

***Data Analysis for BetSports***

**An eSports Betting Company Founded by Chris Jackson**

**ISA 414 Final Project**

**Spring Semester 2019**

Analyses by

Tommy Bojanin, Colin McGreal and Eve Schoenrock

**Abstract**

As recruits of the data science team for the newly formed betting company, BetSports, our core team has been tasked with analyzing and building predictive models on historical data of two of the most popular eSports games, League of Legends (LoL) and Player Unknown’s Battleground (PUBG). Through our in-depth analysis of each data set, we have selected a decision tree model for winning team predictions for LoL and a linear regression model for winning predictions for PUBG as the primary models for outcome prediction. We will provide hypothetical scenarios towards the conclusion of this analysis as it pertains to its added-value for BetSports. Our decision tree model for League of Legends predicts the outcome of a team given in game statistics with a accuracy score of ~98 percent. The linear regression model for PUBG performs with a mean square error of 0.089, and a accuracy score of 84 percent. Given the type of problem that PUBG is, we’re going to focus on the mean square error rather than the accuracy.

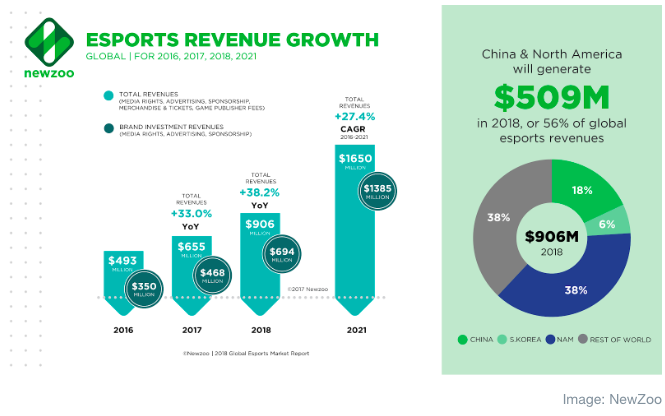
**Introduction**

**eSports Background**

The eSports industry is the hottest growing industry in the video game world and is suspected to pass the $1 billion dollar mark in 2019 (Pei). eSports is the term used to describe the online competitive video gaming world. Players from leagues and teams formed all over the world compete in matches and tournaments that are watched by millions of viewers online. Streaming services such as, YouTube and Twitch allow for audiences to watch the teams face off live. On these streaming sites, popular teams and players build large, dedicated fandoms that ritually watch them compete (Willingham). As of August 2018, a marketing company, Newzoo reported that an audience of almost 400 million people watched eSports in the previous year with the most dedicated fans tuning in from North America, China and South Korea (Willingham).

With such a large audience, eSports attracts many brands and video production companies who pay to sponsor and endorse players as well as tournaments to create brand visibility and target a specific market based on psychographics. Monetary incentives are a large driver of players to dedicate ample time to perfecting their gaming skills. In August 2018, “Newzoo reports brands will invest $694 million in eSports ventures in this year alone” (Willingham).

Not only has the eSports world generated large revenue from a multitude of endorsements and sponsors from brands, but has also sparked a new type of betting industry. Dedicated fans are able to place bets through various online bookmakers using real money or cryptocurrency (Thompson and Rivet). Gamblers can place bets on a variety of different aspects of the game that fall into three types of betting categories. These include match winner, betting outrights and totals. In this first and most simple category, gamblers place bets on the winner of the matchand predetermined odds are provided, which are compiled prior to the start of the match. The betting outrights category can include various types of bets such as when players are eliminated in the game and who will proceed to the latter stages of the game. And the final category, totals, is where gamblers simply place bets on “total maps played, total kills, total points or total rounds” (Thompson and Rivet). With an increasing amount of bets being placed each day, the eSports betting industry is expected to be valued at $1.5 billion by the year 2020 (Thompson and Rivet).

