Making Predictions in Closure of Restaurants within the Next 2 years

Hongju Lee (hongjlee), Jungseo Lee (jungseo), Elizabeth Park (lizpark)

Introduction

The pandemic has changed us permanently. While it seems like the economy is recovering from the pandemic, the economy is still in a fragile state, and we can see that there are still a significant number of restaurants that are struggling to keep their business running. However, it's risky to assume that the negative outcome of restaurants is solely caused by the pandemic as there are many restaurants that are flourishing in business even in our current unstable economy. Our team believes that the lack of information is one of the major reasons that the local business owners were negatively impacted and caused the disparity between top performing restaurants and the low. Through this project we hope to close that disparity and help the business owners to understand and realize which attributes play a critical role for maintaining a successful business.

Our main question of interest was, "will this restaurant still be open for the next 2 years?". Although the question may seem simple and easily solvable, it was more complicated than it seemed. It brought on many additional questions that we must consider to effectively predict the faith of the businesses. An essential question we then asked ourselves was, "are the features about the restaurant itself, which are included in the Yelp API, efficient enough to create a powerful prediction model?" While Yelp API provided us with a good number of attributes in terms of services the restaurant provides, it also brought on some concerns and questions about whether those attributes are sufficient to predict the outcome of a restaurant. Our team strongly believed that there are other important factors that could help determine a restaurant's success on top of the information that the restaurants provide for the customers. Through a critical discussion, demographic and geographic data was added from FBI and Census to obtain more specific information about the location and the overall atmosphere the restaurant is in.

Our team predicted that the location and the overall atmosphere that the restaurant is in can play a vital role in predicting the outcome. We believe that the restaurants that are in highly populated areas will inevitably have a higher customer visiting rate as well as retention rate compared to those that are in less populated areas. This idea was also applied to racial demographics percentage data, median household income data, higher educational degree percentage data because we suspected the restaurants that are in a good and safer neighborhood with higher median income and educational degree, are less likely to be closed as people who live in those areas are less likely financially impacted when an economic crisis hits.

After going through extensive data research, pre-processing, and application and validation of various machine learning models, our team was able to build a prediction model that predicts whether a restaurant will be opened or closed in the following year. Our team strongly believes that it will not only help the local businesses that are struggling to survive but also help the overall economic situation and help revitalize the impacted communities.

Methods

1. Getting data

The dataset is customized by combining the 2020 Yelp dataset, Yelp API, and demographic data. As we are trying to make a prediction model that predicts if a restaurant is likely to close within the next 2 years, we need to build a dataset that contains features in these 2 years and the current opening status.

The base data set is ‘2020 Yelp Academic Dataset_business’. The dataset is obtained from sdkramer10’s GitHub repository as the previous dataset is not officially offered by Yelp. All the original datasets that we used for this project can be found in one of our GitHub repositories[1]. We started with a total of 209,393 data. We selected data that contain “Restaurants” or “Food” as one of its categories to leave data on restaurants only. Next, to build the ‘is_open’ column which is showing the current opening status of the restaurant, we dropped all the data that are already marked as ‘closed’ from 2020 data. Through this process, the total number of data was trimmed down to 56,936 rows. With the remaining data, we try to match the opening status with the 2022 Yelp dataset. Only 675 values returned either ‘close’ or ‘open’ status. This was the point that we found the need of using Yelp API which will give the most recent and correct features of the restaurant.

Although Yelp API is publicly available, registration for API key is required and there is a daily limit to the number of callings that can be made with the key received. We used business_id from the 2020 data to call the current business details. More information about the specific features this API can return could be found in API reference.

We could easily assume demographic features might be another huge factor that can result in the closure of restaurants. From this point, we decided to add some demographic information to finalize our dataset. We have included data from two sources: one from the FBI and one from the Census in order to best capture the characteristics of the particular city where the restaurants were located. From the FBI, we retrieved information on each city’s population, violent crime, property crime, and larceny crime. From the Census, we gathered information on median gross rent (2016-2020), median household income in 2020 dollars (2016-2020), percentage of each race, percentage of people with Bachelor's degree or higher, and percentage of people with health insurance of each state. After rigorous data preprocessing and transformations as described in the next section, we finally got a total of 36,365 restaurants left.

