Elements of Data Processing COMP20008

Assignment 2

Course Coordinator: Chris Ewin

Group 46

Ayush Tyagi, Chiang Ray-En Andre,

Nicholas Sean Phang, Nigel Loh Yesheng

Research Question:

Can supervised learning algorithms like linear regression and decision tree be used to predict 'end of game performance statistics' based on start and mid-game decisions? What are the features most impactful towards these statistics?

Contents

Aim

The aim of this project is to find the relation between endgame results based on various early to mid-game decisions for the popular game “League of Legends”. Our hope is that these results will prove beneficial to mostly the newer demographic, but also help further improve the game sense and knowledge of even the most veteran of players. Thus, the project targets what we believe to be integral variables, such as champions selected and minions killed, to prove that such minute decisions early on can snowball into what inevitably ends up as a win or a loss in a game.

Datasets

In this investigation, we used three datasets: ‘EUmatch.csv’, ‘KRmatch.csv’, and ‘NAmatch.csv’ which contained data from League of Legends matches from the ‘challenger’ rank. Data was collected from challenger games in January 2022 from different regions (as evident in the file names) and to ensure independence, only one randomly chosen player from each game was looked at, and the same game ID was never used in the dataset. It should also be noted that all these datasets were collected in January so some of the data features might be based on the relevant ‘meta’ at the time for example popular champions and their roles at the time.

<u>The features in the original dataset were as follows:</u>

- d_spell - summoner spell on d key
- f_spell - summoner spell on f key
- champion - champion being played
- side - side of map player is on red/blue
- role - role being played out of the 5
- assists - number of assists in a match
- damage_objectives - damage to objectives
- damage_building - damage to buildings
- damage_turrets - damage to turrets
- deaths - deaths in the game
- gold_earned - gold earned in-game

- kda - k/d/a ratio in-game
- kills - kills in-game
- level - level in-game
- time_cc - time crowd controlling others
- damage_total - total damage in the game
- damage_taken - total damage taken in-game
- minions_killed - total minions killed in the game
- turret_kills - turret kills in the game
- vision_score - vision score in game

Preprocessing and Wrangling

For pre-processing we merged the three datasets together to create a larger sample size. This was valid and reasonable as the format for the three datasets was identical, and as our target audience is mainly newer league of legends players who want to improve their game, we believe that the region of the data is not as relevant as players from the highest rank will have similar general playstyle and so combining the datasets seemed to be a useful wrangling tool to increase the reliability and validity of our analyses. After merging the datasets, we decided to remove the rows with any empty columns to clean up our data and ensure consistency, and this was reasonable as due to the large initial sample size, doing so still wouldn't affect the relative size of the sample as much (we still have a large enough dataset for appropriate analyses). Furthermore, columns such as 'd_spell' and 'f_spell' were removed entirely as they seemed irrelevant to our research question. Columns like 'kills' and 'deaths' were also removed as these trends were already captured in the 'kda' attribute.

Analysis Methods

Methods

Based on our aim and target audience, we decided to use linear regression and decision trees as our supervised learning models as linear regression helps to predict ‘gold earned’ which is an end-of-game statistic of high importance in determining the winning team and decision tree help in classifying ‘minions killed’ based on early-mid game decisions which is another important end of game statistic that all players especially beginners should aim to improve on.

Feature Selection

Further feature selection was used to filter out any noise that might implicate our results and choose attributes suitable for classifying the data according to our models. Feature filtering with mutual information (Figure 1) and the use of chi-square tests (for categorical data) were undertaken (results not shown as were not interesting). For the mutual information method, the threshold for the MI value was decided to be 0.4 as that seemed appropriate and the significance level for chi-square tests’ was 0.05.

