Analyzing Yelp Dataset to Predict Restaurant Closure

Final Report

MSBA 5406.31082 - Advanced Applied Analytics

Abidemi Olaoye

Ben Soumahoro

Oluwapelumi Osunrayi

Omotola Adeoye

30th of July, 2020

TABLE OF CONTENTS

EXECUTIVE SUMMARY

People in the United States are spending more money dining at restaurants than ever before, with foodservice sales totaling $770 billion in 2019, a 4.4% increase from 2018 (Stribling, 2020) and projected at $899 billion in 2020 (National Restaurant Association, 2020). This shows the restaurant industry's importance in our society and, as a result, competition is high. Any restaurant that does not gain competitive advantage or adopt strategies to keep the business afloat may suffer a loss of customers to their competitors, which could lead to a decline in revenue and closure (DiPietro, 2016; Gagić et al., 2013).

We acknowledge that restaurants can also fail (close) or succeed (stay open) for a variety of reasons, including food quality, bankruptcy, economic factors, and restaurant environment (DiPietro, 2017). However, we need to analyze other internal factors that influence closure. This is where the Yelp Dataset comes in. This project aims to analyze the Yelp 2019 Dataset to predict the likelihood of restaurant closure based on different attributes. Attributes in the dataset include but are not limited to parking, table service, Wi-Fi, ambiance, price range, wheelchair accessibility, acceptance of credit cards or cash only service, categories (Fast Food, Italian, Mexican, Chinese, American etc.), family accommodations, star ratings, reviews and more. These attributes point to the demand for intangible and tangible restaurant experience by customers. The importance of these experiences and how they affect restaurant or failure will be highlighted through a literature review. After this, attributes that show major significance to closure will be identified with predictive models.

Findings from the analysis show Selection Tree with High-Performance Support Vector Machine as the best model with 81.24% accuracy. The five most important predictors of restaurant closure, as indicated by the Selection Tree in this model, are Review Counts, Chain

Counts, Entertainment, Is Chain (indicating if the restaurant is a chain or not), and Good for Dinner. Therefore, this study recommends restaurants at high risk of closure should provide entertainment such as background music or improve the existing ones, and promotions could be put in place to encourage customers to leave reviews on Yelp after visiting the restaurant. Also, restaurants that experience low traffic at dinner time and are considered "not good for dinner" can tailor their menu options to suit customer needs.

CHAPTER ONE

INTRODUCTION

1.0 COMPANY BACKGROUND

Yelp is a business directory service and crowd-sourced review forum, headquartered in San Francisco, California. It was founded in 2004 by former PayPal employees Russel Simmons and Jeremy Stoppelman, to help people find local businesses like hairstylists, mechanics, dentists, etc. Yelp is so popular among users that between 2009 and 2012, they expanded throughout Europe and Asia (Yelp, 2020). According to Yelp's business website, they had a monthly average of 35 million unique users who visited their mobile app and 178 million unique users across all platforms in the first quarter of 2020. Presently, they have more than 211 million written reviews on their website.

1.1 STATEMENT OF PROBLEM

Every year, Yelp releases an all-purpose dataset for learning in the form of a challenge or competition. This dataset is a subset of their businesses, reviews, and user data for personal, educational, and academic purposes. Having worked with the 2018 Yelp "business" dataset in the past, we discovered that it consists of 33,110 restaurants out of which only 23,118 are open, meaning that 9,992 are marked as closed. Yelp 2019 data also consists of 35,305 restaurants, out of which only 23,867 were open, meaning that 11,438 were closed. As a result, we saw a significant increase in both open restaurants and restaurant closures from the previous year. Our investigations showed that the restaurant industry sees a 1% - 1.2% yearly increase in restaurants, which gives rise to competition (McLynn, 2018), and those that cannot compete

close each year. This also caused us to wonder why and if any of the variables (restaurant attributes) in the dataset are related to said closure. Chang (2013) states that "restaurants do not sell merely food; they also have to sell an experience" (p. 536). Attributes such as ambiance, whether customers were able to find parking or not, Wi-Fi, and others present in our dataset, contribute to the said restaurant experience.

While researching connections between attributes and closure, we found that these attributes are usually mediated by customer perception, satisfaction, and behavioral intentions (Dutta & Venkatesh, 2007; Tripathi & Dave, 2016). If negative, these mediators can reduce revenue and significantly increase expenditure to attract new customers and eventually lead to failure (Ozdemir & Hewett, 2010 as cited in Tripathi & Dave, 2016). From this statement, we deduce that restaurateurs who do not take the statistical significance or correlation between attributes and restaurant closure into consideration by tailoring their dining experiences accordingly could ultimately face closure. Therefore, answers to questions such as these must be understood: What important attributes contribute to the continued operation or shutdown of a restaurant? How do these attributes impact the shutdown of a restaurant?

CHAPTER TWO

LITERATURE REVIEW

As a result of the continuous increase in restaurants, the number of research articles regarding the restaurant industry and its components of operations, service quality, trends, and finances has substantially increased over the past 30 years (DiPietro, 2016). Our literature review addresses restaurant closures by highlighting various factors that influence restaurant closure. This will help us as we proceed in the understanding and analysis of our restaurant industry data.

A majority of researchers have linked restaurant closure or failure to factors such as poor credit management or arrangement, personal use of business funds, insufficient capital, competition, low sales and bankruptcy, slow economic activity, loss of customers, theft, etc. (Assaf et al., 2010; Mao, 2006; Parsa et al., 2005). A group of researchers divided these factors into three groups for easy study: economic, marketing, and managerial perspectives (Parsa et al. 2005). Economic perspective includes "restaurants that failed for economic reasons such as decreased profits from diminished revenues," voluntary and involuntary bankruptcies, foreclosures, etc. (Parsa et al. 2005, pg. 305). Marketing perspective includes restaurants that close at a "specified location for marketing reasons such as deliberate strategic choice of repositioning, adapting to changing demographics, accommodating the unrealized demand for new services and products," etc. (Parsa et al. 2005, pg. 305). Managerial perspective consists of failures resulting from managerial incompetence and limitations such as human resource issues, loss of motivation by owners, technological and environmental changes, etc. (Parsa et al. 2005).

We found that in general researchers give more focus to financial, economic, or external factors. According to Parsa et al. (2005), internal factors such as restaurant concept, experience,

type of operation, food, and service quality are also important determinants of successful restaurants. These internal factors produce higher levels of guest satisfaction and increased return intention, leading to higher sales revenue to avoid financial distress and keep the restaurant open (Gagić et al., 2013; Parsa et al., 2005). However, there is little research acknowledging the direct relationship of these internal factors to restaurant failure; instead, they are usually mediated by customer perception, satisfaction, and behavioral intentions (Dutta & Venkatesh, 2007; Gagić et al., (2013); Tripathi & Dave, 2016). It causes us to believe that service quality and experience alone have less impact on restaurant failure but are crucial determinants of customer satisfaction and future behavioral intentions (Namkung & Jang, (2010) as cited in Gagić et al., (2013). These behavioral intentions are directly related to the profitability of organizations (Luo & Humburg (2007) as cited in Gagić et al., (2013). At the same time, profitability acts as a precondition for surviving restaurant market conditions and achieving restaurant success (Luo & Humburg (2007) as cited in Gagić et al., (2013). This shows an indirect relationship to failure, whereby the consequences of poor customer satisfaction and negative future behavioral intentions reduce revenue but significantly increase expenditure to attract new customers and can eventually lead to failure. These negative future behavioral intentions include an unwillingness to recommend the restaurant, engage in positive word of mouth, and unwillingness to return (Ozdemir & Hewett, 2010 as cited in Tripathi & Dave, 2016).

Service quality and experience are evaluated by key features of restaurants such as environment, food, employee services, etc. and many researchers have tried to determine the most important ones in relation to restaurant success (Chow et al., 2007; Namkung & Jang, 2008 as cited in Gagić et al., 2013). Through extensive literature review, DiPietro (2016) concluded that experiences such as background music, seating arrangements, and interior design/ambiance

are statistically significant predictors of satisfaction and repeat patronage, which contributes to restaurant success. Kong et al., (n.d.) used the Yelp dataset to identify the key features customers look for during their dining experience by looking at each feature's impact on restaurant star ratings. They used models such as Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Tree with Random Forest Model, and Gaussian Discriminant Analysis (GDA) to analyze said features. Through their GDA model which has the highest accuracy at 55.49%, results showed that availability of street parking, ability to make reservations, review count, casual ambiance, noise level, and attire are the six most important features (closely related to high restaurant ratings) in restaurants located in the U.S., U.K., Canada, and Germany. Especially, for the U.S., divey ambiance, the existence of parking lots, and parking valets are the most important.

Similarly, Shellenberger (2017) used correlation analysis, independent sample t-tests, and multiple regression analyses to find several attributes like Wi-Fi, parking, ambiance, and other attributes to be positively correlated with the star rating of a restaurant. Attributes play an essential role in customer satisfaction, which affects restaurant closure, as revealed in our literature review. Despite this, the author's focus was only on finding the best analytical model, so the results were not applicable and could not be related to any literature review.

One paper that provides recommendations is Snow (2018), who advised that chain restaurants that are willing or able to operate in different locations should do so to avoid closure. He went further to advise against over diversifying their menu, recommending that breakfast restaurants should focus on just breakfast, American restaurants on American food, and so on. He also addressed competition highlighted earlier in our literature review and advised that restaurants in low-density areas are more likely to survive because of less competition in the area

(Snow, 2018). However, his analysis of the yelp dataset is different from ours because he had different independent variables like the sum of comments, the gender of commenters, compliment to reviewer ratio, etc. Irrespective of this difference, Snow highlighted the importance of recommendations and explained that it helps restaurants know the exact attributes (parking, serving alcohol, location, etc.) to develop or discard, thereby preventing closure. It could also aid performing restaurants to improve on the attributes they already provide. Lian et al., (2017) used the Yelp dataset for China to determine potential indicators for the long-term survival of restaurants with a focus on Beijing, Shanghai, and Guangzhou. They used Logistic Regression, Gradient Boosted Decision Tree (GBDT), and Support Vector Machine (SVM) to examine predictors. They found that density, competitiveness, and peer popularity are the most important predictors. Determined by Area Under the ROC Curve (AUC) performance, GBDT was the best model for the three cities.

Earlier in the literature review, we established a yearly increase in the number of restaurants. This yearly increase gives rise to constant competition among restaurants, which is also a factor that affects closure. Parsa et al. (2015) listed competition or concentration of competition as one of three key factors that contribute to restaurant failure. They found that failure rates are higher in downtown areas where there is a high concentration of restaurants. To survive in these conditions, restaurateurs need to pay attention to their customers' preferences as these will help provide better and consistent customer experience for competitive advantage (Ching & Bulos, 2019; DiPietro, 2016). Parsa et al. (2015) also found a correlation between ethnicity, gender, and business failure. However, it was backed up by outdated articles from the 80's and 90's which we believe may no longer be accurate based on the advancements of women

and other ethnicities in the restaurant industry today with the likes of Chipotle, Taco Bell, Qdoba, and others.

In conclusion, the quality of experience and service also influences a consumer's review or rating. In his work, Luca (2016) aimed to understand the relationship between consumer reviews and restaurant demand. He concluded that a one-star increase in Yelp rating leads to a 5 - 9 percent increase in revenue for independent restaurants. According to him, this effect is driven by independent restaurants because ratings do not affect restaurants with chain affiliation. Lu et al. (2018), Mao (2016), and Parsa et al. (2005) found large chain restaurants as a feature to be strongly correlated to success. Perhaps chain restaurants could be included as a feature in our analysis to test this theory. The variables in our dataset, such as entertainment, noise level, parking, reservations, etc. are part of a restaurant's service and experience. Variables such as city, postal codes, and state fall under location. We have shown these to be important through the literature review. Therefore, our research and analysis of these variables will help us build predictive models to understand better the correlation between restaurant features and their statistical significance or insignificance related to restaurant failure or success.

Table 1: Literature Review Table

S.N	Author/Title	Data Source / Country of Origin	Number of Samples	Timeline of Data	Subject/ Variables	Data collection and analytical methods	Important Findings Related to your Project
1	Assaf et al., (2011) Evaluating the Performance and Scale Characteristics of the Australian Restaurant Industry	Australia	150 restaurants	2007 - 2008	Restaurant industry, number of full-time equivalent employees, food expenses, beverage expenses, number of seats, total food sales and total beverage sales	efficiency data envelopment analysis (DEA) double bootstrap	Restaurants were operating at a high degree of inefficiency and need to expand production to reach optimum scale of production. Large restaurants in the set were more efficient that small restaurants. Economic trends and competition affect restaurant efficiency. Restaurants need to adopt strategies to increase their levels of operating efficiency if they are to have a viable future in increasingly uncertain and competitive external environment. Inconsistent standards of service delivery contribute to decline in profit margins.
2	Ching & Bulos, (2019) Improving Restaurants'	France	5 Fast	2017	Reviews	Time Series Forecasting using Linear Regression in Waikato Environment for Knowledge Analysis	It is important to consider the customer feedback that is posted online because it contains vital information on what their customers' preferences and experiences

	Business Performance Using Yelp Data Sets through Sentiment Analysis		food chain restaurants			(Weka) machine learning workbench,	and knowing these will help provide the best and consistent customer experience, which will result to increase in profit. However, the use of linear regression in time series forecasting is not very reliable because of limited text reviews in a day.
3	DiPietro, (2016) Restaurant and foodservice research	N/A	160 – 170 Papers	2000 - 2016	Hospitality management, Restaurants, Food and beverage, Foodservice research	Literature Review	Findings echoed others in that food quality and service quality were important, but they also found that the service scape including seating arrangements, background music and fascinating interior design, all helped with creating an environment that was related to customers being highly satisfied. This study encouraged more studies on service scape and environment of the dining experience (Kim and Moon, 2009; Lin and Mattila, 2010; Ryu and Jang, 2008). Ladhari et al. (2008) also assessed dining satisfaction and found that positive emotions and perceived service quality predicted dining satisfaction. Other statistically significant predictors of satisfaction were menu presentation, furnishings in the restaurant, as well as the music being played in the dining environment. Restaurants and service organizations need to create an experience to have a competitive

							advantage over their competition.
4	Dutta et al., 2007 Service failure and recovery strategies in the restaurant sector: An Indo-US comparative study	U.S./India	200	N/A	Restaurants, Customer service management, Customer satisfaction, Consumer Behavior	Literature Review	Failure to deliver service as per customer expectations creates depredation in the customer's psychology which, if left unattended, can ring death knells for the organization. Gro¨nroos (1988) says, customers realize and anticipate that whenever something goes wrong or something unpredictable happens the service provider would immediately and actively take action to control the situation and find a new, acceptable solution. Thus, if the customers feel that the recovery strategies are given importance, they are bound to have a better perception of the organization. Keaveney (1995) reported that if organizations don't adopt recovery strategies it can lead to customer switching over to another service provider.
5	Feng et al. (2015) Determining Restaurant Success or Failure	Yelp.com / U.S.	21,892	2015	Restaurant attributes, Restaurant features, Number of reviews, Ratings	Yelp Academic Dataset. Across cities in the U.K., U.S, Canada and Germany. The study utilized Stochastic Gradient Descent and Back Propagation, Neural Networks, Logistic Regression and Support Vector	For every missing feature in the first study, they replaced them with the midway value between the possible outcomes. For example, Price Range, which has possible values of 1, 2, 3, and 4, would be replaced with 2.5. For the second study, they used every single restaurant, but replaced missing features with the average value for that feature. For example, Price Range would be set

						Machines.	to 1.8 which is the average value across all restaurants. The use of dummy variables also proved helpful. Without dummy variables, the neural network had a 67% success rating. This difference in performance is probably not due to the dummy variables themselves, but due to the fact that as a result of dummy variables, they could use all restaurants. With dummy variables, their neural network outperformed all benchmarks for all three methods which outperformed Logistic Regression and SVM by around 7-8% across the board. Their best neural network classifies restaurants with 83.3% accuracy, while the best performing Logistic Regression had 76.4% accuracy, and SVM had 76.4% accuracy.
6	Gagić et al. (2013) The vital components of restaurant quality that affect guest satisfaction	Serbia	N/A	1987, 1996-1997, 2000-2012	Restaurant; quality; satisfaction; guests	Literature Review	In an increasingly competitive environment, restaurants must be focused on guests using marketing concepts that identify their needs thus leading to their satisfaction and increased retention. Service quality is fundamental component which produce higher levels of guest satisfaction, which in turn lead to higher sales revenue. The main purpose of this study was to examine the quality dimensions that affect guest satisfaction in restaurant

							industry. Food and beverage quality, the quality of service delivery, physical environment and price fairness are analyzed as key components of restaurant experience.
7	Jin & Leslie, (2009) Reputational Incentives for Restaurant Hygiene	U.S.	N/A	N/A	Experience, importance of resources like Yelp and how it influences customer purchase.	Research, Literature Review	How can consumers be assured that firms will endeavor to provide good quality when quality is unobservable prior to purchase? Consider the example of product safety. It is costly for firms to maintain safety, and if they don't, the risk that something will go wrong may be small. As long as nothing goes wrong, consumers will generally never know if the firm exerted appropriate effort. But of course, the cost to consumers in the event of a problem can be severe. In a reputation mechanism, consumers may not observe product quality before making a purchase, but they learn from experience and form beliefs about product quality.
8	Kong et al. (2016) Predicting International Restaurant Success with Yelp	Yelp.com / U.S.	25,071 restaurants from four different countries: the United States, the United Kingdom,	2015	Restaurant attributes, Ratings	1) Mean imputation to fill in missing data values. 2) Two different modes of restaurant classification: binary and multiclass. In the binary case, restaurants with a star rating below 4.0 are classified as 0,	GDA was the best-performing model and the multi-class decision tree performs the worst. Other features corresponding to high star rating include outdoor seating, classy ambience, touristy ambience, waiter service, hipster ambience, garage parking, trendy ambience, Wi-Fi, intimate ambience, good for kids, good for groups, allows

