8-2023

STATISTICAL ANALYSIS AND MACHINE LEARNING TO IMPROVE LEAGUE CHAMPIONSHIP SERIES TEAMS

Alexander Gilles

STATISTICAL ANALYSIS AND MACHINE LEARNING TO IMPROVE LEAGUE CHAMPIONSHIP SERIES TEAMS

A Project

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Information Systems Technology

---

by

Alexander L. Gilles

August 2023

STATISTICAL ANALYSIS AND MACHINE LEARNING TO IMPROVE LEAGUE CHAMPIONSHIP SERIES TEAMS

A Project

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Alexander L. Gilles

August 2023

Approved by:

Dr. Conrad Shayo, Committee Chair & Department Chair, Information and Decision Sciences

Dr. Frank Lin, Member, Reader

ABSTRACT

One area for further study in Esports is the use of advanced analytics from a performance standpoint. This culminating experience project sought to find and implement effective performance analytics techniques, using the most popular Esport (League of Legends) as its subject. The research questions asked are (Q1) How do champions, players, and their associated in-game variables impact the results of League of Legends matches? (Q2) How can machine learning algorithms be implemented to utilize descriptive and predictive analytics for League of Legends most effectively? Additionally, while not an element of the analysis and machine learning model, it is important to discern the importance and scope of the data collected.

The findings are: (Q1) In game variables can be utilized to create descriptive analytics metrics like Champion Matchup Value (CMV), Composition Pace Factor (CPF), or Overall Pace Rating (OPR), and (Q2). The results from the machine learning model focused on correlation and weighting variables, in conjunction with the metrics formulated from answering Q2 can effectively determine a team's chance of winning with an associated confidence rating for that prediction. Additionally, the data collected from OraclesElixir presented a broad set of variables and a substantial number of observations that opened the path to more meaningful analysis than in prior studies. The machine learning model, when fed professional matches of League of Legends saw nearly a 70%

accuracy rating, with a confidence band to determine the likelihood of outcome in each match.

Breaking down the basic statistics into more refined metrics, like CPF or CMV, provides an additional vector from which the game can be analyzed. While some studies aim to use the in-game statistics as they are found, emulating the process of sprots could greatly benefit the world of esports. Creation of advanced analytics allows for a heightened look into how these stats impact games. Additionally, factoring these advanced statistics into a machine learning model which can intake raw in-game statistics, calculate these stats, and utilize them to predict a winner of a match is also beneficial. Many factors come into play with sports, and winners can never be predicted with 100% accuracy, but 70% accuracy is fairly impressive for a model built on original algorithms following brand new advanced statistics. The model also presents a band for how often a certain XFW score results in a winner, thus allowing for more confident predictions with either a high or low XFW score. This model could be further improved to hopefully allow for even more accurate predictions.

Areas for further study would range from data collection method to further expanding on this ML model or building a similar model that explores even more variables. There is more game data available from Riot's API, and even some variables found in the online database, like current-patch, were not included in the final version of this machine learning model but would likely serve only to improve the accuracy of the model.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER ONE:

INTRODUCTION

1.1 Overview

Sports is an incredibly competitive business. With most leagues boasting 30+ teams who all compete for months on end, crowning a singular team the champion of that league. As these leagues grew and evolved, building a winning team became harder and harder. Certain teams lack the funds or the market appeal to compete with industry giants. This paved the way for analytics to make a huge mark on sports as a whole.

In 2002, the Oakland Athletics of Major League Baseball made an improbable run to the most wins during the regular season. Although they had been successful in prior seasons the A's were anticipated to have a significant decline and not be competitive. The team had lost a previous MVP in Jason Giambi and two star players in closer Jason Isringhausen, and outfielder Johnny Damon. Additionally, the team that was constructed before the 2002 season was viewed as incredibly mediocre with a core of players who were not considered desirable. General manager Billy Beane, as well as analysts Bill James, Peter Brand, and Paul DePodesta implemented the use of "Sabermetrics," an analytical approach to find players who contribute to a winning formula. These players would end up costing less than the value they held, and the result was a team with $40,000,000 payroll ($27,000,000 lower than the MLB average that year) keeping up in wins with the highest payroll team in the New York Yankees

($125,000,000). This led to the popular term in sports, “moneyball,” referring to teams looking at analytics to make the best cost-effective decisions in roster construction, rather than relying on reputation, or other conventional methods.

Since the 2002 Oakland Athletics, sports franchises have relied more and more on analytics in their operations. As analytics become more integral to the decision-making process for franchises, the strategies for these analytics become more advanced and diverse. Sports introduces many variables due to the nature of competition, physicality, and human error. For this reason, teams use machine learning and AI in their analysis as this can highlight which variables might need to be focused on and what particular markers are conducive to winning. It is important that all sports use machine learning in their analytics, including Esports.

Esports is a relatively young industry with an incredible opportunity for growth. According to Straits research, the global esports market had an estimated value of $1.178 billion dollars with a projection of $5.743 billion by 2030 (Ryška, 2022 p. 63). This value is a culmination of several specific esports titles and various gaming spaces. This study will be examining the professional League of Legends (LoL) scene for several reasons; it is one of the longest standing professional esports scenes with the first world championship taking place in 2011, the world championship in 2022 was the most viewed esports event with a peak viewership of 5.1 million viewers, and it has global popularity with dozens of professional and semi-professional leagues operating around the world. Additionally, viewership for the world championship has been on the rise

each year, supporting the reports regarding growth of esports as a whole (Gogh, 2023).

1.2 Problem Statement

While it may seem counterintuitive at first considering the restraints presented by video games (possibilities being limited by the constraints of code), esports have an incredibly large number of variables that must be accounted for in their analytics, even more so an esport like League of Legends (LoL). For this reason, it is not immediately obvious where the focus of analytics in LoL should reside. For instance, let's take the National Basketball Association (NBA) as an example; the scope of analytics is relatively limited, which players are on the court, where a shot was taken, what the players are specialized in, etc. League of Legends presents a scope unmatched by traditional sports; there are two teams of players each of whom have their choice of 162 champions to play, each of which can build up to 6 items out of the 200+ in the game, and these are only most surface level decisions in the game.