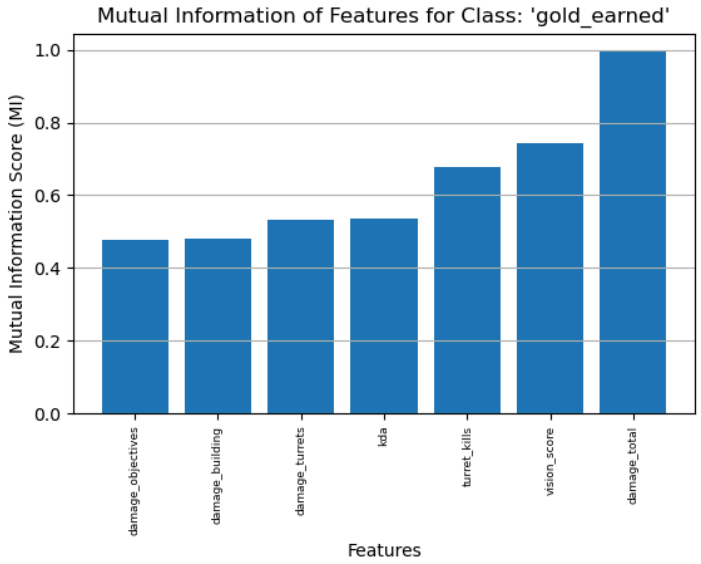

Figure 1.

Training-Test Split

For further analysis, we divided the data into a 70/30 split, for testing and training respectively. The training set was to be used for model fitting the linear regression and decision tree. We used k-fold cross-validation to pick the appropriate linear regression model, and reduce the impact of an 'unlucky split'. Afterwhich, the testing set would be used on the decision tree and linear regression model with the best results based on the performance metrics to produce the most suitable model.

Preliminary Analysis

As our preliminary analysis, we wanted to observe any noticeable trends in the playstyle and decisions made by top players worldwide. Through examining the popularity of champions(10 most picked and least picked), we can immediately see 'meta' champions who are favored by top-rank players and also 'non-meta' players who are discouraged by the best around the world (Figure 2). In this way, instead of researching a large subset of champions, players can follow the top players and assess why some champions are favored over others (usually because they perform better in the current state of the game)

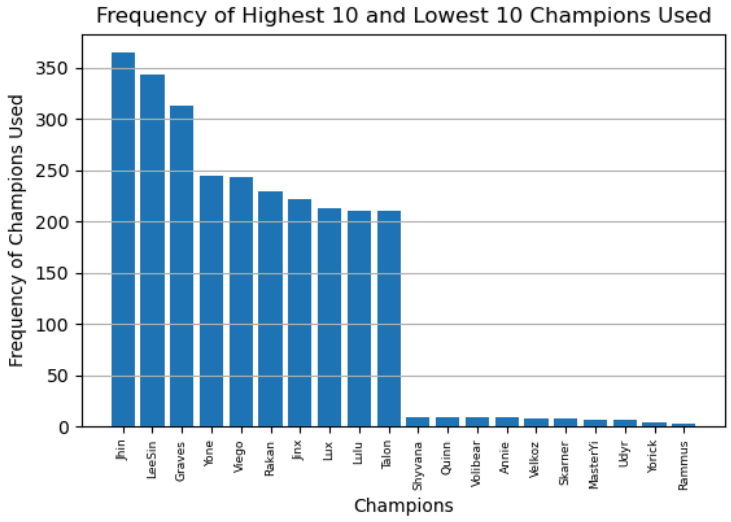

Figure 2.

For the next aspect, we focus on the trend of the amount of ‘Gold Earned’. Gold is the in-game currency of League of Legends. It is used to buy items in the shop that provide champions with bonus stats and abilities, which is one of the main ways to increase their power levels over the course of the game. Therefore, through this analysis, players can give themselves a ‘goal’ to aim for in terms of gold earned by the end of the match, giving them a good indication of where they stand against the top players.

In Figure 3, we used a histogram to display the amount of gold earned over a large range of games from ‘challenger’ players. From the trend, we can observe the average range of gold earned is from 8000 to 12,000 gold earned. There is a symmetric pattern at the average amount of 10,000 gold earned. Therefore, players should aim to get around that much gold earned, depending on their role (more in evaluation).

