Predicting Restaurant Success with the Yelp Dataset and Economic Features

Garrett Pinkston, Ko-Wei Chang, and James Barnson

Department of Applied Mathematics and Department of Engineering Management,

Baskin School of Engineering, University of California Santa Cruz

TIM147: Introduction to Data Mining for Business

Professor Yi Zhang

December 13, 2023

Abstract

Restaurants have some of the highest failure rates of all business types. Predicting failure could prevent economic losses by giving small businesses the chance to turn things around. This project set out to predict which restaurants will fail using the publicly available Yelp Academic, Simple Maps, and Kaggle Household Income datasets. Measures for population, income, number of locations, reviews, and review quality were all tested. The most predictive features were the number of reviews, the slope of the change in reviews over time, the population of the neighborhood, and “interactions” with the reviews. Of the models tested, Random Forest performed the best, achieving 77.86% accuracy. The most significant predictive factors were the number of reviews, positive changes in reviews over time, interactions with reviews, population, and income. All of these features were more predictive than the star rating in reviews. While these results are preliminary, possible recommendations to restaurant owners in light of these findings are to engage with reviews on the platform, encourage honest reviews from customers, and treat negative reviews as useful feedback. Prospective restaurant owners should seriously consider the impact of city population and neighborhood income when selecting a location.

Introduction

Other work has been done to predict restaurant success using the Yelp dataset. A sampling of these were reviewed before formulating and embarking on this project. Four different articles were particularly of note and informed the way we approached this project. They are presented below in order of publication.

As a PhD candidate and intern, Michail Alifierakis found that star ratings were not predictive of remaining open. Features he used included whether it was a chain, restaurant density, review count, rating, and price relative to nearby restaurants. He found Logistic Regression to be the most effective model, identifying the most important attributes: whether restaurants were part of a chain, the number of reviews they had compared to nearby restaurants, and the proximity of other restaurants. He suggests change in population demographics as a key way the analysis could be improved (Michail, 2018).

In 2018, Lu et al. attempted to predict whether a restaurant would still be open in 2017 based on 2016's reviews with a total training set size of less than 2100 restaurants. The features they tested were a combination of review sentiment and business features, such as being part of a chain. Logistic Regression was also used to make the prediction, and their model was able to predict a 17.46% increase over the baseline. They found sentiment had no effect on the accuracy of the model and that the most predictive factor was being part of a chain. (Lu et al., 2018)

Saleh et al. presented at the 2019 International Conference on Advances in Social Networks Analysis and Mining. They predicted restaurant success using the distance between a restaurant's location and places of interest and the cost of living within a zip code. It found that both measures of location significantly correlated with success, with a ratio of 0.81. Success was measured only by reviews and ratings. The datasets they used were the Yelp 2018 dataset and city-data.com for location data (Tayeen et al., 2019).

Dhruv Gargi and Riddha Mathur published a 2022 conference paper focusing on the effects of the COVID-19 pandemic on business survival. The article uses business attributes and COVID-19 features in the Yelp dataset to predict closure in the midst of COVID-19. Population and income by zip code features were included. They tested 6 different types of machine learning models; a key result was that all models performed better when predicting open restaurants than when predicting closed ones (Gargi & Mathur, 2022)

Following this review of the literature, we decided to further explore the influence of reviews, chain size, and economic factors on the continued operation of businesses listed on Yelp. To implement this, we used the Yelp Academic Dataset, other publicly available datasets, and the optimized machine learning models in Python modules. Our goal was to take a similar approach as previous authors but maximize results through the addition of relevant economic data, innovative feature engineering, and rigorous model hyperparameter tuning. In turn, this allows us to directly compare predictive accuracy to the conference papers found in our literature review while investigating the feature importance found in those papers in light of other economic features and features available in the Yelp dataset.

Methodology

For our project, we utilized a combination of 3 publicly available datasets. The Yelp Academic Dataset (Yelp, 2019) is freely available for academic research and consists of 5 tables: Businesses, Checkins, Reviews, Picture Captions, and Users. Out of these tables, we used only the Businesses and Reviews. Since the Yelp dataset lacks economic and demographic information, we sourced those from other freely available datasets. These datasets are the Simple Maps US Zip Code Dataset (Simplemaps, 2023), and the Golden Oak Research Group US Household Income Statistics dataset (Golden Oak Research Group, 2016). Both of these were the freely available versions of the dataset, and therefore more sparse than the paid versions.

