Yelp Data Analysis for Business Survival Prediction

Lincy Chen (lyc26)
Alyssa Yam (aly23)

Table of Contents

Introduction........3
Research Question........3
Data Description........3
Response Variable........3
Predictor Variables........4
Approach........5
Data Preparation........5
K-Means Clustering of Cities........6
Log Transformation of Review Counts........6
Days Open Per Week........6
Label Encoding Ratings........6
Random Forest........7
Results........7
Significance........8
Model Fairness........9
Links and References........9
Appendix........10
Tables and Figures........10
Member Contributions........13

Introduction

Living in this Digital Age, our lives are increasingly shaped by the voices and stories we encounter online. For enterprises big and small, the widespread reach of our opinions has offered businesses the ability to gain valuable insights in understanding customer choice. Several studies have explored the connection between these factors, highlighting the importance of word-of-mouth on the popularity of a business (Zhang et al., 2010). Our project reveals to prospective and current entrepreneurs whether online platforms have a significant influence on their success and provides actionable strategies for leveraging this influence to improve their outcomes.

For example, if we discover that businesses operating in the restaurant industry are less prone to closure when there are more reviews published to their Yelp page overall, this would indicate to restaurant owners that encouraging customers to post a review after being serviced would be beneficial when it comes to maintaining operation. Furthermore, controlling for metrics that may influence business performance is a vital step in operations management; our findings will showcase how businesses can prioritize certain operations, or factors, over others in relation to their online reputation.

Not only is our research question important to business owners, but it could also signal to consumers the value of their Yelp contributions to the community. Ribeiro-Soriano's 2017 study showed that small businesses shape the regional identity and develop local economies. If Yelp reviews play a role in sustaining businesses, encouraging the local community to visit stores and publish reviews could be a crucial factor in gaining the traction that small startups often need.

Research Question

Can we predict the success of a business (open/close) based on Yelp review information, and if so, what factors extracted from the reviews are most predictive of business outcomes?

Data Description

The dataset contains 150,346 rows of relevant business-level information such as location, number of reviews, ratings, opening hours, business attributes, and open status. With this volume of data, we expect our machine learning model to have sufficient information for making accurate predictions.

Response Variable

Operating status `is_open` is a binary variable with values ranging from either 0 or 1, with 0 indicating closure and 1 indicating continued operations. Closures may be a result of unpredictable external factors, shifts in consumer preferences, or poor marketing. In tandem with the following predictors, understanding the reasons behind closures can provide valuable insights for businesses looking to improve their resilience and long-term sustainability.

Predictor Variables

Location Location is measured through a variety of features such as `city` (string), `state` (string), `postal_code` (int), `latitude` (float) and `longitude` (float). Including location in our analysis as a potential important factor would tell us whether a business's location impacts business operations. Further analysis could show whether an area with many businesses clustered together flourish or fail due to the competitive landscape. Our dataset uses data from ~1300 different cities, as plotted in Figure 1.

From Figure 1, we can also see this information has been gathered from 11 major clusters, representing a wide range of US geographies. Given this setting, to reduce the high cardinality of these features, we elect to engineer a new feature that maintains the interpretability and information that longitude, latitude, city, and state provide. A discussion of our approach is continued in the *KMeans Clustering of Cities* section.

Rating Another important variable is `stars`, or average rating of a business. Yelp permits users to rate businesses on a scale from 1–5, with half step increments in between each (1, 1.5, 2, 2.5, … , 4.5, 5). The ratings serve as an ordinal, quantitative measure of satisfaction with the business and if found to be significant, could link customer satisfaction to business performance as well. As shown in Figure 2, this data is discrete and left-skewed, with the mode at 4 stars.

Number of Reviews Tangentially related is `review_count`, another quantitative measure of the number of reviews a business has received. If a business is performing well, we expect it to have more reviews, thanks to it being more popular. This information could help small businesses revise their business strategy and encourage customers to leave reviews on Yelp with every interaction. After an initial assessment of the data, we notice that there are outliers skewing the distribution of reviews with values of over 5000 (see Figure 3a). This may be due to the nature of our dataset, in which more popular businesses tend to attract more customers and reviews, a positive phenomenon. Nevertheless, we decided to log-transform our data as the resulting right-skewed distribution becomes much more apparent in Figure 3b.