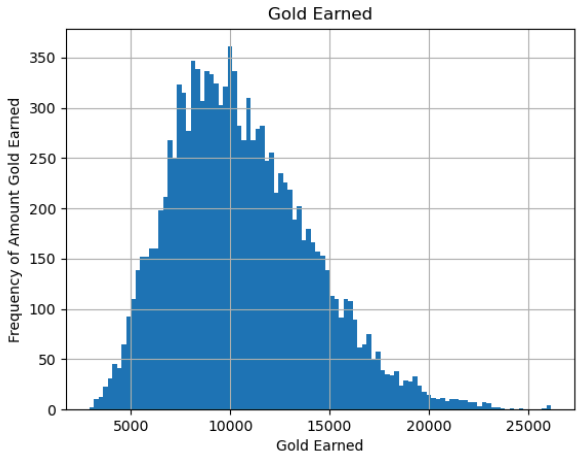

Figure 3.

In League of Legends, controlling minion waves and making decisions around them is important. Minions are the main source of gold and experience, which are critical for players to buy items, level up, and grow throughout the game. For this trend, we focused on the column ‘minions_killed’ in the dataset which is categorized into 2 groups, ‘Few’ and ‘Many’.

For the analysis, we used a bar chart for these 2 categories. We found roughly 7000 players for the group ‘Many’ and slightly more than 5000 for ‘Few’. This is somewhat surprising since you’d expect high-rank players to have more minions killed, but this can be explained by different roles in the game and how they have different goals that may not prioritize gold or minions killed (more in the evaluation section)

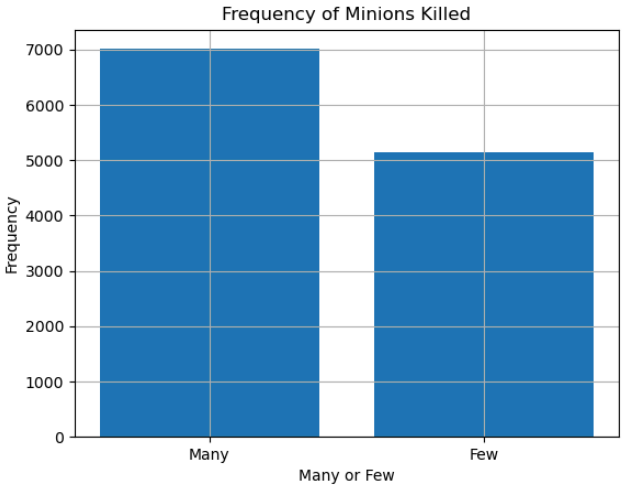

Figure 4.

Modelling

To get a better understanding of whether the features we obtained from our feature selection on gold earned (Figure 1) have a meaningful relationship with the gold earned, we fitted linear regression models. One of these models modelled all the filtered features while the other removed all the ‘damage’ features. We then calculated the R^2, Root Mean Squared Error (RMSE) and Mean Absolute Percent Error (MAPE) by performing a 10-fold cross-validation averaged across the folds to determine the better-fitting model (Table 1). Since the RMSE is heavily influenced by outliers, we also decided to include MAPE which has less of an impact from outliers.

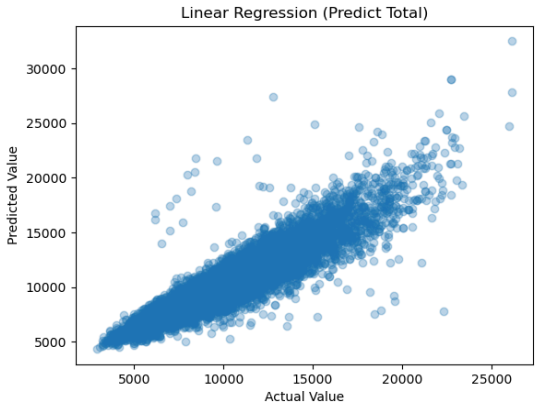

Figure 5a.