Within the Yelp dataset, fifteen attributes are given within the business table. This includes the unique business id (“business_id”); name (“name”); the city, state, latitude, and longitude (“city”, “state”, “latitude”, “longitude”); star rating and review count (“stars”, “review_count”); category tags (“categories”); and whether or not they remain open (“is_open”). Most of these attributes are self-explanatory. “is_open” is the attribute used to annotate whether the business is open or not. Businesses that were open until the dataset release date were marked as 1 in “is_open”, and those which were closed were marked as 0. The Reviews table contains all user reviews up to the release date of the dataset. The reviews table has 9 attributes, and in this analysis, we used “review_id”, “user_id”, “business_id”, “stars”, “date”, and “text”. Review and business datasets could be connected using “business_id”, which means the review.json file stores all the reviews for all the businesses recorded in the business.json file.

Feature Engineering

When beginning our feature selections, we extracted only restaurants from the Businesses table, totaling 52,268 restaurants from approximately 150,000 entries in the business table. Afterward, instances of missing values were dropped; thus, our analysis only captures restaurants that had comprehensive Yelp listings. After all features were joined, all rows with N/A fields were again dropped; further work could be done in the future to impute values in appropriate ways. Our final dataset had 15,217 restaurants. What follows is a description of how each of our features was created.

Population and Population Density. These columns from the Simple Maps dataset were joined with the Yelp Business dataset using postal codes. While the two have very high covariance, joining both allowed comparison within the data with either or both measures of population. The population density was hypothesized to be a better predictor, as zip codes can vary greatly in size; however, testing disproved this hypothesis, and the population was used in the final models.

Median Income. Median income was selected as the most representative income statistic and was taken from the Golden Oak dataset. Simple Maps' dataset had unique, standard 5-digit zip codes like the Yelp Business dataset; however, the Golden Oak dataset uses neighborhood-scale zip codes. These were incompatible with the 5-digit zip codes of the Yelp dataset. Instead, a geopandas spatial join was used to match the nearest neighborhood latitude and longitude with the latitude and longitude in the Yelp Business dataset.

Locations and Chain Size. Michail and Lu et. al. both found that chain size was the most predictive of all features in their work. To control for this, the team decided to include the number of locations and chain size as features in our models. First names of restaurants were preprocessed, then repeats of each name were counted to get the number of locations for each business. These were then one-hot encoded into bins with 1-3 locations not being a chain, 4-20 as a small chain, 21-50 as a medium chain, 51-100 as a large chain, and >100 being a mega chain. Up to 3 locations are not considered a chain to account for possible repeat names of independent restaurants. Moreover, hands-on independent owners overseeing multiple locations in their restaurant business share greater commonalities with single-location business owners than businesses with more locations than a single owner could reasonably manage. Managing 3 locations was assumed to be a reasonable maximum for a highly involved owner.

Review Trends / "Star Slope". Changes in reviews over time could greatly affect the success or failure of a restaurant; a restaurant with many reviews could have a change in quality that results in increasingly negative reviews and ultimately lead to closure. However, it was important to not introduce a measurement of review time into the model. This could bias results by predicting "closed" for restaurants whose most recent reviews were further in the past when the lack of recent reviews is the result of closure rather than a cause. Thus, a measure of the "slope" of changes in review stars was created. This was accomplished using linear regression to model the changes in the review of the restaurant. To do this, we first converted the dates into an integer variable and then applied the MinMaxScaler to the review dates so the slope isn't a large number to use. All the review dates and stars for each restaurant were then looped through and fitted to linear regression to find the slope. The correlation coefficient of the review dates and stars was also calculated in the same way, but it was not used as the slope of the review stars proved to be more vital to model performance than the correlation coefficient.

Other Features. Additional features that were created and included in the final data set were the price of the restaurant and the takeout offerings. Each of these entries came in the form of a dictionary, so this information had to be respectively extracted into a new column and one-hot encoded.

Modeling

After defining our data, we now present the implementation and tuning of machine learning models for our predictive system. Open restaurants dominate the Yelp dataset, and always predicting “open” would achieve a 68.87% accuracy. This provides a baseline model to which we can compare other results. Initially, we chose to examine the results of Random Forest, XGBoost, Support Vector Machine, and Logistic Regression models. These are all common models used in binary classification and would give the best direct comparison against the models selected by other researchers.