OraclesElixir, the most robust website for professional League of Legends stats, tracks around 120 different metrics per match. This only represents the externally available stats, there are additional metrics available via Riot internally. Additionally, the League of Legends ecosystem includes a wide breadth of solo queue, with casual players tallying around 500 million games played in 2022. This data can be useful for matchup data and can potentially shape draft decisions for League Championship Series (LCS) teams. To this end, in order to

create a successful machine learning model, the proper data must first be obtained. The practice of data mining must be explored as it pertains to curating a useful League of Legends database. With the amount of available data, it can be difficult to find which stats are meaningful and in which situations. This demonstrates the necessity to build machine learning models to shape this data into useful analytics.

Previous research in the field has been conducted but, in many ways, needs to be further expanded upon. Tian Wang (2018), who researched champion selection, player performance, and various in game factors, noted that these elements could be more useful in future research through the use of correlation and “weight variables” for these factors (p. 35). Hitar-García et al, (2022) felt as if the amount of data was a limiting factor in the effectiveness of their analysis, using only 7583 observations, and 27 variables, asking further researchers to expand on this limitation (p. 173).

1.3 Study Limitations

This study will only be able to explore the data that has been made available for the public, as the Riot internal stats cannot be found online, nor requested. Additionally, an important set of data comes from scrims between two LCS teams. These matches are private and unavailable for analysis.

1.4 Purpose of Study

The purpose of this study is to use publicly available data for the game League of Legends, and build off previous research, to create a machine learning model designed to help teams set themselves up for winning more in the LCS.

The Project will seek to answer the following questions:

1. How do champions, players, and their associated in-game variables impact the results of League of Legends matches?
2. How can machine learning algorithms be implemented to most effectively utilize descriptive and predictive analytics for League of Legends?

This project adheres to the following structure: chapter 2 is a review of relevant literature, chapter 3 details the research methodology, chapter 4 is data analysis and findings, and chapter 5 is a discussion of the findings, conclusions, and areas for future research.

CHAPTER TWO:

LITERATURE REVIEW

2.1 Scope of Review

In order to proceed with the study, it is necessary to first examine the progress made by others in the same or similar fields. For the purpose of this study there are four fields that must be researched as they apply to the topic: Data Mining, Machine Learning, Traditional Sports Analytics, and Esports Analytics. Following the areas for future studies provided by Wang (2018), and Hitar-García et al, (2022), research was geared towards 1. creating a fundamental understanding of the data needed in sports and League of Legends to identify reliable variables, and 2. understanding the principles of machine learning and how they can be implemented into a sports/esports model. There are several contributors to the field of relevant esports analytics for League of Legends, but most previous research is piecemeal, so this project aims to build upon the foundation of prior research. Some of these researchers include Maymin's (2021) analysis of kills and deaths and Hodge's (2021) research into a game of the same genre as League of Legends, which used a combination of anecdotal experience as well as in-game statistics. This review also entails researching analysis done from relevant fields like sports, specifically the analysis done by Sarlis et al., (2020) regarding the NBA and using end of game statistics to formulate advanced stats, as well as Machine Learning research

from Bottou (2014), and Roh et al., (2021) in order to understand the fundamental principles, goals, and techniques of machine learning. Additionally, research was conducted into the data collection methods for League of Legends despite the use of an online data source. Research into how the data was sourced and the complexity of the data provides a better understanding of the data that will be used for the project.

2.2 Data Source

Hellbach (2021), Jack (2022), put forward their research and code for accessing Riot's API for general rank data as well as individual match data. Understanding how Riot's API works, provides insight into how websites source their data repositories. Understanding the sourcing methods provides a clearer insight into what variables might be necessary and exactly what the online data entails. Jack's work focuses primarily on how to access Riot's API using python. From this he accesses personal match data. He then expands upon this to create a loop to collect data from a list of his most recent matches. Hellbach's work primarily focuses on utilizing the programming language R to tap into Riot's API and save the information found into tables to be used for other purposes later. First, he creates a random sampling of players across the four lowest ranks, then he uses that sampling of players to compile match data. Like J, Hellbach uses a loop to compile the match data for the most recent matches of his sampling of players. Both authors had to manage the rate limit that Riot places on their API for developers. This limits developers to 100 requests per 2

minutes. Due to these limitations, it is more effective to utilize online sources for data, but understanding how that data was sourced is important as well.

For future researchers who are looking to source their own data without scraping the API from the ground up, Maldonis et al., (2022), provides an excellent solution. Maldonis et al., created a github repository called cassiopeia, named after one of the champions in league of legends, designed to help those less versed with the API to access the data within (Maldonis et al., 2022). Cassiopeia is an all-inclusive python framework, allowing its users to easily access information to create third party apps based around league of legends data. Providing pre-built tools allows users to focus more on actual analysis rather than figuring out the intricacies of manipulating the API. This framework is a little more limited as users are not building their own tools from the ground up but may provide a good starting point to avoid unnecessary hassles.

2.3 Sports and Esports

Sarlis et al., use basketball and the NBA as their primary area of focus regarding sports analytics. The idea is to take game statistics that could be gathered after every NBA game and find metrics by which efficiency and success could be measured (Sarlis et al., 2020). This took shape in a couple ways: stats within stats (taking the existing stats and determining the value each had and creating new metrics based on that) and combination efficiency (the effect that different combinations of players based on those stats would affect the chances of winning. These strategies are used here for the NBA but provide a blueprint

for sports analytics in general. They take superficial data and allow teams to make strategic decisions after that basic data has been thoroughly analyzed.

Maymin (2021) used Riot's API to create "advanced stats" based on the typically reported stats and used those stats to calculate success metrics for high level play. Maymin's research focuses on the most easily accessible statistic in League of Legends: kills and deaths. This article reframes the way in which the reader perceives kills and deaths, creating a metric for smart kills, and useless deaths. As stated in the article, total-kills is not as reflective on winning as smart kills (kills that impact other elements of the game in a positive way) and total deaths does not reflect losing as much as useless deaths (deaths that could be prevented via more educated play). This is an incredible first step into advanced analytics for league teams as it shapes the way the game should be played away from feeling and towards analytics.

