Report of Project Stage 2

szha5691 470144491 SHAOWEI ZHANG

Setup

a. Question to answer

This research project is proposed to answer the question “how to increase the win percentage in Dota2 games” based on a dataset from Dota2 games. To be specific, the association between game results and different factors is focused on to answer this question.

b. data

This project is based on a dota2 dataset downloaded from the website “kaggle” and uploaded by Devin. This dataset consists of 18 CSV files and the data used for this project is derived from them. Five features are derived as Table 1 shows:

player_id	match_id	GPM	XPM	KDA	P/F	C/P	result
1	0	347	362	9.00	0.02	0.6	win
2	0	494	659	10.33	0.02	0.6	win
3	0	350	385	3.75	0.09	0.6	win
4	0	599	605	6.75	0.29	0.6	win
5	0	613	762	12.33	0.05	0.6	win
6	0	397	524	2.17	0.01	0.6	lose
7	0	303	369	0.69	0.00	0.6	lose
8	0	452	517	1.25	0.19	0.6	lose
9	0	189	223	0.64	0.00	0.6	lose
10	0	496	456	0.64	0.00	0.6	lose

Table 1 Part of the data derived from dataset

In Table 1, “player_id” represents the sample ID which is the performance of a player in a certain game. “match_id” represents the certain match where this sample comes from. “GPM” is the abbreviation of “Gold Per Minute”, which represents the player’s speed of earning money in a game. “XPM” is the abbreviation of “Experience Per Minute”, which represents the player’s speed of earning experience in a game. “KDA” is the abbreviation of “Kills / Deaths / Assists”, which represents the player’s fighting performance. It is calculated by “(Kills + Assists) / Deaths”. “P/F” represents the ratio between “Push” and “Fight”, which can measure the extent of player’s tendency to “Push”. It is calculated by “Tower_damage / (Tower_damage + Hero_damage)”. “C/P” represents the ratio between “Carry” and “Support”, which can measure the extent player’s tendency to get a more reasonable team. It is calculated by “Carry_number / (Carry_number + Support_number)”. “result” represent the outcome of a game. “GPM”, “XPM”, “KDA”, “P/F” and “C/P” will be referred to as features of player’s performance.

To eliminate the differences between the scales of the features, the features is normalized using Formula 1:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \cdot (1 - 0) + 0 \qquad \text{(Formula 1)}$$

The result is labeled with numbers. “win” is labeled as 1 and “lose” is labeled as 0. The final data for use are showed in Table 2:

player_id	match_id	GPM	XPM	KDA	P/F	C/P	result
1	0	0.16	0.23	0.16	0.02	0.6	1
2	0	0.26	0.42	0.18	0.02	0.6	1
3	0	0.17	0.25	0.06	0.09	0.6	1
4	0	0.33	0.39	0.12	0.29	0.6	1
5	0	0.34	0.49	0.21	0.05	0.6	1
6	0	0.20	0.34	0.04	0.01	0.6	0
7	0	0.14	0.24	0.01	0.00	0.6	0
8	0	0.23	0.33	0.02	0.19	0.6	0
9	0	0.06	0.14	0.01	0.00	0.6	0
10	0	0.26	0.29	0.01	0.00	0.6	0

Table 2 Part of the final data for use

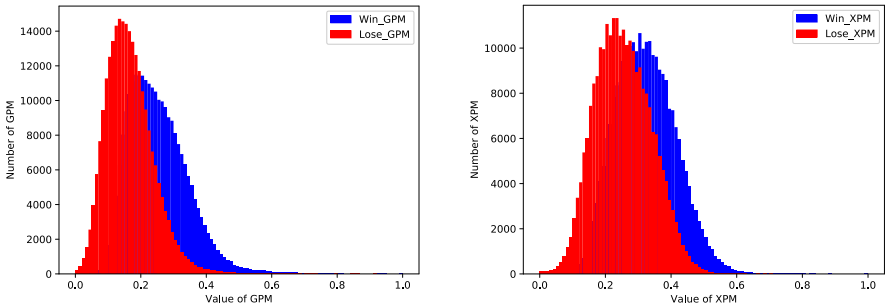
c. Null and alternative hypothesis

This project performs significance tests between features of winning games and losing games. For instance, whether there’s a significant difference between GPM data of winning games and losing games should be tested. If there’s no significant difference, GPM should not be considered as a feature any more.

