Preston Robertson

Dr. Edward Swan

Course Project

April 4, 2022

**Course Project: League of Legends**

**1. Dataset Description**

**1.1 Introduction to the Dataset**

This dataset is on an online competitive game called League of Legends. This dataset was chosen due to a want to challenge myself with this project and force me to learn new techniques and use new techniques that I have learned from my classes this semester. This is due to the unique nature of this dataset. The objective of the dataset is to find the features that will most impact the win, so that it is easier to balance the game. “Balancing” refers to updating the game downgrading strategies that are too strong and upgrading strategies that feel too weak. This document will serve as the preliminary to champion balance by correlating specific stats to a win, so that in the future someone may correlate these stats to champions in the game. This may feel convoluted; however, each game will have 5 winners and 5 losers meaning that a single champions impact on the game is 10%. If a champion is unbalanced, being double the strength of another champion that only raises the 10% to 20%. Due to the law of averages, each champion will still have around 50%-win rate despite being too weak or too strong. That means to properly balance a champion, one must look at the correlation between stats of the champion to the win-rate of each of those stats. This effectively takes out the problem of bias since every game having at least one winner and one loser. This dataset has 23,752 data points and 24 features (or columns). This is a complicated dataset, with several variables requiring several stages of feature modification to run the code. The code is also large enough to have significance. This dataset was specifically chosen since the data is something I am already familiar with, so that allows me to focus on the machine learning techniques.

**1.2 Brief Description of League of Legends**

League of Legends is a part of the MOBA genre of games and is considered the most widely popular game of all time. In its most recent tournaments making more money than the Superbowl, in 2020 League of Legends made 1.7 billion dollars (Wolf, 2020). The objective of the game League of Legends is an online competitive multiplayer game where two teams fight destroy the enemy nexus. Below is the map of the game so it is easier to reference the variables given.



(Figure 1: Map of League of Legends (Gao, 2017))

**1.3 Description of the Map**

The map of league of legends contains 3 paths, each of these are a lane with a corresponding name. These are named Top Lane, Mid Lane (short for Middle), and Bot Lane (short for Bottom Lane). Each lane spawns’ "minions" to help push lanes. These minions are very easy monsters and provide gold if you land the killing blow. Each one of these lanes have 3 towers and an inhibitor. All three of one lanes tower + inhibitor need to be destroyed before you can reach the nexus. The towers provide protection to players by hurting enemy champions when they are in range. The inhibitor provides no uses to ally team; however, if you destroy the enemy inhibitor then in that lane "Super Minions" will spawn. These buffed minions help push to finish the game. There are also forests in the middle of these lanes. This is called the Jungle. In the jungle there are several monsters’ worth gold and that grow stronger as the game goes on. Some monsters, if killed, can even provide special bonuses. All these monsters can be killed by one player. The part that splits the map into two sides is known as the river. This is the equivalent of the half-way line in soccer. In this part of the map Large Monsters spawn that require a group effort to takedown but give huge bonuses.

**1.4 Description of the Gameplay**

The game is known for its complexity and if you want a comprehensive guide, I have provided a link that I think does an excellent job. I will explain the bare necessities needed to follow the data. (Macabaso, 2021). There are 5 players on each team, each player plays a champion. This champion is a character with unique gameplay, stats, and abilities. These 5 players will each fill a specific role: Top Lane goes to the Top Lane, Mid Lane goes to the Mid Lane, the Jungler goes to the Jungle, the Attack Damage Carry (ADC) and Support goes to the Bot Lane. In each of their respective locations each role attempts to make gold and level up. The gold is used to buy items (each with unique effects and every champion pick from the same pool of items), which means the more gold you have the more items you can buy. You get gold by killing different things on the map (for example: minions, monsters, large monsters, and enemy players) and can also be obtained by everyone at a constant passive rate. The more items you have the stronger you are and more capable you and your team are to win the game. Levels occur naturally when you gain experience by being around combat or by killing enemy units (same as gold) this however does not passively increase and must be earned. With each level up your basic stats become stronger which means you are stronger and allows you to level up current abilities or unlock new abilities. There are several other factors to the game, but I think this basic grasp is all that is needed to understand the data.