Hodge et al., (2021) takes a look at how to implement analytics into esports in a broader sense. This article focuses on DOTA 2, a Multiplayer Online Battle Arena (MOBA), which is another popular esport in the same genre of game as League of Legends. The data used in this study were based on several factors: personal experience for high level combinations, map and movement data, and in-game stats. This study incorporates algorithms to best determine winning likelihood based on this data.

2.4 Machine Learning

Bottou (2014) explores the practical and conceptual applications of machine learning as it relates to machine reasoning. Bottou posits that rather

than creating a perfect machine designed around having inferences for every scenario, slowly building more training systems algebraically will be more likely and beneficial. Bottou covers concepts that seemed more novel at the time but have since become typical of machine learning like the concepts of association and dissociation. Bottou's approach, however, is reflective of a solid foundational understanding of machine learning and serves as a starting point postulating the building of ML models.

Roh et al., (2021) explore the overarching concepts behind machine learning, as well as bridging the gap from data mining to the learning process. Machine learning is dependent on data mining, and the process from data collection to final product is important to understand. This article breaks down the process in a tree-like structure, showing the different processes that must be completed to take collected data and shape it into its necessary components for machine learning. The focus of this study is primarily on the use cases of big data for machine learning applications.

Zhou et al. (2017) discuss the benefits of machine learning, and its use cases concerning big data. While this article examines some of the same ideas as the previous one, it dives deeper into the methodology for the algorithms in machine learning. While a large amount of machine learning writing focuses on volume, velocity, and variety of data, this article takes special note of veracity and value of data. Machine learning can be used to measure the veracity and value of data to ensure that further learning is meaningful. Additionally, they

explore deep learning as well as superficial looks at training methods for deep learning.

CHAPTER THREE:

RESEARCH METHODOLOGY

Before answering the research questions the scope of the data that is available in OraclesElixir must be examined. The data that is available in OraclesElixir contains 132 variables across 149,232 observations. For comparison this, vastly out-shadows the 7583 observations, and 27 variables used by Hitar-Garcia et al (2022). These variables will be incorporated into the machine learning model using techniques such as linear regression via the use of correlation as well as classification, typically the use of clustering. The efficacy of this data will be determined as the project begins to take shape.

3.1 Machine Learning/Data Cleaning Goals

Q1. How do champions, players, and their associated in-game variables impact the results of league of legends matches?

The process for answering Q1 will be more directed but also more involved. The answer to this question starts with cleaning the data set, categorizing the variables, and then followed by statistical analysis. The variables that will be relevant for the scope of the machine learning model must be identified. In order to ensure that the data is relevant and useful, it must be categorized. Data relevant to the following categories will be kept and cleaned while the rest are culled:

- Label Champions by Type
  - Game Speed
    - Slow vs Fast
  - Playstyle
    - DPS vs Tank vs Support
- Determine Team Composition per team per game
  - Based on the collection of five champions and their corresponding types
- Determine a champion's various statistics in relation to their specific lane opponent
- Determine individual player's stats
  - Basic Statistics
  - Champion's game speed, playstyle, and combination of the two
- Determine team stats as well as their team composition tendencies.
- Understanding winning contributions
  - Based on several factors centered around all the information gathered above

3.2 Labeling Champions

The champion categories that must be determined are based around the champion's preferred game speed, and the champion's role. Although these categories could be labeled anecdotally, League of Legends players may have

misconceptions about what makes their champion successful. A champion that may be perceived as an early-game champion may scale better than the community gives credit, and this could be reflected by a better win rate in late-game scenarios. A champion who has multiple build paths may have one that skews more towards the tank spectrum which could have a higher win rate despite the damage-oriented build being more popular. These discrepancies grant credence to using a machine learning model in order to properly label champions, so the patterns of winning can be more accurately interpreted. There are two primary methods that will be utilized to classify champions, one will be determining buckets for different champion categories by determining bins, and the other will be utilizing cluster analysis.

First the game pace of all champions will be determined. Both the bin and the cluster methods will be utilized to help determine game pace. For the bin method, first game times will be arranged into 3 buckets, then champions will be evaluated for their win rates within these buckets. Buckets will be arranged via a K-Means clustering system.

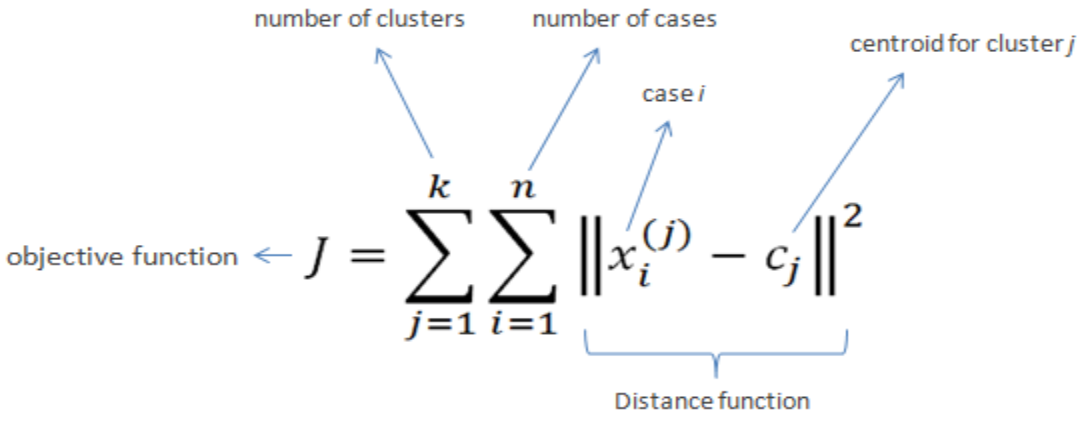


Figure 1 - K-Means Cluster Formula (Sayad, 2015)

Based on these win rates champions will be assigned a game pace value, which will be a numerical representation of their ability to win games during the correlating game length. For example, if a champion wins a high percentage of their games that end in the early-game bin, their early game value (EGV) will be higher. This is done in order to be able to assign a team composition an overall game-pace rating by adding all five selected champions' values together. These values will also be tallied in order to determine a champion's overall performance value, which shows how well they perform through all stages of the game, and if their value in any one category outweighs the weakness of another. Additionally, from these bins, the champions will be assigned a game pace label, which is non-numerical. Depending on the win rate from the first, second, or third bin, champions will be labeled as Early-Game, Mid-Game, or Late-Game champions respectively.