The null hypothesis H_0 is “The two groups of data has no significant difference” and the alternative hypothesis H_1 is “The two groups of data has a significant difference”.

d. Quantification of reliability

This project performs two hypothesis tests to quantify reliability of the features. One-way analysis of variance (ANOVA) and Kruskall-Wallis H-test will be used. Whether H_0 is to be rejected depend on the P-value of these two tests. The distribution of GPM and XPM is showed as examples in Fig. 1:



a) **Distribution of GPM** b) **Distribution of XPM**

Fig. 1 Distribution of features from winning and losing games

From Fig. 1 we can detect the difference between features of winning and losing games. However, we still need quantitative results to support our detection:

	GPM	XPM	KDA	P/F	C/P
ANOVA	2e-4	2e-4	2e-4	2e-4	1e-4
Kruskall-Wallis H-test	2e-4	2e-4	2e-4	2e-4	2e-4

Table 3 P-values of each feature by two hypothesis tests

Table 3 shows that P-values of each feature by two hypothesis tests are all smaller than 0.01, which means there's almost certainly a reliable difference between features of winning games and losing games. There is only a 0.01% to 0.02% chance of falsely rejecting H_0. All of these five features are suitable for this project.

In addition, the result of a game can only be "win" or "lose", so a classification will be performed and the importance of the features will be calculated to answer the original question above. To quantify the reliability of the classifier, Precise, Recall and F-Measure will be used to evaluate the performance of classifiers. Also, False Positive Rate (FPR) and True Positive Rate (TPR) will be used to plot ROC curves which can more efficiently visualize the performance of classifiers. These will be showed in the section of result.

Approach

a. Classifiers

This project will using 3 kinds of classifiers to perform the classification. They are Logistic Regression, K-Nearest Neighbors and Gaussian Naïve Bayes. The n_neighbors parameter of K-Nearest Neighbors is chosen as 5 after optimized. Their corresponding Confusion Matrix will also be plotted to visualize the accuracy of their prediction. Also, the FPR and TPR of their prediction will be calculated and their corresponding Receiver Operating Characteristic (ROC) curves will be plotted to visualize the comparison between them. The classifier with the best performance will be chosen for further research.

After the most suitable classifier is chosen, the importance of each feature will be scored and ranked by using the method "SelectKBest", which is used to rate the relevance of every feature with the target label. The features can be ordered and this result will tell us what we should do to win the game. For example, if "GPM" is the most important feature, then it means dota2 players should try hard to speed up earning money to win the game.

b. Flow chart of the project

The approach of this project can be narrowed down into the flow chat below:

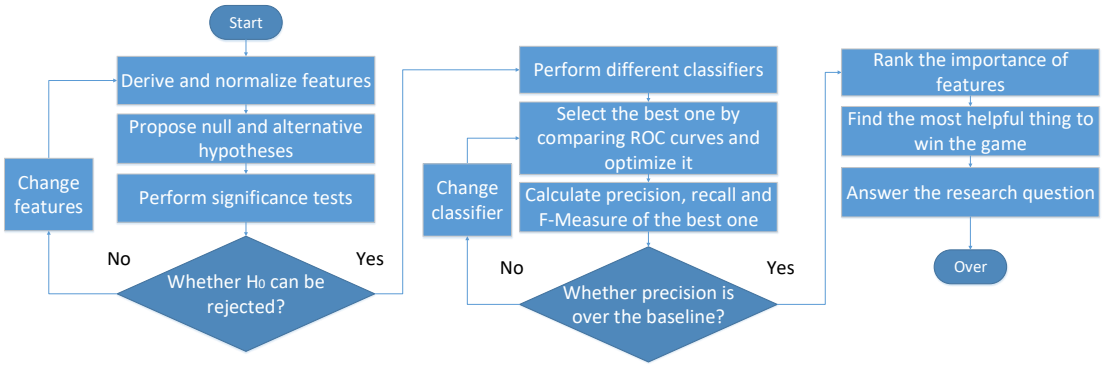
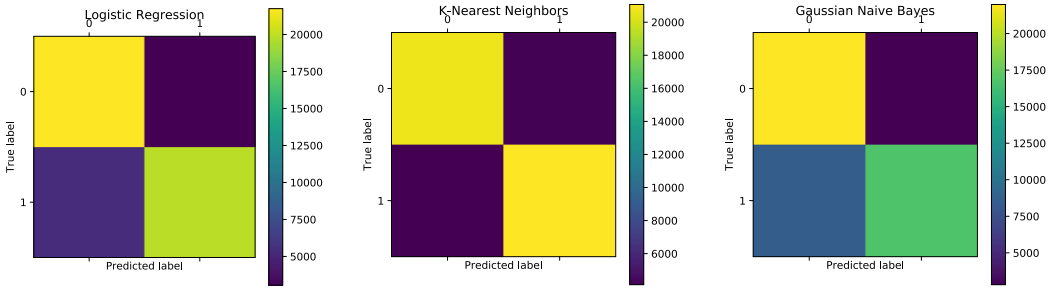