Number of Days Open in a Week The raw feature, `hours`, is a dictionary that maps each day of the week to a time range in 24-hour format. It comprises the most instances of missing records (15.45%) in our dataset, yet conveys valuable information about a business's ability to manage operations and consumer preferences. We typically see small businesses open 5-6 days of the week rather than 7, as corporations tend to have more operational ability. On the consumer side, customers generally enjoy having more choices at their disposal when it comes to meeting their needs and preferences. There are opportunity costs involved in deciding opening hours, so it makes intuitive sense that this feature impacts a business's survival.

Food Industry Lastly, we can use the `categories` column to understand consumer preferences: is the choice to operate in the food industry more important in predicting the closure of a business? In other words, this qualitative feature should reveal whether closures are more prevalent in certain industries than others. For example, if we found that businesses in the food industry are more predictive of closure, then the implication could be that consumers are more influenced by reviews in certain industries than others. Further research could explore at location-level whether different industries are more likely to receive higher reviews than others.

In Figure 5, we look at the proportion of closed businesses in the food industry opposed to the rate of closed businesses in non-food industries. Among businesses not in the food industry, 11% report closure, while this figure more than doubles to 28% for those in the food industry. This preliminary analysis suggests a potential correlation between industry type and closure rates, emphasizing the need for further investigation into the relationship between closures and consumer reviews.

Hence we are likely to be successful in development of this project thanks to its intuitive, well-structured fields, clear target variable, and large volume of data. The features described above are of relevance to real-world business decisions, making our prospective model useful and generalizable when it comes to working with new data.

Approach

Data Preparation

In order to optimize our model performance, we needed to handle missing data, transform variables, and check for multicollinearity among our features. Table 1 shows the incidence of null values in our dataset; fortunately, our data was relatively clean with no duplicates, consistent structure, and minimal missing values. We decided that imputation would not be appropriate in this case as the null rates seemed relatively low in comparison–thus, we decided to drop the missing values.

While checking for multicollinearity among our features is not a major requirement for executing random forest analysis, our goal was to create a reliable model. According to Figure 6, our features fulfilled this criterion as none of the correlations exceeded 0.5.

K-Means Clustering of Cities

As previously mentioned, there were originally ~1300 unique cities in the dataset. Simply throwing this many unique classes into the random forest algorithm would have significantly delayed our ability to refine the model. Instead, we decided to deploy K-Means clustering with k = 11 to obtain labels for general geographic areas while still maintaining the interpretability and information of our data.

11 clusters made intuitive sense to us as the geographic layout of our data in Figure 1 plotted 11 clusters of varying sizes in different parts of the US and Canada. Figure 10 shows the result of our clustering–an almost mirror image of the map we generated in our preliminary analysis. Upon further analysis, we distinguished each cluster label, using the most occurring city, state in each cluster. The results of this mapping are displayed in Table 2.

Log Transformation of Review Counts

In reference to Figure 3a, the occurrence of outliers severely skewed our analysis of review counts. In order to account for these outliers where the number of reviews received by a business far exceeds 5000, we decided to apply a log transformation. While a random forest model is not known to be susceptible to the influence of outliers, in order to optimize for model performance we felt that the log transformation would fairly represent the majority of businesses in the dataset.

Days Open Per Week

Being a dictionary, the original feature, `hours`, was awkward for analysis and input in the random forest model. If there were days where the business was closed, the dictionary would lack keys representing those days. Given this structure, we opted to count the number of days, or keys present in the dictionary, and engineer this new feature to quantify a business's operating schedule. This new feature, `OpenDaysCount`, ranging from [1-7], would also reveal important insights to prospective business owners while also simplifying the analysis and messy structure. Figure 4 illustrates the result of this new feature.

Label Encoding Ratings

In order to preserve the ordinal, discrete nature of the rating information, we decided to encode the `stars` variable. This method was implemented using `scikit-learn`'s LabelEncoder function. As a result, the following was applied based on rating → label: [1.0 → **0**, 1.5 → **1**, 2.0 → **2**, 2.5 → **3**, 3.0 → **4**, 3.5 → **5**, 4.0 → **6**, 4.5 → **7**, 5.0 → **8**].

Random Forest

A random forest model is a powerful supervised learning method that uses multiple decision trees to return real-world inferences and predictions regarding a classification problem. Its ability to handle a large amount of data and predictors with high accuracy assured us that we could depend on its capabilities with our large volume of data and features of varying types. In the context of our research question, where we aim to identify influential factors in business closures and develop an accurate binary classification model, a random forest approach appeared most suitable. During our implementation, we deemed that 100 decision trees was most appropriate given our computational restrictions. Especially since we wanted to apply our model to real-world business decisions, the multiple decision tree characteristic of a random forest reduces the variance of a model, making it ideal for analysis.