			Canada, and Germany.			and restaurants with a star rating of 4.0 and above are classified as 1. In the multi-class case, restaurants are classified from 0 to 5 based on the integer value their star rating. 3) Data classification using models such as Naive Bayes, support vector machines (SVM), decision trees, logistic regression, and Gaussian Discriminant Analysis (GDA) to evaluate the strength of the feature sets we selected. 4) Univariate feature selection with a chi-square scoring functions to choose the most important features.	smoking, and has T.V. In addition, in the North America region, customer satisfaction is positively influenced by the existence of parking lots and parking valet services. We speculated that parking is more important in the U.S. and Canada due to a higher percentage of drivers, whereas in Europe, public transportation is more popular.
9	Lian et al (2017) Restaurant Survival Analysis with Heterogeneous Information	Yelp for China	144,134	2012	Check-ins, reviews, cities, shops	Logistics regression, Gradient Boosted Decision Tree (GBDT), Support Vector Machine (SVM)	Density, competitiveness, and peer popularity are the most important predictors of long-term survival of restaurants
10	Luca, (2016) Reviews,	Yelp Dataset	3,582	2003-2009	Revenue, Ratings	Regression	The impact of consumer reviews on the restaurant industry: (1) a one-star

	Reputation, and Revenue: The Case of Yelp.com	and Washington state department of Revenue/ U.S.					increase in Yelp rating leads to a 5-9 percent increase in revenue, (2) this effect is driven by independent restaurants; ratings do not affect restaurants with chain affiliation, and (3) chain restaurants have declined in market share as Yelp penetration has increased. (4) Consumers do not use all available information and are more responsive to quality changes that are more visible and (5) consumers respond more strongly when a rating contains more information. Consumer response to a restaurant's average rating is affected by the number of reviews.
11	Lu et al. (2018) Should I Invest it? Predicting Future Success of Yelp Restaurants	Yelp.com / U.S.	2014	2016 & 2017	Attributes, Ratings, Reviews	Logistic Regression, Feature ablation study, Unigram and Bigram feature analysis to calculate frequency of positive words.	The balanced accuracy is 67.46%. The result shows that text features failed to have significant indications for the future success of the restaurant, while non-text features, especially business features, do have strong correlation with future restaurant performance. Non-text features are more important in the model. Chain restaurant feature turned out to be the most significant one, and other features, such as trends, nearby comparison, and economic status all have their own influence on the whole model. However, the performance of text features was not so good as expected because word patterns were not analyzed effectively.

12	Mao, Z. (2006) Investigation of the relationship between firm - wise financial factors and firm performance in the hospitality industry	U.S.	256 observations. 56 Restaurants	2000-2004	Dividend Payout, Business Diversification, Geographical Diversification, Liquidity, Solvency, Activity, Growth, Profitability, Size	Regression Analysis, Descriptive Statistics	Liquidity, activity ratio, sales growth, profitability demonstrated significant and positive relations with firm performance. size was found to be a significant positive contributor to restaurant performance. Larger restaurants performed better.
13	Parikh et al., (2014) Motives for reading and articulating user-generated restaurant reviews on Yelp.com	Literature Review, Survey, Yelp Users / United Kingdom	A total of 72 responses were received. The low sample size was likely due in part to Yelp.com blocking additional requests for participation within the study after data	2013	Yelp Users	Literature review was used to highlight the importance of user generated reviews. Then a survey was used to closely examine how customers use user-generated reviews in choosing restaurants and did so by addressing the research question: "How often do USA Yelp.com users seek user-generated restaurant reviews and what factors motivate consumers to seek and contribute reviews?". Analysis included backward	The results indicate that Yelp.com users primarily engage with the Web site for socializing (community membership) and information seeking (finding good restaurants). Therefore, restaurant managers must pay attention to their reviews and ratings on Yelp.com, because customers not only trust such reviews, but they are an important determinant in whether a person visits a restaurant or not. Positive reviews on Yelp.com to help create a positive impression among potential consumers.

			collection had already begun			stepwise linear regressions to examine correlation between variables and Wilcoxon–Mann–Whitney, a non-parametric test, to compare the two groups of categorical data.	
14	Parsa et al., (2005) Why Restaurants Fail	Bankruptcy Filings, Health departmnent Data / U.S. A	2,439	1996-1999	Restaurant ownership turnover	Quantitative and Descriptive statistics	Restaurant failures can be studied from economic, managerial and marketing perspectives. Food quality, Firm size, Location, Staff and Employer training and personality and Chain Restaurants are strongly correlated with business success and failure.
15	Parsa et al., (2015) Why Restaurants Fail? Part IV: The Relationship between Restaurant Failures and Demographic Factors	U.S. Census data for Boulder, Colorado, for 2000 and 2010, and health departmen t records from the Boulder County	118,400, distributed among 5 ZIP codes. 496 restaurant inspection data	2015	Business failure; Boulder; Colorado; restaurants; bankruptcy; insolvency; demographic factors	Descriptive Statistics, Visualization	Location affects restaurant failure. This variable is worth studying critically.

		Health Department					
16	Shellenberger, (2017) Predicting Whether Business is Open or Closed and Suggesting the Good Business Practices	Yelp.com / U.S.	2,265	2016	Restaurant Attributes, Reviews	IBM SPSS: correlation analysis, independent sample t-tests, and multiple regression analyses	IsRomantic, IsIntimate, IsClassy, IsHipster, IsDivey, IsTrendy, IsUpscale, StreetParking, HasParkingLot, HasValet, DoesCater, HasDessert, HasLunch, HasDinner, HasBrunch, AllowsReservations, WheelchairAccessible, NoAlcohol are positively correlated with the star rating.
17	Snow, (2018). Predicting Restaurant Facility Closures	Yelp Dataset / U.S.	36,544	2016-2017	Reviews, Star ratings, gender, longitude	Machine Learning, Applied, Firm A.I., Restaurant, Bankruptcy, Failure, Closures	The study is prescriptive and allows for effective allocation of resources. Knowledge of which restaurants are most likely to close could help management to 1) identify struggling facilities to provide additional assistance to 2) or to identify which facilities to let go of. The model can also be extended to predict many years in advance to assist management to intervene long before the predicted closure.
18	Tripathi & Dave, (2016) Assessing the impact of restaurant	New Delhi	549	2016	Quality of service; Restaurants; Food; Customer	Factor analysis. Structural Equation Modeling (SEM) technique, Questionnaires	A rating prediction model is proposed by combining three factors: user sentiment, user topic similarity, and interpersonal influence. LDA + word2vector model was used to mine

	service quality dimensions on customer satisfaction and behavioral intentions				services; Competitive advantage; Customer satisfaction; Food quality		user interest, which is effective to improve the performance.

CHAPTER THREE

DATA UNDERSTANDING AND PREPARATION

This project's data was initially obtained from the Yelp 2019 dataset challenge and included a JSON data file named "business" with records of 209,393 businesses and 15 variables. The following are steps detailing the essential parts of our data cleaning process:

- First, we converted the file to CSV format and imported it into Juypter Notebook where most of the cleaning was done using Python codes.
- Irrelevant columns such as “Longitude," "Latitude," and “Hours” were dropped.
- While there were no null values in the columns we needed for analysis, we had nested dictionaries variables. We had to remove foreign symbols such as commas and brackets from them to extract the needed information (see Appendix for data cleaning codes).
- After removing foreign symbols and converting to lowercase, we used the "Names" column and the once nested "Categories" and "Attributes" columns to create multiple binary and interval variables representing services and experiences offered at a restaurant. The criteria for creating these variables are that they are beneficial in analyzing the characteristics of open restaurants based on our literature review. For example, we found that "Entertainment" is significant in determining the success of a restaurant; therefore, it was extracted from the nested "Attribute" column, converted to a column of its own, and assigned a 1 for every restaurant that has entertainment. This was done for Parking, Price Range, Delivery, and other variables listed in Table 1 below.
- Using "Categories," we created a binary column called "Restaurant" by assigning a 1 to every business that was categorized as "Restaurant". This allowed us to eventually filter

out any row that was not a restaurant business. We also created an "Ethnicity" column to account for each restaurant's ethnicity. Levels in this variable include American, Chinese, Japanese, Mexican, and Others.

After the steps above, the resulting dataset, which will be used for modeling in SAS Enterprise Miner, is now named "Restaurant" with records of 33 columns and 35,305 restaurants, out of which 23,867 are open, and 11,438 are closed. These restaurants are distributed throughout Arizona, Nevada, North Carolina, Ohio, and Pennsylvania. All variables in our Restaurant dataset are listed and described in the data dictionary below. Variables with the role of "Rejected" will not be used in SAS Enterprise Miner, while those with data source "Extracted" are the dummy variables we created.

Table 2: Data Dictionary (including roles and measurement levels)

Business File				
Column Name	**Description**	**Role**	**Measurement Level**	**Data Source**
Address	Displays street addresses of businesses.	Rejected	Nominal	Yelp Dataset Challenge
Alcohol	Dichotomous indicator of a restaurant that offers alcohol as an attribute (1=Yes, 0=No).	Input	Binary	Extracted
Attributes	Nested column describing certain amenities and services available at each business. For example; parking availability, Table reservations, price range, kid friendly meals, ambience (outdoor seating), availability of alcohol, acceptance of credit cards etc.	Rejected	Nominal	Yelp Dataset Challenge
Business ID	Displays a unique id for each business.	I.D.	Nominal	Yelp Dataset Challenge
Categories	Mentions the specific services that each business provides,	Rejected	Nominal	Yelp Dataset Challenge

	such as Restaurants, Fast Food, Pizza, Mexican, etc.			
City	Lists the city where the business is located.	Rejected	Nominal	Yelp Dataset Challenge
Credit_ card	Dichotomous indicator of a restaurant that has credit card services as an attribute (1=Yes, 0=No).	Input	Binary	Extracted
Entertainment	Dichotomous indicator of a restaurant that has entertainment services (such as Background Music, Live Music, Jukebox and karaoke) as an attribute (1=Yes, O=No).	Input	Binary	Extracted
Good_for_ breakfast	Dichotomous indicator of a restaurant that is "good for breakfast" as an attribute (1=Yes, 0=No).	Input	Binary	Extracted
Good_for_ dinner	Dichotomous indicator of a restaurant that is "good for dinner" as an attribute (1=Yes, 0=No).	Input	Binary	Extracted
Good_for_ lunch	Dichotomous indicator of a restaurant that is "good for lunch" as an attribute (1=Yes, 0=No).	Input	Binary	Extracted
Happyhour	Dichotomous indicator of a restaurant that includes "happy hour" as an attribute (1=Yes: 0=No).	Input	Binary	Extracted
Delivery	Dichotomous indicator of a restaurant that includes "delivery" as an attribute (1=Yes, 0=No).	Input	Binary	Extracted
FastFood	Dichotomous indicator of a restaurant that includes "fast food" as an attribute (1=Yes, 0=No).	Input	Binary	Extracted
Chain_ Counts	Count of all restaurants that are a part of the same franchise and have the same name.	Input	Interval	Extracted
Is_Chain	Dichotomous indicator of a restaurant that has a value of 4 or above in the "Chain_Counts" variable. Indicating restaurant is a chain with 4 or more service	Input	Binary	Extracted

	locations (1=Yes, 0=No).			
Ethnicity	Indicates if a restaurant is labeled as American, Chinese, Italian, Japanese, Mexican, Other.	Rejected	Nominal	Extracted
Is_Open (Target Variable)	Consists of Binary numbers specifying whether that business is functioning (1) or out-of-business (0).	Target	Binary	Yelp Dataset Challenge
Kid_Friendly	Dichotomous indicator of a restaurant that has kid friendly services as an attribute (1=Yes, 0=No).	Input	Binary	Extracted
Name	Displays name of each business.	Rejected	Nominal	Yelp Dataset Challenge
Noise_Level	Nominal indicator of noise level at the restaurant (1 = quiet, 2= average, 3= loud, 4=very loud).	Input	Ordinal	Extracted
Parking	Dichotomous indicator of a restaurant that has parking services as an attribute (1=Yes, 0=No).	Input	Binary	Extracted
Postal Code	Displays zip code of each business.	Rejected	Nominal	Yelp Dataset Challenge
Price_Range	Nominal indicator of the price range per person at a restaurant (1 = under $10, 2 = $11-$30, 3 = $31-$60, 4 = $60 and above)	Input	Nominal	Extracted
Reservations	Dichotomous indicator of a restaurant that has reservation services as an attribute (1=Yes: 0=No).	Input	Binary	Extracted
Restaurant	Dichotomous indicator of a business categorized as a restaurant (1=Yes, 0=No).	Rejected	Unary	Extracted
Review_ Count	Shows the number of online reviews that each business has received.	Input	Interval	Yelp Dataset Challenge
Stars	Shows the number of star-ratings each business has achieved.	Input	Interval	Yelp Dataset Challenge
State	Lists state abbreviations e.g. N.J., OK, NV, etc.	Input	Nominal	Yelp Dataset Challenge
Table_ service	Dichotomous indicator of a restaurant that includes Table	Input	Binary	Extracted

	services as an attribute (1=Yes: 0=No).			
Takeout	Dichotomous indicator of a restaurant that has takeout as an attribute (1=Yes, 0=No).	Input	Binary	Extracted
Wheelchair_accepted	Dichotomous indicator of a restaurant that is wheelchair accessible as an attribute (1=Yes: 0=No).	Input	Binary	Extracted
Wi-Fi	Dichotomous indicator of a restaurant that includes Wi-Fi as an attribute (1=Yes, 0 =No).	Input	Binary	Extracted

3.1 REJECTED VARIABLES

Nested “Names”, "Categories", and "Attributes" columns were rejected during the import process as they were not needed for analysis but had been used to derive our extracted variables. "Restaurant" was also rejected as it only contains one value (Unary) and is not needed for analysis. "Address" was rejected during the import process due to its character or text length and irrelevance. "Postal code" and "City" were rejected during the model configuration process as there are too many levels that would hinder analysis. However, they were used in the exploratory analysis because it shows high Variable Worth as it relates to the target variable “Is_Open”. “Ethnicity” was rejected because many of the restaurants were labeled incorrectly and we are trying to avoid bias in the models.

CHAPTER FOUR

EXPLORATORY ANALYSIS

The data was analyzed from different angles using graphs, charts, tables and summary statistics in order to understand the data and find correlations between the variables.

Our SAS Enterprise Miner (EM) data source was defined using the metadata settings and connected to the StatExplore node to provide preliminary statistics and variable distributions on the target variable and better understand the statistics relating to input variables. We also employed SAS Studio and Tableau alongside E.M. to visualize and explore the input data to observe possible patterns, anticipated relationships, unanticipated trends, check for missing values, and anomalies before building models to gain understanding and ideas.

To begin the exploratory analysis, the target variable's distribution is shown in Figure 1 and Table 2. Here, 67.60% of the target variable represents open restaurants, while 32.40% represents closed restaurants.