Data was randomly split into training and testing, 80% and 20% respectively. Models were optimized through a RandomSearchCV, which tests the model over a randomized grid of specified parameters while utilizing 5-fold cross-validation. This is an extensive way of ensuring the optimal results are achieved for the model and prevents overfitting through cross-validation.

In evaluating all examined models, we employed accuracy scoring, a metric that involves comparing predicted values against true values. Our choice of accuracy scoring was driven by the intention to assess the robustness of our model in comparison to those employed by other researchers. After RandomSearch was performed, we utilized the best combination of parameters found in training to generate predictions using the test set. These predictions were compared and ranked by their accuracy scores. Precision and recall were then calculated for the top-performing model.

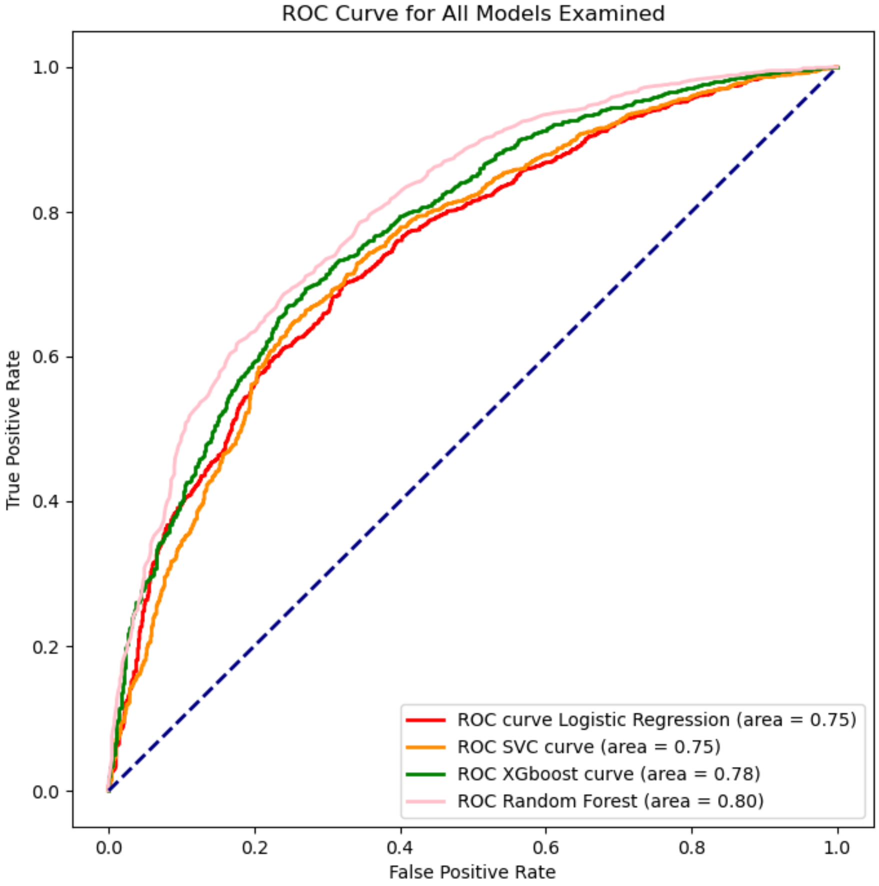
Results

Our Random Forest model achieved a 77.86% accuracy, followed by XGBoost with a 75.85% accuracy, then Logistic Regression with a 73.85% accuracy, and last Support Vector Machine achieved an accuracy of 73.62%. It isn't a surprise that Random Forest and XGboost models had similar performances, as both are bootstrap aggregate methods that utilize sequentially trained decision trees. In order to contextualize the maximum accuracy of 77.86% achieved by Random Forest, we can compare this finding against the baseline accuracy using the table below.

Model	Accuracy	Increase	% Improvement
Random Forest	77.86%	8.99%	13.05%
XGBoost	75.85%	6.98%	10.14%
Logistic Regression	73.85%	4.98%	7.23%
Support Vector Machine	73.62%	4.75%	6.90%
Base: Always "closed"	68.87%	N/A	N/A