Champion playstyle is a vital element of team building in League of Legends. Whether a team decides to draft a balanced collection of champions, or over-index into one aspect can have massive ramifications to how a game plays out. Similar to a game of rock-paper-scissors, certain team compositions may have advantages over one another, but the best way to determine this is via machine learning. Similarly, to how game pace was determined, champion playstyle will be determined via the bin but also cluster methods (Costa, 2019).

The bin method for play style will be a more involved process, as there is not a single metric that determines play style as is the case with game pace.

Play style is broken down into several aspects, some of which won't even be explored in the scope of this project. For this machine learning model, play style will include damage capabilities and tank capabilities. Damage will be measured using a combination of the Damage Per Minute (DPM) stat and the Damage Share stat. For DPM, a number of bins will be created spanning the damage stats across all games, with each champion from each game being retroactively assigned an associated damage value based on the bin in which it appeared. The higher the bin number, the more damage that champion contributed to that game. Once these values have been determined, the mode for damage values for each champion will be assigned to them. Additionally, the damage share stat will be utilized in a similar fashion to the bin method used for game pace. A number of bins will be created based on damage share and the win rates for each champion within each bin will be determined. A second damage value will be assigned to each champion based on the highest win rate within these buckets. This process will be repeated for the tank capability values based on the Damage Taken Per Minute (DTPM) and Damage mitigated stats. To provide context to damage and tank values that have been assigned, clustering will be employed to label the champions as Carry, Bruiser, Tank, or Utility champions (see Figure 4). If a champion has high damage values, but low tank values they are labeled as a Carry. High Damage values with high tank values label a champion as a Bruiser. Low damage and high tank results in a tank designation. Finally, a low value in both damage and tank results in a Utility designation.

3.3 Team Composition

Now that all champions have been assigned various values and labels correlating to how they impact the game, they can contribute to determining a team's composition. From the champion labeling, each team can now be assigned a cumulative tank value and cumulative damage value, as well as the team's balance of game pace. Additionally, the tallying of champion descriptors on each team provides another vector of analysis. The details of a team's composition can extend to metrics outside the scope of this study, including poke compositions, sustain compositions, or dive compositions. From the details that are contained within the study benchmarks can be set for damage and tank values that contribute to winning teams, as well as comparing those values to that of the team compositions.

3.4 Lane Matchups

While overall team composition is an incredibly important aspect of League of Legends, individual lane opponents can drastically shape the outcome of the game. Lane matchups and lane counters are a central figure to the draft phase, and in order to best draft a winning team, these matchups must be considered in conjunction with overall team composition. Each champion will have its stats measured against each lane opponent, and these values will be averaged out. From this analysis it can be determined what champion picks are determined to be winning into any lane opponent. Additionally, using the labels, champion effectiveness will be evaluated into each possible category of lane opponent.

3.5 Player Statistics

Similarly, to how the champion statistics were gathered, now individual player stats will be tallied. Not every player can play every champion in the game, and even the champions they can, will have different levels of mastery. Additionally, some players may be more skilled with certain types of champions than others. From this data it can also be determined which champions players best match-up into as well.

3.6 Team Statistics

In the same way that players' individual stats were collected, so are the statistics of each professional team. Similarly, to players, teams have winning strategies, composition proficiencies, and may not be able to play certain styles. This data can then be utilized in determining team vs team matchups as well.

3.7 Creating a Winning Formula

How can machine learning algorithms be implemented to most effectively utilize the available data for League of Legends?

The process of answering Q2 involves putting together all of the efforts from sections 3.1 to 3.6 and implementing a combination of correlation factors for several of these statistics as they relate to winning, to creating algorithms by weighting these statistics based on those correlations.

Winning in league of legends is a complex prospect that, as other sports do, relies on individual performance on the day. The metrics that were recorded in the above sections will assist in determining winning players, champions, and team composition. There are several factors that contribute to a team winning in

league of legends, and this vast number of variables are constantly changing from game to game. In order to first determine the accuracy of the statistics that have been gathered, it is important to understand the value that each metric has contributed to winning. Take, for example, a game in which a Jungler playing the champion “Wu-Kong,” wins. It is not enough to simply look at the stats and attribute this victory to the Wu-Kong pick. The factors that are included are the player playing the champion, the champions on the rest of the team and the overall team composition, as well the opposing jungler, their champion, and the opposing team composition. For this reason, the first step is to create a metric that determines the relative success as it relates to each individual aspect.

For each player it will be determined how often they tend to win on each champion opposed to their overall win-rate. This will help to determine if a player’s chances of winning go up or down based on the champion they pick. Once this has been determined for all players, with this champion pick, an aggregate score will be tallied. This will reflect how often all players are increasing or decreasing their win-rate in relation to this specific champion. This will be known as the champions Pick above Replacement score (PAR), to determine the rate at which it could potentially help a team win or cause their loss (see Appendix A). This can be seen in figure 2, where *a* represents Player Champion selection, *b* represents Champion, *w* represents wins, and *gp* games played.

$$PAR = \frac{a_w}{a_{gp}} - \frac{b_w}{b_{gp}}$$

Figure 2 - PAR Formula

Additionally, relative values must be assigned to each matchup within the game. This will determine the relative rate at which a champion either wins or loses against their appropriate lane opponent (Champion Matchup Value or CMV). The relative win rates of the two respective teams also come into play, creating a Team Discrepancy Factor (TDF). Finally, by determining in which stages of the game each team's composition had the advantage, and correcting for that individual team's win rate, opposed to the win-rate of that composition, a Composition Pace Factor (CPF) is created for the model as well. Utilizing this the model can appropriately scale wins based on whether that win was expected or not. These metrics will be weighted to determine how much each tends to contribute to the chances of a team winning or losing. From this teams will be able to better draft against their opponents and towards the needs of their players.

