Las Vegas Business Closure Prediction Model

Xinyu Chen
Computer Science
University of Southern California
Los Angeles, CA, US
chen728@usc.edu

Yanan Fei
Computer Science
University of Southern California
Los Angeles, CA, US
yananfei@usc.edu

Yang Wang
Computer Science
University of Southern California
Los Angeles, CA, US
wang824@usc.edu

Fan Zhang
Computer Science
University of Southern California
Los Angeles, CA, US
zhan166@usc.edu

ABSTRACT

This project is a prediction model for local businesses in the Greater Las Vegas Area (Las Vegas, NV, North Las Vegas, NV, and Henderson, NV) to foresee if a new business will succeed or fail based on business features. We split our Yelp Vegas business dataset into two groups based on their "is_open" feature. We combined features extracted from Yelp dataset and some new features such as a business' age, whether a business is a chain or not, and the number of reviews. Our dataset was split into 80% training and 20% test. We examined and analyzed five different machine learning models, including logistic regression, gaussian NB, decision tree, gradient boost, and random forest, to predict the chance of business closure, and compared them based on the accuracy, precision, recall, and F1 score to see which model is the best fit for our prediction.

1 Introduction

According to the survey by restaurantowner.com [1], it requires $275,000 on average to open a new restaurant in the U.S. It is an ambitious undertaking to start a new restaurant business and people need to think carefully before stepping into this business. Thus, predicting success for a new business is key for business investors. Based on Yelp business data in Las Vegas, we construct a model to predict the chance of closure for a new business opened in Las Vegas based on the features.

2 Database

In our problem, first we filtered out businesses and reviews in Las Vegas. After that, we got 35173 businesses and around 2,000,000 reviews.
Used datasets:
Yelp_academic_dataset_review.json→ vegas_dataset_review.json
Yelp_academic_dataset_business.json→vegas_dataset_business.json

However, not all features in our business dataset can be used. We did data preprocessing for our models. Again, our task was to predict a status of a business in Las Vegas. In the original business json file, there was a boolean value called "is_open", which we decided to use it as our ground truth label. Next, we generated several new features, such as "is_chain", "age", "categories", "cluster_similarity", "nearby_count", "price_level", "star_coef". Details were provided as below. After generating and combining our new features, our input data set contained 35173 data samples and each sample had 12 features.

2.1 Features

(1) Is_chain: We assumed a business is part of a chain if the name of this business appeared at least twice. Belonging to a chain is one key factor in the real world to have an impact on the probability of success of a business. If a restaurant is a chain of others, it has a higher probability to be successful. So we decided to use this feature in our prediction model.

(2) Age: From users' review dataset, we extracted reviews for each business in Las Vegas and recorded the earliest review date and latest review date. The age of a business is determined by the difference of the latest and the earliest review.

(3) Categories: We did not use the businesses' categories from the original Yelp dataset. Instead, we relied on Yelp API for a more specific categorization for each business. This feature is one factor in determining the similarity between businesses.

(4) Cluster_similarity: In our business dataset, business's category contained several keywords, like "Chicken Wings, Burgers, Caterers, Street Vendor" or "Shopping, Fashion, Department Stores." It is not that obvious to describe similarities between businesses based on multiple categories. Therefore, we took advantage of clustering and converted categories into an index. We used the centroid of a cluster to represent all categoric keywords belonging to this cluster.

(5) Nearby_count: The number of businesses within the range of one kilometer for each business, extracted from Yelp_academic_dataset_business.json.

(6) Price_level: We got a price level for each business in Vegas from the feature "RestaurantsPriceRange2" in business.json dataset. If a business does not has this attribute, we then assigned the average price level in Las Vegas to it.

(7) Review_count: The number of reviews of each business in Vegas, extracted from vegas_business.json file.

(8) Star_coef: We got all reviews for each business from the review.json dataset. Reviews were then sorted by date and we built a 2-dimensional coordinate system using linear regression model to get the coefficient of the review stars. We can get the stars_coef by combining business_id and the coefficient of the review stars. It's important to get this feature because it will show the trend of review starts for a business, and it will tell us the business is getting better or worse. It is a key factor for whether the business will close or not in the future.

