EODP ASSIGNMENT 2

Problems:

1. Table and figure need label
2. Heatmap - > the outlier successful remove, reduce impact
3. Evaluation

**Research Question**

This project aims to investigate which game features can affect the KDA score in the video games League of Legends and how well we can use these features to predict the KDA score. Moreover, we will determine which game feature have the most significant impact on KDA.

**Target Audience**

The target audience of this report can be people who like playing League of Legends, including professional and amateur players. The research could provide gamers a sight to which aspect they should focus on to increase their KDA. Moreover, the finding also targeted to the Riot Game (Company of LOL), or OPGG (LOL statistical analysis tool), since by comparing how different areas affect the KDA of player, game designer can make a reasonable balance.

**DATASET**

The three datasets from 3 servers, EU match, NA match, KR match from January 2022 has been choose used in this project. The original datasets are authorized by Kaggle, we use the modified version of them.

These datasets were initially store in CSV file format. The EU match dataset contained 5770 summoners records, the NA dataset contained 5759 summoners records and KR dataset has 5696 summoners records. Each dataset contains 21 game factors. The KDA was the response variable in this project.

There are 20 game factors used as explanatory variables and 1 response variable in the investigation:

|  |  |
| --- | --- |
| Features | Description |
| 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 |
| assists | number of assists in match |
| damage\_objectives | damage to objectives |
| damage\_building | damage to buildings |
| damage\_turrets | damage to turrets |
| deaths | deaths in game |
| kda | k/d/a ratio in game |
| kills | kills in game |
| level | level in game |
| time\_cc | time crowd controlling others |
| damge\_taken | total damage taken in game |
| turret\_kills | turret kills in game |
| vision\_score | vision score in game |
| damage\_total | total damage in game |
| gold\_earned | gold earned in game |
| role | role being played out of the 5 |
| minions\_killed | total minions killed in game |

**Pre-processing and wrangling**

1. Combine the three datasets into one

For pre-processing the original dataset, we firstly convert the three csv file into three dataframe, and then we used the concatenation method to combine them into one dataframe. There are 17228 of row in total.

1. Remove Nah value

We notice there are some values missing completely random. Because the sample size is large enough, we decide to remove all the Nah values (missing) and their corresponding row to ensure data's accuracy. 7771 rows left after remove Nah value. Dropping null data is the option rather than filling average data because the remaining observations are adequate for analysis.

1. Convert type function

Moreover, we realized the formats of datasets are inconsistent, the column contains both strings and integer. The designs are not suitable for further analysis; thus, we apply the convert type function to reformat the string data, i.e. champion, side, role and minions\_killed into unique categorical value (integer) for easy to process by machine.

1. Remove Outlier

Use which method? function, which quantile? How many outliers presented/or been removed? Visual graph?

Step 4 is removal of outliers, kda outliers should be removed because it will noise the regression in modelling.

1. Level consideration

Each LOL game duration can last from a few minutes up to one hours. Most accompanied data will considerably increase as the game length increases. The level is a good representative for time duration, we only focus on the successful matches which the match is not end at starting level. We decide to remove the data that have level less than 6, as the ultimate is ready in level 6. The filter the data by game level in each match since the histogram graph shows that most of these matches end in this level range.