CHAPTER FOUR:

DATA ANALYSIS AND FINDINGS

4.1 Goal of Findings

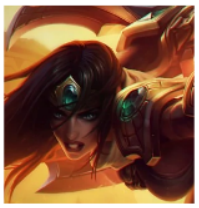
The questions being answered in this project are iterative, so it is important to gain insights from the findings of Q1 before answering Q2. The findings from the process of Q1 will feed the appropriate metrics into the machine learning model to effectively answer Q2.

4.2 Champion Labeling

Q1. How do champions, players, and their associated in-game variables impact the results of league of legends matches?

Q1 findings will be based on the process of labeling champions, analyzing lane, player and team stats, as well as analyzing all prior factors as they correlate to winning. The first step of the process of building this model is to categorize champions. Champions are categorized in two separate fields, their game pace, and their playstyle. For game pace, champions assigned a value for EGV, MGV, LGV, and Overall Pace Rating (OPR). OPR is an aggregate of the pace values at all other stages of the game, weighted based on not only the champion's performance at that stage of the game but also the frequency that games ended in that game state. Champions with a high OPR either have a relatively high performance throughout all stages of the game or perform so well in certain stages that it outweighs their weaker points. Figure 3 shows an example of this

as Sivir (shown top) has a higher EGV and LGV but Rumble (shown bottom) has such a high MGV that it outweighs his negative LGV and he maintains a higher OPR.



EGV	MGV	LGV	OPR
↑ 4.9	↓ 1.3	↑ 1.5	↓ 0.8

EGV	MGV	LGV	OPR
↓ 2.3	↑ 14.8	↓ -7.5	↑ 2.7

Figure 3 - Example of Game Pace Values on Champions

The champions' specific play style is labeled within two specific subcategories: Damage and Tank. Within the damage category champions are assigned a Damage Per Minute Value (DPMV) and Damage Share Value (DSV). DPMV serves as a representation of a champion's overall damage capabilities regardless of the other four members of the team. DSV considers how much of the team's overall damage is dependent on this particular champion. These values will trend in the same direction but have minor differences that may have relevance depending on the champion. For the tank category, champions are assigned a Damage Taken Per Minute Value (DTMV) and a Damage Mitigated Per Minute Value (DMMV). While these serve a similar purpose in game-sense,

function differently overall, and may also have relevance depending on the champion.

Additionally based on the above values, champions are assigned non-numeric descriptive labels as well. Champions are classified as early-game, mid-game, or late-game, based on their EGV, MGV, and LGV. K-means clustering is used in order to classify the champion's playstyle. Using DPMV and DTPM a K-means cluster visualization was created, showing relatively clear delimitations between utility. tank, bruiser, and DPS champions.

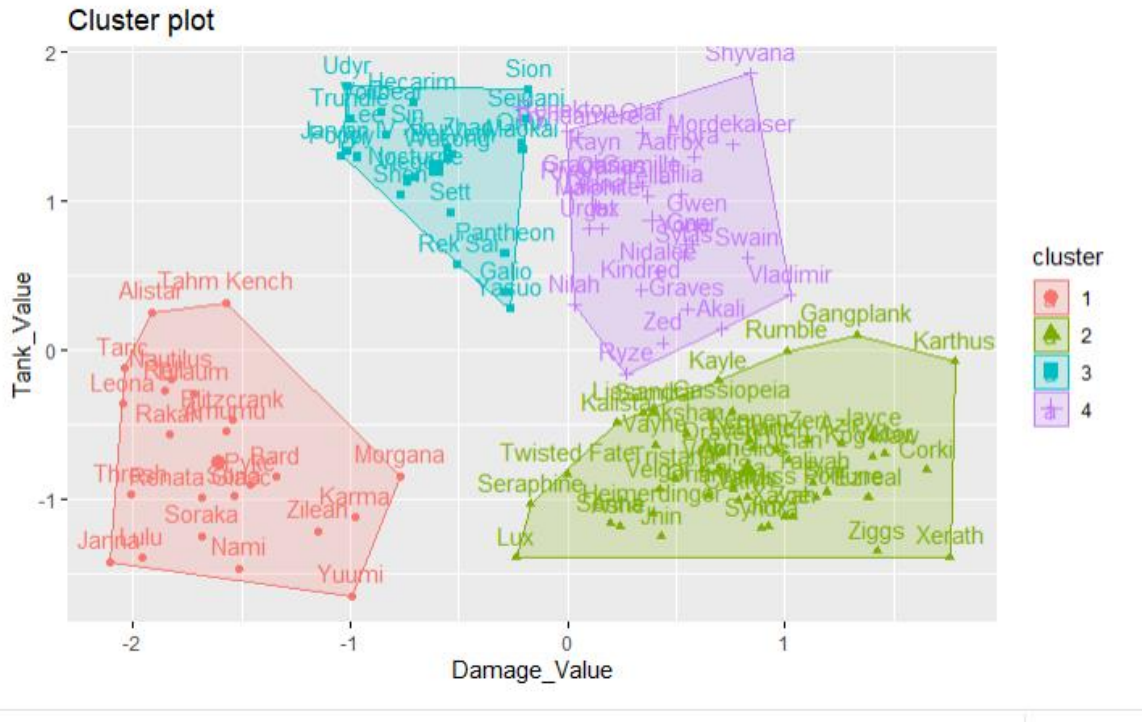


Figure 4 - Champions Sorted into Clusters by Playstyle

4.3 Player, Team, and Lane Findings

Player data collected for this study revolved around the categories from champion labels as well as metrics surrounding win rates. Players have been evaluated based on their efficacy on certain champions, comfort in certain playstyles as well as game pace. Similar analysis was conducted on teams, with their winrrates being calculated based on the champion labels gathered from the first portion of Q1. Lane data combines the combination of lanes and champions by game in order to determine the winning percentages of all champion matchups. As with player and team, these stats relied primarily on the categorical labels that had been generated for each champion. These findings will be combined and utilized as the machine learning model is built for Q2.

