University of California, Santa Barbara

League of Legends

Finding Win Conditions and Categorizing Competitive Region from Gameplay Style

Lina Zeng and Eddie Zhang

DS 100: Data Science Applications and Analysis

Professor Alexander Franks and Professor Kate Kharitonva

June 12, 2020

I. Abstract
In our paper, we will be finding the winning factors to the video game, League of Legends, in their professional matches as well as categorizing these matches by their competitive region in order to see which game variables most affect wins and if regions have different play styles. Our data comes from Kaggle, a public online data science community, that collected data from the official League of Legends website. We utilized PCA, logistic regression, and Altair to analyze and visualize our observations. We found that there is a clear differentiation between LCK, the arguably best region, from the other regions and a distinct difference between the following regions, EULCS and NALCS. Our categorization model has a 48.7% accuracy when predicting the match's region among these three which is significantly over 33% accuracy by guessing. This means that regions have distinct play styles that we can analyze in the future to determine how certain regions constantly perform well internationally. Our win model has an 89.7% accuracy to predict which team wins with only early-mid gold difference and neutral objective variables, indicating that neutral objectives can be pivotal in games even though they are optional and can be ignored.

II. Introduction

A. Goal

Our primary goal of our project is to identify patterns in professional League of Legend matches and predict which region the match comes from as well as identify key factors that contribute to a match win.

B. Motivation

League of Legends is not only the most popular PC game worldwide, bringing in 8 million concurrent players everyday at its peak (Messner), but is also the leading esport in the quickly growing industry. During the 2019 World Championships, the professional tournament brought in more than 100 million viewers (Webb). And more recently during the current pandemic, ESPN--*the* American sports cable company--has shown the North American League of Legends Championship Series (NALCS) in lieu of traditional sports reruns.

There is a large disparity between the competitive regions when we see them face each other on the international stage. If we can map specific properties from matches to a particular region, we can investigate how the regions compare by their game play and dominance. Knowing what contributes to a win could also be significant in improving the team's win rate.

C. Background

A quick 2 minute overview of the game (Riot Games) or a comprehensive game review (Berkovich) may better explain League of Legends, but we'll try as well.

To describe League of Legends simply and relevantly to our project: there are two teams (red and blue) of five champions (players) and their objective is to destroy the other team's base. In order to reach the other team's base, they must destroy structures (towers and inhibitors) that protect the base as well as fighting the opposing team. Additionally, there are neutral monsters to kill that can give advantages to the team to help them grow more powerful and defeat the other team.

These objectives include:
Dragon: they spawn throughout the game and give permanent, but subtle, buffs
Baron: they spawn mid-late game and give temporary, but strong, buffs
Herald: they spawn early-mid game and is a monster to be summoned to help the team until it is killed

Another note is that each player is constantly earning gold passively to purchase items to bring their character's stats up, but can also earn gold by killing monsters, champions, and structures.

On the side of esports and the competitive scene, League of Legends currently has 12 competitive regions. Every year, professional teams play against each other during two regular seasons (spring and summer), and fight for final rankings during each season's playoffs. This all builds up to the biggest competition of the year: World Championships (WC) where the regions send their top teams and they fight on the international stage to claim the title of world champions. The metric to determine which teams are sent have changed over the years, but the main factor is the ranking during the summer season.

D. Dataset
We used the League of Legends dataset from Kaggle compiled by Chuck Ephron. This dataset contains professional game match information from almost every competitive region and four international tournaments from 2014-2018. Every match provides key information that can replicate the state of the game minute by minute. This data is appropriate for answering our question since we have the information to analyze patterns in gameplay and can find correlations between the competitive region and game style as well as winning factors.

The dataset provides 11 out of the 12 regular competitive regions that existed before 2018 as well 4 international tournaments that we are not examining today.

III. Questions of Interest
We have 2 questions of interest:
1) Can we differentiate regions by their style of game play?

2) Can we predict wins by early states of the game, namely through neutral objectives and gold difference?

IV. Data and Methods

Our data comes from Kaggle's League of Legends: Competitive Matches, 2015 to 2018, collected by and maintained by Chuck Ephron. It has a CC0 (Public Domain) License.

