of classifiers such as K-Nearest Neighbors, Random Forest, and Decision Tree varied from 68.33% to 85.17%, depending on which variables were used to train the models.**

***Index Terms*—League of Legends, Machine Learning, Match Result Prediction**

I. Introduction

As e-sports gain popularity and, in some cases, surpass the audience of traditional sports [1] they create career and investment opportunities around the world. According to [2], the US and China, the two largest economies in the world, are also the most lucrative future markets for e-sports. The global gaming market is expected to exceed USD 200 billion by 2023 [2].

League of Legends (LoL) plays a major part in this expansion scenario. An LoL match involves two teams of five players and takes place on a map divided into four sections: three main lanes and a jungle. By conquering secondary objectives such as towers, dragons, and barons, the teams manage to achieve the game's main objective, which is the conquest of the enemy base through the destruction of the opponent's main structure, called nexus.

In official championship matches, teams have victory conditions defined by the different types of game variables, such as economics, involving gold and minions, and secondary objectives, involving conquered towers and barons. LoL is a dynamic and complex game, such that the final result of the match is decided through the best management of the cited variables.

In the professional competitive scenario, understanding the importance of existing variables in the match allows teams to better organize themselves strategically to achieve victory more easily. Thus, in this paper, we proposed four different predictive models based on different data approaches with the same objective of understanding which game variables have the greatest influence on the match's final result. For this, we use the Oracles Elixir (OE)[1] dataset that, through match records, provides us with information about the match, such as the winning side and the gold values for each team.

Through the information provided by the OE dataset, we try to predict the outcome of a match using different types of machine learning algorithms. Four different predictive models were developed. For each of them, we test 4 different approaches with the existing data. In the first approach, we use all available economic variables. In the second approach, in addition to the information previously used, more specific variables like the first baron and first tower obtained in the period from the beginning to the end of the match, called Ingame, were used. Finally, in the third and fourth approaches, we try to predict the outcome of a match by using the exact economic information for minutes 10 and 15 of the match. Unlike the authors in [3], here we use only data referring to the teams. We do not use information about individual heroes and players. The purpose of the present work is to

[1]https://oracleselixir.com/

seek a better understanding of the influence of existing game variables on the final result of the match. As a result, we obtained accuracies between 68.33% and 85.17%, depending on the data approach used.

Through an analysis of the feature importance, we selected the best features for each approach. The purpose of this analysis was to assist in the construction of the proposed models in this paper. In section IV, the feature selection process is described in more detail.

This paper is organized as follows: Section II presents the related works; Section III gives a background of the game League of Legends; Section IV describes the methodology used in this work; Section V shows the results of the experiments and Section VI summarizes our conclusions and discusses possible future works.

II. Related Work

This section surveys previous works with machine learning and multiplayer online battle arena (MOBA) games like League of Legends and Dota 2. In [4], the authors investigate the application of random forest, and logistic regression to predict the chance of a win in an LoL match. For that, they used a dataset composed of 724,817 instances referring to unofficial matches involving the 200 best players from 11 servers in the period from January 2016 to September 2016. The results show an average accuracy of 75% in forecasts, with the gold indicator being the most reliable in generating the forecast models.

In [5], the authors presented two-win predictors for the popular online game Dota 2 based on a dataset of 62,000 matches collected using the Steam Web Platform from 11/20/2015 to 11/22/2015. The information collected for each match was: winning side, duration, experience per minute, gold per minute, kills, assists, and deaths. The first predictor uses linear regression and the random forest classifier with the full post-match data. The second predictor uses logistic regression and the random forest classifier with only hero selection data, reaching an accuracy of 73%.

In [6], the proposed model is a deep learning model based on Bidirectional Long Short Term Memory (LSTM) embedding, which considers a combination of champion statistics for each team. A dataset composed of 4,050 instances from different leagues of the game was used. No further details of the data were provided. Compared with other prediction models, the proposed model obtained a prediction accuracy of 58.07% without additional domain knowledge.