Figure 1: Is_Open pie chart

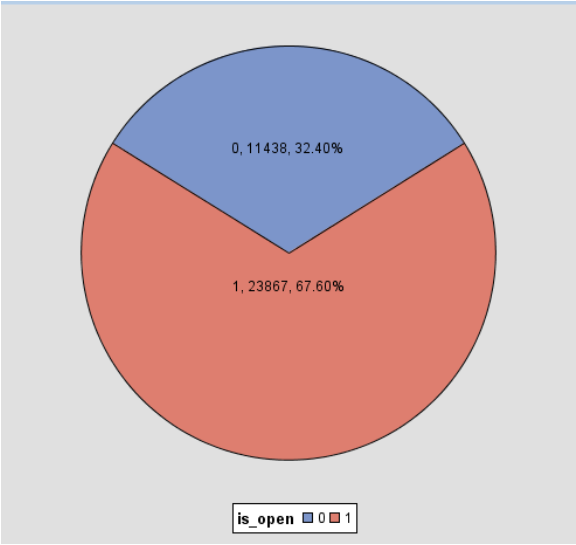
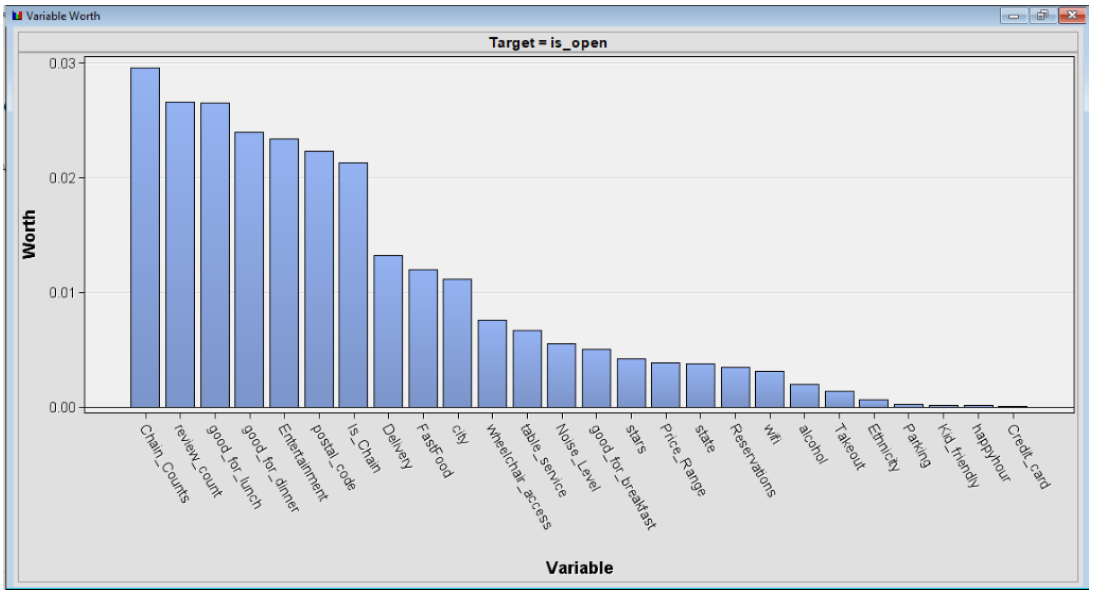


Table 3: Is_Open summary statistics

Distribution of Class Target and Segment Variables					
Data Role = TRAIN					
Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Is_Open	TARGET	1	23867	67.6023
TRAIN	Is_Open	TARGET	0	11438	32.3977

Observing the result of the StatExplore node, the Variable Worth Plot (VWP) in Figure 2 reveals that the top-ten most related (Key) variables to the "Is_Open" target variable are "Chain_Counts", "Review_Count", "Good_for_lunch", "Good_for_dinner", "Entertainment", "Postal_Code", "Is_Chain", "Delivery", "FastFood", and "City" arranged in order of importance. We will focus our exploratory analysis on these ten variables.

Figure 2: Variable Worth Plot



Among our ten key variables, two are continuous, and eight are categorical. We will also be plotting these to understand their distribution.

4.1 CONTINUOUS VARIABLES

Skewness and kurtosis are statistical measures also used to measure a variable's distribution. While skewness provides information about the distortion from the normally distributed bell curve, the kurtosis usually provides insights regarding the presence of outliers or extreme values. By standard, skewness, and kurtosis +/- 3 standard deviations away are considered high. In our case, we can see in Table 3 that both "Chain_Counts" and "Review_Count" skewness and kurtosis are high. We will account for that by transforming the data or by removing some of the extreme values.

Table 4: Interval Variable Summary Statistics

Variable Name	Mean	Std. Dev	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Chain_ Counts	38.26903	107.4537	0	1	1	568	3.838634	14.83487
Review_ count	109.1025	255.7983	0	3	36	10129	10.69444	222.7171

The histograms and the boxplots below show the distribution of our two continuous variables; "Review_Count" and "Chain_Counts" (see Figure 3 and Figure 4). We can see in the histograms that both variables are skewed to the right, which could be explained by the presence of outliers. Also, the box plot of "Review_Count" (Figure 5) shows that most restaurants in the data set have less than 100 reviews, and a great deal of them are above the upper whisker (253 reviews); therefore, considered as outliers. Similarly, the box plot of "Chain_Counts" (Figure 6) shows that most of the sample restaurants have less than eight chains. Restaurants that have more than 19 chains are considered outliers by the box plot.

Figure 3: Histogram of Review_Count

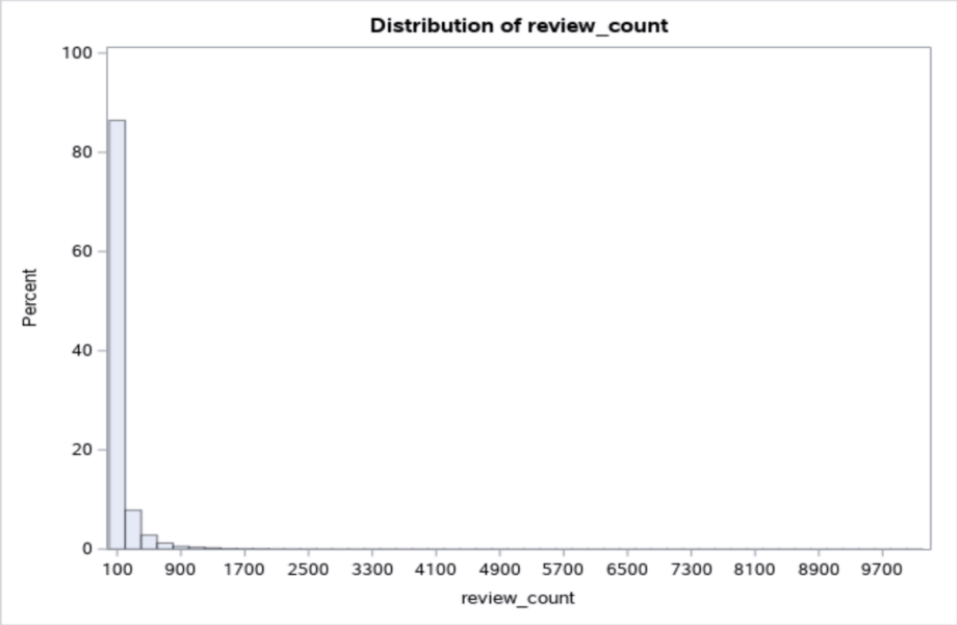


Figure 4: Boxplot of Review_Count

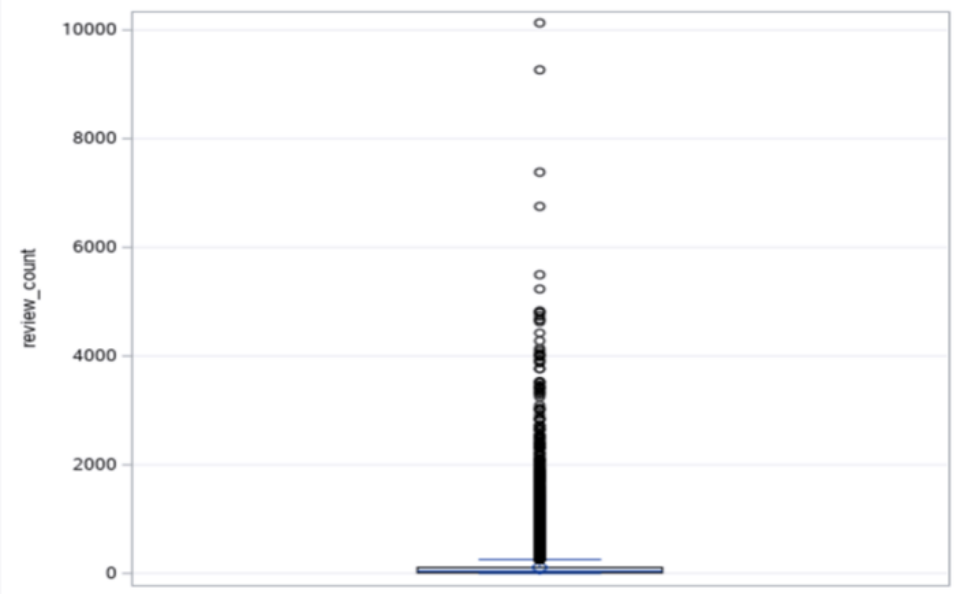


Figure 5: Histogram of Chain_Counts

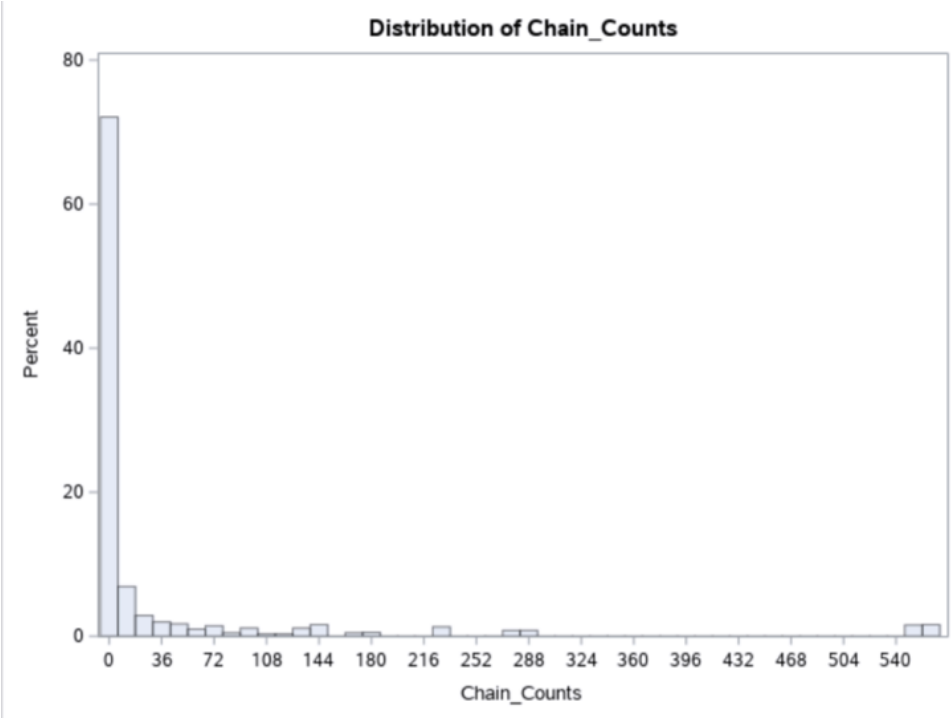
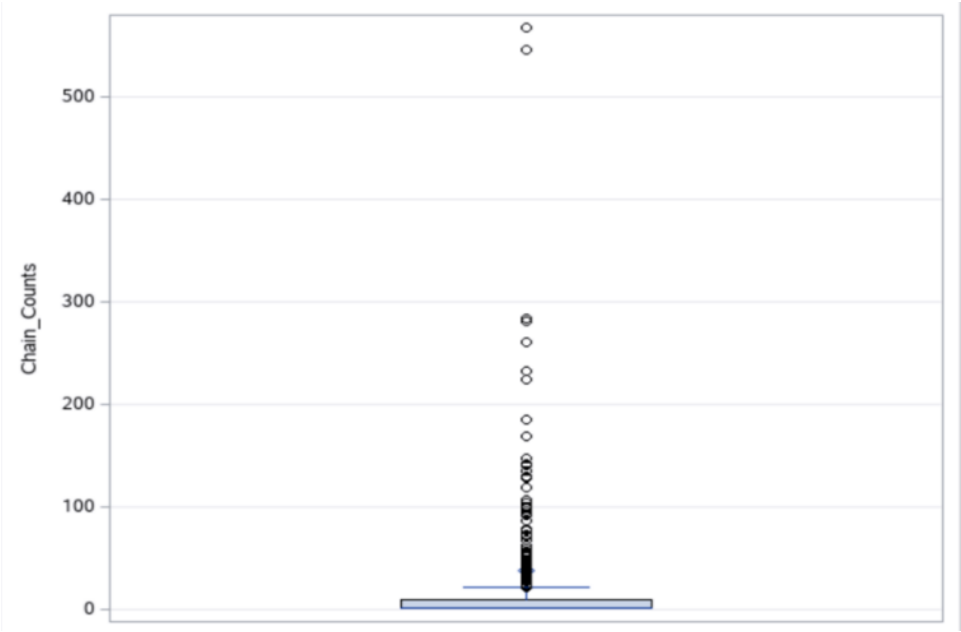


Figure 6: Boxplot of Chain_Counts



4.2 CATEGORICAL VARIABLES

Now, let us look at the eight categorical variables closely related to the target variable. First, here is the summary statistics for nominal variables as a brief introduction (Table 4):

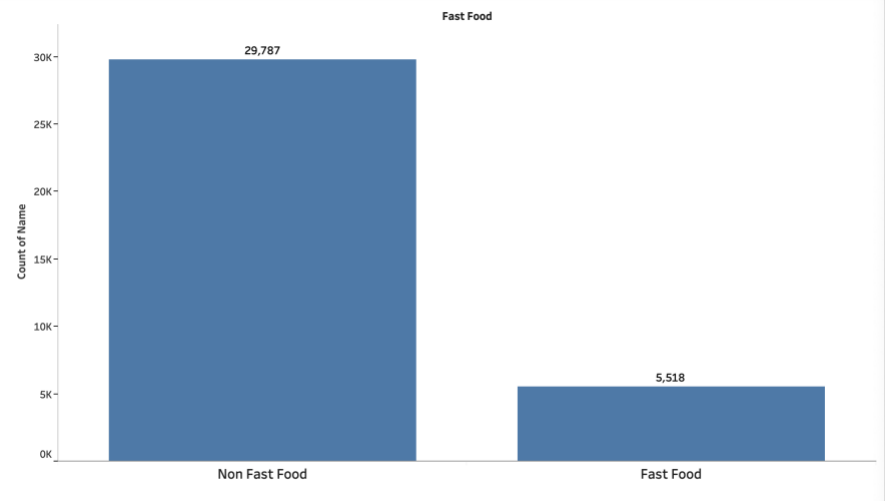
Table 5: Class Variable Summary Statistics

Variable Name	Number of levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Credit_card	2	0	1	90.89	0	9.11
Delivery	2	0	0	65.89	1	34.11
Entertainment	2	0	1	62.74	0	37.26
Ethnicity	6	0	other	40.93	American	32.58
FastFood	2	0	0	84.37	1	15.63
Is_Chain	2	0	0	68.34	1	31.66
Kid_friendly	2	0	1	71.58	0	28.42
Parking	2	0	1	52.10	0	47.90
Price_range	4	0	1	43.51	2	42.05
Reservations	2	0	0	73.07	1	26.93
Takeout	2	0	1	85.16	0	14.84
Alcohol	2	0	0	69.09	1	30.91
Good_for_breakfast	2	0	0	93.07	1	6.93
Good_for_dinner	2	0	0	73.64	1	26.36
Good_for _lunch	2	0	0	69.25	1	30.75
Happyhour	2	0	0	81.20	1	18.80
Postal code	513	28	89109	2.78	85281	1.63
Stars	9	0	4	24.32	3.5	22.60
State	5	0	AZ	34.36	NV	23.64
Table_service	2	0	0	79.48	1	20.52
Wheelchairaccess	2	0	0	80.56	1	19.44
Wi-Fi	2	0	0	64.09	1	35.91
Is_Open	2	0	1	69.87	0	30.13
Noise_Level	4	0	2	49.34	4	32.02

To begin the analysis of categorical variables, let us take a look at the “FastFood” variable. Figure 7 shows that there are 5,518 (15.63%) fast food restaurants in the dataset while

the remaining 29,787 (84.37%) are not fast food restaurants, i.e., dining restaurants, café, buffet and so on.