(9) Star: Open restaurants tend to have a higher rating on Yelp, compared to closed ones. (See Fig. 1 and 2)

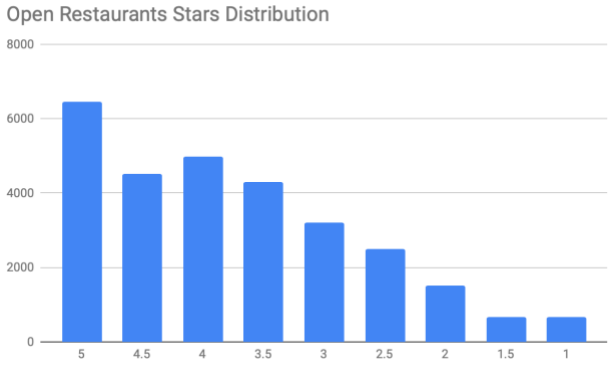

Figure 1 Open Restaurants Stars

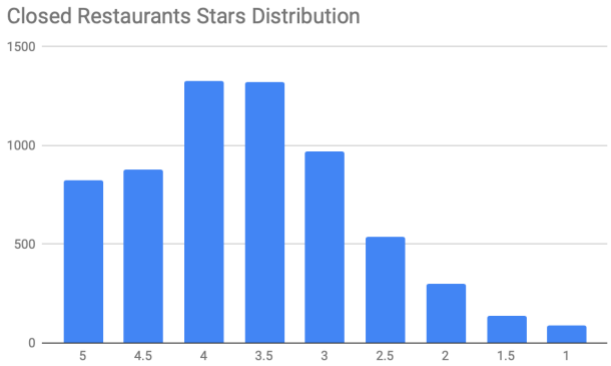

Figure 2 Closed Restaurants Starts

3 Methodology

One of our most significant work was feature extraction. We spent a lot of time on determining which features have the most impacts on our prediction. After features extracted, we applied five classification models on our generated feature sets. Before that, we visualized our data set using TSNE. A visualization example can be found below in Figure 3. Then we split our dataset into a training and a testing set. The ratio of training to test is 8 to 2. For these models, we used 5 classifiers including Logistic Regression, Gaussian NB, Decision Tree, Gradient Boosting and Random Forest. We will compare these

models with their prediction accuracy, precision, recall and f1 score.
Because some of our features were not normalized. We also tried to reduce 1 or 2 insignificant components by using PCA.

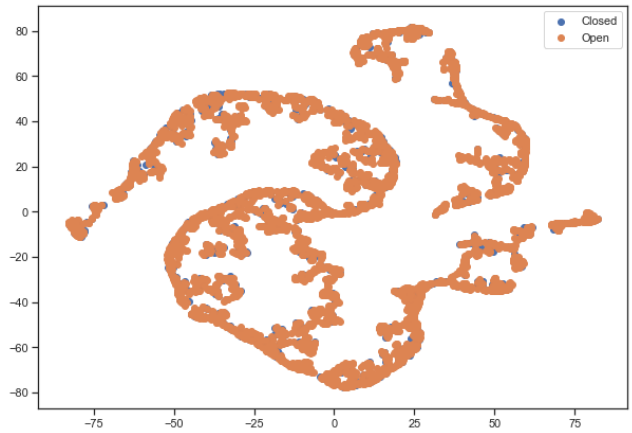

Figure 3 data visualization using TSNE

After training models, we visualized our results and did some analysis. Figure 4 is the ROC curve for models. Figure 5 explained the importance of each features after using Random Forest Classifier.

