Linda Zhang

Class: CS 677

Date:4/25/2022

Final Project

Data Topic: Win Rates with different features in League of Legends

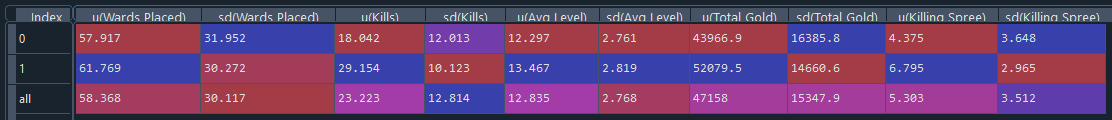
Data Source: <https://www.kaggle.com/datasets/gyejr95/league-of-legends-challenger-ranked-games2020>

Challenger\_Ranked\_Games.csv

The dataset that I picked to talk about is about League of Legends. This dataset had both binary and nominal data. A little bit about the game is that it is a game of five players versus another group of five players. The way to win the game is to capture the enemy Nexus. There are three paths to the nexus protected by two turrets.

When it came to setting up the dataset since it was already in csv format I did not need to do any additional formatting to get it into python. There were also no empty spaces or N/A values that would become an issue later on. The original dataset had around 30k rows and almost 50 columns, and rather than using all of the data set I selected certain columns that I want to focus on. I made sure to select columns that were related to winning but not directly related to winning like turrets destroyed.

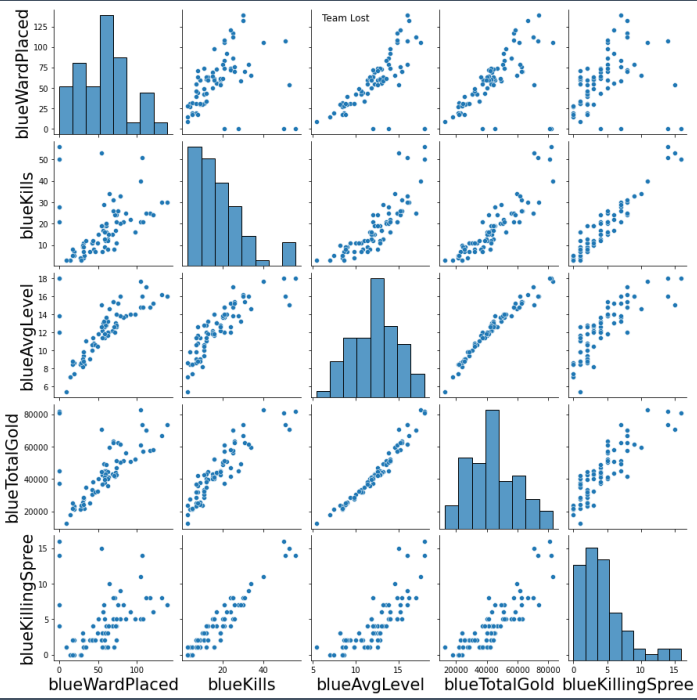
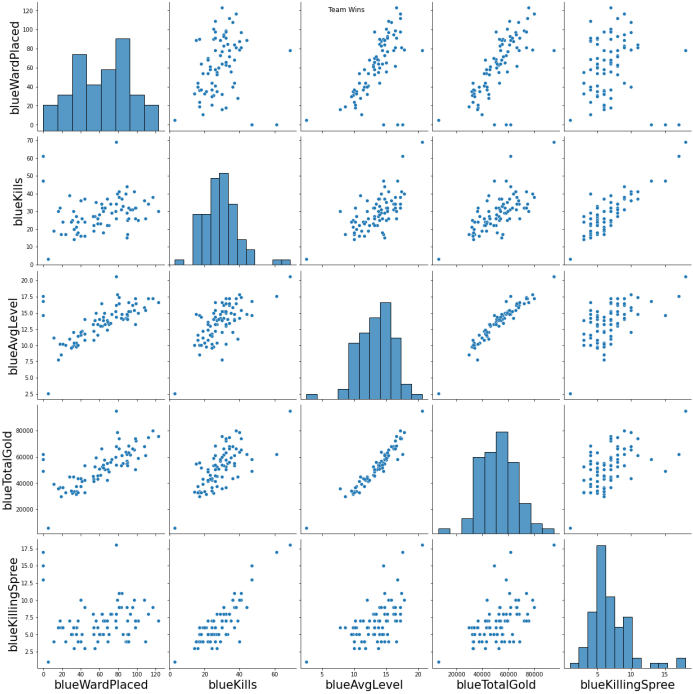
All of the values I decided to use as my features are all nominal data. The only binary data is blue win where that is what i'm going to be trying to predict here. A little bit about the features that I have selected. Wards are used to give vision on the map. Blue kills are the enemy that the team has killed in total. The level is the champion level and everyone starts out as level 1 at the start of the game and the max level that can be reached is level 18. The total gold is the total team gold and they can earn gold by killing different objects in the game which then they could use to buy boosts to help them. Blue killing spree is when a player kills three people in a row before dying.



The first row of data is games that were lost and the second row of data is games that were won.

I found that there are a few interesting points in this dataset. The first thing that stood out to me was that the average wards placed were pretty close to each other, I expected a larger difference for average wards placed. Another thing that surprised me was that the average level of the players were around the same and there is only a difference of one level. I was expecting the levels to be closer to the max level of 18.   
What didn't surprise me about this dataset was that the that the when the team won there was a lot more gold that they had

Winning dataset: Losing Correlation



K-NN



For my first classification, I’m going to be using k-nn. The first step that I did was to find the highest accuracy K and I found out that the highest K is when k=4. When I plug it in and compared it to the test I found out that the machine was more likely to predict a false negative than a false positive.

When it came to feature selection, I first looked at features selection with 4 features, there were two results 0.727 and 0.729. The ones with 0.729 all have wards placed as one of the features. I thought it would be really interesting to test out three features selection. So I kept wards placed and replaced the other two features. I found out that that the highest accuracy rate came from selecting three features of ward placed, kills, and average level with an accuracy rate of 83.8%

Logistic Regression



I have also decided to also use logistic regression to look at the accuracy. What surprised me is that the accuracy for five features do not improve while using logistic regression. When I used feature selection for 4 features the highest one was 77.1%. Using the highest accuracy’s features I tried out the various combinations with three features. I ended up finding out that the highest accuracy for logistic regression happens in two instances. The first one is when all five features are used. The second is when feature 4 is selected where kills, average level, total gold, and killing spree is selected.

Final Thoughts:

I thought it was really interesting that the highest accuracy was using knn with 3 feature selection with an accuracy of 83.8%. In features selection for KNN the most important feature is ward placed, but this does not apply to logistic regression. In feature selection for logistic regression, the one that almost had no effect when removed is ward placed. The highest accuracy using logistic regression was 77.1%.