All related works seek the LoL or Dota 2 matches victory prediction using techniques used in the literature to achieve the proposed solution. For this reason, they have similarities with the present work. The [5] is the most similar approach because the authors presented two-win predictors based on a dataset of official matches. In [4], [5], and [6] general information variables about experience and gold are used. This present paper has the differential of approaching data from different perspectives. In addition to economic and experience variables, this work proposes to use other types of variables, thus we use variables with information about secondary objectives of the game, such as first baron, first dragon, and first tower.

III. League of Legends Overview

LoL is a MOBA game released in 2009 by Riot Games. MOBAs are characterized by their competitive online essence, where two teams are competing using symmetrical map in pursuit of the goal of destroying the main structure of the enemy base. A typical LoL map can be seen in Fig. 1. There are three lanes: top, middle, and bottom represented in yellow, where minions of each team walk to face each other. There is also a jungle between the lanes represented in green, where there are monsters, which yield gold and buffs (beneficial effects) when defeated. The map also contains structures along the tracks. Towers are represented in blue and red dots.

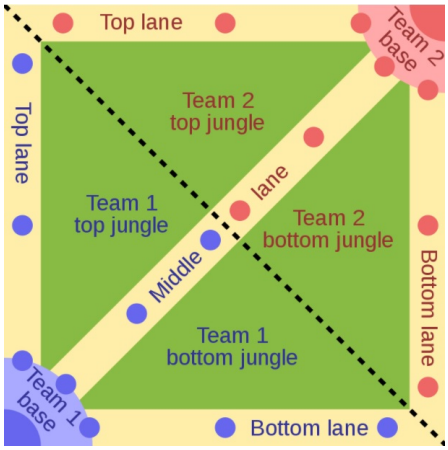

Fig. 1. League of Legends map [7].

LoL matches are formed by two teams with five players each, where each one controls a different character, called champion or hero. Champions are chosen before the match starts so that each champion has unique skills that can be used against other champions, minions, or jungle monsters. A match starts after the team heroes are chosen. Players will seek victory by building an advantage during the game. Champions acquire gold by killing enemies and gold can be used to buy new items that will be responsible for ensuring more damage or more defense.

The main objective of LoL is to destroy a structure within the enemy base, the nexus. To be able to achieve this goal, it is necessary to destroy all the towers and inhibitors in at least one lane. Towers are structures that damage enemies within their radius. Inhibitors are structures that, when destroyed, make the minions of the team that destroyed it stronger and opens the way to the nexus. The game is constantly changing due to constant updates, which seek to establish a balance in the champions' abilities. The current ability balance, which influences which champion is the most useful at each lane, is known as the metagame. Players must always be aware of the current metagame to optimize team composition with the best strategy.

IV. Methodology

The steps of our methodology can be seen in Fig. 2.

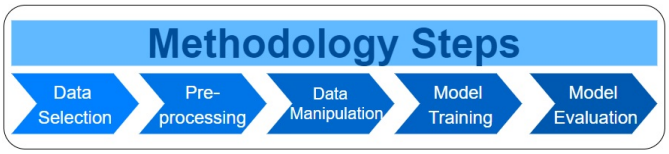

Fig. 2. Methodology steps.

A. Data Selection

The dataset was obtained from the Oracles Elixir (OE) website, where statistics and analyses about matches of official LoL championships are available free of charge since 2015. OE is an independent initiative of the community and has no relationship with the actual company responsible for the game. The dataset has 338,306 instances where the information is distributed by matches so that in each match, there is also a breakdown of the individual information of each player and the complete team. Some variables are time-related, such as the amount of gold and minions, which are registered in the 10th and 15th minutes of the match. Information about teams corresponds to the sum of the values of the respective players, therefore, we chose to use only team variables.

Here we considered matches that took place between the years 2016 and 2020. Our dataset is composed of 18 columns and 56,386 instances containing information about the winning side, gold, the number of minions, the amount of experience, and secondary objectives such as the first dragon, the first baron, and the first tower. Information about gold, minions, and experience is registered on minutes 10 and 15 of the match. Examples of selected features can be seen below:

- result: the team that won the match;
- goldat: the amount of gold in the 10th or 15th minute;
- golddiffat: the gold difference between teams in the 10th or 15th minute;
- xpat: the amount of experience in the 10th or 15th minute;
- xpdiffat: xp difference between teams in the 10th or 15th minute;
- csat: the amount of minions in the 10th or 15th minute;
- csdiffat: minions difference between teams in the 10th or 15th minute;
- firstdragon: the team that defeated the first dragon;
- firstbaron: the team that defeated the first baron;
- firsttower: the team that defeated the first tower.