4.4 Winning Metrics and Correlation

Q2. How can machine learning algorithms be implemented to most effectively utilize the available data for League of Legends?

Four primary metrics serve as the backbone for the culmination of the machine learning model in this study: CMV, TDF, PAR, and CPF. These values will be calculated, then have their correlation towards winning determined.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Figure 5 - Equation for Pearson's Correlation Coefficient (Glen, 2013)

The Champion Matchup Value (CMV), which is derived from the aforementioned lane matchup data, is a relatively simple concept. This value determines which two champions are matched up against each other in their respective lane and determines the win-rate for each in the particular matchup. This value is aggregated across all five positions to determine a team's CMV. CMV was found to have a correlation of 0.297 to winning.

The Team Discrepancy Factor (TDF) is similar in concept to CMV but executed differently. Unlike CMV, TDF does not take into consideration the direct win-rate of one team against another, instead it compares the two-season long win-rates of the individual teams. As TDF approaches zero a team's likelihood of winning increases, whereas a team's chances of winning are dramatically low as it passes 1. TDF was found to have a correlation of -0.442 to winning.

Next is the Pick above Replacement score (PAR). PAR is notable as it takes into consideration the champions that are being picked, as well as how much that particular champion influences a team's chances of winning. Each individual PAR is aggregated into a total Team PAR. Total PAR was found to have a correlation of 0.04 to winning.

Finally, there is the Composition Pace Factor (CPF). CPF takes a team's ability to play certain compositions, and compares that relative to how well other teams play with that composition, as well as the team's overall win-rate. CPF was found to have a correlation of 0.431 to winning.

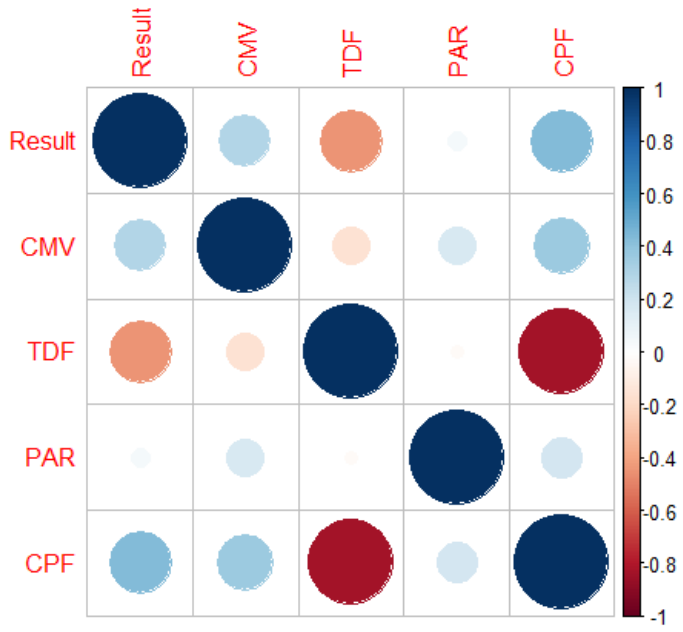


Figure 6 – Correlation Matrix of Advanced Stats and Winning Results

These four data points are calculated with their relative value using their correlation to winning, then added up to an aggregate to give the Xander Factor of Winning (XFW) (see Appendix B).

$$XFW = [(M \cdot M_{\rho}) + (D \cdot D_{\rho}) + (P \cdot P_{\rho}) + (C \cdot C_{\rho})] \cdot 10$$

Figure 7 – Final Aggregate Equation Derived from ML Model (XFW)

XFW when tallied in relation to the opposing team shows the likelihood of winning based upon the ML model. When a team's XFW score is higher than that of an opponent, the model favors that team to win. The degree to which the XFW score is higher determines the confidence the model has for the team winning. When used to test a sample of completed professional games this

particular machine learning model was able to successfully predict nearly 70% of games presented to it.

	Accurate Prediction	WR
1	Yes	0.6964163
2	No	0.3034689

Figure 8 - Final Results for Predicting Games

CHAPTER FIVE:

DISCUSSION, CONCLUSION, AND FUTURE STUDIES

5.1 Overall Results

Through building a machine learning model based around professional league of legends, this study was able to uncover some interesting insights. Through an interactive process and building insights upon new information gathered, the model can provide users with draft strategies far beyond surface level analysis. Some of the methods utilized in building the model ended up being less fruitful than initially suspected. This is to be expected in any sport, as no matter how good a strategy is, it is up to a player's performance on the day to either execute those strategies or ensure their opponents do not overcome the odds. This is especially true with League of Legends which has so many variables from champion select alone that no strategy can even come close to ensuring victory over an opponent. This becomes apparent when correlating most individual stats to winning a game. It is for this precise reason an intricate model is built, in order to inform teams of the strategies that best put them in winning positions.

Q1: How do champions, players, and their associated in-game variables impact the results of League of Legends matches?

5.2 Champion Labeling Insights

The labeling process proved to be fruitful as many insights were gained that run contrary to the theories heading into the project. League is a game of champions with diverse skill sets and abilities. For this reason, this machine learning model had a significant amount of data dedicated to labeling champions based on playstyle. Once champions received these labels however, it became abundantly clear that playstyle was much less of a factor in winning. Nearly all metrics related to damage and tank properties had virtually zero correlation towards winning. The most useful insights that could be derived from the damage and tank stats is that over-indexing in one or the other tended to have a negative correlation to winning. Balanced team compositions obviously were a much better choice, but these compositions ended up having so much parody there was relatively little information that could be gathered from them. Not all is lost when it comes to gaining insights from labelling champions, as game-pace proved to be a much better indicator of winning.

While game pace suffered some similar issues when it comes to parody, there was a noticeable difference in the win rates based on different paced compositions. Early and mid-game drafted teams showed higher win rates than that of their late game counterparts. This is perhaps a reflection of agency in early to mid-game picks as a champion specifically geared towards late-game must survive until their power spike. Early game champions and mid game champions however thrive until that point and may have garnered enough power

to weather the late game storm better than late game champions can survive the early game.