We will be using almost every variable for our exploratory analysis. Most variables are self-explanatory. Ones that are less familiar are each team's dragons, barons, heralds, towers, and inhibitors. These monster variables specify the minute they were killed as well as any special types. The structure variables specify the minute destroyed as well as location.

We mainly preprocessed the gold difference variable on the original variables. Since gold is provided by the minute, we transformed the variables to have early, mid, and late game gold differences (blue - red). We determined early, mid, and late game to be at 16%, 49%, and 83% of the game length respectively. We also removed all of the variables associated with individual players. For context, an average game is often 35 minutes, but can range anywhere from 20 minutes to 120 minutes if played until completion (no forfeits).

This dataset contains only the WC games in 2014. It contains every game from each listed region from 2015 to January 29, 2018 when the dataset was last updated. However, we are missing one major professional region (LPL) from the data. This may be significant as LPL is a top tier region, often one of the top two contenders for world champions. Another important note is that each region's tournament style is not regularized. The amount of teams that compete and the amount of games they play per season are not the same in each region. They may also change per season. Additionally, only three regions have over 1000 games in the data set and the others average around only 300 games due to being introduced more recently to esports and their region's tournament format.

As pertaining to principles of measurement, the data is relevant to our questions of interest. It is not distorted and is quite precise as we know the state of the game to the minute. The main cost was time to collect the data from the developer's website.

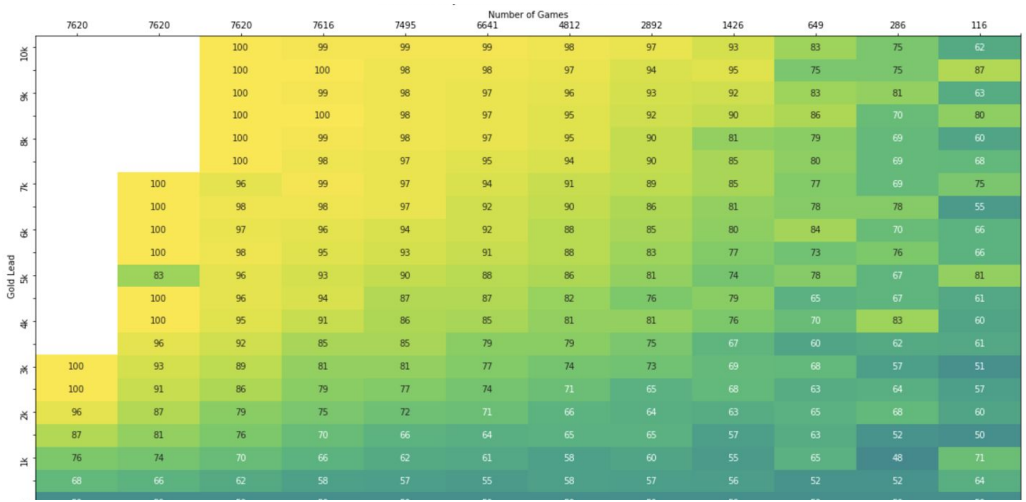


Figure 1: Win Percent by Time and Gold Lead (Ephron)

Chuck Ephron collected this dataset in order "to generate a heatmap of win percent[ages] by gold lead and time using every game in the dataset" as seen in Figure 1 (Kaggle). His analysis did not have any ethical consequences, but further analysis of this dataset could cause harm to those represented in the dataset, with the attached ethical and social costs. Each individual player's name is attached to the dataset and almost every one of their actions and choices in the game are recorded. If analysis was done to defame a player, it could affect the leverage that players have in their negotiations with contracts to teams. Most of these players are also young twenty-somethings that either did not go to college or dropped out. They are vulnerable to manipulation while in pursuit of their dreams to make it in esports.

Although players are likely aware they signed up for the amount of data that would be collected on their game play, they still should be treated as humans. They cannot be reduced down to only numbers and ranks. There is more to a player's worth than just game stats such as team leadership, communication, and cohesiveness. More supportive players may not have the highest stats but may still be an integral part to propelling the team into success. It is easy to misuse and misinterpret the data.

