A Model for Future LCS Success of LCS Academy Players

Cleveland Stephenson
Texas A&M University
College Station, Texas
ctaylorstephenson@tamu.edu

Ryan Bevan
Texas A&M University
College Station, Texas
ryan5905@tamu.edu

Introduction

League of Legends Championship Series (LCS) is one of the highest viewership and prize pool esports. With the recent franchising of the LCS, acquisition of LCS teams by higher profile traditional sports teams such as the Golden State Warriors, and an increase in large sponsorship deals, the LCS is attracting more investment than ever before [7]. Off-season roster changes have the potential to drastically alter the relative strength of a team in competitive League of Legends for better or worse. In addition, player contracts are typically short. These factors lead to a high amount of roster shuffling (all 10 teams had at least 2 out of 5 starters change for the 2020 season [2]).

However, teams are typically conservative in their player acquisition in the sense that they trade with other teams at a much higher frequency than they introduce talent from the LCS Academy which functions as a "minor league" [12] for the LCS. In this paper, we will describe a model for predicting Academy Player success in the LCS which could be utilized by teams to achieve higher return on investment and improve the average team strength by introducing new, potentially superior talent without assuming as much risk as was previously required.

Background

Predicting the success level of people has been done many times before. In its most basic form, conducting an interview is attempting to judge a prospective employee's success level in a small amount of time given specific information. However, this is much more subjective as the information is not cut and dry statistics. MLB (Major League Baseball) teams have invested many resources in an attempt to find younger players to draft. Their goal is to find the most likely players to succeed. There are many statistics that can be used, such as batting average, fielding percentage, earned run average, etc. The factors do not decide the best prospects alone, but they can be a good indicator of how a college player will perform in the big leagues [5].

The primary method of evaluating these statistics is to use a decision forest. Since there are so many different factors to consider, a single decision tree does not make much sense. By splitting up the statistics into different trees and using the overall results to evaluate strength, more accurate numbers can be found. The key is to make each tree distinctive. This is easily accomplished by grouping the statistics. For example, there are many different statistics linked to batting. There's batting average, on base percentage, slugging percentage, etc. These statistics can be in the same tree, while fielding statistics can be in a separate tree. There are plenty different groups of statistics to look at. This has been evolved over many years and well documented in the MLB.

We will use a similar strategy when deciding LCS Academy player's likelihood to succeed in the LCS. By grouping different statistics based on laning, kills, vision score, etc. we can evaluate how Academy players are likely to perform on the big stage.

Data Sets

We used the following data sets to construct a data profile for each of the Academy players considered in this analysis.

oracleselixir.com. Contains LCS and Academy raw statistics as well as computed statistics. Statistics available include: positions, games and win rates, avg kill, avg death, avg assist, kill participation, death share, first blood share, 10 minute differentials (gold, cs, xp), cs per minute, cs share post 15 minutes, damage per minute, damage share, gold per minute, gold share, wards per minute, wards cleared per minute. statistics are available for seasons 2015-2020.

lol.gamepedia. Contains LCS and Academy raw statistics as well as computed statistics. Statistics available are similar to oracleselixir but less structured. Additional statistics include: rosters, roster moves, and runes. LCS data is available as early as 2013 and Academy data is available as early as 2014.

gol.gg. Contains LCS and Academy raw statistics as well as computed statistics. Statistics available are similar to oracleselixir but only relatively recent data is available. Additional statistics include: first blood victim share, vision score per minute, avg game length.

op.gg. Contains solo-queue statistics including: games played, win rate, and cs per minute, avg kills, avg deaths, avg assists, avg damage, avg damage taken. Statistics go as early as 2010 but require preprocessing. Game length is not directly recorded (it can be inferred from cs per minute and cs per game).