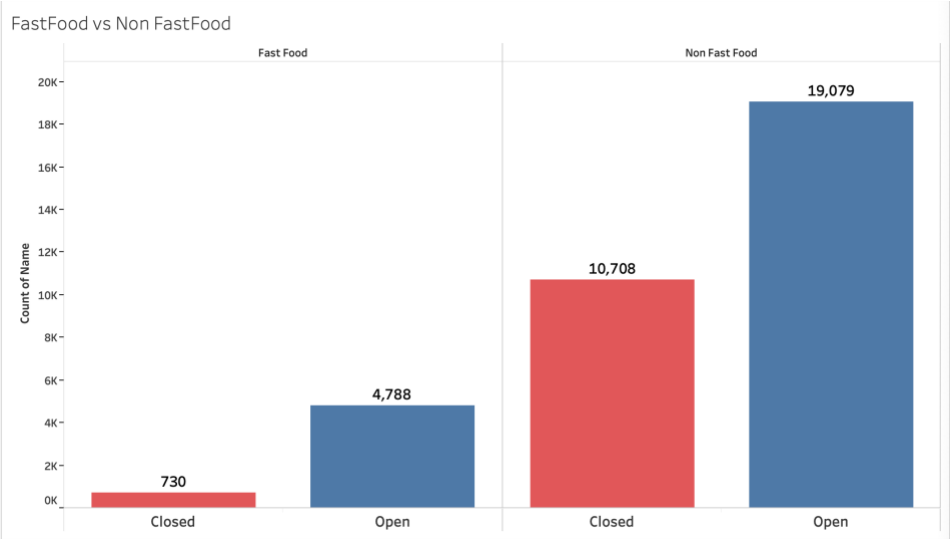
Figure 7: Fast_Food bar chart



Compared to the overall closing rate of 32% of all restaurants, fast-food restaurants are less likely to close. Figure 8 shows the distribution of fast and non-fast food restaurants that are open or closed. For fast food restaurants, 730 (13.23%) are closed and 4,788 (86.77%) are open. For restaurants that are not fast food, 10,708 (35.95%) are closed and 19,079 (64.05%) are open. Through this, we see almost a three times higher rate of closure in non-fast food restaurants than in fast food restaurants. When comparing non fast food closure rates to the overall closing rate of 32%, non fast food restaurants take up 93.6% of closures. QSR and Insula Research estimate that about 50 to 70 percent of fast food sales arrive at drive-thru windows with the remaining percentage distributed through carryout or delivery (McDonnell, 2020). From our understanding of the effects of reviews on customer behavioral intentions, revenue, and restaurant success through our literature review. From our understanding of the effects of reviews on customer behavioral intentions, revenue, and restaurant success through our literature review, the

decreased use of fast food dining facilities and their consistent quality of food (same taste, look, etc.) exempts them from being constantly reviewed or criticized, unlike other restaurants that require dine-in experiences and are constantly evaluated by their services. The results could also mean that fast food restaurants that are usually chain restaurants backed by a big corporate body can survive better in bad business climates. We will investigate this assumption when we address the "Is_Chain" variable.

Figure 8: FastFood by Is_Open bar chart



The next variable we will be exploring is the Entertainment variable. Figure 9 shows that in the dataset, 22,149 (62.74%) restaurants offer entertainment such as background music, television, games, and photo booth while 13,156 (37.26%) restaurants do not offer any entertainment.

Figure 9: Entertainment bar chart

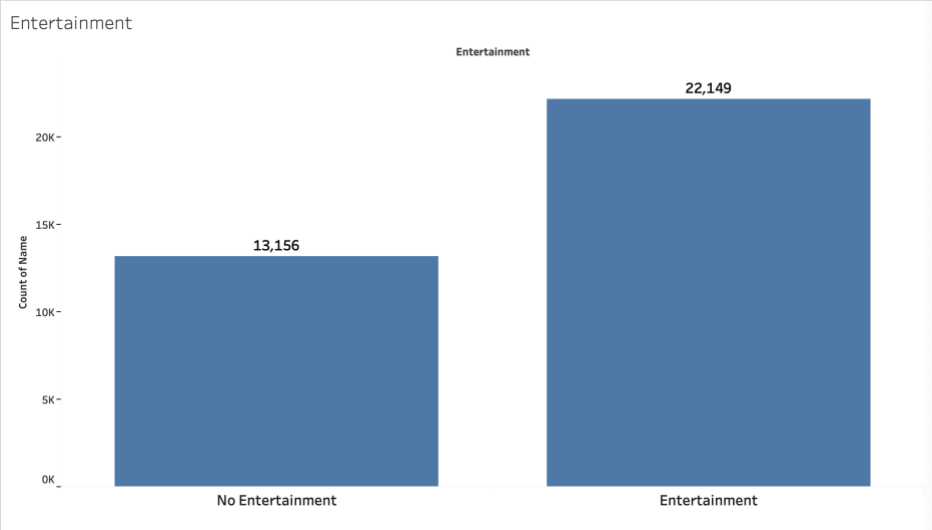
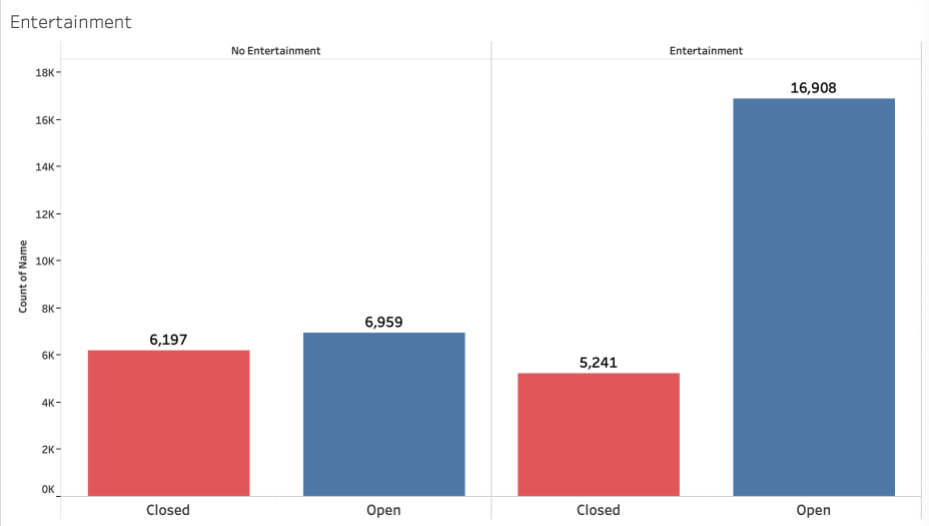


Figure 10 shows restaurants that are open and closed based on if they provide entertainment or not. For restaurants that do not provide any entertainment, 6,197 (47.10%) are closed, and 6,959 (52.90%) are open. For restaurants that provide entertainment, 5,241 (23.66%) are closed and 16,908 (76.34%) are open. There is a higher rate of closure in restaurants that do not provide entertainment. Perhaps this is because experiences such as background music and other forms of entertainment are statistically significant predictors of satisfaction and repeat patronage, contributing to restaurant success (DiPietro 2016).

Figure 10: Entertainment by Is_Open bar chart



Next is the "Is_Chain" variable. As seen in Figure 11, there are 11,179 (31.66%) chain restaurants and 24,126 (68.34%) restaurants not classified as chain restaurants. Figure 12 shows 1,856 (16.60%) of the chain restaurants are closed while 9,582 (39.71%) of non-chain restaurants are closed. Like non- fast-food restaurants, this means that non- chain restaurants close at a higher rate. Also confirming our assumptions about chain restaurants earlier in this analysis, it is true that "large chains have the resources to ride out a protracted shutdown, but independent restaurants" find it harder to survive in a similar climate (Severson & Yaffe-Bellany, 2020).

Figure 11: Is_Chain pie chart

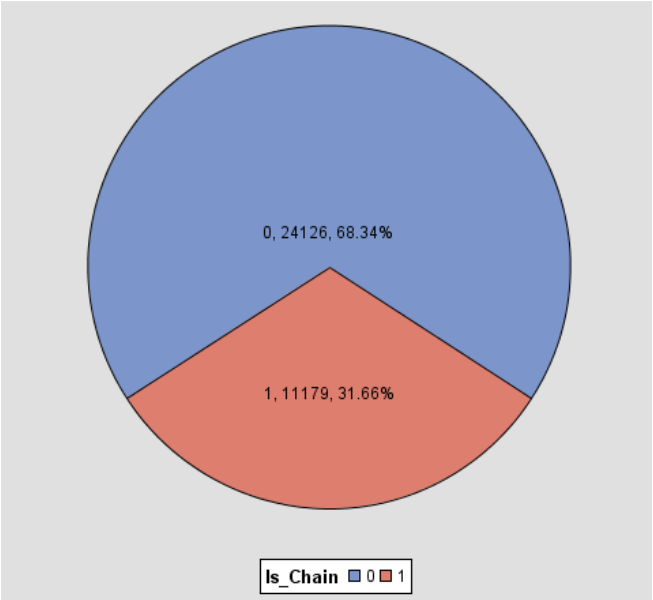
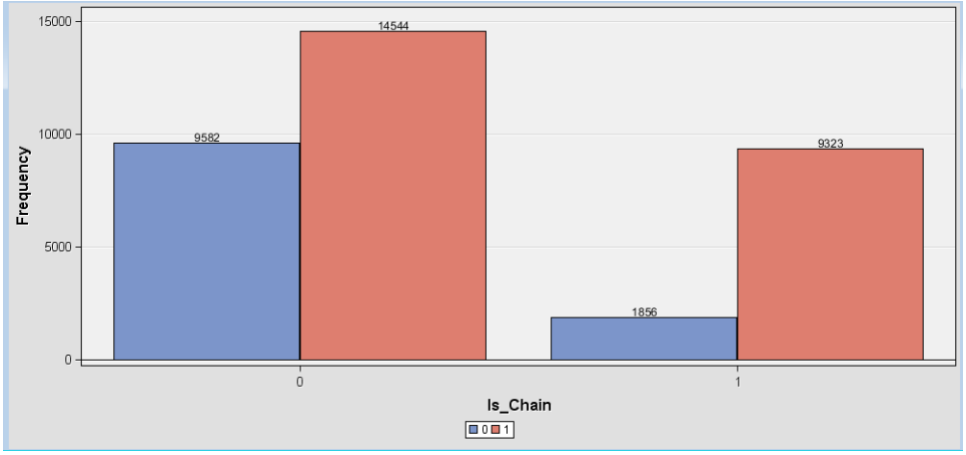


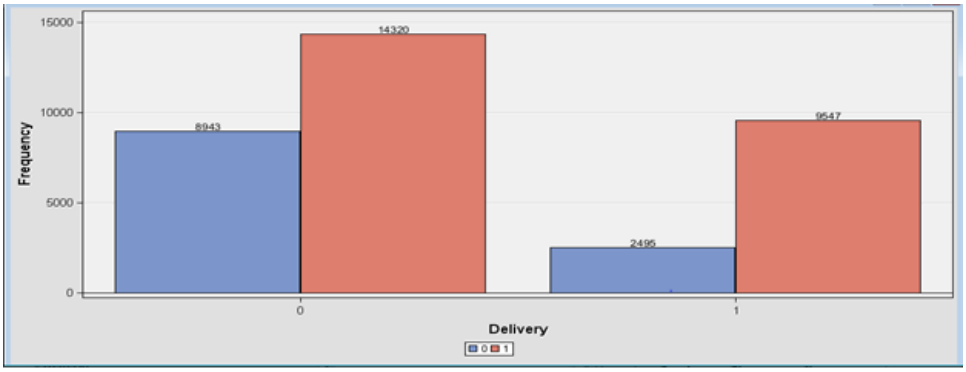
Figure 12: Is_Chain by Is_Open bar Chart



Moving on to the Delivery variable, Figure 13 shows that 20.72% of restaurants that offer delivery are closed (12,042 total delivery restaurants). In comparison, 38.44% of restaurants that do not offer delivery are closed (23,263 total non-delivery restaurants). From this, we assume restaurants that do not offer delivery close down at a much higher rate. Offering delivery in a restaurant is essential today because "the market for food delivery stands at €83 billion, or 1

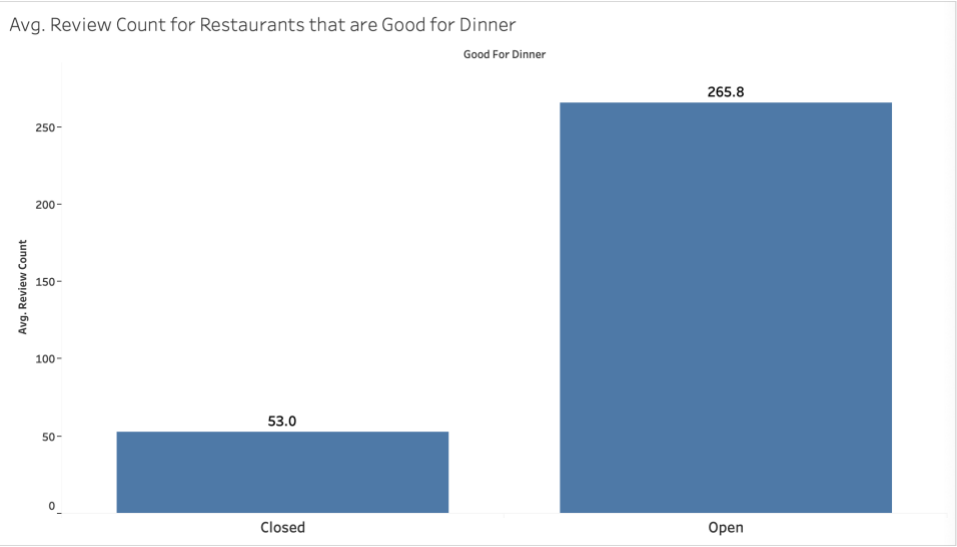
percent of the total food market and 4 percent of food sold through restaurants and fast-food chains" (Wrulich et al., 2020). It is also expected to reach an annual growth rate estimated at 3.5 percent from 2017 through 2021 (Wrulich et al., 2020). Besides, "delivery services are a popular dining option with U.S. consumers, as a November 2016 survey found that 20 percent of respondents use food delivery at least once a week" (Lock, 2020). As a result, we assume that any restaurant that fails to cater to this population may lose customers and much revenue, hence the high rate of failure.

Figure 13: Delivery by Is_Open bar chart



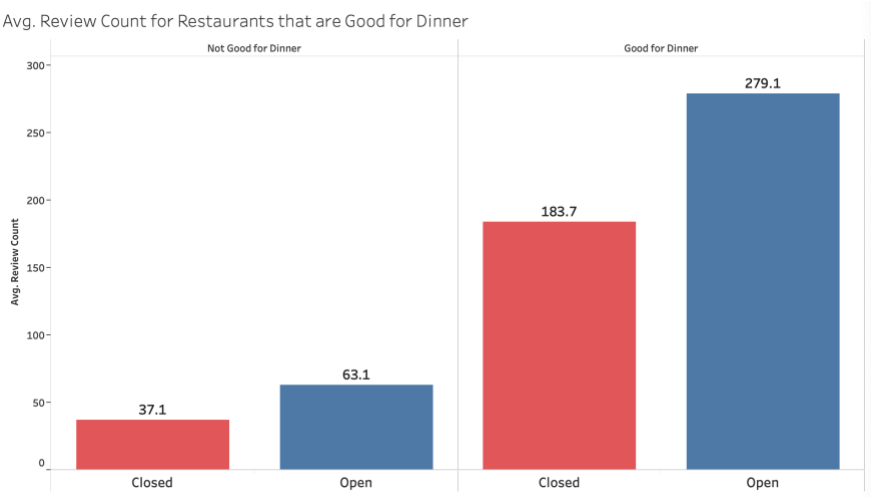
The next variables to be explored are the "Good_for_dinner", "Postal _Code", and "City" variables. In Figure 14, there are 212.8 (265.8 - 53) more reviews on average for restaurants that are good for dinner time compared to restaurants that do not offer or are not suitable for dinner services with an average of 53 reviews. Perhaps more people eat out at dinner time and leave reviews; hence the significant increase in review counts. If these reviews are positive, it could increase new customer patronage and return intentions of old customers for restaurants that are good for dinner.

Figure 14: Average Review_Counts by Good_for_dinner



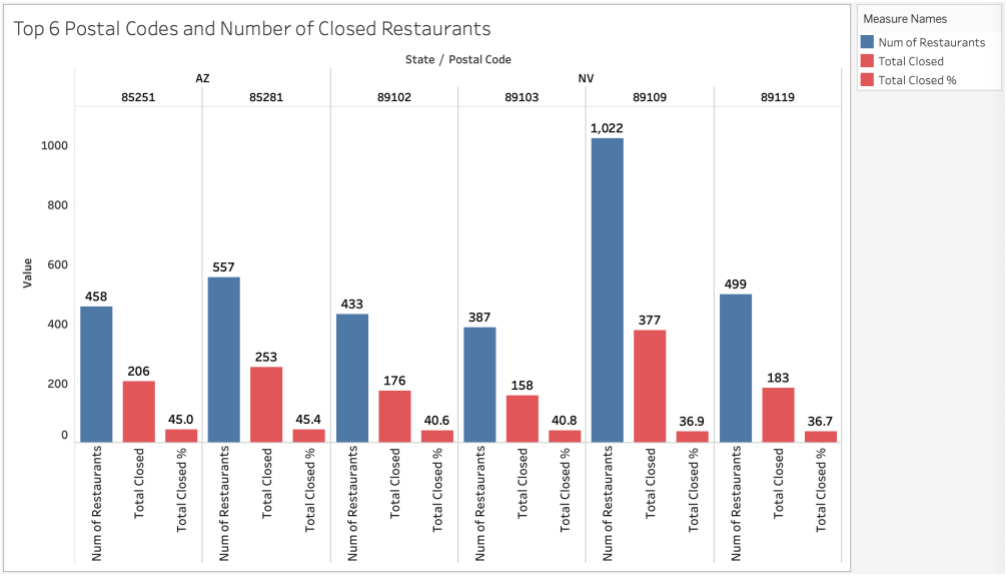
In Figure 15, restaurants that are good for dinner as it relates to the event of the target variable have 242 (279.1 - 37.1) more reviews on average than closed (non-event) restaurants that are not good for dinner. Overall, it looks like high review counts and closed restaurants are mutually exclusive, meaning they cannot occur simultaneously. On the other hand, high review counts and open restaurants are mutually inclusive, meaning they mostly occur together. Therefore, we assume that a higher review count could mean a higher chance of staying open.