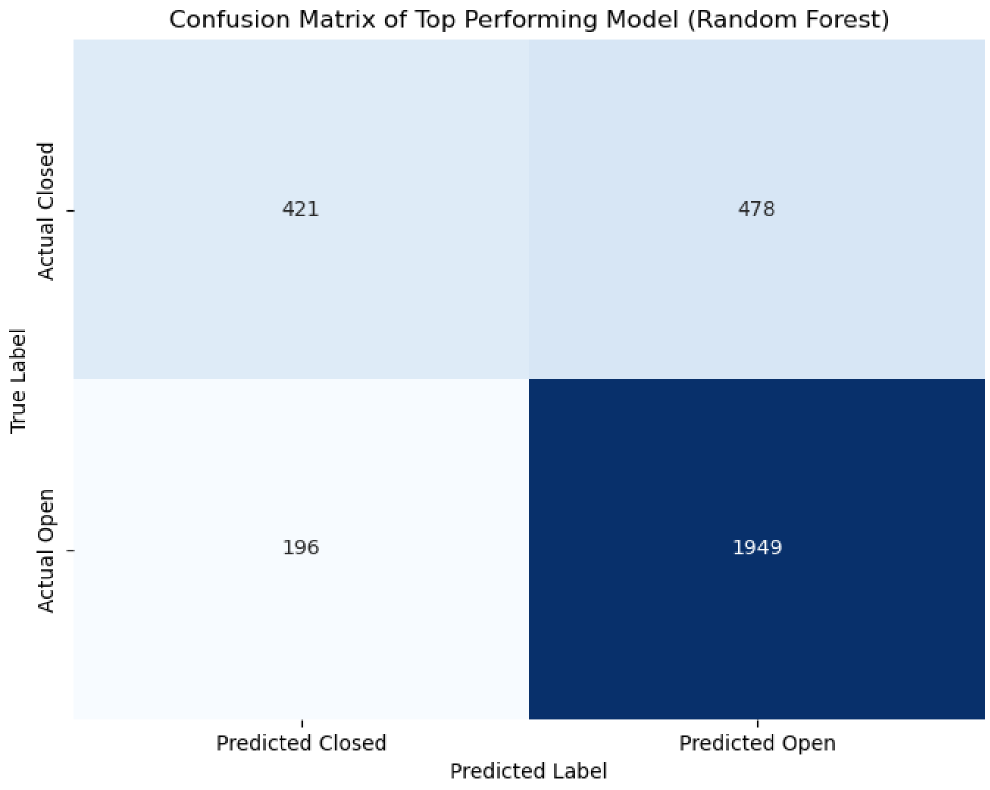
In order to enhance result interpretation, we generated Receiver Operator Characteristic (ROC) curves for each model, which illustrate the trade-off between the true positive rate and the false positive rate. We employed the best model identified through RandomSearchCV and utilized the 'predict_proba' method from scikit-learn to predict the probability of each instance belonging to the positive class. This approach captures the model's confidence for each prediction. Given that the ROC curve function requires model output probabilities, this step was essential. Subsequently, we applied scikit-learn's ROC curve function, which assesses the true positive rate against the false positive rate while incorporating instances with varying confidence levels.

We repeated this process for all four models and then graphed the ROC curves for all plots together. As a baseline, we put the line *y=x* which highlights the results which would be obtained by randomly guessing the class of each instance in the dataset. Additionally, we calculated the Area under the Curve (AUC), an important metric that helps quantify the ROC's overall performance. This AUC output captures the total area under the ROC curve, which serves as a numerical output of the ROC curve. In turn, this helped us quantitatively compare the overall difference of all ROC graphs.