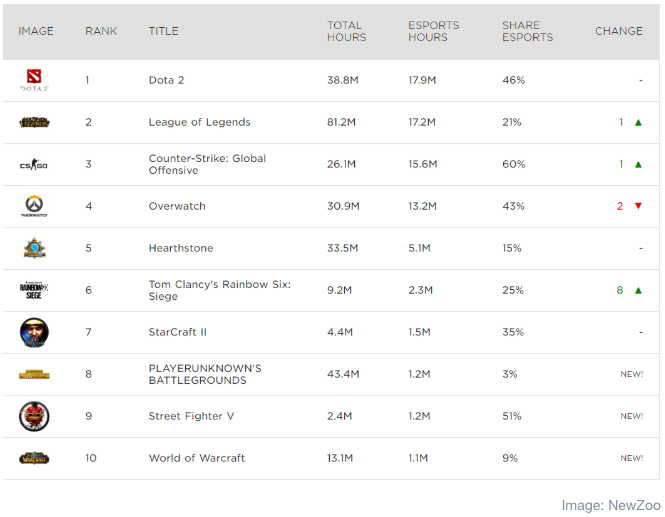


*Figure 1. Infographic displaying eSports revenue growth.*

The figure shown above displays the past total revenues of the eSports industry in the last few years as well as the projected total revenue in the year 2021, which is supposed to increase by a rate of 27.4 percent in a five-year period. The right half of the figure shows the sources of the global eSports revenue in 2018. Note that China and North America contributed to over half of the total revenue - approximately 56 percent.

**eSports Business Problem**

Chris Jackson is an experienced investor who has recently become captivated by the eSports world. He is astonished by the industry’s rapid growth as well as its projected future growth. As a long-time stock market investor, Jackson has decided to begin a completely new business venture and dedicate his time to learning about as well as investing in the rapidly growing eSports industry. Through his research, he has discovered that predictive modeling is a powerful tool that can be used to predict winners as well as other attributes of eSports games. Jackson has decided to launch his own betting agency, BetSports, and plans to utilize predictive models to make data-driven betting decisions. BetSports will be accepting investments from anyone interested who will hopefully receive a great ROI due to betting decisions that will be backed by expert prediction models and historical data analysis. As an investor with no previous experience in the data science field, he has decided to recruit a team of data scientists who will build predictive models for the most popular eSports games. We were each individually recruited by Jackson and assigned to work as a team. We were specifically assigned to analyze as well as build models on the historical data of two of the most popular online games, League of Legends (LoL) and Player Unknown’s Battleground (PUBG).