Figure 15: Average review count for open and closed restaurants that are good for dinner vs. not good for dinner



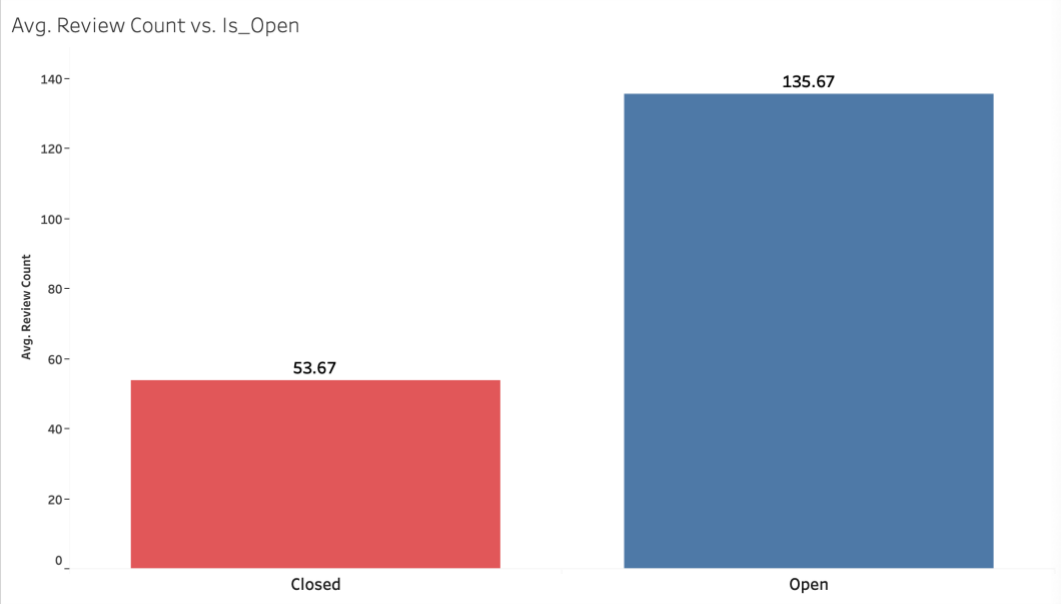
"Postal_Code" and "City" take sixth and tenth place respectively on the VWP. These support our findings in the literature review, which states that location is vital in predicting closure. Las Vegas city is home to more than 41 million visitors each year, and it was rated one of the top ten locations in the country for great food ("Fun Facts | LAS VEGAS", 2020). Therefore, it is no surprise that in Figure 16, four (89102, 89103, 89109, 89139) out of the top six populated postal codes have the highest rate of closure overall. All four postal codes are Las Vegas postal codes, which are probably home to various restaurants who want to profit from the bustling market. Furthermore, there is a positive relationship between the number of restaurants and restaurants that are closed because postal codes with a high number of restaurants also have higher numbers and rates of closed restaurants. This re-establishes another finding highlighting competition where Parsa et al. (2015) mentions that competition is one of three key factors contributing to restaurant failure.

Figure 16: Top six Postal Codes and Number of Closed Restaurants



Lastly, we will discuss some inferential statistics used to determine if there is a significant difference between the means of two or more groups. When we started this project, we were interested in investigating certain relations based on our literature review findings. "Review_Count" for instance, is the second most important variable as it relates to "Is_Open". This points to the literature review where we found that consumer-generated reviews and ratings on sites like Yelp "have become highly influential in directing consumer's choices and purchase decisions" (Parikh et al, 2014, p.162). Reviews of past users can influence prospective customers. Figure 17 shows that open restaurants have more reviews on average. We will use a t-test to measure the significance of the difference between groups of "Is_Open" (Open = 1 and Closed = 0) with regards to average "Review_Count".

Figure 17: Is_Open vs. Review_Count bar chart



4.3 STATISTICAL ANALYSIS

- **T Test for Review_Count versus (vs.) Is_Open**

Hypothesis:

H0: μo- μc = 0 (with O= open and 1= Closed)

Ha: μo- μc ≠ 0

Result:

Figure 18: Summary of Is_open statistics

is_open	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		11438	53.6736	126.5	1.1832	3.0000	5494.0
1		23867	135.7	294.9	1.9086	3.0000	10129.0
Diff (1-2)	Pooled		-81.9926	252.9	2.8761		
Diff (1-2)	Satterthwaite		-81.9926		2.2456		

Figure 19: T Test for Review count vs. Is open result

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	35303	-28.51	<.0001
Satterthwaite	Unequal	34960	-36.51	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	23866	11437	5.43	<.0001

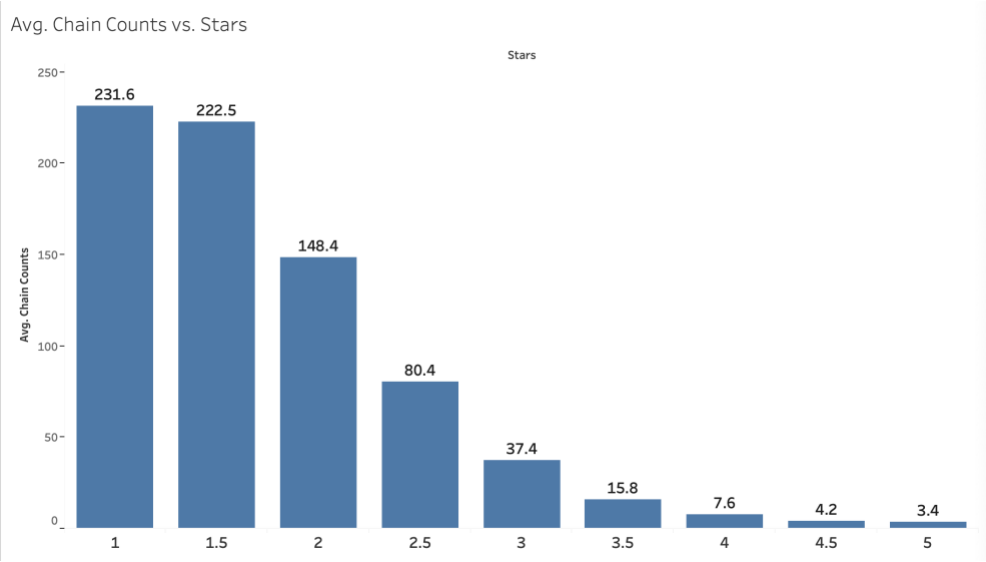
The p-value (<.0001) for the equality of variance is significant (less than 0.05); therefore, the variances are unequal.

Conclusion: There is a significant difference (p-value <0.05) between groups open and closed with respect to mean review count.

Also, a similar analysis was done with "Stars" and "Chain_Counts". Figure 19 shows that the higher the number of chain restaurants, the higher the negative reviews. Restaurants that have a chain count of 8 restaurants or less typically have a higher rating. Perhaps independent or

fewer restaurants are easier to manage than larger chain restaurants. The next step is to conduct an ANOVA test to evaluate the significance of the difference between the various levels of star rating with regards to average "Chain_Counts".

Figure 20: Chain Counts vs. Stars bar chart



- **One-way ANOVA test**

Hypothesis:

H0: Difference in mean = 0

Ha: Difference in mean ≠ 0

Result:

Figure 21: Summary of means of star levels

Level of stars	N	Chain_Counts	
		Mean	Std Dev
1	256	222.472656	222.155164
2	2297	146.286461	178.325278
3	6028	36.808394	97.457261
4	8587	7.531850	41.165797
5	955	3.364398	27.492122
1.5	968	218.816116	204.860563
2.5	3625	79.738207	144.250867
3.5	7978	15.551893	60.546685
4.5	4611	4.193017	29.176775

Figure 22: One-way ANOVA test result for Chain_Counts

Dependent Variable: Chain_Counts Chain_Counts

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	95531172.5	11941396.6	1350.48	<.0001
Error	35296	312099688.3	8842.4		
Corrected Total	35304	407630860.8			

Figure 23: Homogeneity test for Chain_Counts

Levene's Test for Homogeneity of Chain_Counts Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
stars	8	4.407E12	5.509E11	459.22	<.0001
Error	35296	4.234E13	1.1996E9		

Welch's ANOVA for Chain_Counts

Source	DF	F Value	Pr > F
stars	8.0000	535.38	<.0001
Error	3860.2		

- **Post hoc analysis**

After conducting the ANOVA test with "Chain_count" as the continuous variable and star rating as a category, we can safely conclude that there is a significant difference (p-value<0.05) between star rating groups with regards to mean "Chain_count".

Once we conducted the Anova test, we looked at the Least Squares Means for Star Effects table to perform a post hoc analysis. We noticed that most of the groups were statistically significant from one another, hence the reason for the overall ANOVA test results. The only groups where there is no significant difference between groups means regarding chain counts are 1 star and 1.5 stars (p-value 0.9998); 4 stars and 5 stars (p-value 0.9289); 4 stars and 4.5 stars (p-value 0.5716); 4.5 stars and 5 stars (p-value 1.000). There is only a .5 or 1 star difference between stars that do not have a significant difference.

Figure 24: Post hoc analysis

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

stars	Chain_Counts LSMEAN	LSMEAN Number
1	222.472656	1
2	146.286461	2
3	36.808394	3
4	7.531850	4
5	3.364398	5
1.5	218.816116	6
2.5	79.738207	7
3.5	15.551893	8
4.5	4.193017	9

Least Squares Means for effect stars
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: Chain_Counts

i/j	1	2	3	4	5	6	7	8	9
1		<.0001	<.0001	<.0001	<.0001	0.9998	<.0001	<.0001	<.0001
2	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
3	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
4	<.0001	<.0001	<.0001		0.9289	<.0001	<.0001	<.0001	0.5716
5	<.0001	<.0001	<.0001	0.9289		<.0001	<.0001	0.0043	1.0000
6	0.9998	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
8	<.0001	<.0001	<.0001	<.0001	0.0043	<.0001	<.0001		<.0001
9	<.0001	<.0001	<.0001	0.5716	1.0000	<.0001	<.0001	<.0001	

This takes us back to a statement made about the "FastFood" variable. We assumed that non fast food restaurants close as a result of constant criticism, unlike fast food restaurants with

consistent (usually the same) service. When investigated further, it turns out that fast food chain restaurants have 222 chain counts on average and are reviewed less (38 average reviews), but the reviews are mostly negative as seen in Figure 20. While restaurants that are not fast food chains have 1 chain count on average and reviewed more (125 average reviews), the reviews are mostly positive or mixed. This means we can also conclude that restaurants with lower chain counts find it harder to survive with low star ratings (reviews) because they do not have the resources (unlike large fast food chains) to ride out a protracted shutdown (Severson & Yaffe-Bellany, 2020).

As stars reflect customer satisfaction, we were interested in determining if there was a mean difference between various levels of star ratings with the "Is_open" target variable. We predict that there is a significant difference in the star level ratings; we performed a Chi-square test to assess this hypothesis.

Hypothesis:

H0: Difference in mean = 0

Ha: Difference in mean ≠ 0

Result:

Table 6: Stars by Is_open

Table of Stars by is_open			
Stars	0	1	Total
1	69 0.20 26.95 0.60	187 0.53 73.05 0.78	256 0.73
1.5	197 0.56 20.35 1.72	771 2.18 79.65 3.23	968 2.74

2	576 1.63 25.08 5.04	1721 4.87 74.92 7.21	2297 6.51
2.5	1196 3.39 32.99 10.46	2429 6.88 67.01 10.18	3625 10.27
3	2338 6.62 38.79 20.44	3690 10.45 61.21 15.46	6028 17.07
3.5	2977 8.43 37.32 26.03	5001 14.17 62.68 20.95	7978 22.60
4	2619 7.42 30.50 22.90	5968 16.90 69.50 25.01	8587 24.32
4.5	1200 3.40 26.02 10.49	3411 9.66 73.98 14.29	4611 13.06
5	266 0.75 27.85 2.33	689 1.95 72.15 2.89	955 2.70
Total	11438 32.40	23867 67.60	35305 100.00

Figure 25: Chi-square table Stars vs Is_open

The FREQ Procedure

Statistics for Table of stars by is_open

Statistic	DF	Value	Prob
Chi-Square	8	433.4417	<.0001
Likelihood Ratio Chi-Square	8	440.0496	<.0001
Mantel-Haenszel Chi-Square	1	3.3547	0.0670
Phi Coefficient		0.1108	
Contingency Coefficient		0.1101	
Cramer's V		0.1108	

Sample Size = 35305

The chi-square results show that there is a statistically significant difference between the star levels and is_open restaurants. However, one can also determine from Table 6 that a high star rating does not guarantee a higher possibility of a restaurant being open. Restaurants given 1.5 stars have the highest percentage of is_open restaurants at 79.65%. In comparison, restaurants given 2 stars have the second-highest percentage of open restaurants at 74.92%. One can determine that the number of reviews that a restaurant receives has a bigger impact on staying open than the star rating given to the restaurant.

Now that we have explored the variables in the dataset to derive assumptions and insights, the next step is to use various predictive tools in SAS Enterprise Miner to develop models to predict restaurant closure.

CHAPTER 5

MODELING, EVALUATION AND RESULTS

Before modeling the data, Data Partitioning was done to segment the data into subgroups similar to the target (Is_open). This is to avoid over- or underfitting. The training partition was used to build the model, and the validation partition was set aside and used to test the accuracy while we fine-tuned the model. The test partition, although 0, would have been used to evaluate how the model will work on new data but was not necessary. The data was partitioned into a training dataset (50%) and a validation dataset (50%).

Thirty-seven algorithms were created to predict restaurant closure. However, we will be discussing only the Decision Tree, Stepwise Regression, Variable Selection (AOV 16) with Regression (Best Regression), Neural Network, LARS, LASSO, Adaptive LASSO and High Performance Data Mining models.

Table 7: Models created to predict restaurant closure

Selection Tree with HP SVM Poly	HP SVM (Radio Basis Function)
Selection Tree with Neural Network	Stepwise Misclassification Regression
Stepwise Regression with Neural Network	LARS with Regression
Neural Network	LASSO with Regression
LASSO with Neural Network	PLS 0.2 with Auto Neural Network
LARS with Neural Network	Adaptive LASSO with Regression
Adaptive LASSO with Neural Network	PCA with Neural Network
Auto Neural (AOV 16)	PLS 0.2 with Regression
Variable selection with Neural Network (AOV 16)	Variable Selection with Regression AOV16
HP Forest larger	Regression Variable Clustering (Cluster component)
PLS 0.2 with Neural Network	Regression Variable Clustering (Best Variable)

Variable Selection with Regression	Decision Tree
Variable Selection with Neural Network	PCA with Regression
HP Forest	PLS with Neural Network
Adaptive LASSO with Auto Neural Network	PLS with Auto Neural Network
HP SVM Linear	PLS with Regression
Stepwise Regression with Auto Neural Network	HP SVM Sigmoid
LARS with Auto Neural Network	Variable Selection with Auto Neural Network
LASSO with Auto Neural Network	

5.1 DECISION TREE

The simplest type of prediction is Decision Trees. They are also mentioned as classifications because they usually are associated with some type of action, such as classifying a case enrolled or not enrolled. The Decision tree helps in predicting or classifying future observations based on a set of decision rules. In this project, a misclassification tree is used.

A decision prediction can be rated by misclassification or the proportion of disagreement between the prediction and the outcome. For generating the misclassification tree, the maximal tree is pruned to give the best batch of subtrees. SAS Enterprise Miner chooses the model with the simplest model and the best validation assessment. The misclassification tree resulted in 17 leaves. The variables selected for this tree are “Chain_counts”, “Entertainment”, “Review_count “, “Good_for_dinner” , “Price_range”, “Wheelchair_access”, “Noise_level”, “Good_for_lunch “ and “Reservations”. The Variable Importance Plot (VIP) in Table 8 shows the level of importance of the variables. The validation misclassification rate obtained from misclassification tree is 0.225287.