**Data Analyzation, Visualization, and Pre-processing**

**2.1 Further Description of Dataset**

The dataset went through preliminary data analysis that will be discussed in this section. First the feature selected to study is “Win” which is a Boolean variable; this means that the machine learning models used in this experiment will be classification models. Basic analysis was done on the dataset, which can be found in the appendix as Figure 2: Basic Data Analysis of Numerical Features. Looking at this information, that dropping “Damage to Turrets” could increase the accuracy of our model since the standard deviation is higher than the mean. It can also be gathered that this data could cut the farthest datapoints to see an increase in model accuracy since the maximum and minimum are so far from the mean in most features. However, I will not take out these variables since the objective of this experiment is to change as little as possible to not interfere with results.

After the data analysis has been completed, the next step was running checks on each feature to ensure that each were capable for this experiment. A cardinality check and a missing percentage were taken to ensure these qualities. The cardinality was taken since if a feature has far too many unique variables, then the feature will only harm the machine learning algorithms. This missing percentage was taken to ensure no feature had too many variables missing where data analysis on that feature would be impossible. Check in the appendix for Figure 3: Cardinality of Each Feature and Figure 4: Percentage Missing of Each Variable for specifics. This analysis has shown that no features were missing variables and that only one feature has too high cardinality to be included. This feature was the anonymized player name that basically became a second index. This feature will be dropped since it serves no value in this dataset for either machine learning algorithms or game balance.

**2.2 Data Visualization**

The next pre-liminary data analysis that were conducted on the dataset were a heat-map (correlation matrix) and a scatter plot. The first figure shown below is a heat map. A heat map is basically a correlation matrix that shows how much any two feature interacts with one another. The heat map improves upon this idea by also adding then temperature value so it can easily be seen how much to variables correlates. This temperature value can be any value between -1 and 1. Where -1 is a sign of perfect negative correlation and +1 is a sign of perfect positive correlation. The darker the color on the heat map, the “colder” the temperature and vice versa for the brighter colors. Looking that the Figure 5: Heat Map of the Data Set, some values have very high correlation; however, looking at most of these variables it can shown that most features will increase in value as the time moves forward. This gets rid of most of the correlation since these values are interconnected through the feature “duration”. Even with this information, there are some values of note. Such as the high correlation between “Win” and “Win Rate of Champion” (the highest correlation to win) while “Win” and “Champion” do not correlate. This shows the problem discussed earlier in this experiment where the natural bias of the game does not allow for just a normal correlation to be sufficient. Other information to gather from this heat map, is that previously thought important values for a winning team are not very highly correlated to a win. This shows the necessity of such data analysis to the game. Not much else should be deciphered from this heat map since it is helpful when visualizing data, but the most precise with results.

Chart, treemap chart

Description automatically generated

(Figure 5: Heat Map of the Dataset)

The next data visualization done to the dataset was a scatter plot. A scatterplot graphs each feature with respect to each other. This can be very useful for seeing trends visually before any data analysis is done. Look at Figure 6: Scatterplot of the Dataset below to see specifics of each section of the scatterplot. Not much can be gathered from the champion section of this scatter plot since each champion is given their own x value and there is no way to see the x values with the size of this scatterplot. Comparing “Kills” and “Deaths” it is possible to see the problem discussed earlier with the heat map. Looking at both stats, it is easy to see that the amount of gold earned increases no matter what due to each variable being connected through the duration of the game. However, it is possible to see that each variable has its own effect in the scatterplot still, since there is a difference between “Kills” and “Deaths”. The difference being consolidated data and more spread-out data points. That is best information to gather from this scatterplot since the scatterplot is used to visualize the data and not supposed to be used for thorough data analysis.