As per Figure 9, there are more closed businesses than open businesses in our dataset. The advantages of random forests permits us to proceed with our imbalanced class response–the model prioritizes learning splits that more clearly separate the minority class from its counterpart (Chen et al. 2004). Another key feature of random forests uses bootstrap sampling to create multiple subsets of the data for each tree. This sampling technique can help balance the classes in each subset, especially if the minority class is represented in at least some of the subsets. In other words, by considering only a subset of features at each split, this can help prevent the majority class from dominating the splits and allow the model to focus on features that are more informative for both classes.

Results

Table 3: Random Forest Gini Values

Feature	**Gini Importance**
log_review_count	0.5743
city_cluster	0.2024
stars	0.0895
FoodIndustry	0.0769
OpenDaysCount	0.0567

Table 4: Classification Report

	Precision	**Recall**	**F1 Score**	**Support**
0	0.34	0.15	0.21	4701
1	0.83	0.93	0.88	20724
accuracy			0.79	25425
macro avg	0.58	0.54	0.54	25425
weighted avg	0.74	0.79	0.75	25425

Our model was trained to predict whether businesses are open or closed based on several features: city_cluster, OpenDaysCount, FoodIndustry, log_review_count, and stars. The model achieved a resulting accuracy of 78.75% with the following performance metrics of a precision of open businesses of 0.83 which indicates good reliability when the model predicts a business is open, a recall of 0.93 which is considered high and indicates that the model is effective at detecting open businesses, a precision for closed businesses at 0.34 which indicates that model struggles to accurately identify closed businesses without a significant number of false positives, and a recall for closed businesses at 0.34 which indicates a high number of false negatives. The F1 score for open businesses resulted in 0.88 which indicates that there is a balanced performance between precision and recall for open businesses. However, the F1 score for closed businesses is low considering a poor performance and the model failing to balance the precision and recall effectively for closed businesses. Moreover, the ROC-AUC score of 0.63 suggests that the model's ability to distinguish between open and closed businesses is moderate.

In Figure 7, we created a confusion matrix that highlights the significant issue with false positives as many closed businesses are incorrectly predicted. However, the amount of true positives and lower false negatives show how effective the model is in identifying open businesses rather than closed. In Figure 8, it can be observed that the area under the PR curve is 0.88 which in general is a high precision across all levels of recall. This means that our model is reliable in its positive predictions and is close to an ideal of 1.0. Furthermore, in Figure 6, log_review and is_open along with FoodIndustry and OpenDaysCount both have positive correlations where businesses with more reviews are more likely to open and the food industry tends to be open more days in a week. However, stars and FoodIndustry have a slight negative correlation which suggests that lower ratings are found in the food industry than others. Thus, our model demonstrates a robust predictive performance for open businesses.

Significance

The findings provide actionable strategies for businesses to utilize online reviews to their advantage allowing them to increase their chances of success in the long-run and ensuring profitability. For example, the use of our random forest algorithm revealed that features such as long-transformed review count and city clusters are important predictors of how successful and viable a business is as it resulted with the highest Gini importance of 0.5743. This metric underscores the predictive power of customer engagement levels. Overall, these results indicate that businesses with higher review volumes and those located within certain cities as city clusters were another significant predictor at a Gini importance of 0.2024 tend to have a higher probability of survival. These insights can guide business owners to focus on boosting their social media and consider strategies such as relocation or targeted marketing campaigns. To add on, the substantial role of consumer reviews highlights the influence and responsibility of consumers in impacting the businesses' outcomes which could drive more engagement on review platforms. The study's results have significant implications for business owners, consumers, policymakers, and the research community. Methodically, the study advances our understanding of which factors influence business successes and demonstrates how consumer feedback is shaping businesses. This study has deepened our understanding of the factors influencing business success in the digital age and sets the stage for learning more about online consumer behavior and business performance.

Given the performance metrics, the model's results would be considered as strong for predicting open businesses but weak for predicting closed businesses. We can note that the confidence is high for open businesses as seen with the high recall of 0.93 which indicates the model being very effective. Moreover, the precision-recall curve's area under the curve of 0.88 supports this strong confidence. However, there is a weak confidence for closed businesses as seen with the low precision indicating that the model incorrectly labels businesses that are closed along with the low recall. Given these two factors, the confidence in the model's ability to predict closures is weak. We would be willing to use such models in production to change how the company or enterprise makes decisions as it is helpful in giving significant insights in how they can best improve their business. Given the accuracy and performance metrics, our model is reliable and predicts for open businesses very well. However, it is important to note the inaccuracy of predictions for closed businesses. Moreover, there seems to be a bias toward open businesses indicated by the high recall and low precision for closed businesses. Before the model is used, we would need to improve the accuracy of closed business predictions or consider adding more features.