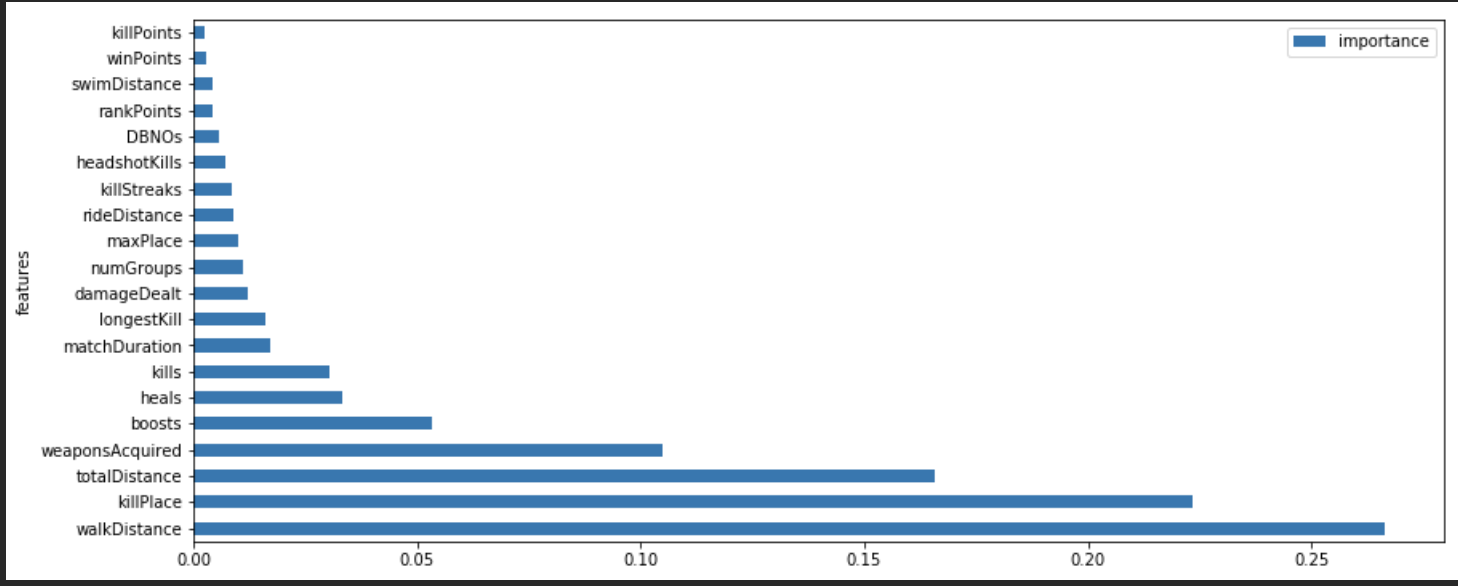


*Table 1. Top 10 eSports Games as of February 2019.*

*Table 1* displays the top 10 eSports games in the world by tournament number of hours watched on the YouTube platform. As it can be seen, LoL takes the number two spot with over 17.2 million hours and PUBG trails close behind at the number eight spot with 1.2 million hours. Our predictive models hope to provide BetSports valuable information for use in betting on the two respective games, applicable to tournament or recreational play.

Our team found it best to take the classification route for the LoL dataset. More specifically, we created a decision tree model for classifying the winning team, team1 or team2, based on numerous variables, which will be discussing in the next section. We found that classification led to the most accurate model for predicting winning team as quite a few of the predictors are able to be optimized in terms of splitting criteria. Furthermore, we determined the best splitting criteria through the use of node impurity measures, gini and entropy. We used both because we wanted gini to minimize misclassification of the data and entropy for data exploration purposes. The final model contains 14 predictors and leads to a predictive accuracy of 96.95 percent on the testing data. This is a very high accuracy measure and further backs up our choice of models. Later in this analysis, we will dive into hypothetical betting scenarios.

For the PUBG data, our team found a regression model to be the most accurate predictor of winners based on historical data collected with a potential outperformance with a neural network given the necessary support. With a target variable of winning place percentile –– 1.0=first place, 0.0=last place –– we found that a regression model best fit this dataset. Instead of using accuracy to test the performance of the model, we used mean squared error. This is due the nature of the problem, which is predicting a float number between 0 and 1 rather than predicting a binary number 1 or 0. It is much more difficult to accurately predict a placement number 0.0-1.0 given a set of features than it is to predict a number and then calculate the mean squared error. Given the PUBG dataset, the features were essentially ‘tertiary’ statistics aggregated by the game.