Additionally, it became apparent that there was a clear method to overcome the parody of drafting similar compositions. Each champion having a specific value representing their strength in each phase (EGV, MGV, and LGV), the arms race in the draft quickly shifts towards beating your opponent in these specific values. Each champion had each stage measured, the total aggregated, and it was determined which team won each phase. This was represented with brackets, and a number inside to represent each phase [#,#,#].

	Comp_Pace	Comp WR
1	0.0.0	0.3927732
2	0.0.1	0.4373500
3	0.1.0	0.4761905
4	0.1.1	0.5368217
5	1.0.0	0.4631783
6	1.0.1	0.5238095
7	1.1.0	0.5626500
8	1.1.1	0.6069330

Figure 9 - Win Rates Based on Different Compositions

These numbers represent the early game, mid game, and late game respectively. Winning any phase is denoted with a 1, and losing is denoted with a zero. This representation of compositions compared to their counterparts proved to be extremely valuable as there were noticeable trends. Teams which tended to have the advantage in mid game drafted compositions, denoted by a 1

in the second digit, outperformed compositions with similar advantages either in the early game or late game. For instance, a team with a 0.1.0 rating typically won more (47.6%), than either 1.0.0 (46.3%) or 0.0.1 (43.7%). Additionally, this pattern remains true when a team has an advantage in 2 of the 3 stages of the game. Mid-game has the highest impact in all these categories, followed closely by early-game but teams drafted with a focus on late-game, while having a disadvantage in the other stages, clearly put themselves in a harder position to succeed.

5.3 Lane, Player, and Team Insights

Data collected based on lane matchups, player statistics and team statistics generally served as the building blocks for creating a machine learning model capable of evaluating winning League of Legends. The information gathered in these sections could still prove to be valuable in many scenarios.

Although not necessarily radically important for this model, there are certainly insights to be gained. These insights would be more useful in a team setting, in a very guided structure. This data would be useful in scouting prospects, scouting opponents, or targeting short-comings for players.

Team data was a bit more useful than player data, as relative win-rates were used in the final machine learning model. Similarly, to players, there are useful insights in the data surrounding playstyle that could be used in a more targeted way. It is important that a team's win-rate, and tendencies be noted as League of Legends operates much more as a collective of all 5 players than

being focused on an individual. Understanding the relative win rates of teams in certain scenarios is important in determining the chances of winning.

Lane matchups were by far the most useful of these categories as they not only provided great insight for determining winning patterns, but also have a great deal to offer for further research. While taken at face value, lane matchup stats and win-rates are valuable, but when combined with the data found from players, and teams, can create a relative value that further reflects on certain champions strengths.

Q2. How can machine learning algorithms be implemented to most effectively utilize the available data for League of Legends?

5.4 Predicting Wins in Professional League of Legends

The above insights all played a valuable role in building a foundation for understanding what determines winning in League of Legends

CMV has a 0.297 correlation towards winning. Although that number may seem low in terms of correlation there are two factors that must be remembered: 1) Sports statistics will typically have lower correlation values, due to the nature of the industry, and dependence on direct competition of the highest level, and 2) this is only a piece of the puzzle when it comes to building a model that can accurately determine winning League of Legends.

TDF is sourced differently than CMV since teams do not play against each other enough to generate enough meaningful data. Although this could change if

access was granted to scrim data, it is unknown if that data would provide accurate insights due to the chaotic nature of scrims. Also, unlike CMV, TDF's correlation to winning is negative, but also much stronger at -0.442.

Due to the nature of Total PAR, its value tends to be low, sometimes close to zero as teams may pick some champions with a positive PAR and some with a low PAR. Due to this variance PAR has the lowest correlation to winning at 0.04. Although it is not as impactful in comparison, it can have some drastic effects on the game, and may move the needle one way or the other.

As stated earlier, although damage and tank values were not very useful in the final stages of the machine learning model, champion pace showed tremendous influence over the game. TDF serves as the final adjustment in the machine learning model as it takes a team's entire composition into consideration holistically. This has the second most significant correlation to winning, barely sitting behind TDF at 0.431.

The correlation of the above metrics forms the basis of the weighting in the machine learning model. For this reason, when the XFW is calculated, it reflects the reasoning stated above for each of these metrics.

A 70% accuracy in predicting winners is a relatively successful benchmark for this model. Considering the nature of sports, it can be tough to predict a winner purely on pre-game values. The model also has a confidence band in its predictions. The X-axis represents the XFW score that a team has and the Y axis represents the likely outcome. From this metric it can be determined which matches are close to a coin-flip, as at a XFW score of 0, the likelihood of winning

is 50%, and decisions could be made in draft in attempt to skew the XFW higher, and thusly give a team a higher chance of winning.

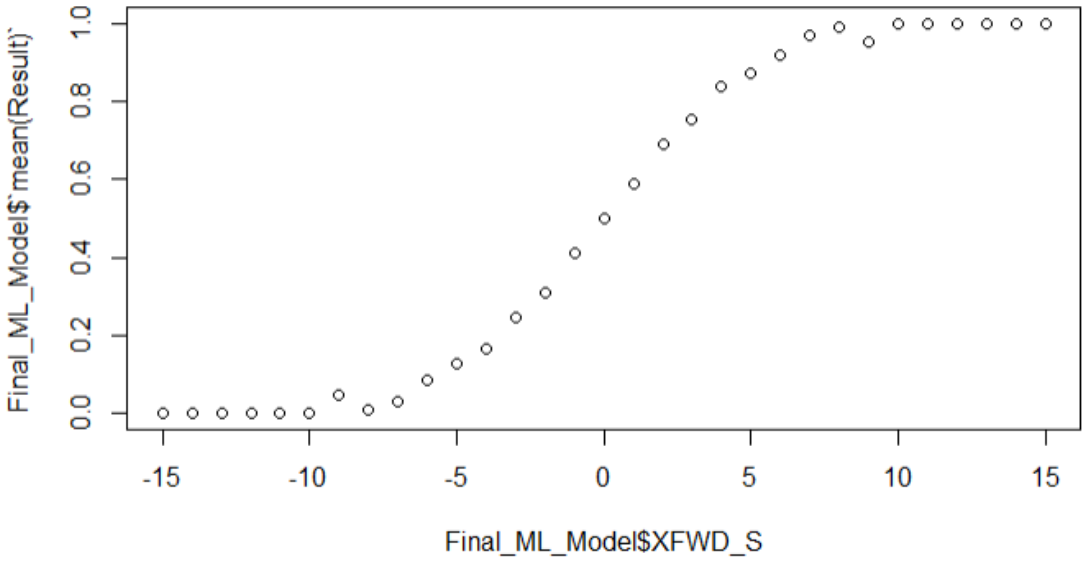


Figure 10 - Win Rate (Y axis) vs XFW score (X axis)