We attempt to differentiate regions between their gameplay by examining each region's game statistics using logistic regression, which is only able to categorize numerical data. We attempt to predict wins by early-mid states of the game, namely through the gold difference, the amount of blue gold, and the amount of red gold. We preprocess our neutral objective variables, reducing it to only the amount each team has with no time stamps. We also remove most of the late game information by removing the gold, kill, inhibitor, and tower variables.

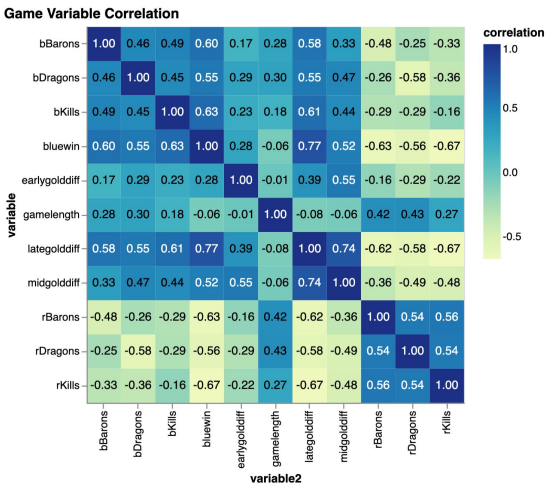


Figure 2: Game Variable Correlation

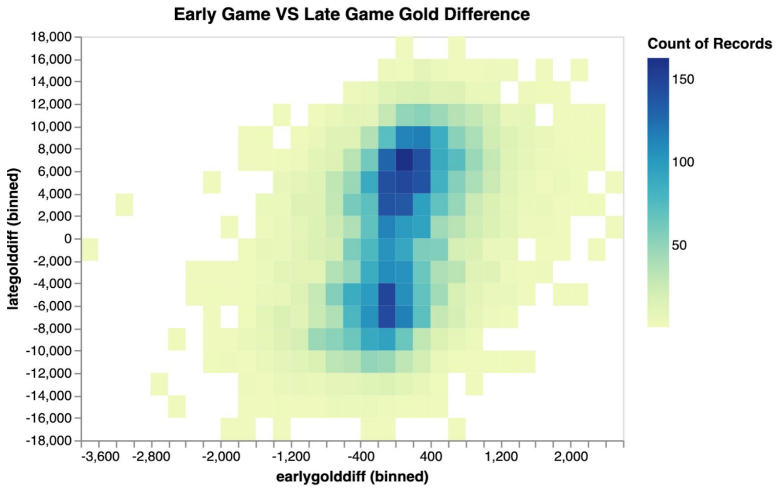


Figure 3: Early Game VS Late Game Gold Difference

We first examined some initial heat maps, trying to find a correlation between any game variable as well as early game and late game gold difference. We can see that the late gold difference had a strong correlation with blue team winning as well as blue kills, dragons, and barons. One correlation we especially wanted to look at is the one between early game and late game gold difference. We wanted to know how much an early game lead could impact the late game. We see that in Figure 3, the correlation is quite low at 39%. In Figure 2, we see that if a team can get a 1.2k gold lead in the early game, they often can keep the momentum and have a late game gold difference as well. However, most games do not have this much of an early game difference. They often stick between 400 gold difference between teams and have a late game gold difference of 7k.

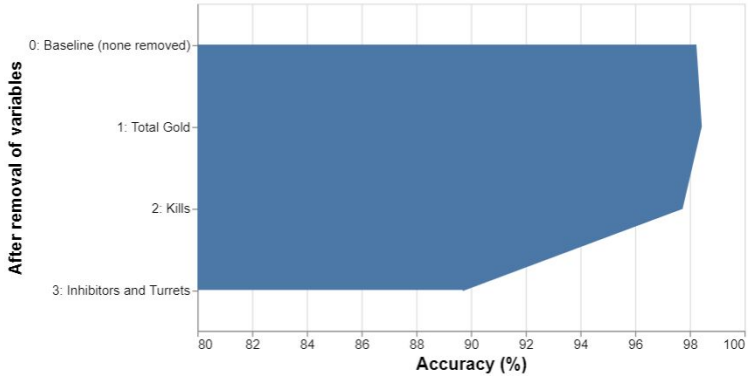
There are no particular outliers nor measurement errors that we can initially see or would find a concern.

