Element of Data Processing

Assignment 2

Group 61

Jinuo Sun
Haohong Liang
Melody Chi
Yufeng Liu

Research Question

This project aims to investigate the correlation between high game performance (measured as kda value) and other in-game data in the video games League of Legends(LOL), including how well one can use given features to predict kda performance. Moreover, this project also contributes to the interpretation of the remaining data characteristics corresponding to high-kda game performance, which could provide an important feedback for further research on the game's ecology.

Target Audience

The target audience of this report could be LOL players, including professional and amateur players. This research could provide players with a global perspective to further understand the logic behind the game, in order to provide general direction and basic guidelines of the gameplay for beginner players; also deepen the understanding of the game for advanced players, and provide data support for further development of game tactics. Moreover, the findings are also targeted to the Riot Game (Company of LOL), or OPGG (LOL statistical analysis tool) as an analysis of game's ecology through data correlations, which could help the game operation team to ensure their design is under a expected balance.

DATASET

Three datasets from 3 servers, EU match, NA match, KR match from January 2022 had been chosen in this project. The original datasets were authorized by Kaggle, and a modified version of them was used.

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 the KR dataset contained 5696 summoners records. Each dataset contained 21 game factors. The kda was the response variable in this project.

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

Table 1: Game features and their description

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	kills/deaths/assists 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, firstly converted the three csv files into three dataframes, and then used the concatenation method to combine them into one dataframe with 17228 rows in total.

2. Remove Nah value
Completely random missing values were noticed in the dataset. As the sample size was large enough, removing was taken upon all the Nah values (missing) and their corresponding row to ensure data accuracy. 7771 rows were left after removing Nah value. Dropping null data was used rather than filling average data as the remaining observations were adequate for analysis.

3. Convert type function
The formats of datasets were inconsistent that the column contained both strings and integers, which was not suitable for further analysis; thus, the convert type function was applied to reformat the string data(i.e., champion, side, role and minions_killed) into unique categorical value (integer) for easily processing by machine.

4. Remove Outlier
Boxplot was used to judge the threshold of the outliers. In the boxplot, numbers lower than the first quartile minus three times interquartile range ($Q1 - 3 \times IQR$) or greater than the third quartile plus interquartile range ($Q3 + 3 \times IQR$) were thought of as outliers and would be removed.

Number of observations decreased from 7771 to 7313 after removal, see figure 1.

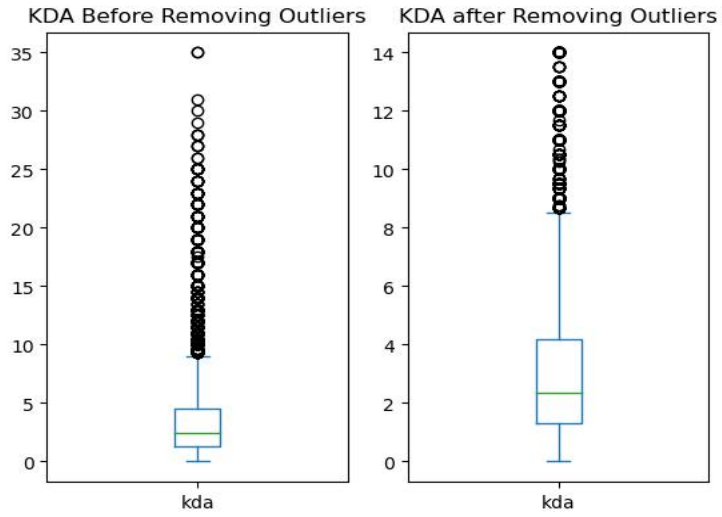

Figure 1: original distribution V.S. distribution without outliers

5. Level consideration

Each LOL game duration can last from a few minutes up to over one hour, while most accompanied data will considerably increase as the game length increases. To have a more accurate analysis, this research focused on matches which were not end at starting level. The level feature was a good representative for time duration, based on which removal was taken upon in-game data that have level less than 6, as the ultimate skill is ready in level 6 which could considerably influence kill and damage. As shown in the Figure 2 histogram, this level 6 filter would not harm the big picture of the analysis as the amount that matches ended below level 6 range was quiet limited.