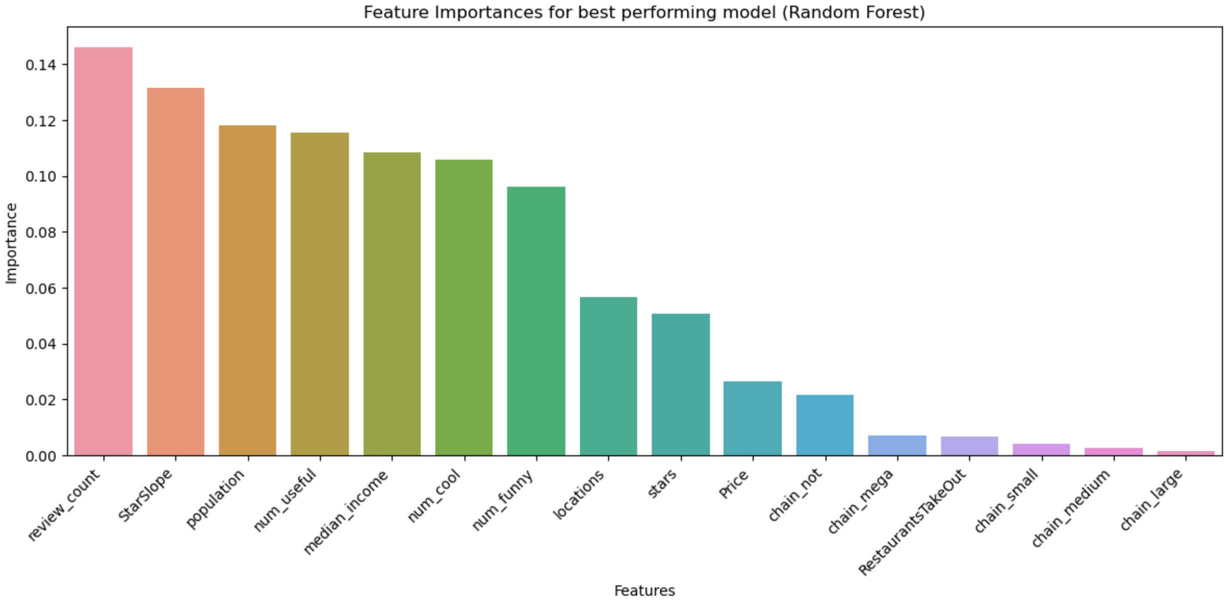
Computing the ROC curve confirmed Random Forest to be our best-performing model through two metrics. First, we can visualize the line in pink (representing the Random Forest ROC curve) occupies the closest arbitrary point towards perfection of (0,1). In other words, it had the greatest increase in true positive rate against false positive rate and maintained this trend throughout the curve. The Random Forest model also achieved the highest AUC of 0.80.

Next, we created a confusion matrix for our Random Forest model in order to further understand the results. This captures pertinent information on all necessary calculations of Accuracy, Precision, Recall, and F1. For brevity's sake, we only examine the results of the Random Forest model here and discuss important metrics such as accuracy, precision, and recall.



Our confusion matrix highlights the strength of our model's predictions. As seen in the ROC graph, the Random Forest model maintained a very high precision (calculated as the number of accurate predictions that a restaurant will be open over the total number of open predictions) of 80.3%. Our recall (calculated as the number of accurate predictions that a restaurant will be open divided by the total number of restaurants that are actually open) achieved a phenomenal result of 90.8%. The presented confusion matrix visually encapsulates all the information found in an F1 recall table. In our context, precision holds paramount significance as it enables prospective investors and landlords to assess businesses they may consider collaborating with. Notably, our model exhibited exceptional precision, surpassing most metrics examined.

After thoroughly examining the model's performance metrics, we can now understand the feature importance of all features inputted for our Random Forest model. This can easily be accomplished by calling scikit-learn's '.feature_importances_' attribute for Random Forest. As our top model, it is imperative to understand the combination of features in order to provide an accurate and comprehensive business strategy.



As visualized above, we can see how the review count has a substantial impact, accounting for a 15% weight in model prediction. Additionally, city population, income, and review trends were large factors in our model’s predictions, achieving weights upwards of 12%. After those, review interactions (such as useful, funny, and cool) achieved significant weights of around 11% each. Next, there is a steep drop off in the importance, as stars and price only reached weights around 5%. Two more engineered features, price and binned data of no chains, played a minor role around 2% impact. Last, the rest of the chain bins and takeout offerings played a minimal role in predictions, all achieving weights of less than 0.5%.

After understanding the importance of the features, we can recommend that restaurant owners improve the frequency of reviews and promote interactions in the reviews by engaging with reviewer comments. Doing this can help the owners gain a clearer understanding of customers' preferences and critiques, enabling them to enhance their restaurant based on the feedback. Prospective restaurant owners should seriously consider the impact of city population and neighborhood income when selecting a location. While this approach may appear trivial, it could significantly underscore concrete business improvements, boost customer satisfaction, and decrease the likelihood of business foreclosure.

Limitations

Population density can significantly vary across a zip code, but Simple Maps did not have multiple values per zip code to be able to match population density more accurately with restaurant location. The free version of the dataset only includes most US zip codes, and the Golden Oak dataset is also an incomplete free dataset; this resulted in many N/A values for population and income measures in our dataset, which were dropped rather than imputed. This represents a significant loss in data and affects the results. The Golden Oak dataset is 6 years old, and income statistics have changed dramatically since then. The accuracy of income statistics could be improved with the newer, updated dataset or time series analysis.