**Analysis**

**Method**

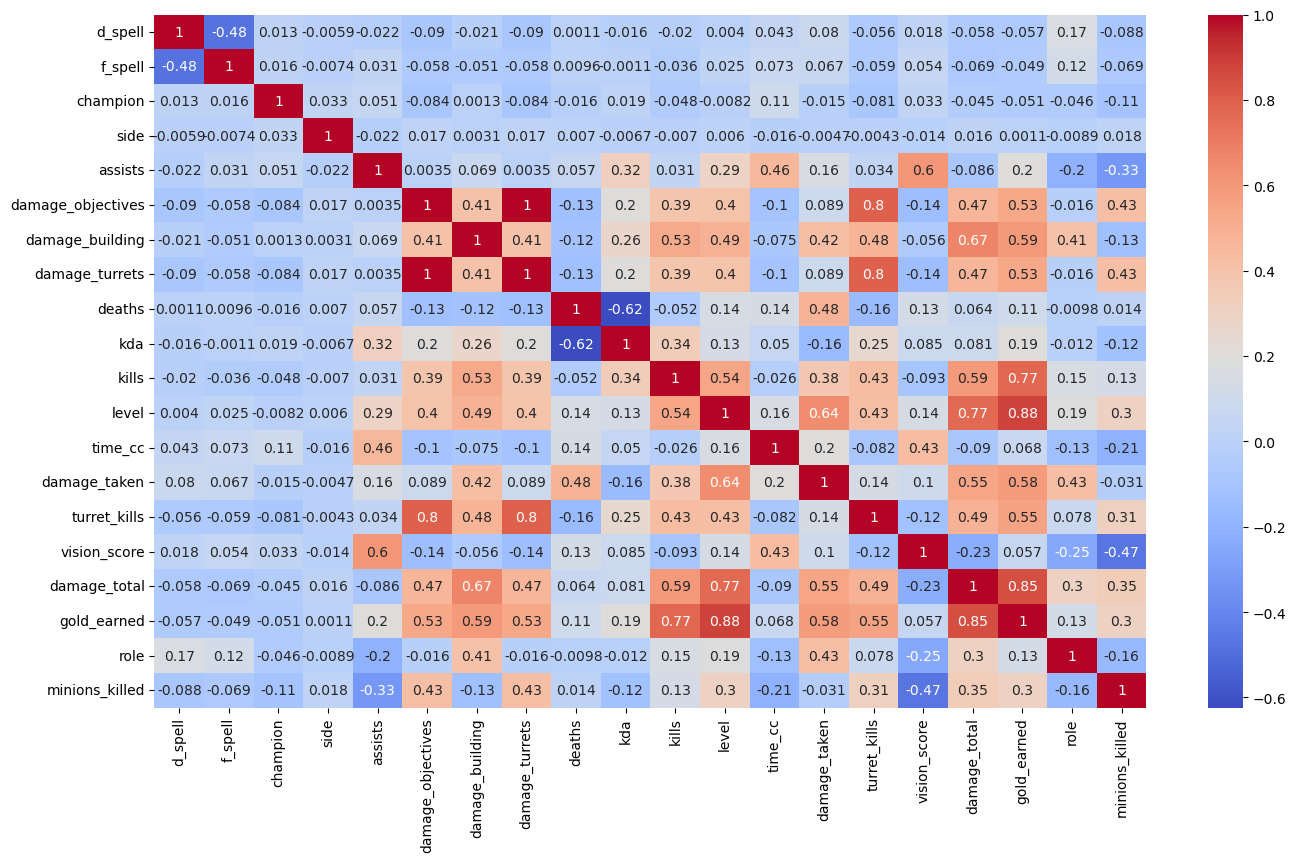
Techniques used in this report include feature selection with Pearson correlation and linear regression modelling, decision tree regression, and random forest regression. Feature selection is implemented in removing unrelated factors and improving the performance of the model. Linear regression is a suitable model in constructing relationships between a factor(kda) and multiple factors. With the model constructed, decision tree regression and random forest regression are used to test how the model built before performs.