Figure 26: Subtree assessment plot of the decision tree

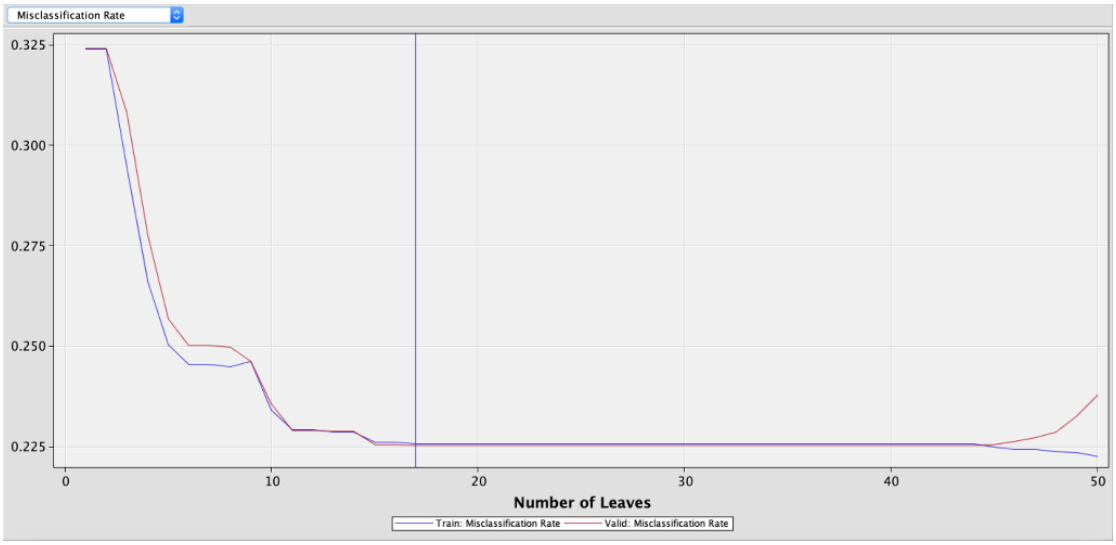


Table 8: Variable Importance of decision tree

Variable Importance Plot				
Variable Name	Number of Spliting Rules	Importance	Validation Importance	Role of Validation to Training Importance
Chain_Counts	2	1.0000	1.0000	1.0000
Entertainment	1	0.9962	0.9637	0.9674
Review_Count	2	0.7247	0.6323	0.8725
Good_for_Dinner	2	0.6855	0.7689	1.1217
Price_Range	1	0.6348	0.7088	1.1166
Wheelchair_Access	3	0.6299	0.6680	1.0604
Noise Level	2	0.4959	0.5299	1.0686
Good_for_Lunch	2	0.4794	0.6001	1.2519
Reservations	1	0.2152	0.2164	1.0056
The other variables have 0 splitting rules and 0 importance				

Figure 27: Fit statistics of decision tree

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
is_open	is_open	_NOBS_	Sum of Frequencies	17652	17653	.
is_open	is_open	_MISC_	Misclassification Rate	0.22564	0.225287	.
is_open	is_open	_MAX_	Maximum Absolute Error	0.944206	0.944206	.
is_open	is_open	_SSE_	Sum of Squared Errors	5602.387	5698.171	.
is_open	is_open	_ASE_	Average Squared Error	0.15869	0.161394	.
is_open	is_open	_RASE_	Root Average Squared ...	0.398359	0.401739	.
is_open	is_open	_DIV_	Divisor for ASE	35304	35306	.
is_open	is_open	_DFT_	Total Degrees of Freed...	17652	.	.

5.2 TRANSFORMATION OF SELECTED VARIABLES

The dataset was mostly cleaned in Python. Hence, there were no missing values, and no imputation of the values was required for this project. Also, most of the variables did not display any skewness except for “Review_counts” and “Chain_counts” which were transformed. Hence, this project required no imputation but the transformation of two variables.

Figure 28: Transformation result

Source	Method	Variable Name	Formula ▲	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	Chain_Counts		17652	0	1	568	39.11432	109.321	3.752594	14.03053	Chain_Cou...
Input	Original	review_count		17652	0	3	9264	109.6016	259.9702	10.99932	223.7574	review_cou...
Output	Computed	LOG_Chain_Counts	log(Chain...	17652	0	0.693147	6.34388	1.731068	1.655816	1.494342	0.884064	Transform...
Output	Computed	LOG_review_count	log(review...	17652	0	1.386294	9.133999	3.664527	1.410104	0.305811	-0.56394	Transform...

5.3 STEPWISE REGRESSION

A stepwise input selection method was used to analyze regression models in this project. A regression offers a different approach to prediction through an association between the target and input variables. The stepwise input selection method allows only those variables to be included in the model with a required level of p-value within the entry cut-off and stay cut-off. The default significance level of 0.05 was used.

However, the results show that all the variables were selected in the model. The Validation Misclassification rate for Stepwise Regression model is 0.199909. On the iteration plot as shown in figure 29, 23 was selected.

Figure 29: Iteration plot for stepwise regression

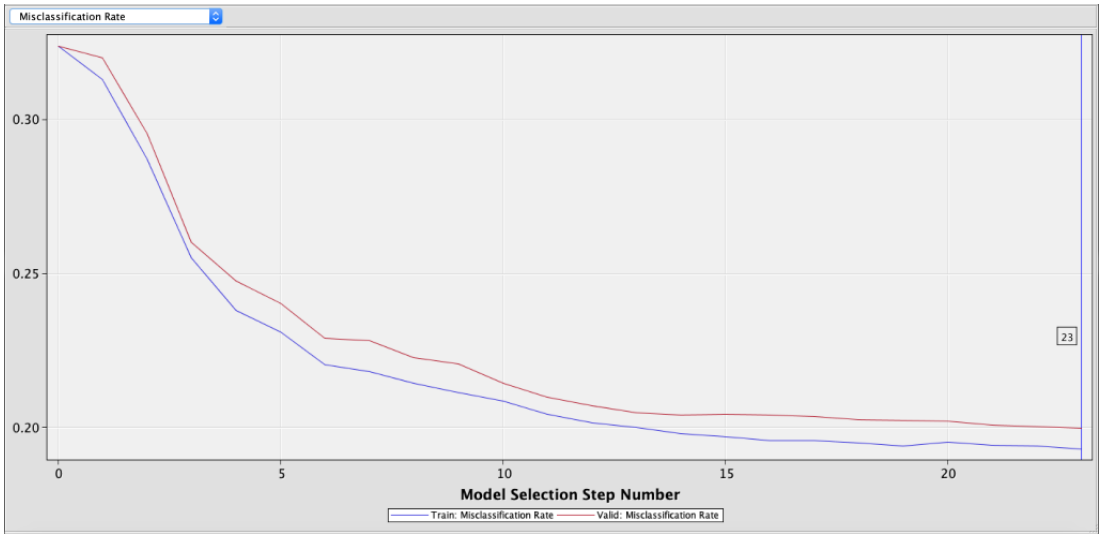
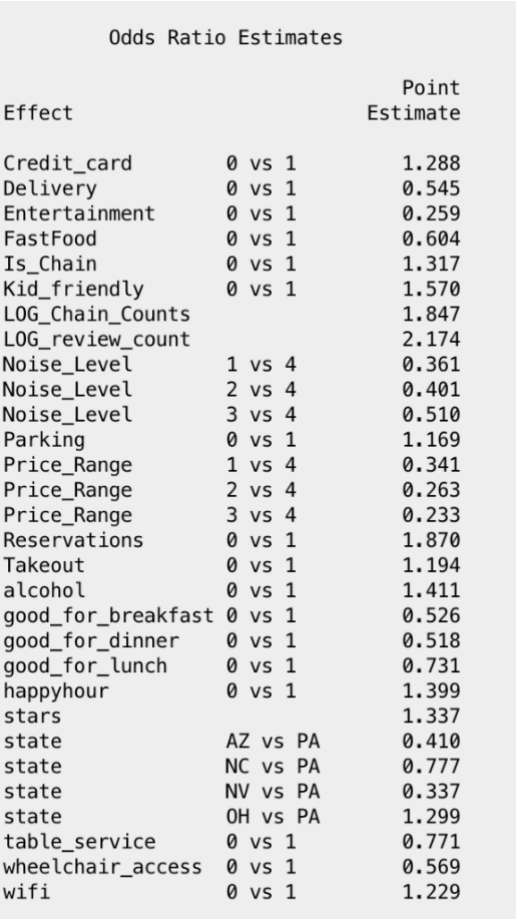


Figure 30: Fit statistics of Stepwise Regression

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
is_open	is_open	_AIC_	Akaike's Information Cr...	15428.15	.	.
is_open	is_open	_ASE_	Average Squared Error	0.139299	0.141784	.
is_open	is_open	_AVERR_	Average Error Function	0.435252	0.441238	.
is_open	is_open	_DFE_	Degrees of Freedom fo...	17621	.	.
is_open	is_open	_DFM_	Model Degrees of Free...	31	.	.
is_open	is_open	_DFT_	Total Degrees of Freed...	17652	.	.
is_open	is_open	_DIV_	Divisor for ASE	35304	35306	.
is_open	is_open	_ERR_	Error Function	15366.15	15578.36	.
is_open	is_open	_FPE_	Final Prediction Error	0.13979	.	.
is_open	is_open	_MAX_	Maximum Absolute Error	0.998162	0.988237	.
is_open	is_open	_MSE_	Mean Square Error	0.139545	0.141784	.
is_open	is_open	_NOBS_	Sum of Frequencies	17652	17653	.
is_open	is_open	_NW_	Number of Estimate We...	31	.	.
is_open	is_open	_RASE_	Root Average Sum of S...	0.373228	0.376543	.
is_open	is_open	_RFPE_	Root Final Prediction Er...	0.373884	.	.
is_open	is_open	_RMSE_	Root Mean Squared Error	0.373557	0.376543	.
is_open	is_open	_SBC_	Schwarz's Bayesian Crit...	15669.29	.	.
is_open	is_open	_SSE_	Sum of Squared Errors	4917.828	5005.841	.
is_open	is_open	_SUMW_	Sum of Case Weights Ti...	35304	35306	.
is_open	is_open	_MISC_	Misclassification Rate	0.193123	0.199909	.

Figure 31: Odds ratio estimates of Stepwise regression



Odds Ratio Estimates

Effect		Point Estimate
Credit_card	0 vs 1	1.288
Delivery	0 vs 1	0.545
Entertainment	0 vs 1	0.259
FastFood	0 vs 1	0.604
Is_Chain	0 vs 1	1.317
Kid_friendly	0 vs 1	1.570
LOG_Chain_Counts		1.847
LOG_review_count		2.174
Noise_Level	1 vs 4	0.361
Noise_Level	2 vs 4	0.401
Noise_Level	3 vs 4	0.510
Parking	0 vs 1	1.169
Price_Range	1 vs 4	0.341
Price_Range	2 vs 4	0.263
Price_Range	3 vs 4	0.233
Reservations	0 vs 1	1.870
Takeout	0 vs 1	1.194
alcohol	0 vs 1	1.411
good_for_breakfast	0 vs 1	0.526
good_for_dinner	0 vs 1	0.518
good_for_lunch	0 vs 1	0.731
happyhour	0 vs 1	1.399
stars		1.337
state	AZ vs PA	0.410
state	NC vs PA	0.777
state	NV vs PA	0.337
state	OH vs PA	1.299
table_service	0 vs 1	0.771
wheelchair_access	0 vs 1	0.569
wifi	0 vs 1	1.229

Some explanations of the odds ratio from our result are as follows:

LOG_Chain_ Counts *1.847*

For each additional restaurant of a chain, the odds of staying open change by a factor of 1.847, or an 84.7% increase.

LOG_Review_Count *2.174*

For each additional review, the odds of staying open changes by a factor of 2.174.

Stars		*1.337*

For each additional unit of star, the odds of staying open changes by a factor of 1.337, or a 33.7% increase.

Entertainment	*0 vs 1*	*0.259*

For restaurants without entertainment, the odds of staying open are 0.259 times lower than the odds of staying open for restaurants with entertainment.

Good_for_dinner	*0 vs 1*	*0.518*

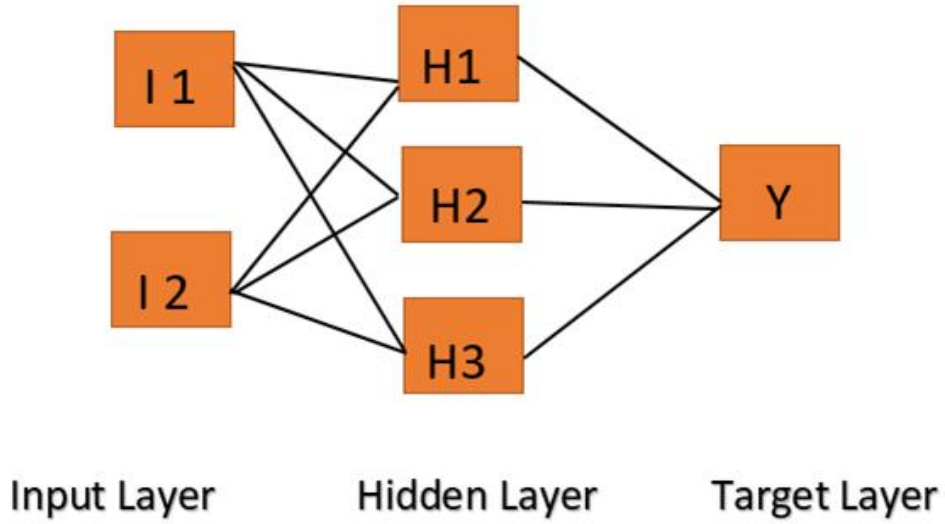
For restaurants that are not good for dinner, the odds of staying open are 0.518 times lower than the odds of staying open for restaurants that are good for dinner.

5.4 NEURAL NETWORK

Neural Network is a natural extension of a regression model. It includes a similar prediction formula to that of the regression model; it has an interesting and flexible addition to model virtually any association between input and target variables. A neural network uses a prediction formula to predict new cases, and a stopped training method to select an optimal model.

As mentioned, a neural network is similar to a regression model on a set of derived input known as hidden layers. The hidden layers or inputs can be considered as regressions and include a hyperbolic tangent, a default link function to shift, and rescale the logistic function. A neural network can be better explained with a Multi-layer perceptron model which arranges neurons in three layers. The first layer is the input layer. The input layer connects to the hidden layer, and the hidden layer connects to the target or output layer. Each part of the diagram has a counterpart in the network equation.

Figure 32: Neural network process



The blocks in the diagram represent the inputs layer, hidden layer, and target layer. The block interconnections correspond to the network equation weights.

Figure 32 shows the iteration plot for neural networks. The validation misclassification rate is 0.189543 for this model and the iteration selected is 25.

