Predicting Yelp's elite

Jay Xiao
University College London

March 18, 2023

1 Introduction

Yelp is a popular online platform that connects people with local restaurants. It gathers a strong community of people who like food and who love to share their experiences. people with high contributions to the community can be selected as an officially recognized elite squad. Broadly speaking, users who want to become elite should write good reviews with appropriate words used, and a suitable length, the more people votes for the review may quantify the quality as well. Additionally, they may also need to contribute actively. It may also depend on the social network, the more elite friends and fans you get may indicates that you are recognized as a good content creator by them.
This report is based on the version 8 yelp dataset by selecting a list of features to accomplish the goal of discerning between elite users and non-elite users. We began by

2 Data set

The 686556 samples taken from the yelp user dataset contain features: date started yelping, number of total votes on all reviews, number of total reviews, list of friends, number of fans, number of total compliments received and average star of the restaurant, elite statue. we will also include the yelp review dataset which contains all of the reviews written by users, each of the samples includes the number of total votes of this review, the stars of the restaurant, and the contents of the review.

3 Features and Preprocessing

3.1 review dataset extraction

The review contents are a list of strings that cannot be directly used by the model, so we first quantify review contents into several features and combine them into the features of user dataset, details of the features used are shown in figure 1.

3.2 Balancing samples

There are about 5.21% of users are elite, which means two classes are severely imbalanced, if we take a naive approach by labeling all samples to non-elite, we would get a 94.79% accuracy He and Garcia (2009). The model may not capture all information to classify elite. Thus, we will take all elite samples and randomly selecting the same amount of non-elite samples, this would form a balanced sample set. Further more, the confusion matrix is introduced to get a more clear view of the models' performance between two classes.

3.3 Principle component analysis

Because the dataset has a large feature set as well as the features we added, any one of them could be correlated. Thus principle component analysis was used, it is necessary to acutely reduce the dimensions and the random noise, but also preserve most of the information contained, and make the model efficientJolliffe and Cadima (2016). Due to the wide range of various concentrations in the dataset, the selected n features' samples $X_1, X_2, ..., X_n$ are standardized by the features' sample mean μ_i and the sample standard variance $\sigma_i, \forall i$. The corresponding correlation matrix is calculated and the principle components are given by:

$$e_{it} = \frac{1}{\sqrt{\lambda_i}} \sum_{k=1}^{N} v_{ik} s_{kt}$$

where λ_i and v_i are the ith eigenvalue and eigenvector. We choose the explained variance to be 95%, which means the total information saved $x_{it} \approx \sqrt{\lambda_1} v_{1i} e_{1t} + ... + \sqrt{\lambda_k} v_{ki} e_{kt}$ are 95%.

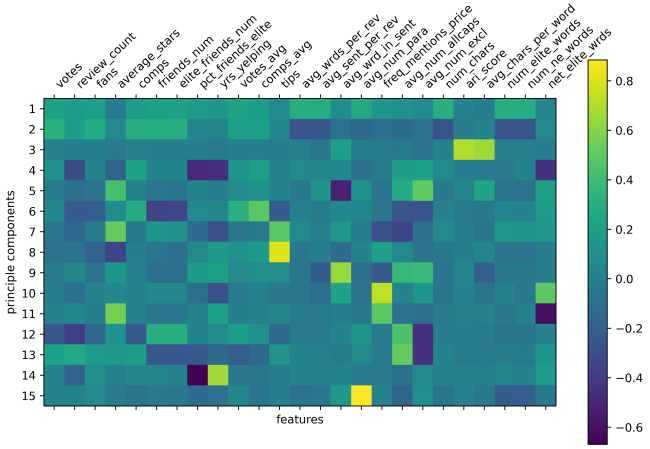

Figure 1: Principle Components from the balanced samples

As we can see the number of dimensions significantly decreased from 25 to 15. And also PCA allows the new dimensions to have no correlation and zero means.

4 Methodology

To analyze the samples, two approaches were investigated: logistic regression, and random forest.