2. Data Preprocessing

Here are some detailed steps that we took to get the final version of data

1) Drop data if there are no ‘is_open’ values
2) Drop if the data was not coming from any one of the US states
3) Make a new column ‘review_change’ from ‘review_count_2022’ minus ‘review_count’
4) Make a new column ‘rating_change’ from ‘rating_2022’ minus ‘stars’
5) Make a new column ‘num_categories’
6) Separate attributes column. This returned 44 new columns. Left only 7 columns that don’t have too many NaN values. ('Parking_street', 'Parking_lot', 'Parking_validated', 'Parking_garage','RestaurantsTakeOut', 'Parking_valet','BusinessAcceptsCreditCards')
7) New columns got sparse data. Reduce the dimension of the attributes column with SVD algorithm. During this process, we labeled NaN or ‘None’ values to a separate label.
8) Combine reduced attribute features with the original data

[1] Hongju Lee:
Jungseo Lee: www.github.com/
Elizabeth Park:

9) Make a new column ‘price_range’ from the attributes’ PriceRange2 value. NaN values filled with median value.
10) Make a new column ‘num_attributes’
11) Make a new column ‘counts’ by getting the number of restaurants that have the same name as the restaurants
12) Merge with demographic data
13) The final data could be found in Table 1
14) Drop the values that don’t have any demographic data (dropped a total of 8 rows)
15) Minmax scaling

3. Data Description

Name of feature	**Description**
stars	Star rating of the restaurant in 2020
review_count	Number of reviews in 2020
Review_change	Changes in the number of reviews in 2022 since 2020
Rating_change	Changes in the star rating in 2022 since 2020
Num_categories	Number of categories
Price range	Price range of the restaurant
0	First value from singular value decomposition from the values of the attributes
1	Second value from singular value decomposition from the values of the attributes
2	Third value from singular value decomposition from the values of the attributes
Num_attributes	Number of attributes/details describing restaurant
counts	Number of restaurants in same names
population	Population of state
Property_crime	Property crime includes the offenses of burglary, larceny-theft, motor vehicle theft, and arson (FBI)
Larceny_theft	Larceny-theft is the unlawful taking, carrying, leading, or riding away of property from the possession or constructive possession of another (FBI)

Median_gross_rent	Median gross rent from 2016 to 2020
Median_household_income	Median household income (in 2020 dollars) from 2016 to 2020
Bach_deg_edu	Percentage of people with Bachelor's degree or higher
Insurance_und65	Percentage of people under age 65 with health insurance
White_perc	Percentage of population who are white alone
Black_perc	Percentage of population who are Black or African American alone
Asian_perc	Percentage of population who are Asian alone
Hispanic_perc	Percentage of population who are hispanic or latino alone
Is_open	Current status of the restaurant

Table 1. Description on dataset features

4. Model

Using sklearn’s classifications models, we fitted our dataset on logistic regression, ridge regression, support vector machine, k-nearest neighbor, random forest, gradient boost, ada boost, XGBoost, and multi-layer perceptron. We have selected these models based on our experiences throughout the course with assignments and kaggle competitions. We have set the models’ hyperparameter to its default settings to later on tune the parameters after model selection. To develop the models, 75% of the dataset was used to train the model and the rest 25% was used to test the model.

Evaluation and Analysis

There are 6 different metrics commonly used to evaluate the classification models: 1) Accuracy 2) Confusion Matrix 3) Precision 4) Recall 5) F1-score 6) ROC-AUC

1) Accuracy is a metric to get the number of data that are correctly predicted

$$Accuracy \ = \frac{TP + TN}{TP + TN + FP + FN}$$

2) A confusion Matrix is a cross table that shows the number of data classified to each category

	Predicted False	Predicted True
Actual False	TN	FP
Actual True	FN	TP

3) Precision is a good metric when you would like to minimize the Type Ⅰ error. (Reject the null hypothesis when it is actually true)

$$Precision = \frac{TP}{TP + FP}$$

4) Recall is a good metric when you would like to minimize the Type Ⅱ error. (Accept the null hypothesis when it is actually false)

$$Recall = \frac{TP}{TP + FN}$$

5) The F1 Score is a good metric to select a model with considering both Precision and Recall.