Diagram, timeline

Description automatically generated

(Figure 6: Scatterplot of the Dataset)

**2.3 Data Pre-Processing**

Now that the dataset has thoroughly been analyzed, it is now time to process the data to prepare it to be put through machine learning algorithms. This set-up will be rather short since the point of analysis is to be thorough and not exclude many features or samples. Since time is not an issue with this dataset it is not a problem to perform feature reduction on the dataset. This is important to keep track and make the machine algorithm does not lose value of any one feature. Despite the Principal Component Analysis method mentioned in the project proposal, it will not be implemented for those reasons stated above.

First data-preprocessing done on the dataset was using a “One Hot Encoder”. This algorithm splits up string values into their own column and turning it a Boolean value. For example, in this dataset, Position was changed into 4 new features with capability of displaying all 5 unique values in that feature. This was used on the features: Team, Position, and Win. This method can quickly add up the amount of feature a dataset has, so I thought it was important to include the lowest cardinality string variables to allow more the most accurate predictions without having 100+ features. After this is applied to the dataset, there are now 27 features (including the features removed previously). The rest of the string values were processed through a string indexer which turns all string variables into numbers. This has the benefit of being easily recognizable to the machine but has the downside of the computer thinking the value is a continuous integer. It is necessary to do before running some of the chosen machine learning algorithms. Ideally, One Hot Encoding would be the best option if it were not due to the cardinality of some features in this dataset. The next piece of data preprocessing was adding an interaction term, Gold per Minute. This stat combined gold and duration of the game to see if the rate of the gold gathered had an impact on win. Finally, the data was split into testing and training data which was split 20% and 80%, respectively.

**Appendix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature Name: | Kills | Deaths | Assists | Gold Earned |
| Count | 23752 | 23752 | 23752 | 23752 |
| Mean | 5.011 | 4.965 | 7.434 | 10223.871 |
| STD | 3.977 | 2.859 | 5.375 | 3479.974 |
| MIN | 0.00 | 0.00 | 0.00 | 666.00 |
| MAX | 29.0 | 20.00 | 40.00 | 25133.00 |
| Feature Name: | **Damage to Champions** | **Damage to Turrets** | **Damage Received** | **Vision Score** |
| Count | 23752 | 23752 | 23752 | 23752 |
| Mean | 15166.688 | 2685.702 | 20085.445 | 27.511 |
| STD | 9393.446 | 2965.702 | 10096.737 | 19.051 |
| MIN | 0.00 | 0.00 | 0.00 | 0.00 |
| MAX | 91579.00 | 52378.00 | 122378.00 | 153.00 |

(Figure 2: Basic Data Analysis of Numerical Features)

|  |  |
| --- | --- |
| Feature Name: | Data Types |
| Unnamed | int64 |
| Champ | object |
| Team | int64 |
| Gold Earned | int64 |
| Kills | int64 |
| Deaths | int64 |
| Assists | int64 |
| Win | bool |
| Total Damage Dealt to Champions | int64 |
| Damage Dealt to Turrets | int64 |
| Vision Score | int64 |
| Total Damage Taken | int64 |
| Turret Kills | int64 |
| Inhibitor Kills | int64 |
| Total Minions Killed | int64 |
| Champ Level | int64 |
| Wards Placed | int64 |
| First Blood Kill | bool |
| First Blood Assist | bool |
| WR general | float64 |
| WR champ | float64 |
| Duration (in Seconds) | int64 |
| Position | object |
| Role | object |

(Figure 11: Data Type for Each Feature)

|  |  |
| --- | --- |
| Feature Name: | Cardinality |
| Unnamed | 23752 |
| Champ | 148 |
| Team | 2 |
| Gold Earned | 10602 |
| Kills | 29 |
| Deaths | 20 |
| Assists | 39 |
| Win | 2 |
| Total Damage Dealt to Champions | 15129 |
| Damage Dealt to Turrets | 7304 |
| Vision Score | 137 |
| Total Damage Taken | 15768 |
| Turret Kills | 10 |
| Inhibitor Kills | 5 |
| Total Minions Killed | 361 |
| Champ Level | 18 |
| Wards Placed | 83 |
| First Blood Kill | 2 |
| First Blood Assist | 1 |
| WR general | 79 |
| WR champ | 198 |
| Duration (in Seconds) | 1863 |
| Position | 5 |
| Role | 5 |