Data Description

We first collected data from Academy LCS players from 2014 to the present. The data included 16 statistics such as the player's win rate, number of kills, and the number of champions they can play. These statistics all show some metrics by which we can measure a player's in-game success. We also added other pieces of information to the data that we felt were relevant. First, we included which team the players were on during their Academy tenure, as some teams have better talent-building infrastructure than others. We also added the player's position because, due to the way the game is played at the professional level, some players necessarily lead to better K/D/A ratios or gold earned in game. Further, it is common for players to play for an Academy team for several years. In order to integrate the data, we kept variables for the player's first year and their last year. We also took the difference of these statistics for each year and divided them by the number of years they've played Academy, which produces a measure of their average improvement per year.

To clean the data, we first removed players that only played a handful of Academy teams. We chose to do this because professional teams will often bench a player on their starting team temporarily if they are sick or can't play for another reason in a particular week. Further, some players are added later in the split and so we don't have a full set of data for them. We selected a cutoff point of 10 of the 18 games so that we had data for the player for at least half of the season. We felt this was the best way to ensure we were getting players who were true Academy players.

Once this data was cleaned, we needed to label the data. To do so, we scored players based on how well they did if they went professional. This method is described in the next section.

Scoring the Players

To label our data, we created a success formula based on different accomplishments players might achieve while playing pro. The higher the score a player receives, the higher their estimated value. The score is made of the following components:

LCS Titles: 25/20/10

The first aspect of a player's score is whether they have been on a team that placed well during a regular LCS split. Each player on the team would receive either 25, 20, or 10 points based on whether their team received first, second, or third. We chose to include this team-based aspect because a good team must essentially be composed of good players and also players that play well together. Further, the ability to play well with a team is an important skill for a player and very desirable by LCS teams. We chose the scale of 25-20-10 because LCS tournaments are tier-based, where the first and second place team are in the top tier and the third-place team is in the second tier. We wanted to capture a player's ability to successfully play for a top-tier team but also award players who were close to achieving this goal over players who were further.

World Championship Progress: 20 to make the tournament, 10 for each subsequent round

The second aspect of a player's score comes from their progress at the World Championship. For LCS teams, making Worlds and performing well is rare, as other regions are more competitive. Teams that make worlds are composed of players that play well domestically, so we award each player on these teams 20 points. For teams to make further progress, their players must be both excellent team players and individually talented. These skills are strong indicators of success for players. As such, for each round a team makes it to in Worlds, each player earns an additional 10 points.

Weekly MVP: 10 points

Players have the opportunity to earn an MVP title each week during a normal split. This is decided by a Twitter poll that the LCS publishes each week. This poll shows who the LCS scene considers the most skilled for that week. Players that receive this honor more than once will receive points for it each week they earn it. This category allows players who perform very well but perhaps don't achieve some of the team statistics to still earn points based on their solo performance.

Highest K/D/A: 15/10/5

This statistic records the highest kill/death/assist ratio for each split. Players with good K/D/A ratios were able to progress their team's position without dying at the hands of the enemy team. We awarded points for first, second, and third place, which would earn 15, 10, and 5 points respectively. Even though the player's position may influence their K/D/A, the best handful of players have always been near the top regardless of their position, so we felt it was an appropriate measure.

MVP: 50

An MVP is named once a split. Between 30 and 40 coaches, commenters, and past players vote on the MVP among all the players in a split. An MVP player will typically be someone who not only had impressive personal statistics but were able to play well with their team. We chose to give this honor 50 points because it is the highest individual honor in the LCS and signifies great success.

Player of the Game: 1

The final aspect of the score is the Player of the Game honor. This metric is also decided by a Twitter poll, which is settled for each game between any two teams during the split. We awarded a single point for this honor because it is relatively minor, especially if the two teams are mismatched in skill. However, we felt it was important to include in order to differentiate between players who otherwise have similar statistics.