4.1 Logistic regression

Logistic regression is a binary classification model that can be used to analyze the relationship between the elite and non-elite classes. based on the 16 principle components, we will take a weight vector $\mathbf{w} = (w_1, w_2, ..., w_{16})$ For the yelp dataset $\mathbf{x}$ with labels $\mathbf{y} = 0, 1$, the logistic regression model uses the sigmoid function to map the predicted values onto the range [0,1], which can be interpreted as the probability of the positive class. It reduce the influence of the samples that are far away from the decision boundary.The sigmoid function is a non-linear mapping defined as follows:

$$\sigma(\mathbf{x} \cdot \mathbf{w}) = \frac{1}{1 + e^{-\mathbf{x} \cdot \mathbf{w}}}$$

The logistic regression model estimates the coefficients $\mathbf{w}$ using maximum likelihood estimation.

$$L(\mathbf{w}) = \prod_{i}^{n} \sigma(\mathbf{x} \cdot \mathbf{w})^{y_i} (1 - \sigma(\mathbf{x} \cdot \mathbf{w}))^{1-y_i}$$

The right-hand side is the probability of the whole training set. This involves finding the values of the coefficients $\hat{\mathbf{w}}$ that maximize the likelihood of observing the given data, assuming a logistic distribution of the dependence variable.

$$\hat{y} = \begin{cases} 1 \quad if \quad \sigma(\mathbf{x} \cdot \hat{\mathbf{w}}) \geq 0.5 \\ 0 \qquad\qquad otherwise \end{cases}$$

4.2 random forest

A random forest consists of M decision trees, where each tree is built on a random subset of the training set with a random subset of the features. One advantage of random forest is we can extract feature importance to get which feature have the most influence on the elite selection. And it also improves robustness compare to decision tree.

To build each tree, we recursively partition the training set into subsets based on the values of the features. At each step, we choose the feature j and the threshold k that minimize a certain impurity measure, here we use cross-entropy loss:

$$L(\mathbf{y}) = \frac{1}{N} \sum_{i}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

The resulting tree is a binary tree, where each internal node represents a decision based on a feature and threshold (j, k), and each leaf node represents a prediction.

The prediction of a new sample x, is obtained by passing through each tree to get a set of M predictions, then aggregating these predictions by taking the majority vote for classification.

5 Result

The training set and the test set will be split into 75% and 25%. The final performance of a model is tested on 5 randomly selected balanced samples, and measured the average performance by the confusion matrix.

5.1 Model performance

Because random forest is a non-parametric model, we cannot use method like in-sample likelihood to compare the performance with the logistic regression model. We will use Root mean square error instead to test the training set performance:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}$$

Table 1 is the results coming from averaging the calculation of the confusion matrix. Two logistic regression models got similar RMSE, whereas Random forest model fitted more better, especially the model without apply PCA.

5.2 out of sample accuracy

All models have high recall score than precision score, which means all model's predicting power of elite samples are more strong. Principle component analysis did not improve the performance for this particular classification problem, as we can see that both methods applied principle component analysis reduce accuracy, random forest classifier is more affected. The random forest without principle component analysis model shows dominating accuracy compare to others.

Table 1: Models overall average performance

Model	precision score[1]	recall score[1]	f1 score[1]	accuracy[1]	RMSE[2]
Logistic regression (applied PCA)	0.9224	0.9247	0.9235	0.9236	0.2772
Logistic regression	0.9229	0.9273	0.9251	0.9250	0.2722
Random Forest (applied PCA)	0.9195	0.9491	0.9341	0.9328	0.2606
Random Forest	0.9502	0.9778	0.9638	0.9630	0.1940

[1] These four performance scores were calculated from the test set.
[2] The root mean square error was calculated from the training set.

5.3 class-based comparison

We are going to use receiver operating characteristic(ROC) curve and area under the ROC Curve score for each model as the comparison criterion. Due to the weak performance on the Principle component analysis samples, we will only compare two models that using the original data. An ROC curve is a graph showing the performance of a model under the change of classification threshold from $[0, 1]$ and the AUC score is the integration of the ROC score. From figure 2, we can see that both models

close to the perfect model, and the AUC score for random forest classifier model is slightly higher, it is not surprised to see the results based on the previous performance table. Overall, we will choose the random forest model without principle component analysis as our optimal model.

