Comparisons of linear and classification methods on analysis and prediction of question earnings on the website Quora

1st, February 2022

I. Introduction

This report will analyze and predict the income from the Quora Partner Program: how much money can be earned from each posted question on the website Quora.

- Quora is a social question-and-answer website based in California, United States. To increase the advertisements revenues, Quora has devised a strategy called Partner Program in 2018, where users in this program are paid for regularly posting new questions. This helps increase the website's traffic volume.
- My Partner Program was in the Japanese language, so this report paper will only investigate the income scheme of the program in Japanese, which is probably inapplicable to programs in other languages.
- The structure of this report has 6 parts. The "Problem Formulation" section describes the purpose of this report, the datapoints, features, and label. The "Methods" section describes the dataset details, feature selection, the ML models, and their loss functions. The "Results" section shows the models' results and the comparison between their performance. The "Conclusions" section finalizes which model performs better and how questions should be formulated. The "Code Appendix" section attaches the Python code.

II. Problem Formulation

The problem in question is: given certain characteristics of the posted questions, how much income the question is likely to earn?

- To answer this question, this paper aims to systematically gather the necessary data and make a general analysis and prediction towards the income scheme of the Quora Partner Program.
- For each question datapoint, there are four features because they are easy to be collected in a highly automated manner, which is illustrated in the picture below

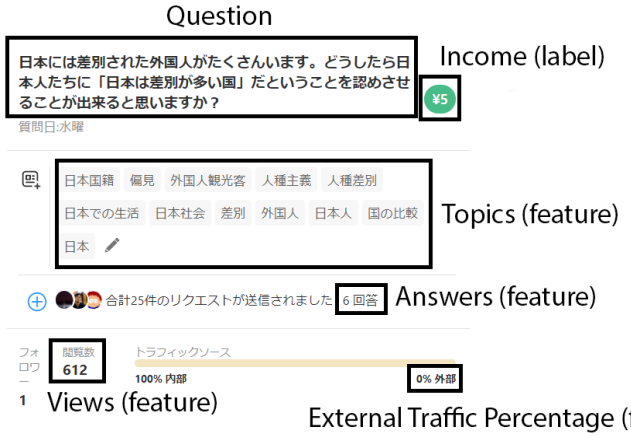

Figure 1: A posted question (datapoint) on the Quora website with many different features

Summary of datapoints, features, labels, and their data type

Datapoint: A single question with five features and one label

Features: There are 4 features: answers (integer), views (integer), external traffic percentage (integer), and topics of the question (string)

Label: Income of the question (integer - currency in Japanese yen). The income information is not available when the question is posted. It is only known one month after the question posting

III. Methods

3.1 Dataset

- **Source of the data:** The dataset is not available publicly because the income data is strictly exclusive to me. Nevertheless, the questions can be found at the URL provided below. Although there are over 7000 questions I have posted, I only choose the top 500 earning questions to study their properties and patterns. There is no missing data in any fields. After receiving the data file Quora has sent me, I proceed to filter the data so that it is readable by Python libraries, as well as manually input the features.

Link to the account on Quora: https://jp.quora.com/profile/Nguyen-Thanh-Luan

3.2 Feature Selection

I plot the data visualizations between the earnings label and the individual features. From the scatterplots in the annexed code, there seems to be a linear relationship between the views feature and earnings label, while there are no clear relations between the earnings and external traffic percentage. The answer feature seems to be weakly correlated with the label, but for the purpose of detailed analysis, it will also be chosen. The topics also correlate with earnings, because some certain topics are likely to earn more than the others.

=> Selected features: The answers, views, and topics features are chosen for this ML problem

3.3 Model (hypothesis space) selection

Due to the nature of the data points, I will separate the model selections into two types: the linear methods and the classification methods.