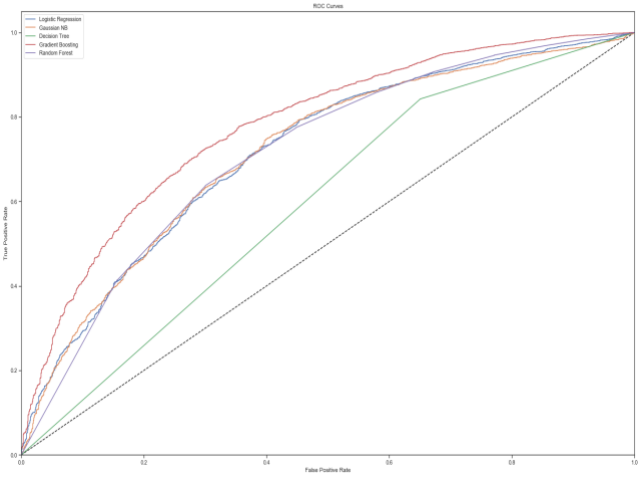

Figure 4 ROC curves and model comparison

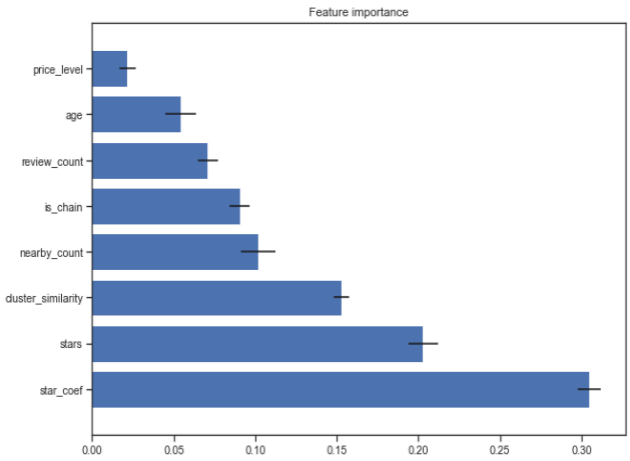

Figure 5 Feature importance

4 Result and Discussion

4.1 Model Comparison

Figure 4 shows one of our results. According to the ROC curves, our Gradient Boosting performed the best based on our data.

models	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.8172	0.8203	0.9946	0.8991
Gaussian NB	0.8043	0.8428	0.9354	0.8867
Decision Tree	0.7525	0.8539	0.8416	0.8477
Gradient Boost	**0.8326**	**0.8431**	**0.9773**	**0.9053**
Random Forest	0.8021	0.8585	0.9080	0.8825

Table 1 Results of models

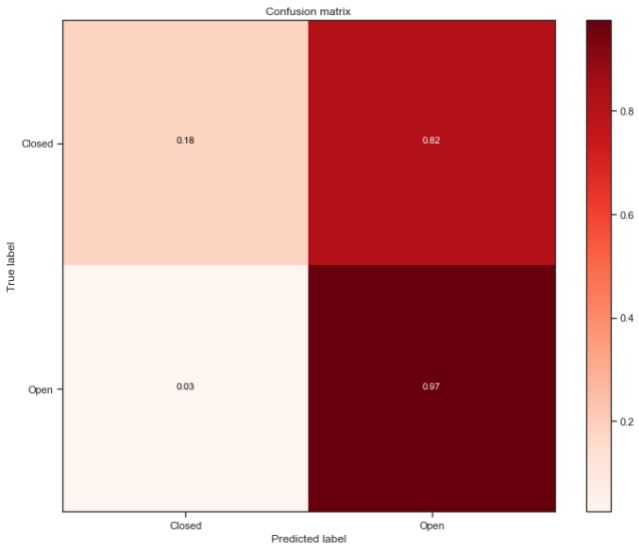

Figure 6 Confusion Matrix of Gradient Boosting

4.2 Feature Ranks

In Figure 5, "star_coef" and "stars" are more important to decide if one business could succeed or not, which means customers' evaluation is first matter in this area.

4.3 Using Principal Component Analysis

With PCA algorithm, we reduced components from 8 to 7, and used above models again to train and test data. The result didn't have significant changes, so the eight features all had matter to the result.

models	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.8166	0.8205	0.9932	0.8987
Gaussian NB	0.8030	0.8408	0.9366	0.8861
Decision Tree	0.7491	0.8491	0.8434	0.8462
Gradient Boost	0.8313	0.8433	0.9750	0.9044
Random Forest	0.7955	0.8538	0.9052	0.8787

Table 2 models Results With PCA

APPENDIX

https://github.com/yananfei-Bette/INF553-Yelp-Project

REFERENCES

[1] How Much Does it Cost to Open a Restaurant, https://www.restaurantowner.com/public/Survey-How-Much-Does-it-Cost-to-Open-a-Restaurant.cfm

[2] A. H. Jahromi and M. Taheri, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features," *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, Shiraz, 2017, pp. 209-212.

[3] O. Irsoy, O. T. Yildiz and E. Alpaydin, "Budding Trees," *2014 22nd International Conference on Pattern Recognition*, Stockholm, 2014, pp. 3582-3587.

[4] Stanford, J.H. (2010). Greedy Function Approximation: A Gradient Boosting Machine.

[5] GERTJAN J. BURGHOUTS (TNO, Intelligent Imaging, Oude Waalsdorperweg 63, 2597 AK, The Hague, The Netherlands) International Journal of Pattern Recognition and Artificial Intelligence 2013 27:04

[6] Kärkkäinen T., Saarela M. (2015) Robust Principal Component Analysis of Data with Missing Values. In: Perner P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2015. Lecture Notes in Computer Science, vol 9166. Springer, Cham

[7] Aly R. (2014) Score Normalization Using Logistic Regression with Expected Parameters. In: de Rijke M. et al. (eds) Advances in Information Retrieval. ECIR 2014. Lecture Notes in Computer Science, vol 8416. Springer, Cham