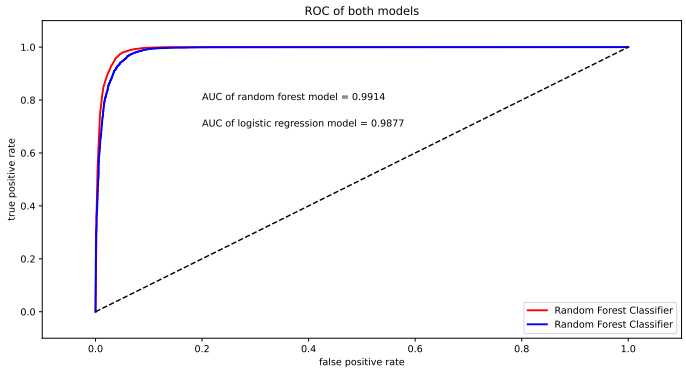

Figure 2: ROC curves of both models

5.4 Feature importance

Gini importance is a common way to measure the feature importance of random forest classification model, it is directly derived from the Gini index Breiman (1996). The Gini index measures the node split impurity of a feature in a node of decision tree.In this problem we have two classes, so for the ith decision tree in RF at node k, the Gini index is given by

$$GI_k^i = 2 \times p(1-p)$$

where p is the probability of of an element being classified for a distinct class. For a specific feature's importance at node k in this tree is calculated by minus the Gini index of the following two child nodes. The Gini importance value of a feature in a single tree is calculated by adding up the amount of all nodes' Gini index if the feature j appear at nodes in a set M. Then the overall importance for this feature is by averaging all importance calculate from each decision tree in the random forest.

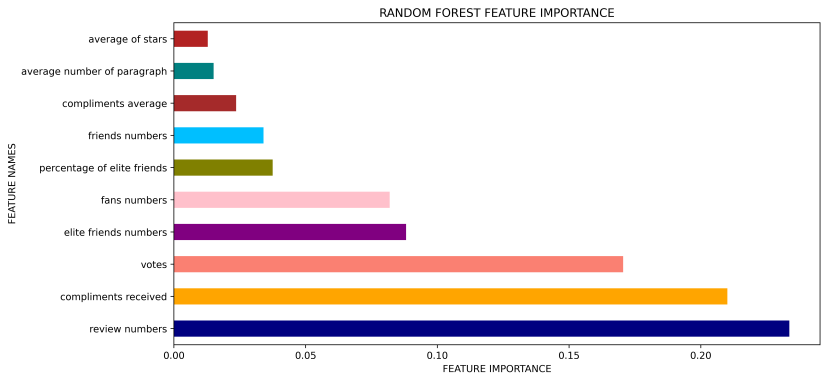

Figure 3: Top 10 feature importance from the Random forest

6 Discussion and Conclusion

The resulting optimal model shows 96.30% accuracy with 97.78% true positive rate. It is not very surprising that user's contribution, popularity and activity are the most important in determining his elite statue. The number of reviews has the largest importance value, which means the more active you post your review, the more yelp think you contribute to the community. The recognition of user's reviews by voting and complements is also important, so user have to write attracting and high quality review as well. we can also see that the number of elite friends a user have is more important than the number of friends in general. The conclusion is quite clear, it is not only good for a user to know what is the best way to do to become an elite, but also good for the company to evaluate a user's overall performance in the community.
Further analysis should consider:

1. Hyperparameter tuning method should be included to find optimal model.
2. Why logistic regression model performed so poorly in our setting and how to improve the model?
3. Will the model perform better if the samples we choose has different elite ratio?

References

Breiman, Leo. 1996. Bagging predictors. *Machine learning* **24** 123–140.

He, Haibo, Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**(9) 1263–1284. doi:10.1109/TKDE.2008.239.

Jolliffe, Ian T, Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* **374**(2065) 20150202.