Methods

Because of our data's structure, we chose to use a random forest to model our data. Several of our statistics were perfectly multicollinear. For example, the Kill/Death/Assist ratio is directly composed of the Kills, Deaths, and Assists statistics. However, we didn't want to exclude any of these statistics from our model because we didn't have a theoretical argument for why one would be better than the other. We thought it equally likely that having a high number of kills would predict professional success as having a good K/D/A would. Random forests are robust to this multicollinearity because they do not attempt to determine the effect on the output from a specific input as opposed to the other inputs. Further, we had nearly 50 input features and only about 75 players we were able to collect statistics on. As such, we wouldn't be able to support all of these variables in other model types because we simply didn't have enough data to. Thus, a random forest is robust to both the collinearity in our data as well as our abundance of input data.

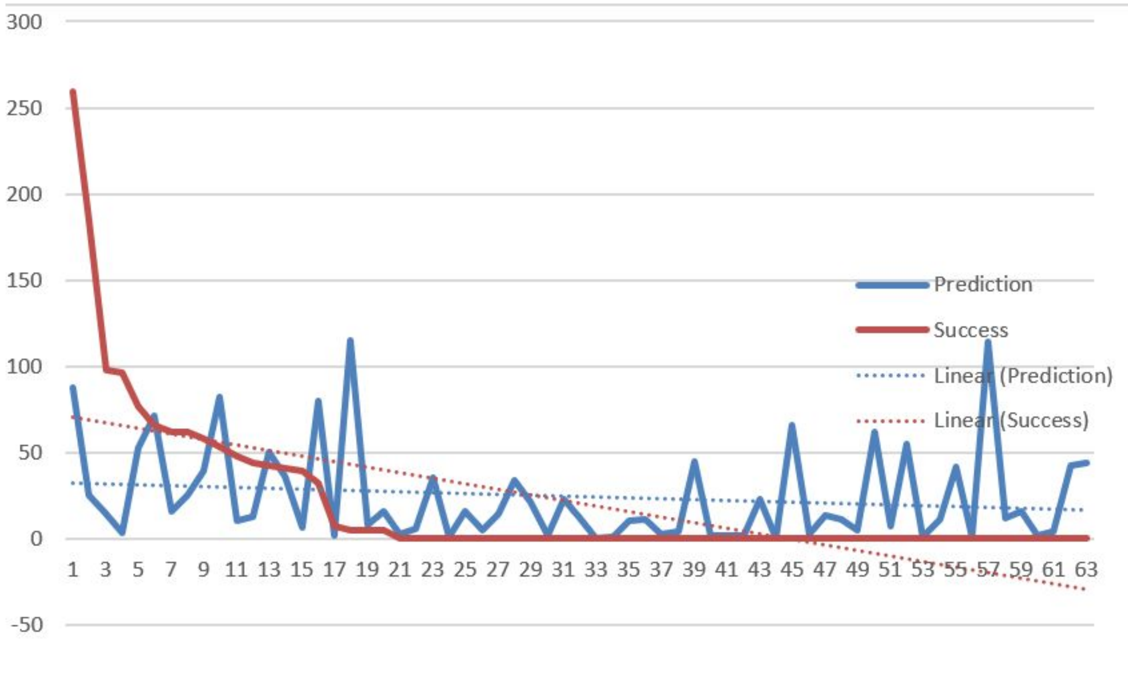
Before putting the data into a random forest, we encoded two variables - the player's position and what team they played on. Next, we ran the data through a PCA, which served to cut down on some of the multicollinearity in our input data. Our best explaining components explained 22.1%, 12.4%, and 10.9% of the variance. We didn't attempt to derive the linear combinations of each component but we expected they would accomplish a similar effect discussed above in the baseball example. Instead of grouping statistics into comparable categories and choosing the best from the category, we relied on the PCA to condense some of the input data and reduce its dimensionality. We tried varying the number of dimensions to reduce to and settled on 8. Once these input modifications were complete, we fed the data into the random forest model.