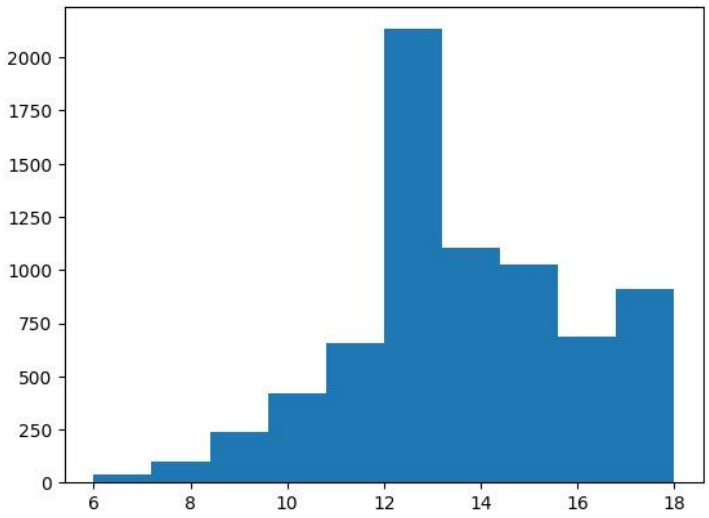

Figure 2: Level distribution in dataset

Analysis

Method

Techniques used in this report were composed of: feature selection based on Pearson correlation Heatmap and back elimination; modelling using linear regression modelling, decision tree regression, random forest regression and feature importance given the goal and datasets, as these techniques are fittable for dealing with continuous data and investigating relationships between factors. For visualizing the result, box plots and heatmaps were used.

Training test split

For further analysis and prediction, the data were randomly split into training sets 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, cross-validation has been done to minimise the impact of an " unlucky" split.

Preliminary Analysis of feature selection

The heatmap had been used as a visualisation technique to provide a preliminary understanding of how each game factor related to kda. In figure 3 and figure 4, the Pearson correlation heatmap shows the linear relationship between each factor, where the warm colour represents features that were positively related, and the cool colour represents features that were negatively related. As the heatmaps show, the configuration of correlation before and after were similar, implying removal of outliers only had a slight impact.

