Restaurant Popularity and Failure Prediction

Peng Zhao
Informatics Institute
University of Missouri
Columbia, MO 65211
peng.zhao@mail.missouri.edu

Danlu Liu
Department of Computer Science
University of Missouri
Columbia, MO 65211
dltb9@mail.missouri.edu

Abstract

Consumers are interested in finding good businesses and business owners hope to know how to improve their popularity. Many factors are associated with the success of a business and it is often infeasible to manually sort out the relationships of these factors. Artificial intelligence has seen many applications in improving our daily life, such as intelligent personal assistants, recommendation systems, etc. In this study, we hope to add our contribution by predicting the popularity and failure of restaurants. We have developed 2 models to predict a restaurant's popularity and closure using supervised machine learning techniques. Naïve Bayes, decision tree, and random forest algorithms were selected to build the candidate models. The 2 final models have moderate AUC of 0.7 and 0.75. The first model can help users to predict the popularity of a restaurant with little effort. The second model can help restaurant managers to predict their risks of failure so that they can make wise decisions of improvement.

1 Introduction

1.1 Background

Yelp is a popular American multinational local business review publisher founded in 2004. For many people, it is the first site to visit before making a dining or shopping decision. As of March 31, 2017, Yelp has accumulated 127 million reviews [1]. Every month, Yelp has 84 million, 73 million, and 26 million unique visitors from desktop, mobile web, and mobile app, respectively [1]. The reviewed businesses are across various categories, including shopping, restaurants, home and local services, beauty and fitness, arts and events, health, auto, nightlife, travel and hotel, and other miscellaneous categories. The top 3 most reviewed categories are shopping (22%), restaurants (18%), and home & local services (13%). Yelp users can share their personal consumer experience with a business by posting reviews and rating the business by stars ranging from 1 to 5 with half star possible. Each review can receive feedbacks from other users, such as being useful, funny, or cool. Inappropriate reviews, such as wrong information, threats, and promotional materials can be reported to Yelp through a report function. Users can also friend with each other through the Yelp social network.

Although Yelp has a large amount of local business reviews, currently users are only allowed to rank businesses by the relevance to their search keywords, by total number of reviews, or by star-rating. Ranking by total number of reviews can be risky because a business may be ranked top by receiving a lot of complaint reviews. Users will have to scrutinize the reviews and determine if they are complaint reviews. Ranking by star-rating is also not a good idea because it can be biased if a business has very few number of reviews. Although an in-house algorithm is used to calculate the star-rating for a business by weighting the stars received from different reviews, the algorithm cannot handle bias. For example, if a new business receives only one rating, saying 5 stars, the final star-rating of the business will be 5 even though it has only one review. In most cases, users need to choose a business by considering both the review count and star-rating, which can be time consuming. This situation becomes more pressing when users are hungry and seeking for a good restaurant or when they have flat tires on road and seeking for a good auto shop nearby. In other words, the current ranking functions provided by Yelp are not smart enough and can diminish users' experience when they need to do a decision promptly.

What's more, Yelp does not provide suggestions to local businesses regarding how to become successful. From a business's point of view, it is important to know if it will fail in the future. If businesses can predict the risk of failure, they can improve their services to reduce the potential risk. This can help to promote the success of local business.