5.5 Future Studies

There are several areas for further study in League of Legends analytics. This particular model could be expanded upon utilizing certain elements such as patch and pick order to create a better live model for drafting purposes. Creating a model that can utilize data from high-level solo queue that considers how different it is from stage play could also be incredibly helpful. There are in game stats that this model did not utilize such as dragon control and baron control that could also be useful.

APPENDIX A

PICK ABOVE REPLACEMENT (PAR) SCORE

The pick above replacement (PAR) score is derived from an individual's total wins on a specific champion, games played on a specific champion, and the total wins and games played by that champion overall in competitive play. The individual's wins are divided by the games played to find the individual's win rate on that champion. The same is done with the champions overall wins and games played. The difference between the two calculations is the PAR.

APPENDIX B

XANDER FACTOR OF WINNING (XFW)

The Xander Factor of Winning (XFW) is derived from a team's CMV, TDF, PAR, and CPF, and multiplying each of those values by their associated winning correlation. This effectively weighs their contribution to winning. These values are then added up and multiplied by 10 to obtain the XFW. This formula is to be used with both teams participating in a single match.

REFERENCES

Bottou, L. (2014). From machine learning to machine reasoning: An essay. machine learning, 94(2), 133–149. https://doi.org/10.1007/s10994-013-5335-x

Glen, Stephanie (2013). "Correlation coefficient: Simple definition, formula, easy steps" https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

Goukgh, C. (2023, February 14). League of legends championships viewers 2022.Statista. https://www.statista.com/statistics/518126/league-of-legends-championship-viewers/#:~:text=The%20competition%20draws%20not%20just,5.15%20million%20peak%20concurrent%20viewers.

Hellbach, F. (2021, May 24). Retrieving data from the RIOT API for League of Legends in R. https://rpubs.com/. Retrieved February 15, 2023, from https://rstudio-pubs-static.s3.amazonaws.com/773313_1239374979da45659947571649f15ead.html

Hodge, V. J., Devlin, S., Sephton, N., Block, F., Cowling, P. I., & Drachen, A. (2021). Win prediction in multiplayer esports: Live professional match prediction. IEEE transactions on games, 13(4), 368–379. https://doi.org/10.1109/TG.2019.2948469

J. A. Hitar-García, L. Morán-Fernández and V. Bolón-Canedo, "Machine learning methods for predicting League of Legends game outcome," in IEEE transactions on games, vol. 15, no. 2, pp. 171-181, Februrary 2022, doi: 10.1109/TG.2022.3153086.

J, J. (2022, July 30). Becoming a LOL analyst: An introduction to using the Riot Api. Medium. Retrieved March 3, 2023, from https://medium.com/the-esports-analyst-club-by-itero-gaming/becoming-a-lol-analyst-an-introduction-to-using-the-riot-api-bb145ec8eb50

Jason Maldonis, Rob Rua, Eric Carmichael, Johannes Christ, Anton Pohli, Paaksing, Francesco Zoffoli, Fabien Culpo, samgho, Mert Kutay, rdk31, xEc, Hammaad K., Artem Kholodov, Guillaume DESSAIN, hawk93, Khorne, Michal Baumgartner, Jeff Putlock, … Thomas Pynchon. (2022). meraki-analytics/cassiopeia: v5.0.3 (v5.0.3). Zenodo. https://doi.org/10.5281/zenodo.7178532

L. M. Costa, A. C. C. Souza and F. C. M. Souza, "An approach for team composition in League of Legends using genetic algorithm," 2019 18th Brazilian symposium on computer games and digital entertainment (SBGames), Rio de Janeiro, Brazil, 2019, pp. 52-61, doi: 10.1109/SBGames.2019.00018.

Maymin, P. Z. (2021). Smart kills and worthless deaths: eSports analytics for League of Legends. Journal of quantitative analysis in sports, 17(1), 11–27. https://doi.org/10.1515/jqas-2019-0096

Ming-Syan Chen, Jiawei Han and P. S. Yu, "Data mining: an overview from a database perspective," in *IEEE transactions on knowledge and data engineering*, vol. 8, no. 6, pp. 866-883, Dec. 1996, doi: 10.1109/69.553155.

Roh, Y., Heo, G., & Whang, S. E. (2021). A survey on data collection for machine learning: A big data - AI integration perspective. *IEEE transactions on knowledge and data engineering*, *33*(4), 1328–1347. https://doi.org/10.1109/TKDE.2019.2946162

Ryška, Jaromír. "Development and forecast of esports industry." (2022).

Sarlis, V., & Tjortjis, C. (2020). Sports analytics — Evaluation of basketball players and team performance. Information systems (Oxford), 93, 101562–. https://doi.org/10.1016/j.is.2020.101562

Sayad, S., K-means. (n.d.). Retrieved April 1, 2023, from https://www.saedsayad.com/clustering_kmeans.htm

Wang, Tian. *Predictive analysis on esports games: A case study on League of Legends (lol) esports tournaments.* 2018. https://doi.org/10.17615/ez9n-t517

Xue, L., Song, P., Rai, A., Zhang, C., & Zhao, X. (2019). Implications of application programming interfaces for third-party new app development and copycatting. *Production and Operations Management*, *28*(8), 1887–1902. https://doi.org/10.1111/poms.13021

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. Neurocomputing (Amsterdam), 237, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026