Model Fairness

To evaluate the model's fairness, it was seen that the equal of opportunity for open businesses (the True Positive Rate) is *0.93*, the equal of opportunity for closed businesses (true positive rate) is *0.15*, the predictive equality for open businesses (false positive rate) is *0.85*, and the predictive equality for closed businesses (false positive rate) is *0.07*. As a result, the true positive ratio is 0.1613, the false positive ratio is *0.0824*, and the equalized odds ratio is *0.0824*. Because the true positive rate for open businesses is much higher at *0.93* than for closed businesses at *0.15*, the model is significantly better at identifying open businesses correctly than it is at identifying closed ones. For predictive equality, the false positive rate for open businesses is very high at *0.85* which suggests that the model frequently incorrectly predicts businesses that are open, but the false positive rate for closed businesses is considered very low at *0.07* which indicates that the model doesn't frequently incorrectly predict businesses as closed. Based on these metrics, the model appears to be unfair as the difference between the true positive rate and false positive rate between the two classes suggest a bias towards predicting businesses that are open. The equalized odds ratio is considered significantly lower than the acceptable range for a model to be considered fair. The model unfairness could be contributed by the class imbalance in the training data or the model parameters and features potentially not adequately capturing the characteristics of closed businesses. Overall, fairness is an important consideration for our model, so revising the feature selection until there is less of a bias towards open businesses would be beneficial.

This study does not have the possibility of producing a Weapon of Math Destruction as our predictions do not have a direct influence on business decisions that could create harmful feedback loops. These predictions are intended not for the public but for Yelp to help businesses. The results of our predictions have the possibility to provide targeted support to businesses considered at risk including offering grants or sponsorship content, for example. Hence, the predictions wouldn't harm anyone as it is only used to determine which businesses are open and closed, factors such as cities and ratings being associated with it, and how to help create profitable strategies for those businesses to use more online reviews, increase success, and customer satisfaction.

Links and References

- Github link: https://github.com/AlyssaYam/ORIE4741-Project
- Yelp Dataset derived from: https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset/data
- Chen, Chao, and Leo Breiman. "Using Random Forest to Learn Imbalanced Data." *ResearchGate*, Jan. 2004, www.researchgate.net/publication/254196943_Using_Random_Forest_to_Learn_Imbalanced_Data.
- Ribeiro-Soriano, D. (2017). Small business and entrepreneurship: their role in economic and social development. Entrepreneurship & Regional Development, 29(1–2), 1–3. https://doi.org/10.1080/08985626.2016.1255438
- Zhang, Ziqiong, et al. "The Impact of E-Word-of-Mouth on the Online Popularity of Restaurants: A Comparison of Consumer Reviews and Editor Reviews." ResearchGate,

International Journal of Hospitality Management, Dec. 2010, www.researchgate.net/publication/223407272_The_impact_of_e-word-of-mouth_on_the_online_popularity_of_restaurants_A_comparison_of_consumer_reviews_and_editor_reviews.

Appendix

Tables and Figures

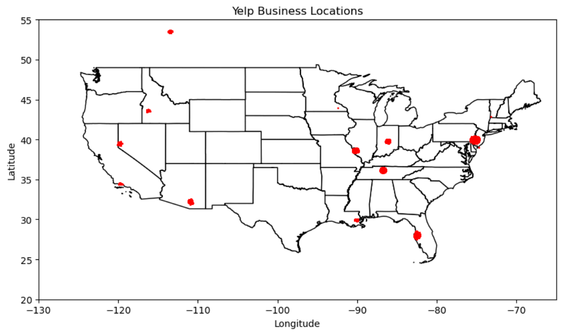

Figure 1: Each point represents the location of a business captured in the dataset.

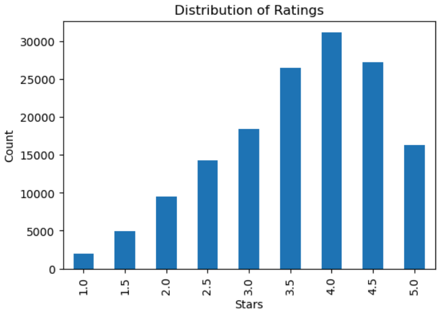

Figure 2: A business-level bar plot of average star ratings.