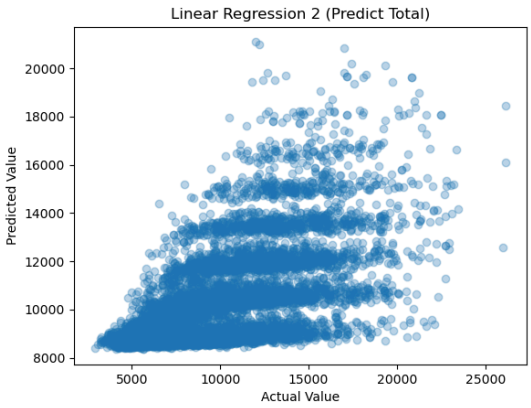

Figure 5b.

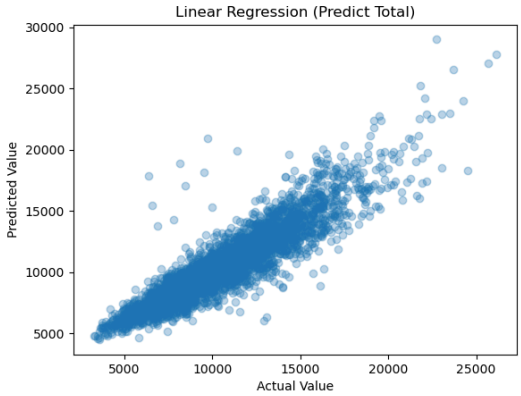

Figure 5c.

Table 1: Goodness of fit of models

	R^2	Root Mean Squared Error	Mean Absolute Percent Error	K-fold Score
Model 1	0.826	1465.269	10.769%	0.825
Model 2	0.316	2905.929	24.430%	0.314
Testing Model 1	0.831	1418.123	10.590	-

We also used decision trees to attempt to make predictions about the amount of ‘minions killed’ based on early-game decisions like champion selected and the role picked for playing the

champion. In this way, beginners can examine the current ‘meta’ of the game by analyzing if a champion and role combination tends to lead to ‘many’ or ‘few’ minions killed.

Discussion

In this section, we look at our graphs and models and discuss the shape and importance of the model. From our feature selection graph (Figure 1), we can tell that the features selected had a high correlation to gold earned. Our first linear regression graph (Figure 5a) supported this finding as it appears to form a linear trend with gold earned (although not too linear), with a mean absolute percentage error of around 10%. This is further proven by the residual plot that showed an even scatter of points around the 0 line which indicates a good fit (although the variance seems to increase around the tail of the model) (Figure 6a)

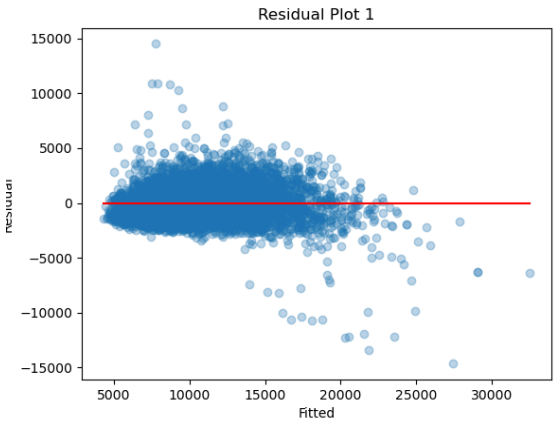

Figure 6a.

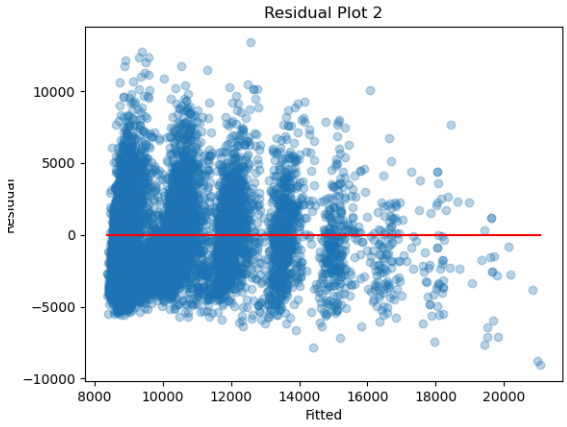

Figure 6b.