Figure 33: Iteration plot for the neural network

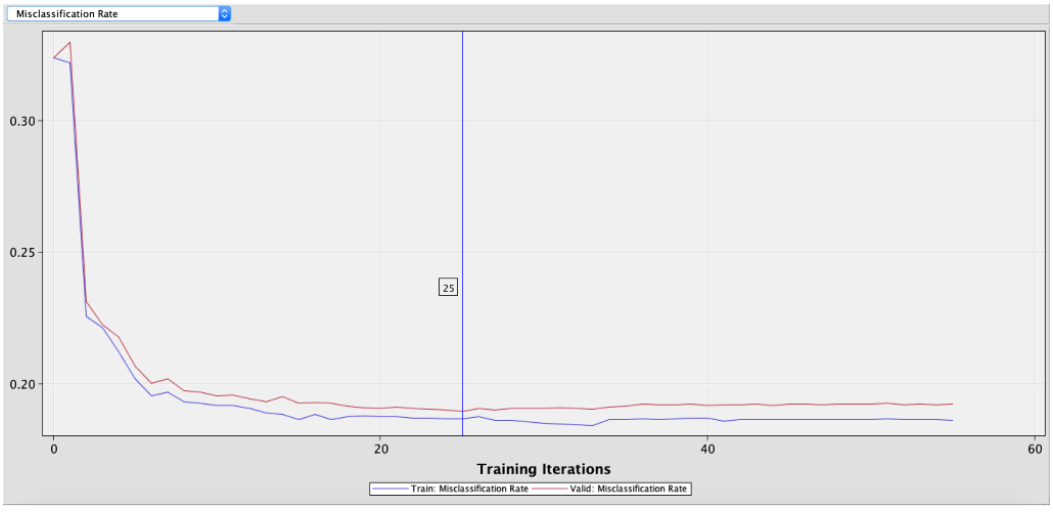


Figure 34: Fit statistics of the neural network

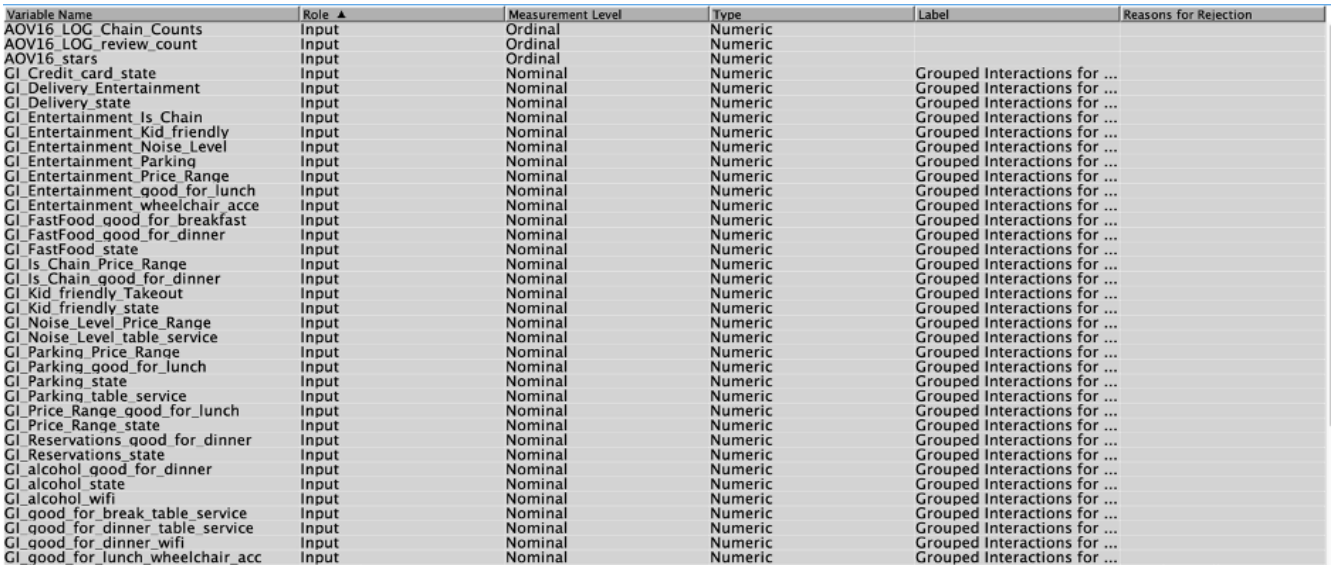
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
is_open	is_open	_DFT_	Total Degrees of Freed...	17652	.	.
is_open	is_open	_DFE_	Degrees of Freedom fo...	17555	.	.
is_open	is_open	_DFM_	Model Degrees of Free...	97	.	.
is_open	is_open	_NW_	Number of Estimated W...	97	.	.
is_open	is_open	_AIC_	Akaike's Information Cr...	14726.02	.	.
is_open	is_open	_SBC_	Schwarz's Bayesian Crit...	15480.54	.	.
is_open	is_open	_ASE_	Average Squared Error	0.131574	0.13456	.
is_open	is_open	_MAX_	Maximum Absolute Error	0.994086	0.995127	.
is_open	is_open	_DIV_	Divisor for ASE	35304	35306	.
is_open	is_open	_NOBS_	Sum of Frequencies	17652	17653	.
is_open	is_open	_RASE_	Root Average Squared ...	0.362732	0.366824	.
is_open	is_open	_SSE_	Sum of Squared Errors	4645.099	4750.76	.
is_open	is_open	_SUMW_	Sum of Case Weights Ti...	35304	35306	.
is_open	is_open	_FPE_	Final Prediction Error	0.133028	.	.
is_open	is_open	_MSE_	Mean Squared Error	0.132301	0.13456	.
is_open	is_open	_RFPE_	Root Final Prediction Er...	0.36473	.	.
is_open	is_open	_RMSE_	Root Mean Squared Error	0.363732	0.366824	.
is_open	is_open	_AVERR_	Average Error Function	0.411625	0.420893	.
is_open	is_open	_ERR_	Error Function	14532.02	14860.05	.
is_open	is_open	_MISC_	Misclassification Rate	0.186721	0.189543	.
is_open	is_open	_WRONG_	Number of Wrong Clas...	3296	3346	.

Stepwise Regression with Neutral Network: The Neural Network node was connected to the Stepwise Regression model in order to see if it will result in a better misclassification rate. However, the result of the Neural Network remained the same.

5.5 VARIABLE SELECTION (AOV16) REGRESSION

Variable selection is a variable reduction technique that can help remove irrelevant variables. When connected to a regression like in our case, it can significantly improve the model's prediction performance. The variable selection node created 3 AOV16 variables and 34 interaction variables.

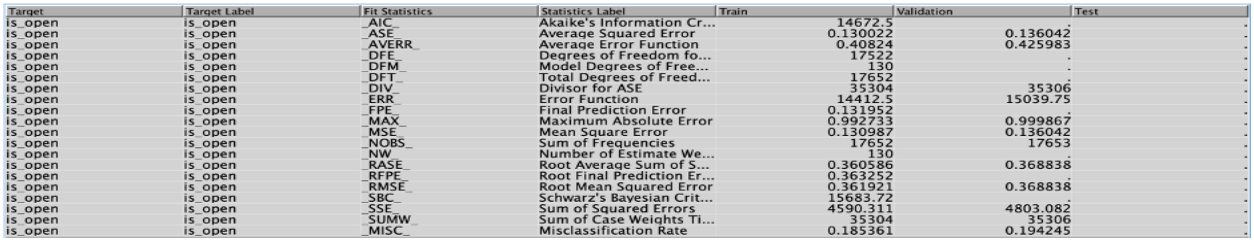
Figure 35: Variables selected by the variable selection node



Variable Name	Role ▲	Measurement Level	Type	Label	Reasons for Rejection
AOV16_LOG_Chain_Counts	Input	Ordinal	Numeric		
AOV16_LOG_review_count	Input	Ordinal	Numeric		
AOV16_stars	Input	Ordinal	Numeric		
GI_Credit_card_state	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Delivery_Entertainment	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Delivery_state	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Entertainment_Is_Chain	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Entertainment_Kid_friendly	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Entertainment_Noise_Level	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Entertainment_Parking	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Entertainment_Price_Range	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Entertainment_good_for_lunch	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Entertainment_wheelchair_acce	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_FastFood_good_for_breakfast	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_FastFood_good_for_dinner	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_FastFood_state	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Is_Chain_Price_Range	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Is_Chain_good_for_dinner	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Kid_friendly_Takeout	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Kid_friendly_state	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Noise_Level_Price_Range	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Noise_Level_table_service	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Parking_Price_Range	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Parking_good_for_lunch	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Parking_state	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Parking_table_service	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Price_Range_good_for_lunch	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Price_Range_state	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Reservations_good_for_dinner	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_Reservations_state	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_alcohol_good_for_dinner	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_alcohol_state	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_alcohol_wifi	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_good_for_break_table_service	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_good_for_dinner_table_service	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_good_for_dinner_wifi	Input	Nominal	Numeric	Grouped Interactions for ...	
GI_good_for_lunch_wheelchair_acc	Input	Nominal	Numeric	Grouped Interactions for ...	

We use AOV16 option to help detect potential nonlinear relationships with the target variable. When activated, this option bin interval variables into 16 equally spaced groups (AOV16). We then connected a regression node to the variables selection node. As a result, we observed a reduction of the misclassification (0.194245) rate compared to the other competing regression.

Figure 36: Fit statistics of the regression



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
is_open	is_open	_AIC_	Akaike's Information Cr...	14672.5		.
is_open	is_open	_ASE_	Average Squared Error	0.130022	0.136042	.
is_open	is_open	_AVERR_	Average Error Function	0.40824	0.425983	.
is_open	is_open	_DFE_	Degrees of Freedom fo...	17522	.	.
is_open	is_open	_DFM_	Model Degrees of Free...	130	.	.
is_open	is_open	_DFT_	Total Degrees of Freed...	17652	.	.
is_open	is_open	_DIV_	Divisor for ASE	35304	35306	.
is_open	is_open	_ERR_	Error Function	14412.5	15039.75	.
is_open	is_open	_FPE_	Final Prediction Error	0.131952	.	.
is_open	is_open	_MAX_	Maximum Absolute Error	0.992733	0.999867	.
is_open	is_open	_MSE_	Mean Square Error	0.130987	0.136042	.
is_open	is_open	_NOBS_	Sum of Frequencies	17652	17653	.
is_open	is_open	_NW_	Number of Estimate We...	130	.	.
is_open	is_open	_RASE_	Root Average Sum of S...	0.360586	0.368838	.
is_open	is_open	_RFPE_	Root Final Prediction Er...	0.363252	.	.
is_open	is_open	_RMSE_	Root Mean Squared Error	0.361921	0.368838	.
is_open	is_open	_SBC_	Schwarz's Bayesian Crit...	15683.72	.	.
is_open	is_open	_SSE_	Sum of Squared Errors	4590.311	4803.082	.
is_open	is_open	_SUMW_	Sum of Case Weights Ti...	35304	35306	.
is_open	is_open	_MISC_	Misclassification Rate	0.185361	0.194245	.

The model had an accuracy of 86.1 percent, which is the best among all the other competing regression models. It also did a great job of predicting true positives with a sensitivity of 89.16, which is an important metric for our purpose.

5.6 LARS, LASSO AND ADAPTIVE LASSO

The LARS node was employed for the use of variable selection, Model-fitting and Prediction. LASSO was used for selection as well but based on a version of ordinary least squares and Adaptive LASSO node was used to apply weights to the parameters in the LASSO constraint. Each node mentioned so far was subsequently connected to a Regression, Neural and Auto Neural node to get the best result. LARS and LASSO models performed the same each time but Adaptive LASSO outperformed them with the AutoNeural node. Of the nine connected nodes, the LARS and LASSO Neural Networks performed best with the same error rate of 19.04%, accuracy of 80.96% and specificity of 63.36%. 67 was selected in their iteration plots which can be seen in Figure 37. The fit statistics can be found in Figure 38.

Figure 37: LARS and LASSO with Neural Networks iteration plot

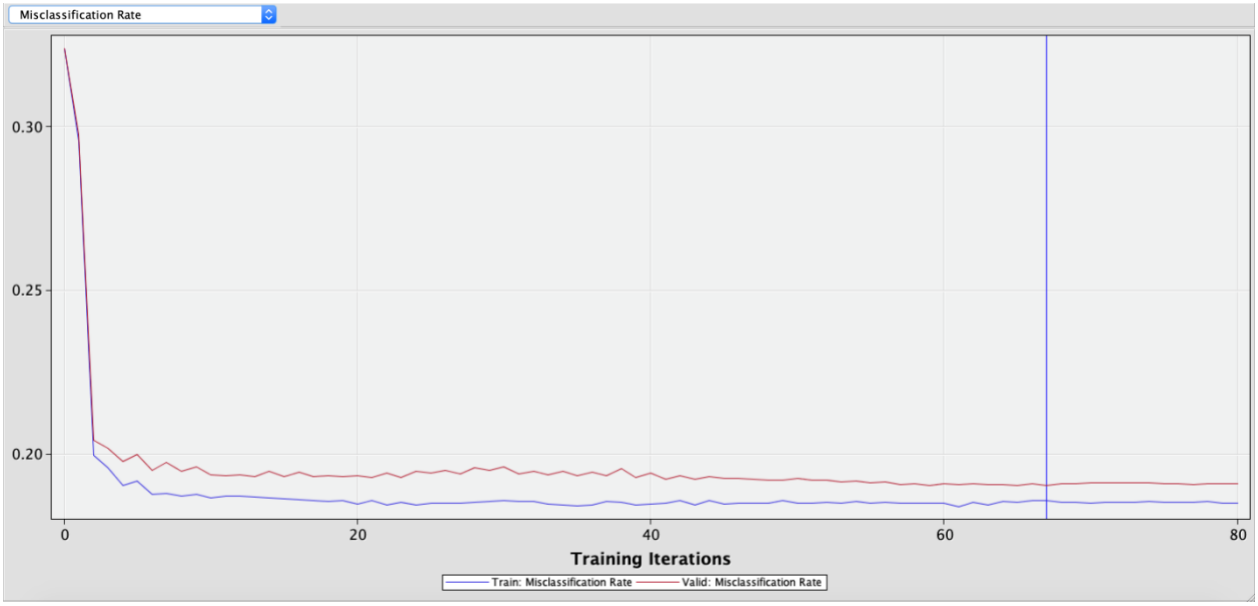


Figure 38: Fit statistics of LARS and LASSO with Neural Network

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
is_open	is_open	_DFT_	Total Degrees of Freed...	17652	.	.
is_open	is_open	_DFE_	Degrees of Freedom fo...	17558	.	.
is_open	is_open	_DFM_	Model Degrees of Free...	94	.	.
is_open	is_open	_NW_	Number of Estimated W...	94	.	.
is_open	is_open	_AIC_	Akaike's Information Cr...	14650.55	.	.
is_open	is_open	_SBC_	Schwarz's Bayesian Crit...	15381.74	.	.
is_open	is_open	_ASE_	Average Squared Error	0.130606	0.135535	.
is_open	is_open	_MAX_	Maximum Absolute Error	0.996353	0.991893	.
is_open	is_open	_DIV_	Divisor for ASE	35304	35306	.
is_open	is_open	_NOBS_	Sum of Frequencies	17652	17653	.
is_open	is_open	_RASE_	Root Average Squared ...	0.361395	0.368151	.
is_open	is_open	_SSE_	Sum of Squared Errors	4610.93	4785.197	.
is_open	is_open	_SUMW_	Sum of Case Weights Ti...	35304	35306	.
is_open	is_open	_FPE_	Final Prediction Error	0.132005	.	.
is_open	is_open	_MSE_	Mean Squared Error	0.131306	0.135535	.
is_open	is_open	_RFPE_	Root Final Prediction Er...	0.363325	.	.
is_open	is_open	_RMSE_	Root Mean Squared Error	0.362361	0.368151	.
is_open	is_open	_AVERR_	Average Error Function	0.409658	0.424159	.
is_open	is_open	_ERR_	Error Function	14462.55	14975.36	.
is_open	is_open	_MISC_	Misclassification Rate	0.185701	0.190393	.
is_open	is_open	_WRONG_	Number of Wrong Clas...	3278	3361	.

5.7 HIGH PERFORMANCE DATA MINING

High-Performance Support Vector Machine: This part of our analysis used the High-Performance Support Vector Machine node (HP SVM). HP SVM supports binary targets and is used for classification and regression tasks. The complexity of its calculations does not depend on the dimension of the input space; therefore, it avoids the curse of dimensionality.

The first step in building our HP SVM was to create four SVM nodes. The first was named HP SVM Linear and set to a Linear Kernel. The second was named HP SVM Poly and set to Active Set Optimization Method with a Polynomial Kernel. The third was named HP SVM RBF and set to Active Set Optimization method as well, with a Radial Basis Function Kernel. The fourth was named HP SVM Sigmoid, also set to Active set but with a Sigmoid Kernel. All other properties remained at their default settings. Of the four, HP SVM Poly performed best, so we connected it to a Selection Tree (ST HP SVM Poly) as its variable selection method. This was done because Selection Tree helps to select inputs for flexible predictive models. It also performed the best out of all variable selection methods, pushing the Selection Tree Neural Network to the top of the model comparison.

As expected, ST HP SVM Poly is the best model with the highest accuracy, ROC index and lowest error rate. It consists of 7363 support vectors, with 6919 of them on the margin. Its validation accuracy is also the highest at 81.24%, with a misclassification error rate of 18.76%, a sensitivity of 90.64%, and 61.64% specificity.

Table 9: Results from the ST HP SVM Poly

Description	Train	Validation
Number of Observations Read	35305.0	NaN
Number of Observations Used	17652.0	17653.0
Number of Input Interval Variables	3.0	NaN
Number of Input Class Variables	20.0	NaN
Number of Input Class Variable Levels	47.0	NaN
Norm of Longest Vector	22.765625	NaN
Number of Support Vectors	7363.0	NaN
Number of Support Vectors on Margin	6919.0	NaN
Maximum F	4.4975127	NaN
Minimum F	-5.435420	NaN
Accuracy	0.830897	0.812440
Error	0.169103	0.187560
Sensitivity	0.920647	0.906394
Specificity	0.643582	0.616434

Table 10: Fit statistics of the ST HP SVM Poly

Target	Fit Statistics	Statistics Label	Train	Validation
is_open	_ASE_	Average Squared Error	0.165817	0.170858
is_open	_DIV_	Divisor for ASE	35304.0	35306.0
is_open	_MAX_	Maximum Absolute Error	0.843299	0.921397
is_open	_NOBS_	Sum of Frequencies	17652.0	17653.0
is_open	_RASE_	Root Average Squared Error	0.407206	0.413350
is_open	_SSE_	Sum of Squared Errors	5854.019275	6032.346959
is_open	_DISF_	Frequency of Classified Cases	17652.0	17653.0
is_open	_MISC_	Misclassification Rate	0.169102	0.187560
is_open	_WRONG_	Number of Wrong Classifications	2985.0	3311.0

High-Performance Forest Larger: HP Forest Larger is an ensemble of classification or regression trees used to overcome the instability a single tree brings. Two HP Forest nodes were connected directly to the data partition node. One was named "HP Forest Default" and set to its default properties, while the other was named "HP Forest Larger" and set to 200 Maximum Number of Trees and 0.8 proportion of observations in each sample. The "HP Forest Larger" model was the better of the two with a validation misclassification rate of 0.193678 while "HP Forest Default" has 0.196341. The "HP Forest Larger" selected all the variables with "Review_count" being the most important variable. The VIP can be seen in Figure 39.