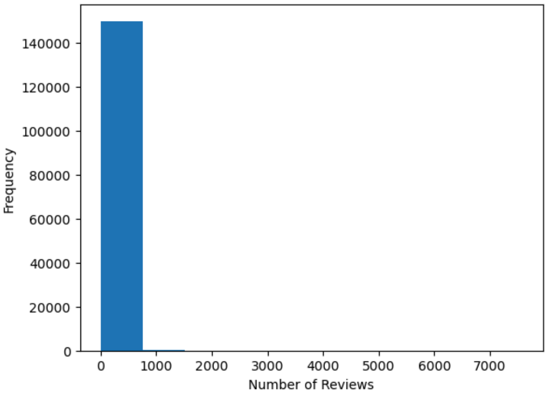

Figure 3a: A histogram representing the number of reviews each business has. Outliers obscure the underlying distribution.

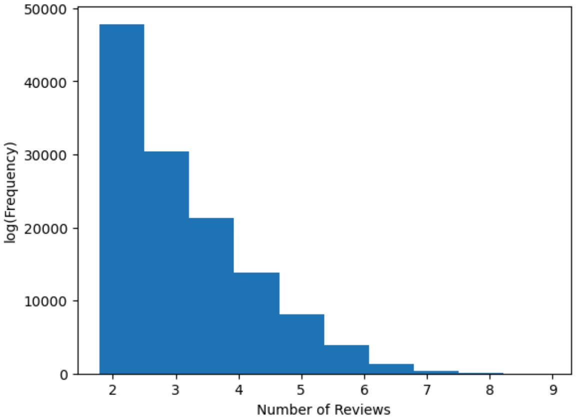

Figure 3b: The log transformation of Figure 3a reveals the underlying distribution by reducing the impact of outliers.

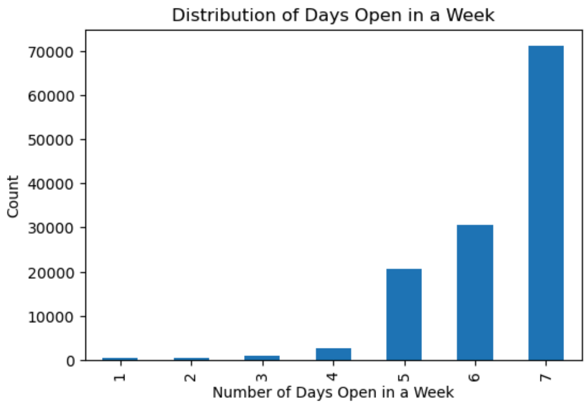

Figure 4: Each bar corresponds to a different number of days open while the y-axis represents the number of businesses belonging to each category.

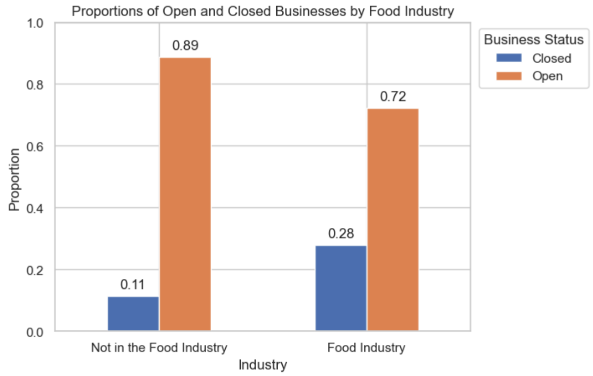

Figure 5: A grouped bar chart illustrating the proportion of closures of businesses within the food versus non-food industries.

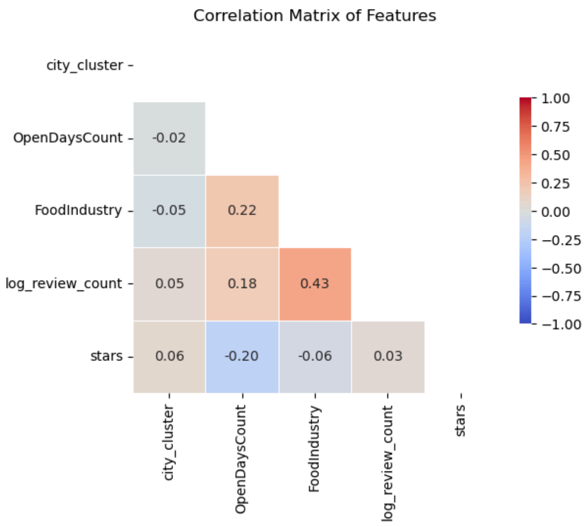

Figure 6: A correlation matrix of the features used in our random forest algorithm.