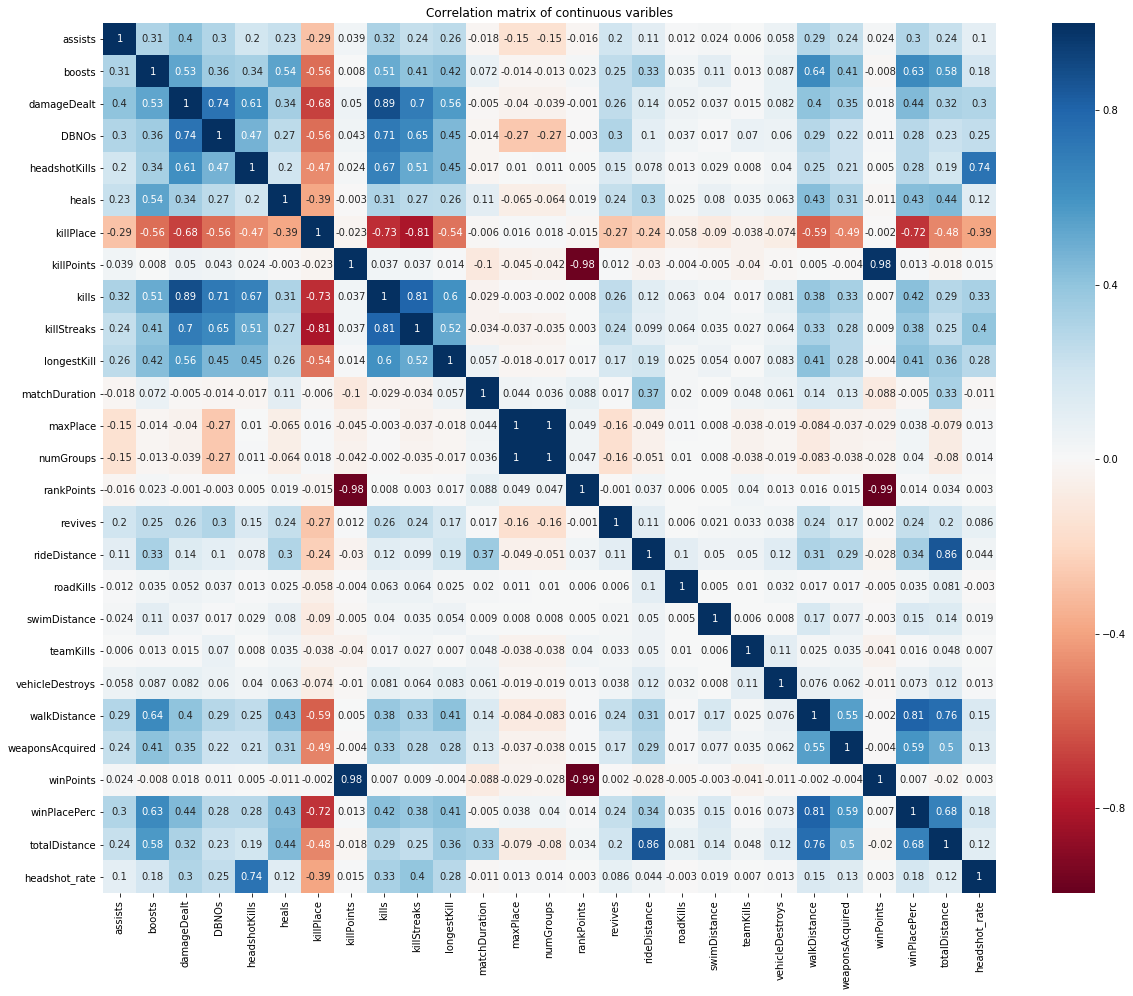
Primary and secondary statistics are much more difficult to collection given that examples of them are raw skill at the game, game strategy and sense, and the ability to combine these to outplay opponents. Examples of the statistics the team has aggregated are walk distance, kills in the game, and guns picked up. Given these statistics we were able to predict which features played a more important role in predicting the placement of the player inside the game. The outcome of the data analysis is staggering. The most influential feature on the list is the amount of distance a player has ran. Initially this was confusing and we believed it to be incorrect, later thought invoked that it makes sense. The more a player runs, the longer he is alive in the game, so the higher chance he has to survive. 