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

6) ROC – AUC
Receiver Operation Characteristic Curve (ROC) shows how True Positive Rate (TPR) is changing accordingly when False Positive Rate (FPR) is changing. When the curve is closer to the end of the y-axis, the ROC score is higher which represents the better performance of the model. Area Under ROC (AUC) is the area under the ROC curve. The maximum AUC is 1.

Due to the imbalance in our data, solely using the accuracy score to evaluate the model would not be appropriate in this case. Using ROC-AUC scores might be a great idea as it considers both TPR and FPR. Among the classification models we have implemented, random forest had the highest ROC-AUC score with 0.845, followed closely by XGBoost with a score of 0.842. Using GridSearchCV, we tuned the best random forest model with max_depth of 64 and n_estimators of 600, resulting in a small improvement to the score of 0.851. Table 2 shows the score results for our final random forest model.

ROC-AUC	Accuracy	Recall	Precision	F1 score
0.851	0.905	0.991	0.909	0.948

Table 2. Random Forest Performance for Predicting Restaurant Closure

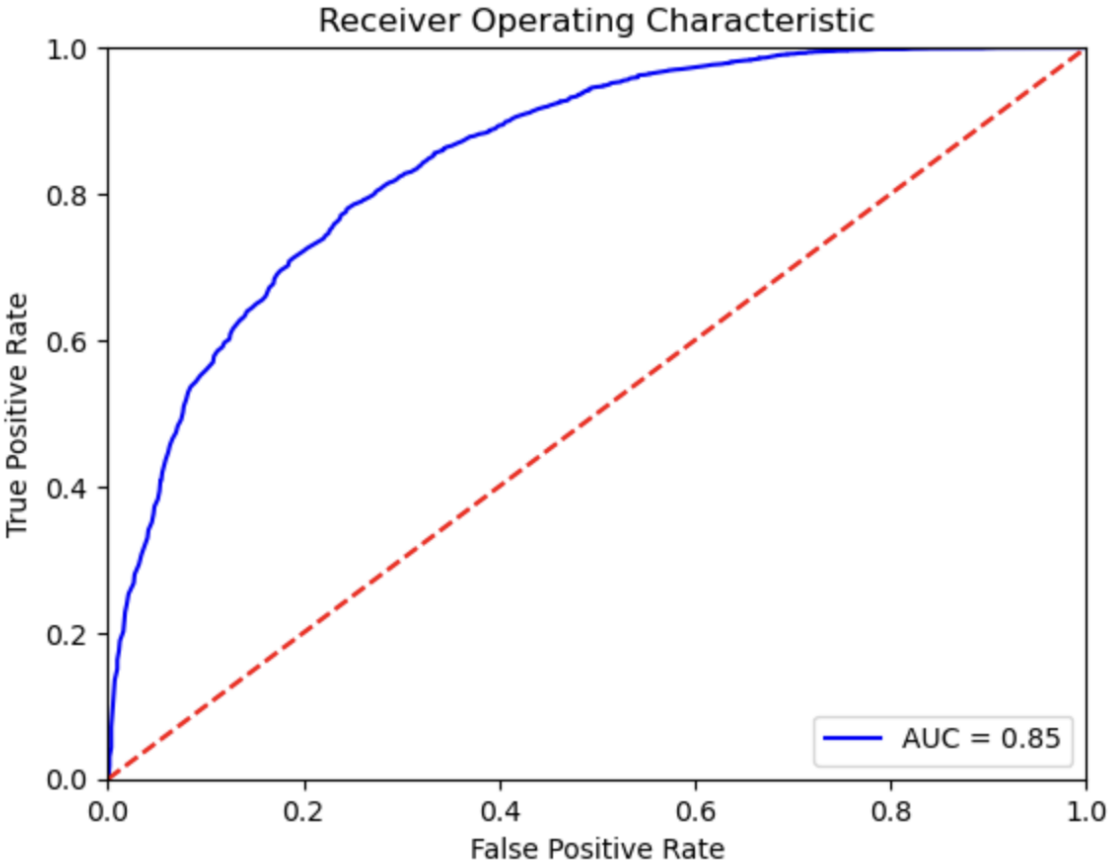

Figure 1. ROC-AUC Curve on Final Random Forest Model

Figure 1 is the ROC-AUC curve of the prediction result with our tuned random forest model. As 0.5 suggests no discrimination to make a prediction (dotted red line), the ROC curve in blue color infers the model brings an excellent prediction performance.

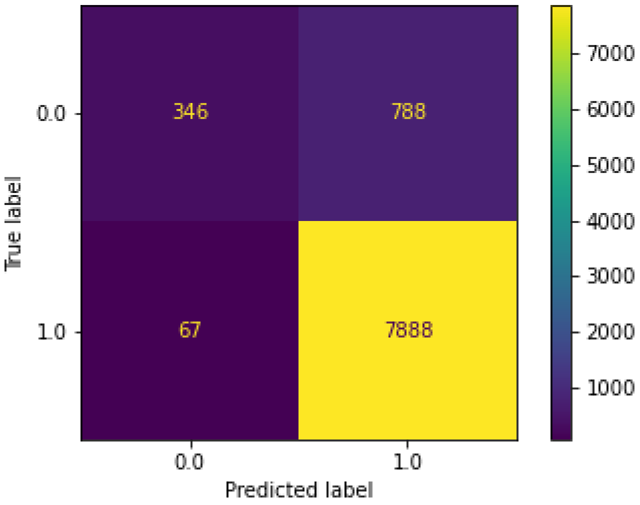

Figure 2. Confusion Matrix from Final Random Forest Model

Figure 2 depicts the confusion matrix resulting from our tuned random forest model. There were 346 TN cases, 788 FN cases, 67 FP cases, and 7888 TP cases. As the results show, our model was not best in avoiding false negatives. This primarily may be due to the imbalanced classes. There are far more open restaurants (label 1) than closed restaurants (label 0) in our dataset.

Related work