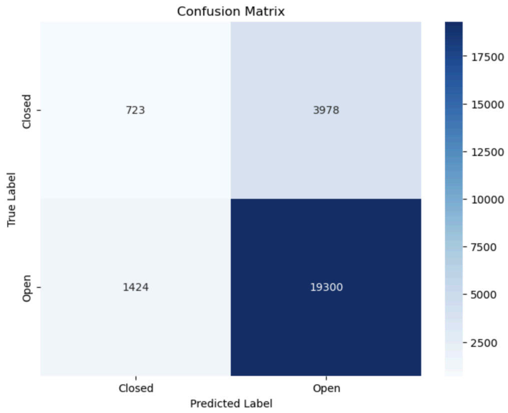

Figure 7: A confusion matrix of the random forest algorithm's performance, depicting actual versus predicted classifications.

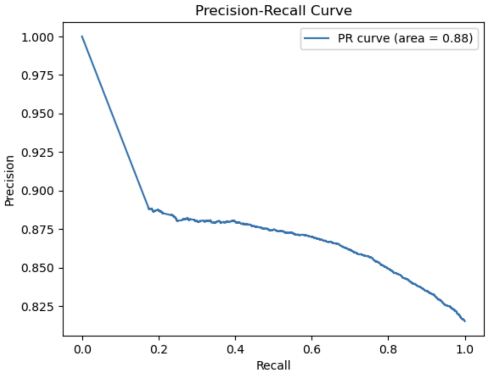

Figure 8: A precision-recall curve for the random forest model.

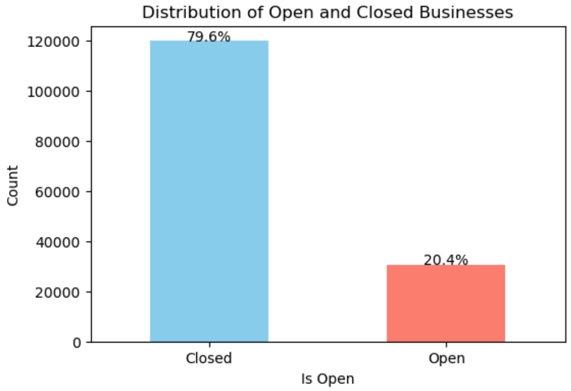

Figure 9: A bar chart depicting the makeup of the binary response variable.

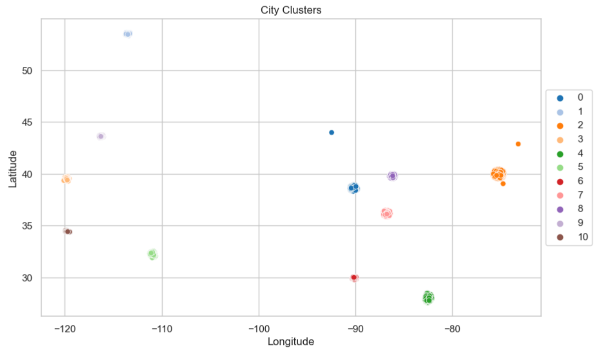

Figure 10: K-Means 11 cluster-visualization using each business's latitude and longitude data

Table 1: Incidence of null values

Feature	**Percent Missing from the Dataset**
business_id	0.000%
name	0.000%
address	0.000%
city	0.000%
state	0.000%
postal_code	0.000%
latitude	0.000%
longitude	0.000%
stars	0.000%
review_count	0.000%
is_open	0.000%
attributes	9.142%
categories	0.069%
hours	15.446%

Table 2: City Cluster Mappings

Cluster Label	**City**	**State**
0	St. Louis	MO

1	Edmonton	AB
2	Philadelphia	PA
3	Reno	NV
4	Tampa	FL
5	Tucson	AZ
6	New Orleans	LA
7	Nashville	TN
8	Indianapolis	IN
9	Boise	ID
10	Santa Barbara	CA

Member Contributions

Lincy Chen

Code: Data cleaning, exploratory data analysis, data visualizations, feature engineering, model building

Report: Introduction, Research Question, Data Description, Approach, References, Table 1-3 and Figures 1-5, 9, 10

Alyssa Yam

Code: Performance metrics, confusion matrix, correlation matrix

Report: Results, Significance, Model Fairness, Table 4, Figures 6-8