*Figure 2. Importance of features included in our model.*

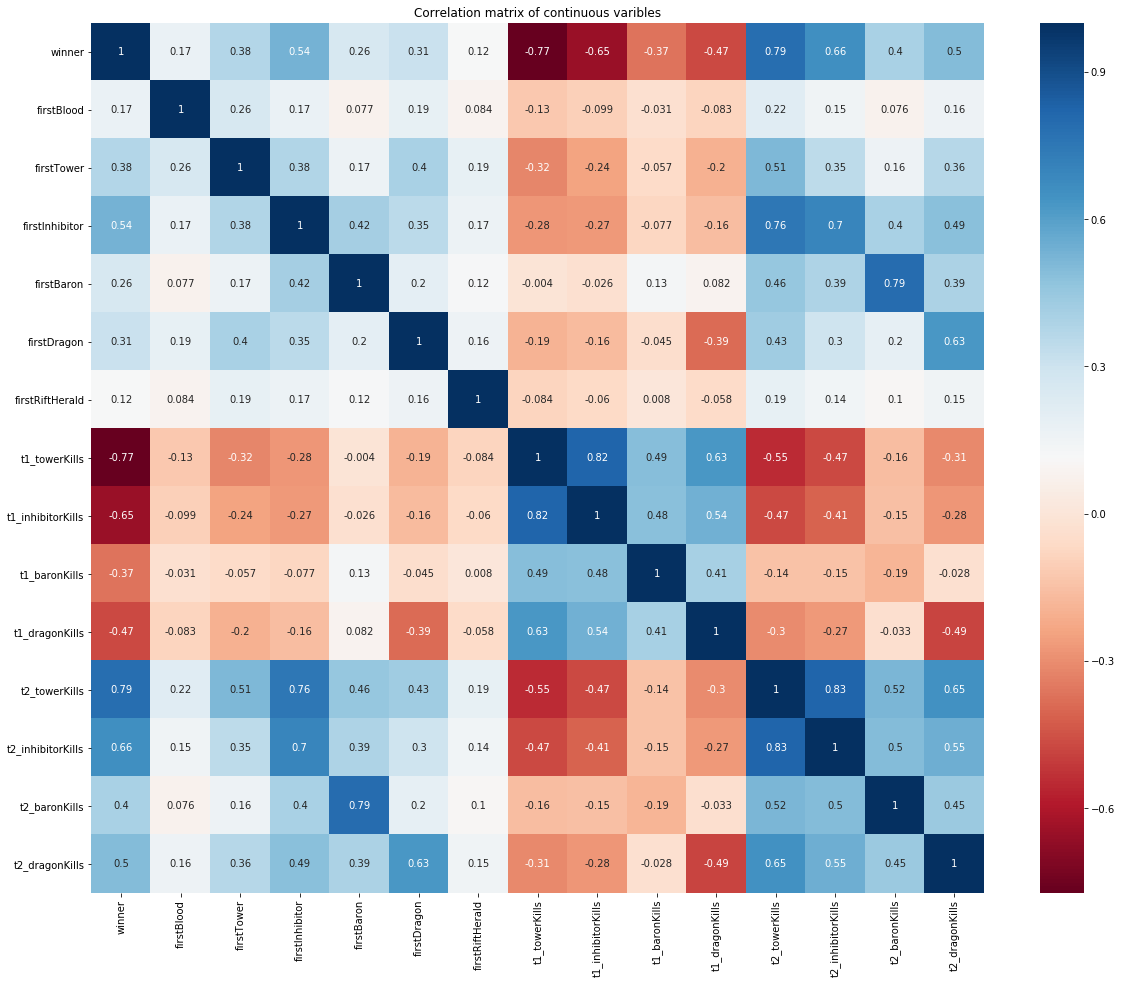
Here is a scenario to explain the importance: player A’s strategy is to kill as much players as he can, as he is an above average player who has confidence in his ability, player B is scared of confrontation in the game so he hides until there are two players remaining. Player A has put himself in harm's way much more often than player B, so if he loses the game (player B kills him), the importance of the amount of kills he achieved doesn’t matter as the amount of distance the player has ran. This statistics also is highly influenced by the playerbase of the game. The player base of most video games can be explained by a highly left skewed normal distribution (players are Y axis, player skill is X axis) where most players are average skill or unskilled at the game. These players are more likely to hide in the game (unskilled player) rather than to go out and try to achieve as much kills as possible (skilled player).

All in all, these statistics can be used in predicting the outcome of a game with professional players as long as the we have the statistics of those players. They can also be used in giving players advice on how to improve their in game ability and raw skill. For example, given a players statistics, we can see if they’re run distance correlates with the distance of people more likely to win games and give them advice such as ‘drop in less confrontational areas in the game’ or ‘you are dropping too far from the main areas of the map, try dropping closer to the middle’.

**Data Collection and Preparation**

Our team scraped Kaggle for historical game data for each respective game, LoL and PUBG. We found a 50,000 as well as a 450,000 historical game dataset for LoL and PUBG, respectively. We explored each dataset determining target variables of winner –– team1 = 1, team 2 = 2 –– and winPlacePerc –– percentile of team placed (1=win, 0=last place) –– for LoL and PUBG. Our team performed some data cleaning as some anomalies were present. For example, in the LoL data, we dropped games less than three and a half minutes as these were not complete games. Furthermore, in the PUBG data, we filtered out hacker-like behavior such as having a 100 percent headshot rate and killing almost 50 percent of players in the lobby. We also removed major anomalies across the board in statistics, these include but are not limited to, running too much in game and picking up copious amounts of weapons. We then created correlation matrices to explore the relationships between variables and see what predictors may influence the target variables the most. 