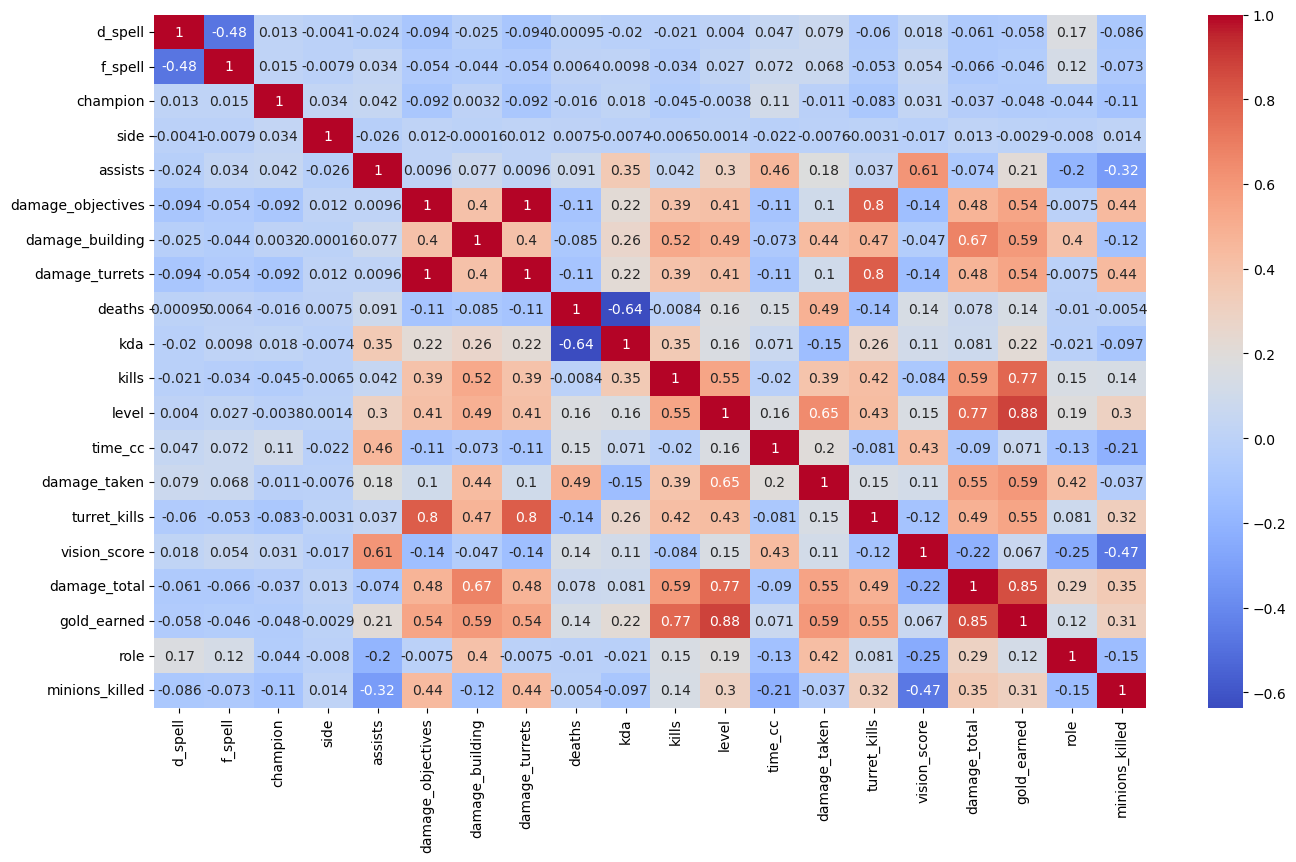
**Training test split**

For further analysis and prediction. The data was randomly split into training set and test sets by a ratio of 8:2. The training set was used for preliminary analysis and model fitting, while the test set was used to assess how well the training set model fit generalised. In addition, ?? validation has been done to minimise the impact of an " unlucky" split.

**Preliminary Analysis**

In this stage, Pearson correlation between each factor is calculated and drawn in a heatmap. Two heatmaps represent the correlation before and after outlier removal respectively.





As the graphs show, the configuration before and after are similar, implying removal of outliers are intact for the data. Then, based on background information, features kills, deaths and assists should be removed since those factors are the elements in calculating kda ratio. After that, the removal of the least six correlative factors (damage\_taken, minions\_killed, role, d\_spell, side, f\_spell) is executed because their correlation is too low.

**Feature selection**

1.Using heatmap (visual technique -> 1. how feature correlated to each other 2. outlier remove）

2. Pearson correlation -> remove less 0.01

3. Back Elimination

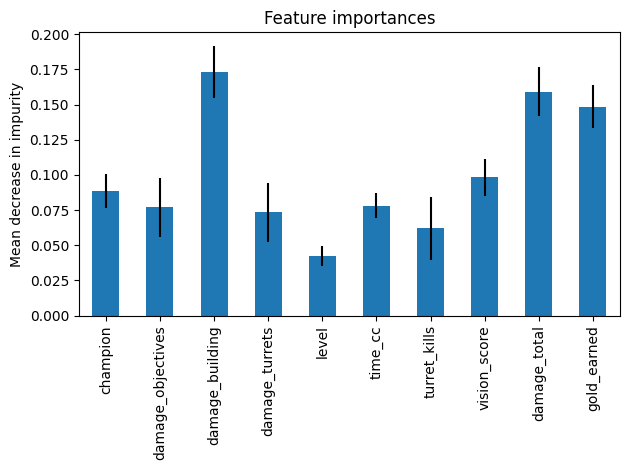
**Model**

1. Linear regression

3. Decision Tree

4. Random forest

5. Feature importance

To get the high accuracy result, lots of graphs are drawn. They are heatmap, scatter plot, barchart. From all those graphs, the low relative factors and high relative factors are shown directly. The low relative factors are  f\_spell; d\_spell, side; role; damage\_taken; minions\_killed). The top three high relative factors except kills, death and assists are  damage\_building, damage\_total and gold\_earned. These three are not only relative to kda but also relative to each other. This means if one of these three factors are high, the other two will be more likely to be high.

However, there is a bit datas that shows that there are high kda and those three factors are low. That’s the reason the EU, KR and NA are joined together. More data will get more credible results and there will always be outliers.

**Evaluation**

Before the graph, the assumption is time\_cc and vision\_score should be high relative, but this is not shown from the result. The result of why people who have high kda will also have high damage\_building, damage\_total and gold\_earned is due to this reason. Normally people like to stay in the turret and only few people have a chance to kill the turret when there is another champion under the turret. So only the people who have high gold\_earn can buy more and better weapons. Then, they will have the ability to kill more champions. If there are not another side champions under the turret, they will destroy more turrets. So that’s how high kda players have high damage\_building, damage\_total and gold\_earned.

**Limitation**

Main limitation should include the following two points:

1. How is our model? good or bad? which model might give a better prediction that can apply in the future.

This project has several limitations. Firstly, these models displayed flawless R2 values, particularly the linear regression, indicating that the model might not be valid. Other models that might be better for this research, such as artificial neural networks, which often perform better than most models.

2. How does the feature affect / limit the research?

Secondly, the results are limited because only a subset of potential features is taken into consideration. Since several crucial features are not included in the dataset, this could lead to incorrect results. There might be a deeper underlying relationship between features, however we only test whether they are correlated and select the most predictive ones but not correlated to other features. Prior to conducting further analysing, we can normalise the feature for improvement.