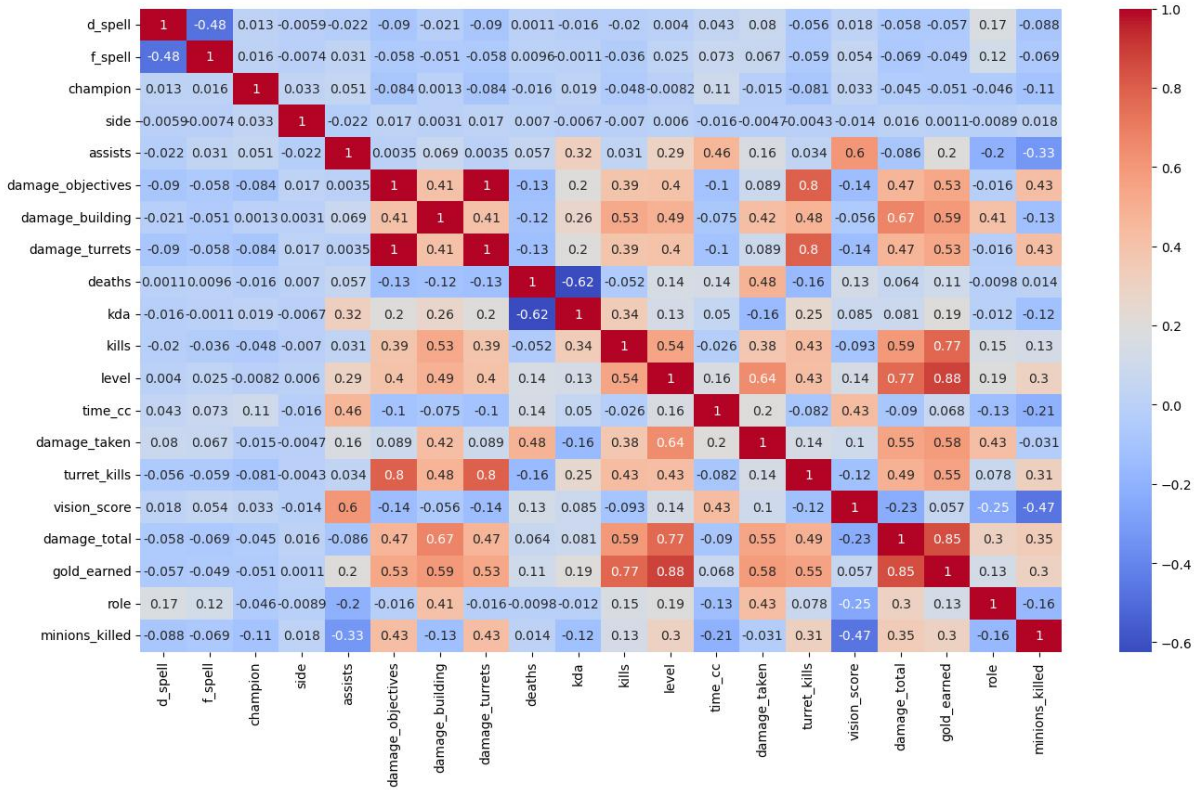

Figure 3: Heatmap for initial dataset

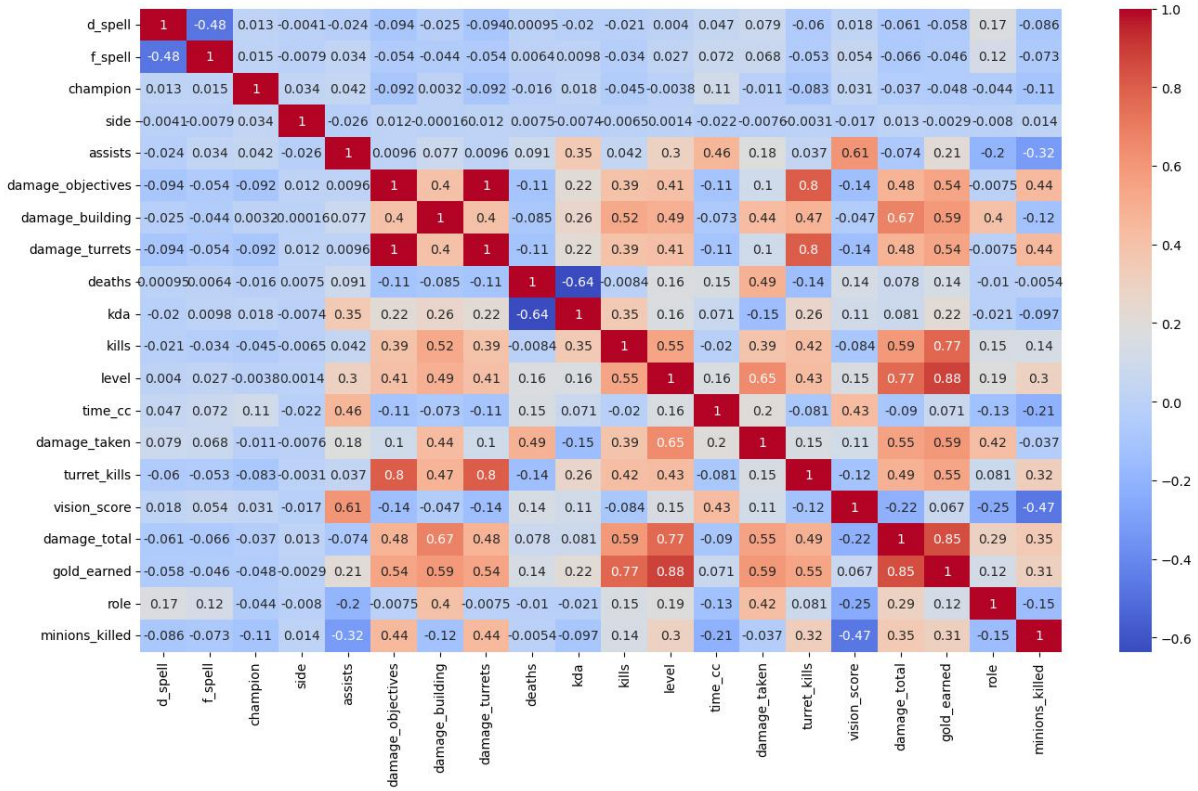

Figure 4: Heatmap for dataset after removing outliers

Then, based on background information, features kills, deaths and assists should be removed since those factors were the direct elements for calculating the kda. By using the