*Figure 3. Correlation matrix of all continuous variables in PUBG data.*



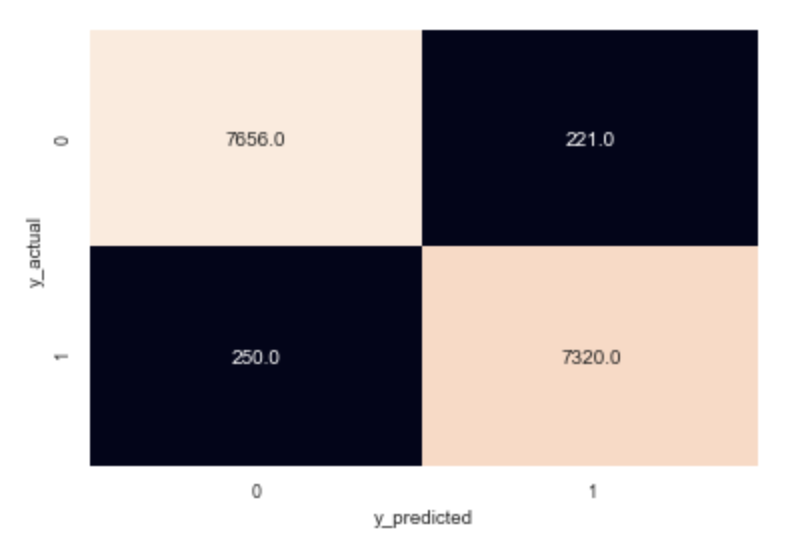
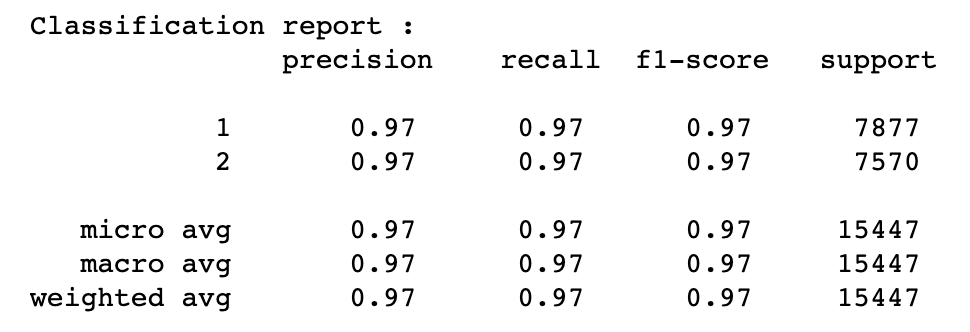
*Figure 4. LoL Correlation Matrix of Highly-correlated Continuous Variables*

As can be seen with the PUBG correlation matrix in *Figure 2*, walk distance, weapons acquired, and damage dealt have high positive correlations toward the target variable of win place percent. We argue that these variables will be highly important in our final model. An important thing to note from figure 3 is that having opposite correlations on one team compared to the other is OK, this is due to the nature of a correlation matrix, where two teams who have the same value are being compared to the same outcome.

From these correlation matrices and summary statistics, we moved onto normalizing the PUBG data in order to more accurately produce a linear model for use in predicting the winners. We then went on and created a decision tree classification model for predicting winning teams for use in LoL games.

**Data Analysis and Evaluation**

As previously mentioned, we conducted a decision tree classification model for predicting the winning team in LoL games. When evaluated against the test data, we found an accuracy measure of 96.95%. More results are detailed below: as you can see of the very high precision and recall scores. Simply put, the decision tree was effective in returning the most accurate data with most of it being relevant rather than irrelevant.