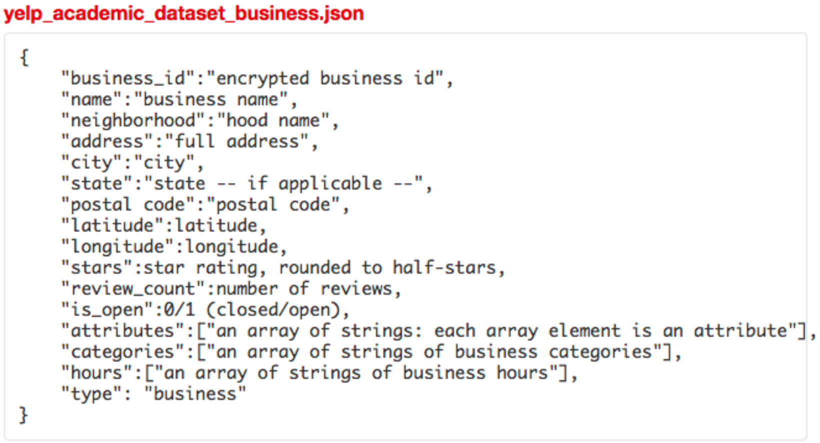
1.2 Aims of this study

In this work, we have 2 aims. The first aim is building a model to predict a restaurant's popularity by a new unbiased popularity metric. The second aim is building a model to predict if a restaurant will be closed in the future. We decided to model restaurant because it is much closer to everyday life and has large amount of data (18% of Yelp reviews are for restaurants [1]).

2 Methods

2.1 Data source

The data is from the 2017 Yelp Dataset Challenge [2] with 4.1 million reviews and 947,000 tips by 1 million users for 144,000 businesses. The original dataset contains 6 different data tables in JSON format, including business, review, user, check-in, tip, and photos. In this study, the business data is used. Figure 1 shows the attribute information of the business data. The attributes include "business_id", "name", "neighborhood", "address", "city", "state", "postal code", "latitude", "longitude", "stars", "review_count", "is_open", "attributes", "categories", "hours", and "type". The value of "attributes" has a dictionary structure with different attributes being the keys. The value of "hours" also has a dictionary structure with "Monday" to "Sunday" being the keys and the business hours of each day being the values. The value of "categories" is a list of categories for a given business.



yelp_academic_dataset_business.json

```
{
    "business_id":"encrypted business id",
    "name":"business name",
    "neighborhood":"hood name",
    "address":"full address",
    "city":"city",
    "state":"state -- if applicable --",
    "postal code":"postal code",
    "latitude":latitude,
    "longitude":longitude,
    "stars":star rating, rounded to half-stars,
    "review_count":number of reviews,
    "is_open":0/1 (closed/open),
    "attributes":["an array of strings: each array element is an attribute"],
    "categories":["an array of strings of business categories"],
    "hours":["an array of strings of business hours"],
    "type": "business"
}
```

Figure 1. Structure of the business data.

2.2 Data preprocessing

2.2.1 Data format transformation

The JSON data has been converted into a feature matrix in CSV format using a Python script provided along with the original dataset. The resulting CSV file has 16 attributes and 144,073 instances.

2.2.2 Data cleaning

Since most instances have missing values in "neighborhood", the "neighborhood" attribute was removed. The "type" attribute was also removed because all the instances have the same value "business" for this attribute. Then the instances with missing values were removed. The resulting data after data cleaning has 14 attributes and 90,079 instances.

2.2.3 Attribute expansion

The original "attributes", "categories", and "hours" attributes have aggregated values and cannot be directly applied in machine learning. These aggregated values have been unfolded and transformed into a sparse feature matrix with Python scripts. Then the instances with the category of "restaurant" has been filtered out. The resulting data after attribute expansion and filtering has 671 attributes and 33,867 instances.

2.2.4 Create a popularity attribute

A new popularity attribute has been created by multiplying the review count and the star count using a Python script. In this case, the review count and the star count will be scaled by each other. Because the review count has a scale of 0 to infinity and the star count has a scale of 1 to 5, they need to be normalized prior to multiplication. The formulas of popularity and normalization are as follows:

$$Popularity = Normalize(review\ count) \times Normalize(star\ count)$$

$$Normalize(\underline{X}) = \frac{\underline{X} - Min(\underline{X})}{Max(\underline{X}) - Min(\underline{X})}$$

2.3 Modeling

2.3.1 Feature reduction

The attribute expansion step has created a sparse feature matrix and many attributes have very skewed distributions. It is necessary to perform feature reduction to remove useless variables. Because the attributes are categorical, principle component analysis [3] is not applicable. The feature selection has been performed with Information Gain [4] method in Weka [5] with respect to the class attributes "is_open" and the newly created "popularity" for the 2 models. Information gain is the expected reduction in entropy by partitioning the dataset with respect to an attribute and it measures the importance of the variable in discriminating different classes [4]. After this step, the numbers of features of the data for predicting popularity and closure have been reduced to 56 and 72, respectively.