Pearson Correlation value in the heatmap, removals were taken upon the data that had an absolute PC value of less than 0.05. Therefore, f_spell, side, champion, d_spell and role had been removed from the dataset as they were considered to have an extremely weak correlation to our target label in kda. As these features with PC values not presenting a strong correlation could lead to a possible mediocre performance of the model.

Furthermore, backward elimination was used to choose the essential features from the data through gradually removing less important features and left the most important ones eventually. P-value threshold of 0.05 was used as a measure of how likely a feature would be necessary by calculated and then compared. Using backward elimination, nine features were selected to survived for further modelling and analysis eventually. Three features 'champion' 'damage_objectives' 'damage_turrets' were removed due to relatively low importance.

Modelling

Linear regression models were utilized to gain a better understanding of whether a factor had a considerably close association with kda. Nine features were selected to fit the model from the training set, and R^2 values from both training set and test set were recorded. Additionally, 5-fold cross validation was performed and calculated under same metrics to generate an average for gaining more reliable results. Also, RMSE had been calculated to help understand the stability of model.
For this particular dataset and our analysis goal, linear regression could be too frail that might be easily affected by collinearity, therefore regression tree was chosen for a potential more fittable analysis. Similarly, R^2 values, 5-fold cross validation and RMSE had been recorded. Figure 5 shows a visualized regression tree.
Moreover, the random forest regression could even gain a more accurate and reliable analysis as it operates a set of regression trees for accurating the output. One hundred decision trees was run in this research. Similarly, R^2 values, 5-fold cross validation and RMSE had been recorded.
Last but not the least, impurity-based feature importance model was applied to find the determinant features.

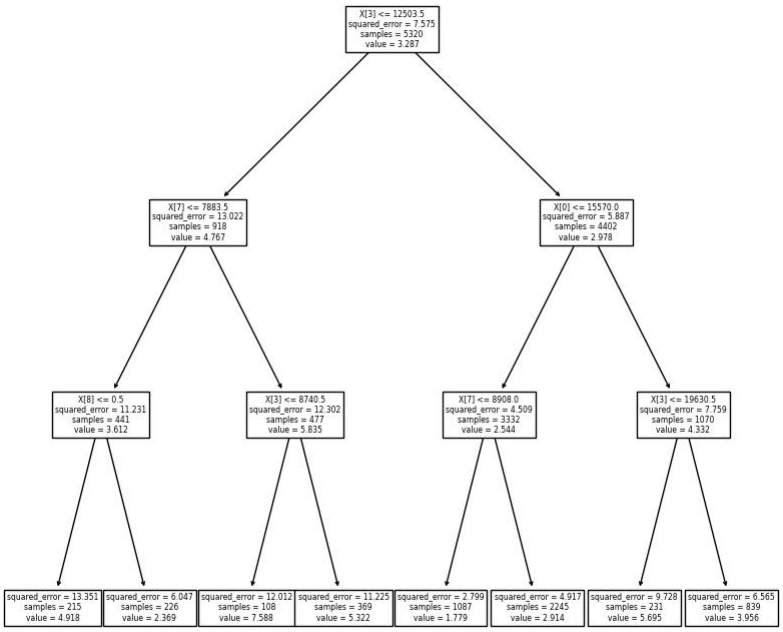

Figure 5: Decision tree model

Discussion

In this section, it will interpret the models and discuss the significance of the results. Refer to the Regression Models table below.

For Linear Regression Model, the Train (R^2) and Test (R^2) score are both low, indicates the correlation coefficients of each feature are low, this pointed out a weak linear relation exists between features and kda. Besides, The RMSE value is not large but still shows that there could be ± 2.246 value between the predict and true value. These finding also supported the low Pearson Correlation value that gained in the preliminary analysis section. Therefore, the Linear Regression cannot be the satisfied model.