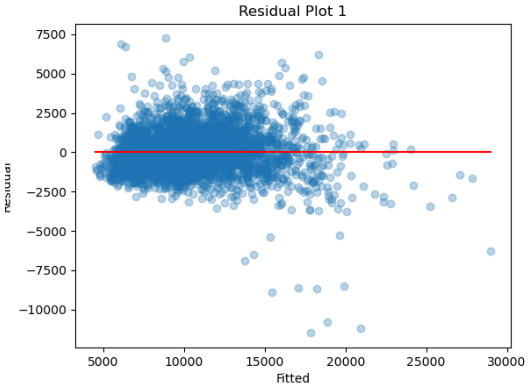

Figure 6c.

However, we felt that any damage done should not have that high of relation with gold earned as in the game League of Legends, dealing damage does not give any gold. As such, we hypothesized that the gold earned does not depend on the damage dealt and will be mainly driven by features such as ‘kda’. Our hypothesis was rejected however as seen in the graph for the linear regression 2 (Figure 5b), as the graph obviously is not linear and does not show any obvious trends, having a MAPE of around 25% and a higher R^2 score than the first model (Table 1).

The residual plot for this model (Figure 6b) also shows a large variance in the data, having results spread far wider and more irregular than for the first linear model (Figure 6a). As such, we can deduce that despite not having any observable impact in the game, the damage dealt does affect the gold earned at the end of the game and therefore we chose our first model as being more appropriate for the data.

Testing our model 1 with the testing data, we can report the testing data follows a similar trend as testing data (Table 1) Figures 5c and 6c show the linear regression graph and also the residuals respectively which also have a similar distribution to the training of the first model (figures 5a and 6a). Thus, we can say that the model is suitable for the dataset.

For our decision tree, we found an accuracy score of 0.93 in predicting the category of ‘minions killed’ based on the champion picked and the role played. As decision trees are made from numerical data in python, we first had to encode our ‘champions’ and ‘roles’ into numerical values through ‘one hot encoder’ and then modelled.

Evaluation

This section will discuss the limitations of our research and how we may look to improve our results and findings in the future. Firstly, our dataset lacked precise clarifications of the roles played, constituting a major portion of the game's gameplay and decision-making. As such, our research may be biased against certain roles that do not view the features we selected as important such as the support role, which usually won’t prioritize minions killed and gold earned etc.

Secondly, since the data for “minions killed” is limited to just “few” and “many”, we are unable to assess to what extent “minions killed” contribute towards “gold earned”. Thus, the link between “minions killed” and “gold earned” may not be as significant as we believe it to be. A clear definition for “minions killed” would be more appropriate, or a split into a larger number of groups would clarify the data much more.

Thirdly, since our data stems from January, the information our research is on may be outdated, especially for the decision tree which is based on champions played and roles. However, we do not believe this limitation has as large of an impact on our linear regression model as they are based on evergreen stats that tend to remain unchanged regardless of which champion is played. Finally, since linear regression models are based on the amalgamation of variables, it is unclear whether the results are based on the combined impacts of the variables or just a single variable. Models such as individual linear models might assist in showing the relation between gold earned and the various features.

Furthermore, another potential limitation of our analyses is that we didn’t really deal with outliers. However, this was a conscious choice as we believe that top-rank players are generally consistent and we have a large sample size; therefore the outliers might be helpful.

References

Suter, Andrew. “LoL Challenger Soloq Data (Jan, Kr-Na-Euw).” *Www.kaggle.com*, 1 Jan. 2022, www.kaggle.com/datasets/andrewasuter/lol-challenger-soloq-data-jan-krnaeuw.

Zach. “K-Fold Cross Validation in Python (Step-By-Step).” *Statology*, 4 Nov. 2020, www.statology.org/k-fold-cross-validation-in-python/.