- The linear map will be used for modeling the features number of views and number of answers since in the data plots, they appear to be linearly related to the income. The two applied linear methods are Huber Regression and the Ridge Regression. Their hypothesis space will therefore be linear maps.
- The classification method will be used for modeling the question topics. For question topics, it contains 10 discrete topics and thus it is sensible to use classification methods for this feature. The two applied classification methods are logistic regression and support vector machine (SVM). These classification methods by definition use linear maps for their hypothesis space.

3.4 Loss functions selection

- For the ML models, their respective loss functions are chosen as follows:

- Huber Regression: The Huber loss functions (sklearn.linear_model.HuberRegressor)

There are a few extreme outliers in the dataset that will severely affect the result of the regression (Some questions have abnormally large earnings). This loss function is chosen because the Huber loss function is robust against outliers and thus provides a better prediction than a normal Linear Regression. [1]

- Ridge Regression: MSE loss regularized with L2-norm (sklearn.linear_model.RidgeRegression)

The label appears to have considerably large variance and the Linear Regression is expected to perform badly on validation sets. The Ridge Regression trades off an increase in bias with a decrease in variance so that the predictor would yield better results by introducing a penalty term [2]

- Logistic Regression: Logistic loss (sklearn.linear_model.LogisticRegression)

Logistic loss indicates how close the prediction probability is to the corresponding actual label. Particularly, the logistic loss is sensitive to outliers and there are many outliers in my dataset, which makes this loss function a good testing one for my model. [3]

- Support vector machine: Hinge loss (sklearn.svm.SVC)

There are many classes (10 topics) and therefore the boundaries are unclear. The hinge loss incorporates a margin from the boundary into the cost calculation. Thus, it will penalize the small margin if the margin from the decision boundary is not large enough, even if the observations are classified correctly [3]

3.5 Training and Validation Set

There are 500 datapoints in the dataset. First I split it into two sets: train-validation and testing sets of ratio 90% - 10%. Subsequently, I would split the train-validation set into training and validating sets using K-fold cross-validation with K = 5 as it is the standard commonly used fold number. Cross-validation usually delivers better results than a simple train-test split because it prevents overfitting. Moreover, on small datasets, the extra computational costs of cross-validation are not significant. [4]

IV. Results

4.1 Linear Methods

For the linear methods, the mean squared error in sklearn.metrics.mean_squared_error will be used to calculate the training and validation errors. Mean squared error (MSE) of a regression method measures the average squared differences between the actual and estimated values [5]. The best-chosen model minimizes the average validation error the most in the cross-validation. The obtained errors are:

	Training MSE Error		Validation MSE Error	
Features	Huber Regression	Ridge Regression	Huber Regression	Ridge Regression
Views	188288	181006	109014	102863
Answers	420880	326376	161739	198376

4.2 Comparison and selection of linear methods

From the MSE error table above, the validation error of the Ridge Regression is smaller for the views feature. On the other hand, the Huber Regression validation error is smaller for the feature Answer.

=> The chosen linear method is the Ridge Regression for feature Views and Huber Regression for feature Answers. These are the slope and intercept of their respective chosen model.

	Views	Answers		Views	Answers
Slope	0.016	9.249	Intercept	131.123	149.271

4.3 Classification Methods

For the classification methods, the accuracy score in sklearn.metrics.accuracy_score will be used to calculate the accuracy of the classifier models. Accuracy score computes the ratio of the accurately predicted features [6] Since there are 10 different groups and the accuracy of the model would be low, I merge the topics into three groups: Group 0 (News/Culture/Politics/Business/Career), Group 1 (Health/Psychology/Life/Relationship) and Group 2 (Entertainment/Technology/Education/ Fashion)

Similarly, the chosen accuracy is from the classifier model that maximizes the average validation accuracy in the cross-validation. The obtained accuracies are:

	Average Training Accuracy		Average Validation Accuracy	
Features	Logistic Regression	Support Vector Machine	Logistic Regression	Support Vector Machine
Topics	0.388	0.401	0.339	0.342

4.4 Comparison and selection of classification methods

From the accuracy score table above, both the training and validation accuracy of Support Vector Machine are greater than those of the Logistic Regression.