For decision tree regression, the train (R^2) value is 1.0, however, after doing 5-fold CV, the score is -0.2402, which represents that the model is overfitting in the train set, indicates the model didn't perform well on a single decision tree. Through RMSE, the regression tree performs even worse than the linear regression models. Therefore, decision tree regression model also cannot have statistical significance.

Consider the last model random forest regression, the train (R^2) value has increase to 0.9159 which is within the expected performance. In addition, compared with the previous two models, the RMSE value in random forest regression is lower which also represents that there has smaller value difference between the predicted value and the true value. However, although the test (R^2) value has increase to 0.4399, but the difference between train and test set value is up to 0.476, it denotes the train set is inflated.

In conclusion, it is random forest regression model is the most meaningful model that could better demonstrate the relationship between features and kda.

Table 2: Regression model table

Model	Train (R^2)	Test (R^2)	5-fold CV (R^2)	RMSE
Linear Regression	0.3237	0.3249	0.3200	2.2246
Decision Tree Regression	1.0	-0.2648	-0.2024	3.0452
Random Forest Regression	0.9159	0.4399	0.3867	2.0263

Below the figure 6 shows that the relationships between feature importances and the mean decrease in impurity. Simply can get that the top three high relative factors except kills, death and assists are damage_building, damage_total and gold_earned, while damage_building has the highest value 0.1725 which represents that it has the strongest connection to kda than any other features. Meanwhile, the feature level has the lowest value 0.044 which performs weak relationships with kda.

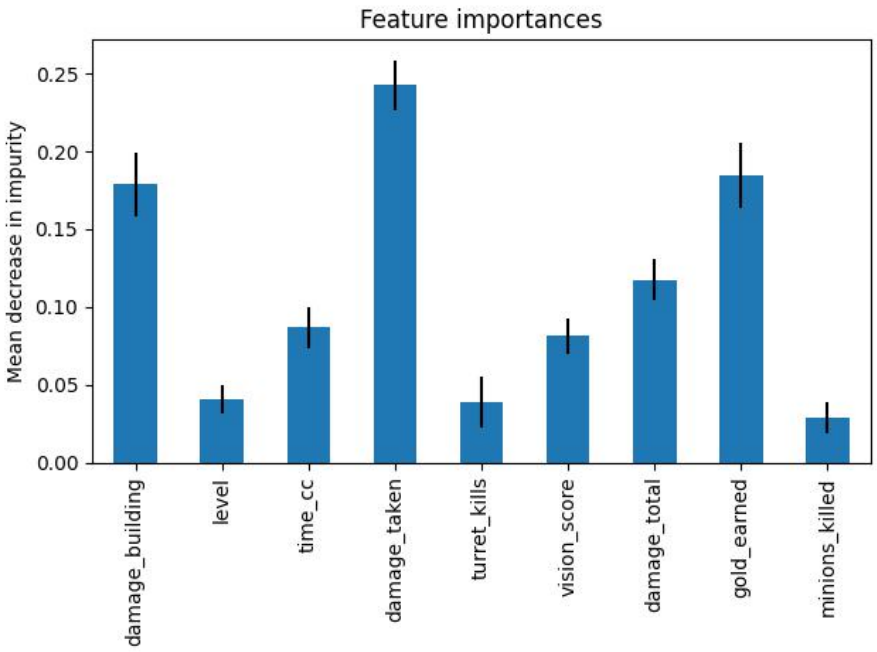

Figure 6: Feature importance with mean decrease in impurity.

Limitation

Several limitations during research could exist in this project. Firstly, these models displayed flawless R^2 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.
Secondly, the results are limited because only a part of potentialn correlated features were provided in the dataset. Since several crucial features are not included in the dataset, this could lead to inaccurate or inappropriate analysis.
Furthermore, deeper underlying relationship between features could exist, which might lead to imperfect multicollinearity issue, resulting in regression coefficients being estimated

imprecisely or having large standard errors, and therefore might mistakenly lead to statistically insignificant regression coefficients. Prior to conducting further analysing, we can normalise the feature for improvement.
Last but not the least, we assume three datasets to be equivalent, while we do not exactly know whether they have the minor difference among regions or not. These could also be taken into further consideration in future research by pre-analysing the given datasets.