ORIE 4741 Midterm Report:
Advice for players to improve their game experience in a LOL match

Jin Cui(jc3362), Yifan Zhu(yz2522), Yuqi Hong(yh854)

November 2019

Understanding Data in Greater Detail

The data used for analysis is from Kaggle dataset, which provide different elements of each match of the game and also the result of the game. For example, the champions selected by players, different roles played by players, damage/heal made by champions, kills/deaths of each player, etc.. To be specific, seven csv files are provided for analysis, which are:

- Teambans.csv: information about the team ban, which contains 'match_id', 'team_id', 'champion_id', 'ban_turn'
- Champs.csv: information about champions used by player, which contains 'name', 'id'
- Participants.csv: information about participants, which contains 'id', 'match_id', 'player', 'champion_id', 'ss1', 'ss2', 'role', 'position'
- Matches.csv: information about each match and game version, which contains ['id', 'game_id', 'platform_id', 'queue_id', 'season_id', 'duration', 'creation', 'version'
- Stats_1.csv + Stats_2.csv: which contains specific detailed information about each game match, for example, 'win', 'trinket', 'kills', 'deaths', 'assists', 'largest_killingspree', etc..
- Teammates.csv: information about teammates' performance, which contains 'first_blood', 'first_tower', 'first_inhib', 'first_baron', 'first_dragon', 'first_harry', etc..

To make full analysis of what elements may make a team win a match, we will try to implement as many as data we have acquired to make the prediction from different aspects.

Our Goal

The goal of learning from League of Legends (LOL) is to give some suggestions to improve players game experience. For example, gamers would know the most important factors that may help them win the games and they also may know the elements that may help them acquire pentakills.

Plan to Avoid Overfitting and Underfitting

Avoid Underfitting:

- We tried to use more features and trained the model by large dataset to avoid underfitting

Avoid Overfitting:

- We checked the correlation between every feature and the interested outputs and determined the features with the high correlations and dropped the features with the low correlations.
- We used regularization and cross validation to avoid overfitting

- ## Test the Effectiveness of the Models We Develop?

1. Check the difference between training and test accuracy. It is always an obvious and fast method to check whether our model is overfitting or not. For our model, the difference is almost zero.
2. Use validation to evaluate the model. For example, cross validation is always the first choice.
3. Record how long it takes for model to train on data set. It is also important that our model is of high training efficiency.
4. Use other metrics to evaluate the model, such as F1-score, recall, precision, ROC-AUC, etc..

- ## Cleaning up, Features and Examples

Cleaning raw data

- There are two datasets (stats1 and stats2) include all matches information with same columns, so we combine them together into one csv file.
- We merge all information from participants, champions, matches and stats by using join method.
- Add a new column which presents the players' role, let it only includes five roles: TOP, MID, JUNGLE, DUO_SUPPORT and DUO_CARRY.
- Add a new column which presents the label of different teams by transferring player's label.
- Remove all duplicate roles in one team.

In the cleaned dataset, we have 1486362 examples with 76 features.

- ## Data is Missing or Corrupted

There are not much data missing or corrupted, when we were working on merging different datasets, there were a little duplicate or invalid examples needed to be dropped.

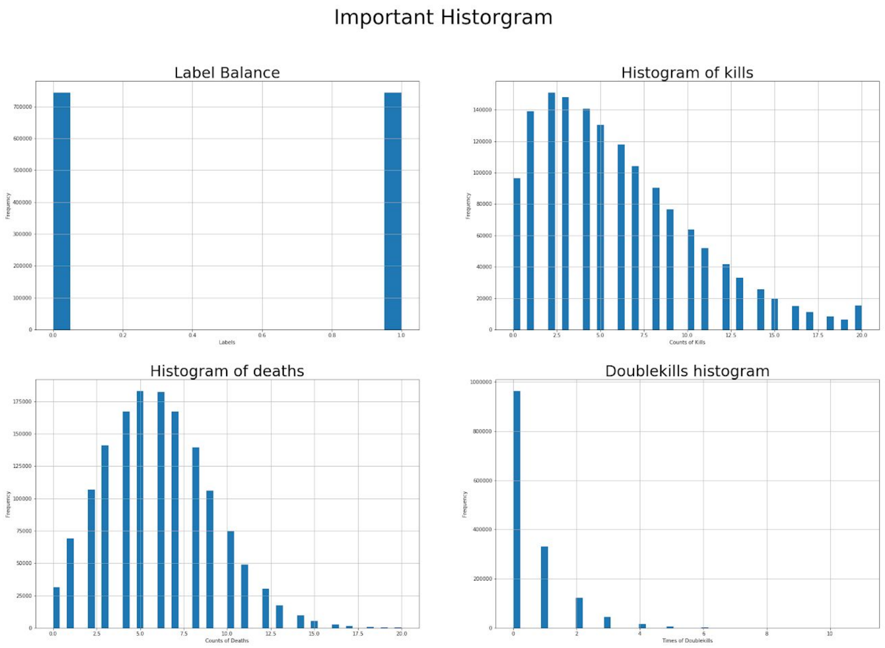
- ## Features (and Transformations) to Use

We used correlation matrix to help us determine which feature should select. First we select all numeric columns, then we plotted the correlation matrix to see how many features are highly correlated, then selected them from high to low scores. Finally we selected 23 features to be used for prediction.

- ## Histograms or Other Descriptive Statistics

There are four histograms we plotted below.
The top left plot presents the label of game result, so there are only two bars, one at 0 which stands for lose, another at 1, which stands for win.
The top right plot presents the histogram of kills, we can find around 2 kills has the highest frequency, and the number of kills most stay from 0 to 10.
The bottom left plot presents the histogram of deaths, we can find around 5 and 6 death has highest frequency, and the plot is close to normal distribution.

The bottom right plot presents the histogram of double kills, we can find 0 double kills has highest frequency, and it keeps decreasing for larger times of double kills.



- **Run a few preliminary analyses on the data**

  Features selected for the baseline model are based on the correlation matrix, which is presented by a heatmap. Features have relatively large correlation(both positive and negative) with targets are used for training the model.

  For preliminary analysis, logistic regression with L1 norm is used for getting the first sense about how well the data is presented. Tool used is based on scikit-learn linear model - logistic regression. Parameter ‘solver’ is tuned as ‘liblinear’ to make the training process faster. Cross validation is not used here, but will be implemented to prevent overfitting for future analysis.

  The result of baseline model is much better than the team expected. The training set accuracy is 85.483%, while the test set accuracy is 85.482%. While the test accuracy is pretty high, there is evidence showing that our baseline model is not overfitting.

- **what remains to be done, and how you plan to develop the project over the rest of the semester**

  We only used part of data to analyze the correlation with the win rate and some basic statistical analysis. In the following weeks, we will use more data to find the relationship between input features and how to get more kills (pentakills) in the game. We will also checked how other factors besides champions we picked affect the matches by using the same way, which we did until this part.