B. Pre-processing

Columns with non-relevant information such as Id, URL, and data involving post-game information were discarded, and instances that had missing information were removed. The data were scaled using the robust scaler method [8], which uses the quartile interval for data scaling. The target variable is balanced in our dataset, thus it was not necessary to perform any data balancing technique.

The correlation coefficients between all inputs and output were analyzed. In the testing stage, we sought to analyze the learning of algorithms fed with and without inputs with weak correlation to the output, i.e., Pearson's correlation coefficient lower than 0.3. We found that the best situation for Approaches 1 and 2 (described below) occurred with the removal of the weak and less important attributes. However, for Approaches 3 and 4, the best situation occurred when they were maintained.

C. Feature Importance

Here we analyze the importance of all features. We obtained the importance of the features with the decision tree method from the scikit-learn library [8]. The score, which varies from 0 to 1, is calculated based on the usefulness of each attribute in the tree building process.

To improve the construction of our model, we analyzed the learning of algorithms trained with and without the less important features. For Approaches 1 and 2, we found that the best situation occurred with the removal of the less important features, i.e., scores lower than 0.1 and 0.01, for the first and second approaches, respectively. However, for Approaches 3 and 4, the best situation occurred when they were maintained.

D. Data Approaches

Here we considered four different data approaches that used the features indicated in the Approach column from Table I. The first approach consisted of understanding the influence of the main economic variables in the game. In this analysis, we use all economic information, such as the number of minions and gold value, referring until the 15th minute of the match to predict the match's outcome. In the dataset, the match is composed of two instances, one for each team. The final result of the match is indicated by the result column where the winning team takes the value 1 and the losing team takes the value 0. Thus, we verified the impact of the economic factors on the final status of a match. A new dataset comprised of 9 columns and 56,386 instances was obtained, as shown in Table I.

LoL is a complex game, and several variables can influence the final result of a match. Thus, the second approach consists of feeding the predictive model with the set of variables that change during the match. Therefore, in addition to the information from approach 1, this analysis uses more specific information, such as the team that obtained the first baron and the first tower. We sought to verify the result of the match using information from in-game variables. This resulted in a new dataset comprised of 18 columns and 56,386 instances which can be seen in Table I.

In the third approach, we reduced the information used in the first one. We selected only variables with values corresponding exactly to the tenth minute of the match, such as the number of minions and gold value. The led to a new dataset comprised of 7 columns and 56,386 instances which can be seen in Table I.

In the fourth and last approach carried out in this work, the objective was to show the influence of the economy at exactly

TABLE I
COLUMNS USED IN THE DATA APPROACHES

Approach	Features	Description
1, 2, 3	goldat10	gold amount in the 10th minute
1, 2, 3	csat10	minions amount in the 10th minute
1, 2, 3	golddiffat10	gold difference in the 10th minute
1, 2, 3	csdiffat10	minions difference in the 10th minute
2, 3	xpat10	experience amount in the 10th minute
2, 3	xpdiffat10	experience difference in the 10th minute
1, 2, 4	goldat15	gold amount in the 15th minute
1, 2, 4	csat15	minions amount in the 15th minute
1, 2, 4	golddiffat15	gold difference in the 15th minute
1, 2, 4	csdiffat15	minions difference in the 15th minute
2, 4	xpat15	experience amount in the 15th minute
2, 4	xpdiffat15	experience difference in the 15th minute
2	firstblood	the team that got the first kill
2	firstdragon	the team that got the first dragon
2	firstherald	the team that got the first herald
2	firstbaron	the team that got the first baron
2	firsttower	the team that got the first tower
1, 2, 3, 4	result	match result: 1 = victory and 0 = defeat

the 15th minute of the match on the final result. A new dataset comprised of 7 columns and 56,386 instances were obtained, which can be seen in Table I.

E. Machine Learning Algorithms

Three different classifier methods from scikit-learn library [8] were considered in our experiments: K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF). We used the cited methods with python language (version 3.8.3) to implement our models.