The number of locations for each restaurant is complicated by repeat names of restaurants that are not associated with each other and by variations in names from location to location in chains. Sophisticated data preprocessing techniques, such as spatial mapping, would be required to effectively separate independent restaurants under the same name. This limitation was partially mitigated by categorizing 1-3 locations as not a chain and by normalizing the names of restaurants. Time-intensive research would be required to identify chain restaurants whose location names significantly differed from each other and categorize them correctly.

Accuracy for larger chains was also severely impacted by the limitations in Yelp's dataset. Therefore this analysis does not measure the true effect of location numbers, and it cannot be a true control for locations overall. With more than 50 locations in the dataset, the representation of chain size becomes extremely inaccurate. Additionally, usage of the Yelp Academic dataset solely caters to metropolitan areas. Great care should be taken when applying the results of this model outside of urban areas.

Conclusion

We set out to predict with high accuracy whether a restaurant would remain open and to see if the results of previous work held up with additional features provided by Yelp and economic factors. The best model tried was Random Forest, achieving an accuracy of 77.86%, a recall of 90.8%, and a precision of 80.3%. Our findings were that the inclusion of population, income, and additional review-related features greatly decreased the impact of chain size on the predicted success of restaurants. Review-related features such as the number of reviews, the slope of change in reviews as of the time of the most recent review, and interactions with the reviews were all by far the most significant predictors of success. The number of locations and the star rating were important factors, but not as significant as in previous findings. Independent restaurant owners will be reassured by these findings and should focus their efforts on encouraging an active Yelp page and understanding user feedback.

Contributions

Garrett Pinkston	Built and Tuned all 4 examined models (Random Forest, XGBoost, SVM and Logistic Regression) Calculated review interactions, price and whether they offered takeout Created graphs of ROC curves and confusion matrix of top model. Also made a graph of feature importance and word cloud (seen in the presentation). Created USA contour graphs of Income, Restaurant Distribution, and interactive restaurant map of Florida's Restaurants Performed initial text exploratory data analysis (seen in presentation) Model & Results of research paper
Ko-Wei Chang	Created review star function to find the overall review star slope for the restaurants for feature engineering Created visualizations for locations and chain size for the restaurants (seen in the final presentation) Worked on references, feature engineering, in-text citations of research paper Wrote and edited limitations slides for the presentation Helped come up with ideas of how to make suggestions to restaurants that may not succeed Final proofreading of the research paper
James Barnson	Merged population and income features and wrote corresponding paragraphs. Created locations & chain size features and wrote corresponding paragraphs. Analyzed code and results of modeling for key figures, statistics, insights, implications, and limitations.

	Wrote abstract, introduction, literature review article summaries, limitations, and conclusion sections. Edited the Modeling and Results sections for clarity and flow. Created presentation slide deck structure, wrote and edited contents of many slides, and created results table in slides.

References

Alifierakis, M. (2018, January 8). Using Yelp Data to Predict Restaurant Closure. Medium. https://towardsdatascience.com/using-yelp-data-to-predict-restaurant-closure-8aafa4f72ad6

Dhruv Gargi, & Mathur, R. (2022). Using Machine Learning to Predict Business Survival in COVID-19. Smart Innovation, Systems and Technologies, 316, 249–261. https://doi.org/10.1007/978-981-19-5403-0_21

Golden Oak Research Group. (2016, April 16). US Household Income Statistics[Data set]. https://www.kaggle.com/datasets/goldenoakresearch/us-household-income-stats-geo-locations

Lu, X., Qu, J., Jiang, Y., & Zhao, Y. (2018). Should I Invest it? Proceedings of the Practice and Experience on Advanced Research Computing. https://doi.org/10.1145/3219104.3229287

Mangal, S., Behl, H., & Venkatesh, P. (2019, January 23). Restaurant-Success-Predictor-based-on-Yelp-Dataset. GitHub. https://github.com/goelshivani321/Restaurant-Success-Predictor-based-on-Yelp-Dataset/blob/master/Project_report.pdf

Tayeen, A. S. Md., Mtibaa, A., & Misra, S. (2019). Location, location, location! Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. https://doi.org/10.1145/3341161.3345334

Simplemaps. (2023). US Zip Codes Database [Data set]. https://simplemaps.com/data/us-zips

Yelp. (2019). Yelp Dataset [Data set]. https://www.yelp.com/dataset