=> The chosen classifier method is the Support Vector Machine for feature Topics

4.5 Testing Set, testing errors, and accuracy

As stated above, the size of the testing set comprises 10% of the 500 datapoints or 50 datapoints, which has not been used in the cross-validation. The testing set helps provide an unbiased performance of the final fit model on the training set. The testing errors and accuracy have been finalized in the table below.

Features	Huber Regressor Error	Ridge Regressor Error		Logistic Regressor Accuracy	SVM Accuracy
Views	1218286	1218836	Topics	0.4	0.4
Answers	2305027	1705803			

V. Conclusion

5.1 Summary, result interpretations, and question formulation predictions

- Ridge regression on Views - Earnings: according to the Ridge Regression model, for every 1000 views a question receives, it earns an additional income of roughly 16¥. Since a question must gain user traffic to earn money, the intercept 131¥ of the regression line has little meaning because the number of views is 0.
- Huber regression on Answers - Earnings: according to the Huber Regression model, for every 1 answer a question receives, it earns an additional income of roughly 9.3¥. A question can earn money without answers on Quora and thus, for the top-earning questions with no answers, they earn 149.2¥ on average
- Support Vector Machine on Topics - Earnings: the predictions determined by SVM are given as:

Earnings: [0 20…160 180 200 220…1820 1840 1860 1880…1940 1960…2280 2300 2320 2340…10000]

Topic : [1 1 … 1 1 0 0 … 0 0 1 1 … 1 2 … 2 2 0 0 … 0]

By the categorization, it appears that group 0 and 1 topics dominate the lower part of income. On the other end, group 0 topics have the highest chance of reaching a high income. For group 2 questions, they may earn as much as group 1. Besides, group 2 is likely to earn a sufficiently large income like group 0.

- Question formulation prediction: Since views and answers are features that cannot be controlled, I can only depend on the question topics. It appears that Japanese people are usually interested in their own culture, daily life tips, recommendations on appearance, relationships, and work. A good strategy could be composing questions about hotly debated social problems, latest news, and product recommendations.

5.2 Results optimality and flaws

- Optimality: since the training and the validations errors of the Ridge Regression model are close for the views feature, the model does not overfit the training data, which can be attributed to the K-fold cross-validation. For SVM classifier, it performs moderately well at around one correct prediction out of three.
- Flaws: There are great discrepancies between the training and validation errors of the Huber Regression model for answers feature, suggesting that the model overfits the training data. For testing results, the testing MSE errors of both linear models are remarkably greater than their training and validation errors, which stems from the extreme outliers included in the testing set and thus results in poor performance.

5.3 Future directions and improvements on the ML methods

- First, more datapoints should be collected for the dataset. Because the chosen datapoints only represent the top 500 earning questions, they may not truly represent the whole trend of the 7000 posted questions.
- Secondly, there could be a better prediction if linear regression is a combination of various features. In that case, Lasso Regression can help to determine which features are important in explaining the earnings.
- Thirdly, there could probably not exist any linear relationships between the income and the view and answer features. In other words, the income is purely random. A more elaborate ML model will be needed to discover such a relationship if it truly exists.

VI. References

[1] Calculate Huber Loss using TensorFlow 2, available at
https://lindevs.com/calculate-huber-loss-using-tensorflow-2/
[2] Ridge Regression and Lasso Regression, available at
https://discuss.boardinfinity.com/t/ridge-regression-and-lasso-regression/6208
[3] Understanding Hinge Loss and the SVM Cost Function. Available at
https://programmathically.com/understanding-hinge-loss-and-the-svm-cost-function/
[4] Cross Validation. Available at:
https://www.kaggle.com/code/dansbecker/cross-validation/notebook
[5] How to Calculate Mean Squared Error (MSE) in Python. Available at:
https://vedexcel.com/how-to-calculate-mean-squared-error-mse-in-python/
[6] Accuracy Score. Available at
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

VII. Code appendix