Because the topic we are trying to address, the survival of businesses, plays an important role in understanding our economic situation, it has been under people's attention for years. There are related publications that tackled this topic and took a similar approach using the Yelp API, as Yelp has been opening its own dataset challenge for quite a time, however, these publications may seem similar in an outer shell but different when investigated deeper. The past winners for Yelp challenges mostly analyzed and built nlp models using the public yelp reviews dataset. For instance, the sixth challenge winners, Li et al. proposed a new model called Topic Regularized Matrix Factorization (TRMF), combining topic modeling with a Latent Dirichlect Allocation and matrix factorization in order to recommend users' potentially preferred business by predicting the star rating they would give it.

Many of the publications focused solely on using Yelp API data without any additional supporting data from outside sources. For instance, Tao and Zhou, published a similar study to look into rather online consumer reviews signal restaurant closure using deep learning based time series analysis. The authors used a hybrid classification method and a novel triple word embedding model to extract data and integrated deep learning and time series analysis techniques to predict the business closure.

In addition to the official publications, there have also been blogs that utilized Yelp API to predict the outcomes of businesses. In 2018, Michail Alifierakis used feature engineering to predict business closure on the long term using reviews and features based on relative performance to surrounding restaurants for 2013 yelp dataset. The author used a simple linear logistic regression model, resulting in precision of 91% but recall of 61%. Another blog author named Scott Kramer analyzed restaurant closure impacted by COVID-19 using quantitative analysis on various aspects such as price, hours of operation, cuisines, and types of services offered. He concluded that low-cost restaurants that are dependent on lunch patrons were the most impactful of COVID-19, leading to closure. However, the analysis only uses restaurants closed from June to November of 2019.