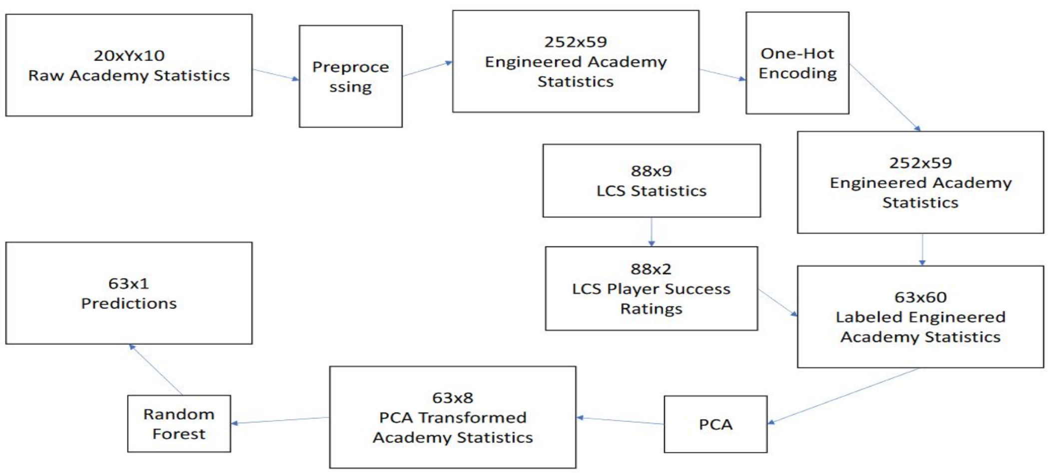
Results

We had hoped to achieve a test accuracy of at least 25%. The average LCS player salary is around $300,000 and the Academy average is around $75,000. A 25% accuracy would make signing 4 predicted successes breakeven on average with signing a proven player [9]. We were concerned that this goal would be unreasonably high due to data limitations and non-quantifiable factors in success such as ability to verbally communicate clearly with team members in game and coachability. Ultimately, our concerns appear to be well-founded as our test accuracy is quite low.

Our model seems to have partially learned criteria for what makes a player likely to succeed. Most of the high scoring players in reality were scored highly by the model. An interesting note though is that some of the highest raw errors come from the highest scoring players (> 100 "Success") perhaps because they are outliers who have overperformed in one area of our scoring system such as making it deep into the world's tournament.

Prediction	Error	Success
87.4597	171.5403	259
24.8878	160.1122	185
14.17	83.83	98
3.2205	92.7795	96
52.6894	24.3106	77
71.4228	5.4228	66
15.4062	46.5938	62
25.2911	36.7089	62
39.394	18.606	58
82.1539	29.1539	53
10.3878	37.6122	48
12.3997	31.6003	44
50.3235	8.3235	42
35.7336	5.2664	41
6.1874	32.8126	39
80.0793	48.0793	32
1.6659	5.3341	7





Discussion

We believe that an accuracy of 25% is still achievable despite our model's failure to hit that benchmark. There are several techniques we could have applied to improve the accuracy of our model. For example, for future models we would consider dividing the success score of each player by the number of years they have played in the LCS, as we feel our current model might be penalizing players who are new to the professional scene and haven't had time to accumulate points. For example, the player at index 18 in Figure 1 has a high predicted score, especially compared to the actual score. The player at that index, Johnsun, was a rookie this year and his team performed worse than predicted. However, he is widely considered to have been successful on his team. Our current scoring system penalizes him on both the dimensions of his time in the scene and his team's underperformance.

This team underformance is part of a larger issue we encountered with our model. This issue is that the success score of the players was composed of several high-value items rather than more numerous individual statistics. For future models, we would consider a more complex model that took more achievements into account. Our current method left the majority of players with zero points. As such, we appeared to be better at predicting standout players rather than more average successful players. However, since we didn't have a reliable method of tallying more average players, it was difficult for our model to make the distinction between an average player and a standout player.

Another significant issue we would like to fix if more data were available is the disparities between position. We attempted to take this into account in our input data, but the relationship is probably more complex than simply encoding the position would be able to account for. A player's position affects almost all of their statistics. Ideally, we would like to train a separate model for each position, as we feel that this would allow us to capture these complexities.