The tests were systematized to contemplate the entire methodology presented in this work. Regarding the machine learning algorithms, the Grid Search CV tool [8] was used to search for the best configuration of the hyperparameters in each of them to obtain the best result. The hyperparameters were optimized using the values provided in Table II. All models were evaluated using 5-fold cross-validation, using 42 as the initial value of the random seed. Our models were compared using the average accuracies and corresponding standard deviations after 5-fold cross-validation.

TABLE II
GRID SEARCH VARIATIONS

Algorithm	Parameters	Parameters variations
KNN	n_neighbors	3, 5, 7, 10
	metric	euclidean, manhattan, chebyshev
DT	max_depth	3, 4, 5, 6
	criterion	gini, entropy
RF	n_estimators	10, 100, 150, 200
	criterion	gini, entropy

V. RESULTS

In this section, we present the best results obtained for each approach using the previously described methodology.

A. All economic variables

Gold is one of the main indicators of a team's power in an LoL match. It allows players to strengthen themselves through its use. Therefore, we verified the impact of this factor on the final result of a match. According to the correlation and feature importance analysis, it was observed that the gold difference in the 10th minute (0.43 correlation with the target, 0.14184 decision tree importance score) and in the 15th minute (0.53 correlation, 0.23251 importance score) are the economic variables of greatest impact in this approach. We found similar results for the tested algorithms. The best result found had 73.47% accuracy value with the DT algorithm. Table III shows the results obtained.

TABLE III
RESULTS - APPROACH 1

Algorithm	Accuracy	Standard deviation
KNN: metric= chebyshev, n_neighbors= 10	71.23%	0.0105
DT: criterion= entropy, max_depth= 4	73.47%	0.0050
RF: criterion= entropy, n_estimators= 200	72.95%	0.0096

B. Ingame variables

This analysis consisted of understanding the importance of the set of variables that change during the whole match for the team's victory. In an LoL match, there are other important indicators besides gold, such as the first tower and the first baron that provided a significant improvement in the model's learning and accuracy. The first baron (0.67 correlation coefficient, 0.27225 importance score) and the gold difference in the 15th minute (0.53 correlation coefficient, 0.12079 importance score) are the more important variables here, according to the correlation and feature importance analysis. The best result found had 85.17% accuracy value with the RF algorithm, similar to the DT result with 84.89% accuracy value. Table IV shows the results obtained for this approach.

TABLE IV
RESULTS - APPROACH 2

Algorithm	Accuracy	Standard deviation
KNN: metric= manhattan, n_neighbors= 10	71.72%	0.0106
DT: criterion= entropy, max_depth= 4	84.89%	0.0088
RF: criterion= entropy, n_estimators= 200	85.17%	0.0085

C. Economic variables in the 10th minute

This analysis consisted of understanding the influence of the economy at exactly the 10th minute of the match on the final result. According to the correlation and feature importance analysis, it was observed that the gold difference in minute 10 (0.43 correlation coefficient, 0.22633 importance score) is the economic variable of greatest impact in this approach. This was the approach with the lowest result. All results presented

in Table V were similar, the best being 68.33% accuracy value with the DT algorithm.

TABLE V
RESULTS - APPROACH 3

Algorithm	Accuracy	Standard deviation
KNN: metric= chebyshev, n_neighbors= 10	65.57%	0.0096
DT: criterion= gini, max_depth= 3	68.33%	0.0073
RF: criterion= entropy, n_estimators= 200	67.64%	0.0093

D. Economic variables in the 15th minute

The fourth and last approach of this work consisted of understanding the influence of the economy at exactly the 15th minute of the match on the final result. According to the correlation and feature importance analysis, it was observed that the gold difference in the 15th minute (0.53 correlation coefficient, 0.27142 importance score) is the variable of greatest impact in this approach. Here we found similar results with the first approach. The best result found had 73.76% accuracy value. In Table VI, we can observe all the results obtained.

In order of best accuracy, it is possible to list approaches 2, 4, 1, and 3.