There are also publications that focused on predicting consumer businesses in retail. This specific source, released by faculties of Economics and Business at the University of Groningen, also tackles the effects of COVID-19 and predicts the performance of businesses specifically in the retail field. The authors analyzed how COVID-19 has changed the competitive landscape of retail and long-term changes of consumers' shopping behaviors.

The three main parts that differentiates our project to the existing implementations of predicting restaurants' success outcomes are updated data subsequent to COVID, additional features and utilization of multiple machine learning models to find the best model. Our team improved these existing implementations by adding additional predictors from various sources such as FBI, Census.gov and used demographic data to get the population density of each city as well as crime rates, household income and educational level by city which radically improve our model's accuracy and prediction outcome. We also investigated further into the original Yelp data to extract additional predictors by performing careful calculations on the main data which also significantly improved the accuracy of our model.

Discussion and Conclusion

The period of relative isolation, social distancing and economic uncertainty changed the way we behave and impacted the global economic sentiment entirely. The insecure state has caused many businesses to close and discontinue their business for a period of time which has brought more attention to discover the factors that determine the outcome of businesses. We tackled this problem by investigating it through multiple viewpoints and extracted data from various sources to strengthen our model's accuracy. Through this paper, we strongly believe that we effectively showed how restaurant closures can be predicted with business attributions and its environment.

Our best model random forest stood out in performance with accuracy of approximately 91% and ROC of 85%. The results prove that the model can be used to predict restaurant closure in practice effectively. Figure 3 shows the Top 15 feature importance calculated using random forest. The most impactful feature was review_change, which is the difference in number of reviews in 2022 since 2020. Second came review_count, the number of reviews in 2020. What caught our attention the most was num_attributes, which was the number of attributes Yelp provided or restaurants manually included describing the restaurant. For example, attributes included whether they accept Credit Cards or not, whether they offer take outs and/or delivery, and availability for parking. We computed this variable in the first place because we assumed that the higher the value for num_attributes is, the more popular the restaurant is or the more attention the restaurant received from the management and the restaurant itself.

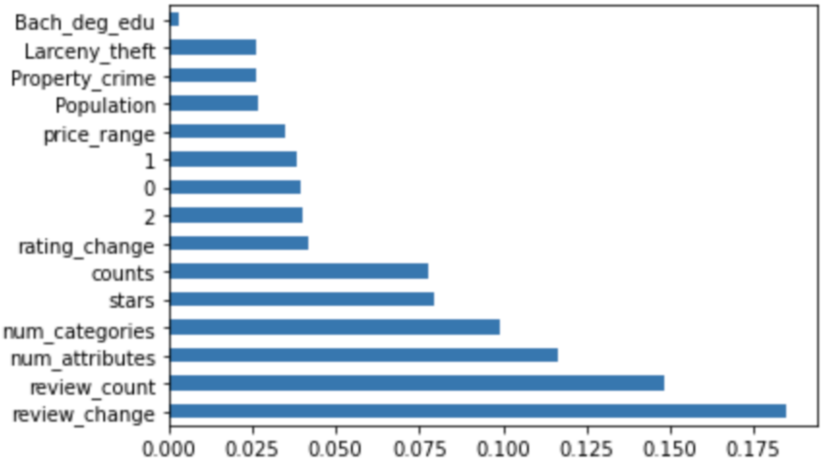

Figure 3. Random Forest Top 15 Feature Importance

This research can be continued in several directions to address its limitations. To mitigate the impact of heavy class imbalance, applying benchmark algorithms such as RFQ, SMOTE-RF, SMOTEBoost, and RUSBoost may be promising. SMOTEBoost and RUSBoost are both widely used to resolve class imbalance problem. SMOTEBoost is an oversampling method based on the SMOTE algorithm (Synthetic Minority Oversampling Technique), which uses k-nearest neighbors to create synthetic examples of the minority class. Such a method allows more weight for the minority class. RUSBoost achieves the same goal by performing random undersampling (RUS) at each boosting iteration instead of SMOTE. There are also models that are specifically applied from random forest, which was our best model. RFQ is a random forest model that applied q*-classifier proposed by O'Brien and Ishwaran. The inclusion of q*-classifier maximizes the sum of the true positive and true negative rates, reducing class imbalance. Kotipalli and Suthaharan also proposed a new mathematical model using random forest classification approach including its adoption with SMOTE. Their approach explains the relationship between true positive classification rate and the imbalanced ratio between the majority and minority classes.