Fig. 2 Flow chart of the project

Results

a. Classifier selection

This project will use Logistic Regression, K-Nearest Neighbors and Gaussian Naïve Bayes for comparison. Fig. 3 shows the prediction of these classifiers with confusion matrixes:



a) **LR** b) **KNN** c) **GB**

Fig. 3 Confusion matrixes of three classifiers

Fig.3 shows that all these three classifiers have good performance. After comparison, Logistic Regression and K-Nearest Neighbors have a better accuracy. To select the best one, we need to plot their ROC curves in Fig.4:

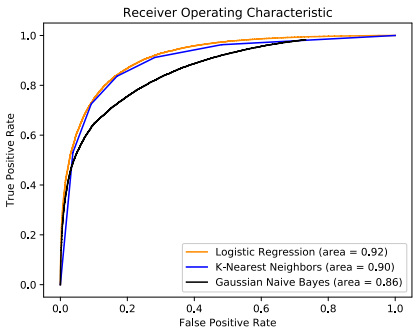


Fig. 4 ROC curves of three classifiers

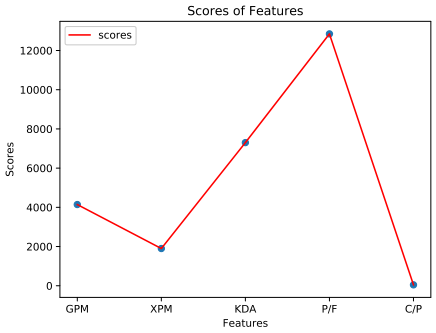


Fig. 5 Feature scores

Fig. 4 verifies the conclusion above that Logistic Regression and K-Nearest Neighbors have a better performance. There's almost no difference between Logistic Regression and K-Nearest Neighbors according to their ROC curves. Logistic Regression is slightly better so this project will using it for further research.

b. Optimization for Logistic Regression

This project uses "grid_search" method to optimize the parameters. One with C values in [0.1, 1, 1e1, 1e2, 1e3, 1e4, 1e5, 1e6, 1e7], penalty in ['l1', 'l2'] and class_weight in [None, 'balanced'] is considered. The result shows that the best parameters is {'C': 1000, 'class_weight': 'balanced', 'penalty': 'l2'}. The calculation of precision, recall and F-Measure will be based on these parameters.

c. Precision, Recall and F-Measure

This project will use a 10-fold cross validation to calculate the precision, recall and F-Measure (f1-score). The result is showed in Table 4:

Result	precision	recall	f1-score	support
lose	0.80	0.88	0.84	24808
win	0.87	0.79	0.82	25192
average/total	0.83	0.83	0.83	50000

Table 4 10-fold cross validation results

The data contains 25000 “win” results and 25000 “lose”, so the baseline of both result prediction should be 25000 / (25000+25000) = 50%. The average precision of the prediction is 83%, which is largely bigger than 50%. So we can use Logistic Regression to classifier the game result and its corresponding feature ranking can be considered as reliable.

d. Feature ranking

A method called “SelectKBest” is chosen to rate the relevance of every feature with the target label. After ordering, the most important feature can be known:

	GPM	XPM	KDA	P/F	C/P
score	4142.08	1898.55	7301.59	12848.11	45.23

Table 5 scores of features

According to Fig. 5 and Table 5, the feature “P/F” gets the highest point. It means the extent of player’s tendency to push matters most for the result. Players always trying to push will tend to get a higher winning rate.

e. Critical analysis

The biggest problem is the dataset itself. This dataset has timeliness. The dataset itself is derived from dota2 data a few months ago. The version of dota2 is being updated all the time. We can apply the conclusion of this project to recent games but we can’t guarantee the accuracy for future versions. Besides, three classifiers may be still not enough. Classifiers like Randomforest or SVM (although too time-consuming) should also be tested. The average precision of Logistic Regression is 83%. It’s high but not high enough to convince everybody. We need to keep testing other classifiers with better performance.

f. Lessons learned

To beigin with, I have learned how to mining useful data for my project. Sometimes the data are not so straightforward and I have to calculating with the existing data. For example, the “P/F” feature in my project is calculated by the data of “Tower_damage” and “Hero_damage”. Besides, I have learned that I should always try to learn the background knowledge of a certain project before analysing the data. For instance, if I do not know this game, I will never figure out what “Tower_damage” and “Hero_damage” mean for the final result of a game. I may not consider “P/F” and I may ignore the most important feature.

In spite of the drawbacks of this project mentioned in critical analysis, I will still recommend this as a solution to my original question. The first reason is that I have considered 2 hypothesis tests and 3 classifiers to get the best solution. The result should be comparatively convincing. The second reason is that I have some experience of playing Dota2 and to my experience, “trying to push” is indeed the most important thing we should do to win the game.