2.3.2 Algorithms

The supervised machine learning algorithms selected for this study are Naïve Bayes, decision tree, and random forest because they can work with categorical attributes. Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem [6] by assuming the predictors are independent [7]. It calculates the posterior probability of being one class given the predictors. Decision tree is a family of tree structure-based predictive algorithms with leaves being class labels (classification)/probabilities (regression) and branches being conjunctions of features leading to the class/probability assignment [8]. Random forest is an ensemble algorithm operated by constructing many decision trees and outputting the mode of the classes (classification) or mean prediction (regression) of the decision trees [9].

2.3.3 Evaluation

The dataset has been randomly split into a training set and a test set (4:1). The training dataset is used to model the relationship between variables and the class attribute and the test dataset is used to measure the performance of the model. The area under the receiver operating characteristic curve (AUC) [10] has been used as the performance metric to compare models in this project. The receiver operating characteristic curve (ROC) is a graphical representation of a binary classifier's performance as the discrimination threshold is varied [11]. AUC measures the model's ability of discrimination and can be interpreted as the probability that the model will rank a randomly selected positive sample higher than a randomly selected negative sample.

3 Results

3.1 The model to predict a restaurant's popularity

The class attribute of this model is "popularity". The model's performance is shown in Figure 2. It can be seen the Naïve Bayes model has a moderate performance with AUC of 0.7. The random forest model has AUC of 0.66. The decision tree model has a poor performance with AUC of 0.56, which is close to random guess. The Naïve Bayes model has been selected as the final model for restaurant popularity prediction.

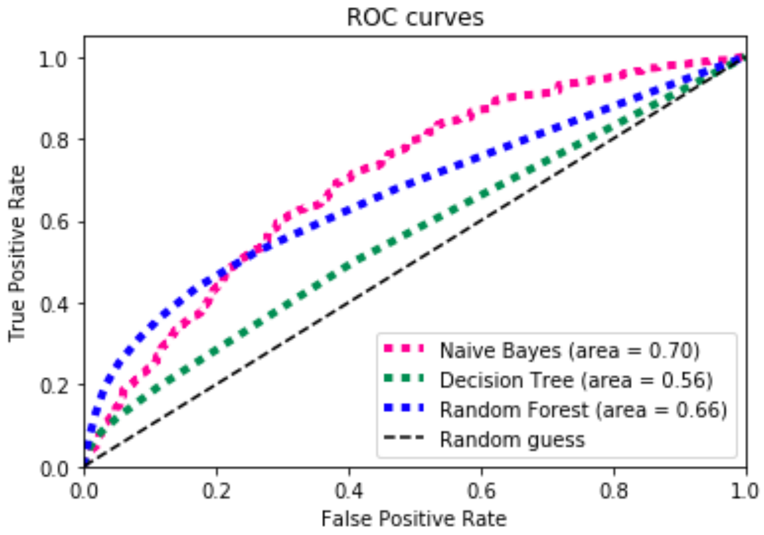


Figure 2. ROC curves for the candidate models to predict a restaurant's popularity.

3.2 The model to predict if a restaurant will be closed in the future

The class attribute of this model is "is_open". The model's performance is shown in Figure 3. It can be seen the random forest model has a moderate performance with AUC of 0.75. The decision tree model has AUC of 0.66. The Naïve Bayes model has a poor performance with AUC of 0.57, which is close to random guess. The random forest model has been selected as the final model to predict if a restaurant will be closed in the future.