As Figure 3 implies, review appears to be a strong subject of matter in determining the closure of restaurants. So, incorporating our model with further analysis on review may be another plausible direction for the future. In addition, as our results are promising, exploring by joining more information such as restaurant inspection scores or COVID 19 may be interesting to compare results with our model.

Through this project, we learned the importance of data reliability. Our team has invested much time to make sure we have good quality of data because the unqualified data could lead to inaccurate results. According to IBM research publication, many researchers and practitioners focus on improving the quality of models while investing very limited efforts towards improving the data quality and reliability. The authors mentioned, "One of the crucial requirements before consuming datasets for any application is to understand the dataset at hand and failure to do so can result in inaccurate analytics and unreliable decisions." Because we combined multiple data from various sources, we made sure that we are joining the data without any data gaps, anomalies, and duplicates and ensure the data we put into our models are accurately structured and correct.

References

- O'Brien R, Ishwaran H. A Random Forests Quantile Classifier for Class Imbalanced Data. Pattern Recognit. (2019, Jun). ;90:232-249. doi: 10.1016/j.patcog.2019.01.036. Epub (2019, Jan 29). PMID: 30765897; PMCID: PMC6370055.
- Kiranmayi Kotipalli and Shan Suthaharan. (2014). Modeling of class imbalance using an empirical approach with spambase dataset and random forest classification. In Proceedings of the 3rd annual conference on Research in information technology (RIIT '14). Association for Computing Machinery, New York, NY, USA, 75–80. https://doi.org/10.1145/2656434.2656442
- Anna Vasilyeva. (2018, May 8). "Using SMOTEBoost and RUSBoost to deal with class imbalance" *Medium*. https://medium.com/urbint-engineering/using-smoteboost-and-rusboost-to-deal-with-class-imbal
- Alifierakis, Michail.(2018, June 13). "Using Yelp Data to Predict Restaurant Closure - Towards Data Science." *Medium*,, towardsdatascience.com/using-yelp-data-to-predict-restaurant-closure-8aafa4f72ad6.
- Jain, Abhinav, et al. (2020, Aug.). "Overview and Importance of Data Quality for Machine Learning Tasks." *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &Amp; Data Mining*, ACM, Aug. 2020, https://doi.org/10.1145/3394486.3406477.
- Kotipalli, Kiranmayi, and Shan Suthaharan. (2014). "Modeling of Class Imbalance Using an Empirical Approach With Spambase Dataset and Random Forest Classification." *Proceedings of the 3rd Annual Conference on Research in Information Technology - RIIT '14*, ACM Press, 2014, https://doi.org/10.1145/2656434.2656442.
- Kramer, Scott. (2021, Dec. 31). "Analyzing COVID-19 Restaurant Closures With Yelp Data." *Medium*. https://medium.com/13-fund/analyzing-covid-19-restaurant-closures-with-yelp-data-f9116c7d563a
- O'Brien, Robert, and Hemant Ishwaran. (2019, June). "A Random Forests Quantile Classifier for Class Imbalanced Data." *Pattern Recognition*, vol. 90, Elsevier BV, pp. 232–49. https://doi.org/10.1016/j.patcog.2019.01.036.
- Tao, Jie, and Lina Zhou. (2022). "Can Online Consumer Reviews Signal Restaurant Closure: A Deep Learning-Based Time-Series Analysis." *IEEE Transactions on Engineering Management*, Institute of Electrical and Electronics Engineers (IEEE), pp. 1–15. https://doi.org/10.1109/tem.2020.3016329.
- Verhoef, Peter C., et al. (2022, July). "Reflections and Predictions on Effects of COVID-19 Pandemic on Retailing." *Journal of Service Management*, Emerald, https://doi.org/10.1108/josm-09-2021-0343.