(Figure 3: Cardinality of Each Feature)

|  |  |
| --- | --- |
| Feature Name: | Missing Percentage |
| Unnamed | 0.0000% |
| Champ | 0.0000% |
| Team | 0.0000% |
| Gold Earned | 0.0000% |
| Kills | 0.0000% |
| Deaths | 0.0000% |
| Assists | 0.0000% |
| Win | 0.0000% |
| Total Damage Dealt to Champions | 0.0000% |
| Damage Dealt to Turrets | 0.0000% |
| Vision Score | 0.0000% |
| Total Damage Taken | 0.0000% |
| Turret Kills | 0.0000% |
| Inhibitor Kills | 0.0000% |
| Total Minions Killed | 0.0000% |
| Champ Level | 0.0000% |
| Wards Placed | 0.0000% |
| First Blood Kill | 0.0000% |
| First Blood Assist | 0.0000% |
| WR general | 0.0000% |
| WR champ | 0.0000% |
| Duration (in Seconds) | 0.0000% |
| Position | 0.0000% |
| Role | 0.0000% |

(Figure 4: Missing Percentage of Each Feature)

**References**

Biau, Gérard, and Erwan Scornet. “A Random Forest Guided Tour.” TEST, vol. 25, no. 2, 2016, pp. 197–227., https://doi.org/10.1007/s11749-016-0481-7.

H. Samet, "K-Nearest Neighbor Finding Using MaxNearestDist," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, pp. 243-252, Feb. 2008, doi: 10.1109/TPAMI.2007.1182.

Gao, Gege & Min, Aehong & Shih, Patrick. (2017). Gendered design bias: Gender differences of in-game character choice and playing style in League of Legends. 10.1145/3152771.3152804.

Li Zhang, Weida Zhou and Licheng Jiao, "Wavelet support vector machine," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 34, no. 1, pp. 34-39, Feb. 2004, doi: 10.1109/TSMCB.2003.811113

Macabasco, Agilio. “Absolute Beginner's Guide to League of Legends.” Mobalytics, Mobalytics, 1 Dec. 2021, <https://mobalytics.gg/blog/absolute-beginners-guide-to-league-of-legends/>.

Mukherjee, Saurabh, and Neelam Sharma. “Intrusion Detection Using Naive Bayes Classifier with Feature Reduction.” Procedia Technology, vol. 4, 2012, pp. 119–128., https://doi.org/10.1016/j.protcy.2012.05.017.

Noble, William S. “What Is a Support Vector Machine?” Nature News, Nature Publishing Group, Dec. 2006, https://www.nature.com/articles/nbt1206-1565.

Oshiro, Thais Mayumi, et al. “How Many Trees in a Random Forest?” Machine Learning and Data Mining in Pattern Recognition, 2012, pp. 154–168., https://doi.org/10.1007/978-3-642-31537-4\_13.

Peterson, Leif E. “K-Nearest Neighbor.” Scholarpedia, 2009, <http://scholarpedia.org/article/K-nearest_neighbor>.

Schapire R.E. (2013) Explaining AdaBoost. In: Schölkopf B., Luo Z., Vovk V. (eds) Empirical Inference. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41136-6\_5

Wolf, Jacob (September 18, 2020). "League 101: A League of Legends beginner's guide". ESPN. Archived from the original on January 11, 2021. Retrieved December 23, 2020. "Each team has 11 turrets. Every lane feature two turrets in the lane, with a third at the end protecting the base and one of the three inhibitors. The last two turrets guard the Nexus and can only be attacked once an inhibitor is destroyed."

Xuchun Li, Lei Wang, Eric Sung, AdaBoost with SVM-based component classifiers, Engineering Applications of Artificial Intelligence, Volume 21, Issue 5,2008,Pages 785-795,ISSN-0952, 1976, https://doi.org/10.1016/j.engappai.2007.07.001.https://www.sciencedirect.com/science/article/pii/S0952197607000978