V. Analysis, Results and Interpretation

We use the scikit-learn library to do our predictive analysis. We use this for classification with logistic regression when predicting which region the game came from as well as win conditions.

A. Win Prediction Model

Figure 4: Win Prediction Accuracy after Variable Removals



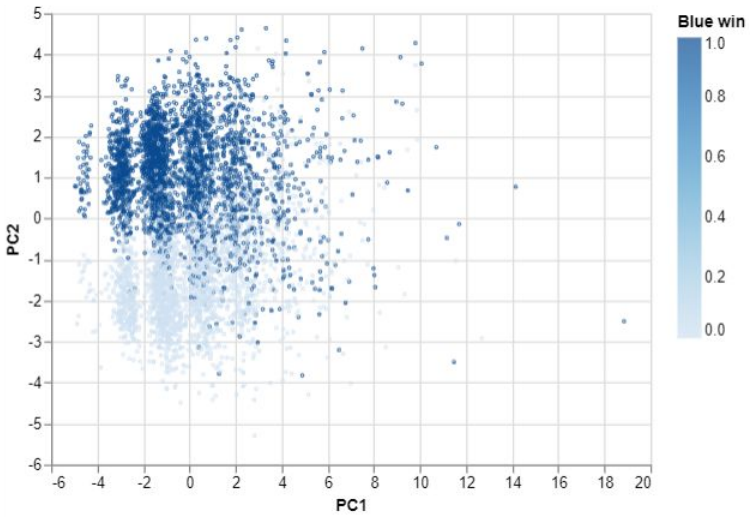
We first began to predict which side would win with every game variable: game length, early, mid, late and total gold, towers, inhibitors, dragons, heralds, of both teams as well as gold difference at the three points of the game. We see a 98.2% accuracy as shown in Figure 4. When we removed the total gold, we actually saw an increase in accuracy which was surprising: 98.4%. We suspect that total gold gives irrelevant information since the difference of gold paints more of a picture of who is winning. We then removed kills from each team in addition, and saw less than 1% in decrease at 97.7%. Again, we believe the total number of kills is not as relevant versus the difference between kills. When we removed only one structure (inhibitor or tower), we also saw barely a difference in accuracy. However, when we removed both, the accuracy dropped to 89.7%. We suspect this is because the number of either structures can be representative of the other structure. For example, at least one inhibitor and the two defending towers must be destroyed in order to start attacking the base and end the game.

Figure 5: Correlation Coefficients of Game Variable and Win (Blue Side)

	gamelength	earlygoldblue	midgoldblue	earlygoldred	midgoldred	earlygolddiff	midgolddiff	bDragons	bBarons	bHeralds	rDragons	rBarons	rHeralds	Year
0	0.456919	-0.003758	-0.064918	0.001802	-0.311468	-0.020671	0.641232	0.649796	1.117982	0.038407	-0.633143	-1.254191	-0.037293	0.04978
1	-0.456919	0.003758	0.064918	-0.001802	0.311468	0.020671	-0.641232	-0.649796	-1.117982	-0.038407	0.633143	1.254191	0.037293	-0.04978

From further examination of correlation coefficients in Figure 5, we see that dragons and barons have the largest impact on game win or loss. They are .65 and 1.12 respectively for blue team dragons and barons contributing to blue side win. As mentioned, these neutral objectives give the team stat increases for the game. If the teams' skill levels are relatively equal, neutral objectives would be the differentiating factor. They are especially important during late game when death timers (how long players stay dead until they revive) are much longer. Kills in the late game are then essential to shift the favor to one side and give that team an advantage for the final push to win the game. For example, a 5 v 4 team fight is very hard to win for the team with a player down. If the team is already ahead, they may have just secured the game. Hence, dragons and barons especially in the late game could be a great indicator of which team will win.