*Table 2. LoL Decision Tree Classification Report Figure 3. LoL Classification Matrix*

The predictors that were selected for use in the final model are listed as follows:

first\_blood, first\_tower, first\_inhibitor, first\_Baron, first\_Dragon, first\_RiftHerald, t1\_tower, t1\_inhibitor, t1\_baron, t1\_dragon, t2\_tower, t2\_inhibitor, t2\_baron, t2\_dragon. Summaries of these variables are listed in the [kaggle description](https://www.kaggle.com/datasnaek/league-of-legends). When it comes to using our model for production use, consider the following hypothetical case.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| feature | first\_blood | first\_tower | first\_inhibitor | first\_Baron | first\_Dragon | first\_RiftHerald | t1\_tower |
| value | 1 | 1 | 2 | 1 | 1 | 1 | 10 |
| feature | t1\_inhibitor | t1\_baron | t1\_dragon | t2\_tower | t2\_inhibitor | t2\_baron | t2\_dragon |
| value | 2 | 1 | 4 | 7 | 2 | 1 | 1 |

*Table 3. Hypothetical predictor values.*

These values are pulled from knowledge of a common LoL game. Under this scenario;

first team win probability is % [93.61702127659575]

second team win probability is %: [6.382978723404255]

These are the winning team percentages. Given this particular scenario, team1 has an approximate 94% winning probability. The analysts at BetSports can use this model to predict future game outcomes, hopefully for a considerable profit. For major LoL events, the analysts can determine how teams have historically scored in each one of the predictors from our model and determine the winning probability of specific teams head-to-head. This is an extremely capable decision tree model that is sure to perform in practice.

Similarly to the PUBG data, this model can be used to advise up and coming professional teams performance in game. For example, if a team has great synergy and in game strategy but performs poorly in the early game. I can advise the team to try and capture objective rather than team fight so they can outperform the opposing team in the early game and use their early game power to overpower the opposing team.

The PUBG data followed a more varied approach. The team realized that the PUBG data contained almost twice the amount of relevant features than the League of Legends data. This may result in the linear models to overfit the data when trained on the four million rows. So we decided to use a neural network to improve our model and lower the mean squared error even further. Initially the model performed very poorly, ~5 times worse than the linear model. Within 30 iterations of the neural network, it matched the mean squared error of the linear model. While this network was run on laptop with GPU that supports tensor operations, BetSports can use their server infrastructure to run dedicated machine learning models on data, with performance and logic improvements in the code, this model will outperform the linear models.

**Conclusion**

As was discussed in the previous section, these two models are sure to lead to accurate predictions for use in eSports betting. The models are created through training on thousands of various scenarios. The more accurate the assumptions, the better the results. If for instance top LoL eSports teams “Unicorns of Love” and “Flash Wolves” are playing in a future tournament and BetSports wants to place a large bet, they can use our model as well as in-house data from their analysts to figure out the inputs to the model. If these teams have met in the past, their data can be mined through the LoL API, just as how this model’s data was collected. It can then be aggregated to determine optimal inputs for this future match where bets can be placed. If all goes according to plan, BetSports can use their infrastructure to support our models and run them in real time across newly played League of Legends and PUBG games.

**References**

Pei, Annie. “Here's Why Esports Can 'Crack into the Mainstream' on Its Way to Potentially

Hitting $1B in 2019.” CNBC, 23 Jan. 2019,

www.cnbc.com/2019/01/20/heres-why-esports-can-become-a-billion-dollar-industry-in-2019.html.

Pannekeet, Jurre. “Global Esports Economy Will Top $1 Billion for the First Time in 2019.” *Newzoo*, 12 Feb. 2019, newzoo.com/insights/articles/newzoo-global-esports-economy-will-top-1-billion-for-the-first-time-in-2019/.

Thompson, Emma, and Coin Rivet. “What Is Esports Betting?” *Yahoo! Finance*, Yahoo!, 8 Apr. 2019, finance.yahoo.com/news/esports-betting-130033176.html.

Willingham, AJ. “What Is ESports? A Look at an Explosive Billion-Dollar Industry.” *CNN*, Cable News Network, 27 Aug. 2018, www.cnn.com/2018/08/27/us/esports-what-is-video-game-professional-league-madden-trnd/index.html.