We also see that two players, Darcoch and Ryu at indices 4 and 7 respectively, were predicted to perform poorly but actually performed quite well. These players represent a class of players that we should perhaps treat differently. Both of these players were starters in the early years of the LCS when play was less competitive and there were fewer players on the scene. These players were benched to Academy for several seasons when the scene got more competitive and were able to re-earn their way onto teams. It's possible that their previous experience helped them perform and also gave them practice playing at a starting level. It also shows that they were a safer bet than their stats may have shown.

A final piece of data that would have helped our analysis is the salary data of the players. A player's salary is an important measure of their success as it reflects the value of that player to the team. However, this data is often not published at an individual level. If teams have more of an interest in this kind of modeling, they might be persuaded to publicize this kind of data in the future.

Ultimately, these issues with the model show our model for what it is - a preliminary study for modeling LCS players. We are confident that industry and other researchers will be able to improve on our model and help teams select players from Academy. We hope that these models will pick players that will help the LCS improve and become competitive on a global scale. A more reliable method of selecting players from Academy may encourage LCS teams to invest more in their Academy players, allowing them to bring up talent from the region without relying on drafts from other regions.

References

[1][n.d.].Naive bayes classifier. Retrieved March 7, 2020 from https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[2]lol.gamepedia2019.Roster Swaps/2020 Preseason/North Amer-ica/Current Rosters.lol.gamepedia.Retrieved March 8, 2020 fromhttps://lol.gamepedia.com/Roster_Swaps/2020_Preseason/North_America/Current_Rosters#LCS_Starting_Rosters

[3] Ben Steenhuisen Adam Summerville, Michael Cook. 2016.Draft-Analysis of theAncients : Predicting Draft Picks in DotA 2 Using Machine Learning. UC SantaCruz, Falmouth University, DatDota.

[4]Gokhan Ciflikli. 2017.Dealing with Correlated High-Dimensional data. Attest.Retrieved March 7, 2020 from https://www.gokhan.io/post/fuzzy-forest/

[5]Edd Gent. 2018.How AI is helping sports teams scout star players. NBC News.Retrieved March 5, 2020 from https://www.nbcnews.com/mach/science/how-ai-helping-sports-teams-scout-star-players-ncna882516

[6]Salma Ghoneim. 2019.Fuzzy logic and how it is curing cancer. Towards DataScience. Retrieved March 5, 2020 from https://towardsdatascience.com/fuzzy-logic-and-how-it-is-curing-cancer-dc6bcc961ded

[7]Darren Heitner. 2018.A Look Inside Riot Games, From $320,000 Player Salaries ToUsing Esports As A Catalyst For Sales. SportsMoney. Retrieved March 8, 2020from https://www.forbes.com/sites/darrenheitner/2018/05/02/a-look-inside-riot-games-from-320000-player-salaries-to-using-esports-as-a-catalyst-for-sales/#71c434ab2c6a

[8]Vaibhav Kumar. 2018.Random forests and decision trees from scratchin python.Towards Data Science.Retrieved March 8, 2020 fromhttps://towardsdatascience.com/random-forests-and-decision-trees-from-scratch-in-python-3e4fa5ae4249

[9]Kurt Lozano. 2019.LCS players have an average salary of over US$300,000. ONEeSports. Retrieved March 8, 2020 from https://www.oneesports.gg/lol/league-of-legends-lcs-players-have-an-average-salary-of-over-us300000/

[10]u/dataScienceThrow1. 2019.Building An NFL Draft Model Using Machine Learning.Retrieved March 5, 2020 from https://www.reddit.com/r/nfl/comments/b27abi/oc_building_an_nfl_draft_model_using_machine/

[11]Tony Yiu. 2019.Understanding Random Forest. Towards Data Science. RetrievedMarch 7, 2020 from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[12]Gabriel Zoltan-Johan. 2017.How NA LCS' Academy System can earn top marks.The Score eSports. Retrieved March 8, 2020 from https://www.thescoreesports.com/lol/news/14521-how-na-lcs-academy-system-can-earn-top-mar