Figure 6: PCA for Professional League of Legend Match Wins



We decided to do a quick PCA analysis to visualize our data.We use this as a quick way to see if this task will be at all possible, or if the data is so clumped together that it would be pointless to analyze. The original data has 23 dimensions and with PCA, but we were able to visualize around 60% of the variance with just 2 dimensions. From our previous data exploration and analysis, we can infer that the prime principal component is the structure objectives (inhibitors and towers) because they had such a large impact on our model's accuracy.

B. Region Categorization Model

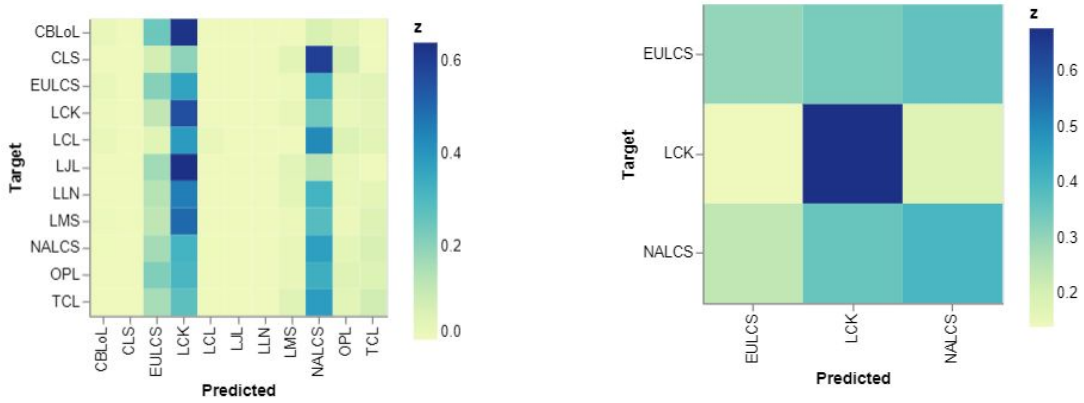


Figure 7: Confusion Matrices for All Regions and Top 3 Regions

At first we let our logistic regression try to categorize every region for each match. It had a 23.9% accuracy but we can clearly see that it best predicted 3 regions: LCK, EULCS, and NALCS. We then further examined the data per region and realized that the most accuracy came from the regions that had the most game data. These three regions all had over 1000 games, the next region only had 778 games, which is a substantial difference. More data allows for more correlation.

We then decided to take the 3 regions with the most games and only use those matches to predict the region because the dataset had much less data on the other regions. We can see that our regression model could predict the games quite well at 48.7% accuracy. LCK was the region that was predicted with most accuracy. Not only does it have the most data, but it also has the most distinct data, indicating a unique gameplay compared to the other regions. EULCS and NALCS were more similar in their game play. This can be reflected in the international results. LCK has the most titles for world championships and is almost unanimously considered the best region above all the other regions (except maybe for LPL which is not represented in this dataset.) EULCS and NALCS on the other hand are often considered among the same tier below LCK and LPL and have similar playstyles.

VI. Conclusions and Future Work

Our final model predicts games that belong to the top 3 regions provided by the dataset with 48.7% accuracy. It also predicts game wins quite easily from just early and mid game gold difference and neutral objectives with 89.7% accuracy. This means that regions have different gameplay and that neutral objectives are significant to winning the match. They can sway the

game or secure the game. This is probably why live analysts during the game emphasize “baron power plays.”

During our project, our biggest problems came from transforming our data. We had to observe each variable and see what was relevant to our dataset and remove unnecessary data. This included removing international tournaments that were grouped with the region variables and transforming the objective variables into numerical data. This not only cleaned our data so our model could use it but also removed extra information (such as time and place of the final tower destroyed) that would make our model predict too easily.

Also as more data has been collected in the past 2 years, we can reuse our logistic regression model to predict the other regions and not just the three with decent accuracy. After recognizing that LCK has a distinct playstyle from our model predicting its games quite well, we can examine the games of the top region in the future. What makes them often dominate the other regions year after year? This would be a good question of interest to investigate in the future.