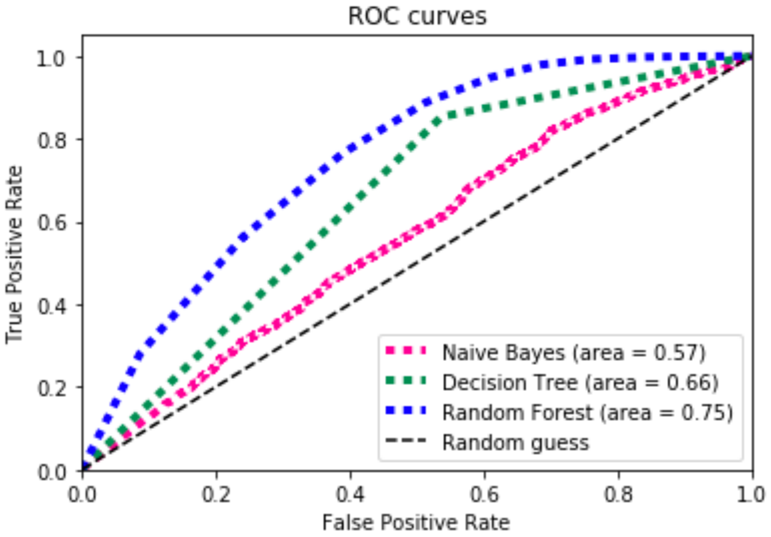


Figure 3. ROC curves for the candidate models to predict if a restaurant will be closed in the future.

4 Discussions

As shown above, the random forest algorithm is stable in both cases. It generally has high classification performance without heavy computation. It is solid for imbalanced data and data with missing value, and good for data with up to thousands of attributes [12]. In both cases, the random forest algorithm outperforms the decision tree algorithm. A possible explanation is that the random forest algorithm is an ensemble of multiple decision trees constructed with random rows and columns of the original data and is more stable than a single decision tree, which is sensitive to variations in the dataset. Another observation from the result is that the Naïve Bayes algorithm has drastic difference in both cases. The Naïve Bayes classifier is generally easy to train but it sacrifices the accuracy to some extent due to the assumption that the features are independent. A possible explanation to the discrepancy in the 2 models is that some of the attributes might be dependent when predicting if a restaurant will be closed in the future and Naïve Bayes algorithm is not suitable in this case.

There are many future directions in this work, such as using association mining algorithms to extract the frequent attribute patterns of popular restaurants to study why they are successful or building a restaurant recommendation system based on users' similarity.

5 Conclusion

The application of machine learning techniques in local business prediction has many potential and the outcome of this study is meaningful. In this study, 2 models have been developed to predict the popularity and closure of a restaurant. The popularity metric created in this work is more reliable than review count and star count themselves independently. The first model can help users to predict the popularity of a restaurant with little effort. The second model can help restaurant owners to predict their risks of failure so that they can make wise decisions of improvement. The methodology used this study can be generalized to broader domains, such as predicting health care providers' revenue with ZocDoc data, etc.

Acknowledgments

We want to acknowledge the generous data sharing from Yelp.

References

1. An Introduction to Yelp Metrics as of March 31, 2017 [Internet]. 2017 [cited 2017 Oct 5]. Available from: https://www.yelp.com/factsheet
2. Yelp Dataset Challenge [Internet]. 2017 [cited 2017 Oct 5]. Available from: https://www.yelp.com/dataset_challenge
3. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6 1901 Nov;2(11):559–572.
4. Kullback S, Leibler RA. On Information and Sufficiency. The Annals of Mathematical Statistics 1951 Mar;22(1):79–86.
5. Frank E, Hall MA, Witten IH. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques." 4th ed. Morgan Kaufmann; 2016.
6. Bayes T, Price R. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. Philosophical Transactions of the Royal Society of London 1763 Jan 1;53:370–418.
7. Domingos P, Pazzani M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning 1997;29:103–130. PMID: 10575050
8. Quinlan JR. Induction of decision trees. Machine Learning 1986;1(1):81–106.
9. Tin Kam Ho. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition IEEE Comput. Soc. Press; 1995. p. 278–282.
10. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983 Sep;148(3):839–843.
11. Green DM, Swets JA. Signal detection theory and psychophysics. New York, New York, USA: John Wiley & Sons, Inc; 1966.
12. L. Breiman, "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)," (in en), pp. 199-231, 2001/08 2001.