TABLE VI
RESULTS - APPROACH 4

Algorithm	Accuracy	Standard deviation
KNN: metric= chebyshev, n_neighbors= 10	71.66%	0.0149
DT: criterion= gini, max_depth= 5	73.76%	0.0051
RF: criterion= entropy, n_estimators= 200	73.37%	0.0121

VI. CONCLUSION

The aim of this paper was to predict victories in competitive League of Legends matches. Understanding which variables are most influential in the final result of an LoL match is of great importance for the professional game scenario. We predict the teams' victory successfully with the four proposed approaches, and we obtained results with accuracy that varied between 68.33% and 85.17%. Our best result, shown in Table IV, was obtained in the second approach using the RF method.

In the professional scenario, the various existing variables are very well controlled by the teams. Thus, when two teams face each other looking for victory, what will guarantee the positive final result to one of them is better control of the variables.

Through the results of the first two approaches, it is possible to analyze that, although gold is one of the main strength indicators of a team during the match, there are also other indicators, such as the first tower and the first baron, which when inserted in Approach 2 provided a significant improvement in the model's learning and accuracy.

In addition, analyses involving only gold were less successful in the 10-minute stage than in the 15-minute stage. Thus, the performances obtained in the first ten minutes of the game, although important, have less impact in determining the victory than subsequent periods.

Finally, it is worth highlighting the importance of this study for the betting market. Electronic sport is responsible for the movement of a large financial volume in the virtual bookmakers [9]. Therefore, as future work, it would be interesting to test new algorithms or implement neural networks to seek the evolution in the results of predictions.

Another interesting alternative would be to develop an application to assist in the process of collecting detailed data of interest and in the graphical visualization of the collected values. Finally, since LoL is a game in constant evolution, it would also be useful to perform an analysis based on seasonal data and data from individual heroes and champions to get more detailed results.

REFERENCES

[1] J. Lynch. "As NFL ratings drop, a new internet study says young men like watching eSports more than traditional sports." Business Insider, Entertainment, 2017. https://www.businessinsider.com/nfl-ratings-drop-study-young-men-watch-esports-more-than-traditional-sports-2017-9 (accessed May 25, 2021).

[2] AJ. Cortese, W. Chen. "The US and China lead global gaming, but how are they developing the next generation of talent?" Kr Asia, Insights, 2020. https://kr-asia.com/the-us-and-china-lead-global-gaming-but-how-are-they-developing-the-next-generation-of-talent (Accessed May 25, 2021).

[3] Z. Yang and Y. Wang and P. Li and S. Lin and S. Shi and Shao-Lun Huang "Predicting Events in MOBA Games: Dataset, Attribution, and Evaluation," 2020.

[4] R. T. Souza. "Aplicação de algoritmos classificadores para previsão de vitória em uma partida de League of Legends," Advisor: M. N. Cortimiglia, 2017, p. 21, Graduation conclusion work - Production engineering, Federal University of Rio Grande do Sul, Porto Alegre, 2017. https://lume.ufrgs.br/handle/10183/179708 (Accessed May 25, 2021).

[5] N. Kinkade, L. Jolla, and K. Lim, "Dota 2 win prediction," Technical report, University of California, San Diego, Tech. Rep., 2015.

[6] C. Kim and S. Lee, "Predicting Win-Loss of League of Legends Using Bidirectional LSTM Embedding," KIPS Trans. Softw. and Data Eng, vol 9, pp. 61–68, Feb 2020.

AJ. Cortese, W. Chen. "The US and China lead global gaming, but how are they developing the next generation of talent?" Kr Asia, Insights, 2020. https://kr-asia.com/the-us-and-china-lead-global-gaming-but-how-are-they-developing-the-next-generation-of-talent (Accessed May 25, 2021).

[7] S. Berkovich. "Understanding League of Legends Data Analytics" Medium, Snipe.gg, 2019. https://medium.com/snipe-gg/understanding-league-of-legends-data-analytics-c2e5d77b55e6 (Accessed May 5, 2021).

[8] F. Pedregosa et al, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol 12, pp. 2825–2830, 2011.

[9] K. Wimer. "Betting is esports' biggest and most underappreciated opportunity" Venture Beat, Games Beat, 2019. https://venturebeat.com/2019/06/03/betting-is-esports-biggest-and-most-underappreciated-opportunity/ (Accessed June 8, 2021).