Figure 39: "HP Forest Larger" Variable Importance Plot

Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Valid: Gini Reduction	Valid: Margin Reduction	Label
review_count	981	0.013733	0.027465	0.01043	0.023996	0.008143	0.021858	review_count
state	737	0.005079	0.010158	0.00313	0.008275	0.003452	0.008616	state
Noise_Level	727	0.004336	0.008671	0.00278	0.007134	0.003433	0.007906	Noise_Level
stars	659	0.005057	0.010114	0.00321	0.008238	0.003832	0.008762	stars
Price_Range	646	0.009427	0.018854	0.00857	0.018108	0.008861	0.018353	Price_Range
Chain_Counts	645	0.013470	0.026940	0.01174	0.025447	0.010959	0.023519	Chain_Counts
wheelchair_access	608	0.007651	0.015302	0.00711	0.014697	0.007365	0.014661	wheelchair_access
alcohol	578	0.004442	0.008884	0.00398	0.008395	0.003903	0.008250	alcohol
Delivery	561	0.004249	0.008498	0.00358	0.007806	0.003469	0.007881	Delivery
Kid_friendly	544	0.003509	0.007017	0.00290	0.006371	0.002233	0.005537	Kid_friendly
Reservations	512	0.003211	0.006421	0.00234	0.005496	0.002987	0.006224	Reservations
good_for_dinner	501	0.010949	0.021899	0.01070	0.021641	0.011628	0.022567	good_for_dinner
Entertainment	492	0.019943	0.039886	0.01929	0.039336	0.016824	0.035292	Entertainment
Parking	455	0.001481	0.002962	0.00063	0.002112	0.000918	0.002469	Parking
table_service	423	0.003150	0.006300	0.00268	0.005824	0.002684	0.005703	table_service
good_for_lunch	418	0.014026	0.028053	0.01400	0.028044	0.013581	0.027076	good_for_lunch
wifi	411	0.001088	0.002177	0.00037	0.001482	0.000240	0.001239	wifi
Is_Chain	312	0.012013	0.024026	0.01173	0.023765	0.011343	0.022484	Is_Chain
FastFood	268	0.002894	0.005788	0.00260	0.005497	0.002092	0.004624	FastFood
good_for_breakfast	251	0.001197	0.002393	0.00085	0.001983	0.001107	0.002376	good_for_breakfast
Takeout	242	0.000648	0.001297	0.00026	0.000883	0.000457	0.001118	Takeout
Credit_card	195	0.001068	0.002136	0.00081	0.001907	0.000848	0.001989	Credit_card
happyhour	185	0.000399	0.000798	-0.00006	0.000338	0.000057	0.000479	happyhour

5.8 EVALUATION

The models that were created were compared with the model comparison node to identify the best model. In this project, we found the Selection Tree with High-Performance Support Vector Machine to be the best model. It has the lowest misclassification rate of 0.18756, a ROC index of 86.3%, and 81.24% accuracy. A summarized version of the validation misclassification rate, ROC index, specificity, sensitivity, and accuracy of the top ten models, as well as other selected models with results we found interesting can be found in Table 11. The top ten models are highlighted and the other models are in no particular order.

Sensitivity is the proportion of truly positive cases that were classified as positive and specificity is the proportion of truly negative cases that were classified as negative.

Table 12 shows the classification for these selected models. For this restaurant closure prediction, false positives are the best metric to evaluate the model. The lower the number of false positives, the better the model is. False positive is when the model predicts that a restaurant will remain open even though it is closed.

Table 11: Fit statistics of the selected models

Model	Validation Misclassification Rate	ROC Index	Accuracy (%)	Sensitivity (%)	Specificity (%)
Selection Tree HP SVM Poly	0.18756	0.863	81.24	90.64	61.64
Selection Tree with Neural Network	0.189543	0.863	81.05	90.39	61.56
Stepwise Regression with Neural Network	0.189543	0.863	81.05	90.39	61.56
Neural Network	0.189543	0.863	81.05	90.39	61.56
LASSO with Neural Network	0.190393	0.861	80.96	89.40	63.36
LARS with Neural Network	0.190393	0.861	80.96	89.40	63.36
Adaptive LASSO with Neural Network	0.190732	0.863	80.93	89.63	62.78
Variable selection with Auto Neural (AOV 16)	0.193112	0.859	80.69	88.54	64.32
Variable selection with Neural Network (AOV 16)	0.193225	0.861	80.68	89.16	62.99
HP Forest Larger	0.193678	0.857	80.63	92.65	55.56
Variable Selection with Regression AOV16	0.194245	0.86	80.58	89.12	62.74
Stepwise Misclassification	0.199909	0.849	80.01	88.84	61.59

Regression					
Decision Tree	0.225287	0.799	77.47	85.61	60.49
HP SVM Linear	0.198833	0.847	80.12	89.42	60.72
HP SVM (RBF)	0.199739	0.83	80.02	90.63	57.90
HP Forest Default	0.196341	0.856	80.37	92.45	55.16
HP SVM Sigmoid	0.37767	0.396	62.23	88.44	7.55
LARS with Regression	0.200136	0.849	79.99	88.83	61.54
LASSO with Regression	0.200136	0.849	79.99	88.83	61.54
Adaptive LASSO with Regression	0.200816	0.849	79.92	88.75	61.49
LARS with Auto Neural	0.199173	0.848	80.08	89.78	59.86
LASSO with Auto Neural	0.199173	0.848	80.08	89.78	59.86
Adaptive LASSO Auto Neural	0.197304	0.849	80.27	89.79	60.40

Table 12: Event classification of the selected models

Model	False Negative	True Negative	False Positive	True Positive
Selection Tree with HP SVM Poly	1117	3526	2194	10816
Selection Tree with Neural Network	1147	3521	2199	10786
Stepwise Regression with Neural Network	1147	3521	2199	10786
Neural Network	1147	3521	2199	10786
LASSO with Neural Network	1265	3624	2096	10668
LARS with Neural Network	1265	3624	2096	10668
Adaptive LASSO with Neural Network	1238	3591	2129	10695
Variable selection with Auto Neural (AOV 16)	1368	3679	2041	10565
Variable selection with Neural Network (AOV 16)	1294	3603	2117	10639
HP Forest larger	877	3178	2542	11056
Variable Selection with Regression AOV16	1298	3589	2131	10635
Stepwise Misclassification Regression	1332	3523	2197	10601
Decision Tree	1717	3460	2260	10216
HP SVM Linear	1263	3472	2247	10670
HP SVM (RBF)	1118	3312	2408	10815
HP Forest Default	901	3155	2565	11032

HP SVM Sigmoid	1379	432	5288	10554
LARS with Regression	1333	3520	2200	10600
LASSO with Regression	1333	3520	2200	10600
Adaptive LASSO with Regression	1342	3517	2203	10591
LARS with Auto Neural	1220	3424	2296	10713
LASSO with Auto Neural	1220	3424	2296	10713
Adaptive LASSO with Auto Neural	1218	3455	2265	10715

5.9 SCORE DATA

The contribution of SAS Enterprise Miner to model implementation is a scoring recipe that is capable of adding predictions to any data set structured in a manner similar to the training data.

After training and comparing predictive models, the best model, Selection Tree HP SVM Poly is selected to represent the association between the inputs and the target.

We used the 2018 Yelp dataset challenge for scoring. From figure 40 and 41, we can see that the model performed fairly well.

Figure 40: Summary statistics for Score data

Class Variable Summary Statistics

Data Role=SCORE Output Type=CLASSIFICATION

Variable	Numeric Value	Formatted Value	Frequency Count	Percent
I_is_open	.	0	11379	34.3673
I_is_open	.	1	21731	65.6327

Figure 41: Summary statistics for Validated data

Data Role=VALIDATE Output Type=CLASSIFICATION

Variable	Numeric Value	Formatted Value	Frequency Count	Percent
I_is_open	.	0	4643	26.3015
I_is_open	.	1	13010	73.6985

Figure 42: Actual data target distribution

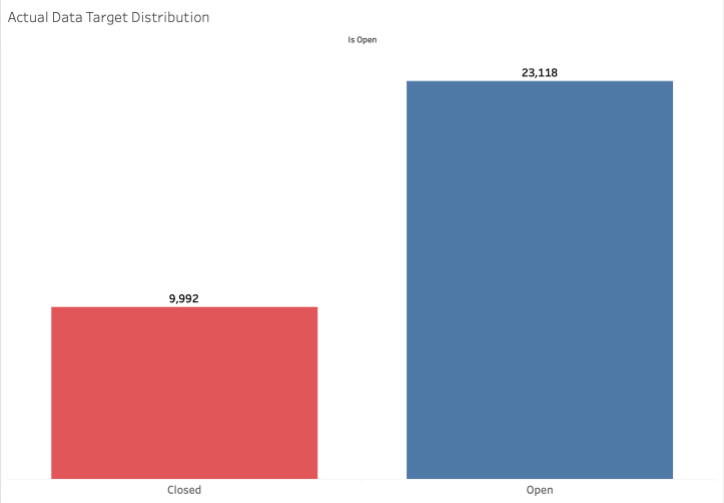
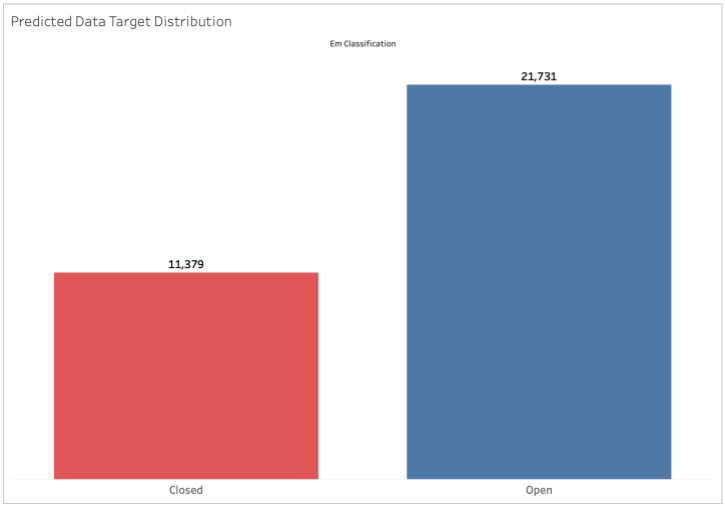


Figure 43: Predicted data target distribution



CHAPTER SIX

CONCLUSIONS, DISCUSSIONS AND RECOMMENDATIONS

6.1 LIMITATIONS OF THE STUDY

The original dataset is limited in the type of variables it encompasses. It has more text and binary variables than it has quantitative variables. Hence, we converted some text to binary to increase the number of variables in our analysis.

Also, during exploration we considered examining other restaurant types that were not categorized as fast food (such as mom-and-pop shops). As a team, we decided that would not be beneficial to the overall project because there are over 600 different types of restaurants. Focusing on a few levels would skew the data and cause a bias towards one or more categories.

We avoided this issue with the 'Entertainment' variable by grouping multiple attributes into one single entertainment category. Splitting all the different attributes would have caused the column to have a bias or too many levels.

A similar limitation occurred while we extracted information for the 'Ethnicity' column. Information in the Yelp dataset is user imputed which can lead to human error. We discovered not every restaurant was labeled according to its restaurant type (i.e. not all "American" restaurants were labeled as American). Restaurants that should not have been categorized as "Other" were incorrectly labeled causing a bias in the data. Once we determined this was a limitation, we rejected the 'Ethnicity' column and did not continue the exploration of other restaurant types.

Furthermore, customers are more likely to leave a review if they experienced a bad service from a business. This extremity may result in the unreliability of some user generated variables in the dataset or the production of false positives and false negatives. To prove this point, a study

conducted by dimensional research in 2013 showed that 95% of customers share bad experiences while a lesser 87% share good experiences with others (Dimensional Research, 2013).

6.2 CONCLUSION AND RECOMMENDATIONS

The best model for predicting restaurant closure is Selection Tree with High-Performance Support Vector Machine Poly. Based on the Selection Tree Variable Importance Plot (VIP), the five most important variables for predicting restaurant closure include Review Counts, Chain Counts, Entertainment, Is_Chain, and Good_for_dinner.

Therefore, we have the following recommendations:

1. Independent restaurants at risk of closing can expand their business to grow in different locations or join a franchise to avoid closure. If this is not feasible, restaurants can also adopt the following recommendations.
2. Restaurants should provide entertainment such as background music, live music, and T.V. or improve the existing ones. Experiences such as background music are statistically significant predictors of satisfaction and repeat patronage, which contributes to restaurant success (DiPietro, 2016).
3. Promotions could be put in place to encourage customers to leave reviews on Yelp after visiting the restaurant. A restaurant owner who gave 50% off to customers to give him 1 star yelp reviews in a bid to protest against Yelps rating system, succeeded in attracting more customers as a result of this promotion and increase in the number of Yelp reviews ("The restaurant owner who asked for 1-star Yelp reviews", 2020).
4. Restaurants that experience low traffic at dinner time and are considered "not good for dinner" can tailor their menu options to best suit customer needs.

REFERENCES

Assaf, A. George, Margaret Deery, and Leo Jago. "Evaluating the Performance and Scale Characteristics of the Australian Restaurant Industry." Journal of Hospitality & Tourism Research 35.4 (2011): 419-36. Web.

Ching, M.R., & Bulos, R.D. (2019). Improving Restaurants' Business Performance Using Yelp Data Sets through Sentiment Analysis. In Proceedings of the 2019 3rd International Conference on E-commerce, E-Business and E-Government (ICEEG 2019). Association for Computing Machinery, New York, NY, USA, 62–67. DOI:https://doi.org/10.1145/3340017.3340018

Chow, H. S., Lau, V. P., Lo, W. C., Sha, Z., Yun, H. (2007). Service quality in restaurant operations in China: decision-and experiential-oriented perspectives. International Journal of Hospitality Management, 26(3), 698-710

Dimensional Research. (2013, April). Customer Service and Business Results: A Survey of Customer Service From Mid-Size Companies [PDF File]. Retrieved from https://d16cvnquvjw7pr.cloudfront.net/resources/whitepapers/Zendesk_WP_Customer_Service_and_Business_Results.pdf

DiPietro, R. (2016). Restaurant and foodservice research. International Journal of Contemporary Hospitality Management, 29(4), 1203-1234. doi:http://dx.doi.org.vortex3.uco.edu/10.1108/IJCHM-01-2016-0046

Dutta, K., Venkatesh, U., & Parsa, H.G. (2007). Service failure and recovery strategies in the restaurant sector: An indo-US comparative study. International Journal of Contemporary

Hospitality Management, 19(5), 351-363. doi:http://dx.doi.org.vortex3.uco.edu/10.1108/09596110710757526

Feng, J., Kitade, N., & Ritterz, M. (2015, March 18). Dartmouth College. Retrieved March 20, 2020, from https://www.cs.dartmouth.edu/~lorenzo/teaching/cs174/Archive/ Winter2015/Projects/fin als/fkr.pdf

Gagić, S., Tešanović, D., & Jovičić, A. (2013). The vital components of restaurant quality that affect guest satisfaction. Turizam, 17(4), 166-176.

Jin, G. Z., & Leslie, P. (2009). Reputational incentives for restaurant hygiene. American Economic Journal.Microeconomics, 1(1), 237-267. doi:http://dx.doi.org.vortex3.uco.edu/10.1257/mic.1.1.237

Kong, A., Nguyen, V., & Xu, C. (2016). Predicting International Restaurant Success with Yelp. Retrieved from http://cs229.stanford.edu/proj2016spr/report/062.pdf

Lian, J., Zhang, F., Xie,X., & Sun, G. (2017). Restaurant Survival Analysis with Heterogeneous Information. Retrieved from https://www.semanticscholar.org/paper/Restaurant-Survival-Analysis-with-Heterogeneous-Lian-Zhang/a85590325ebe7abba 2a6c3915111a783cce8cc

Lock, S. (2020). Food delivery industry in the U.S. - Statistics & Facts. Statista. Retrieved, from https://www.statista.com/topics/1986/food-delivery-industry-in-the-us/.

Lu, X., Qu, J., Jiang, Y., & Zhao, Y. (2018). Should I Invest it? Predicting Future Success of Yelp Restaurants. Retrieved from https://jiamingqu.com/files/Yelp%20Prediction.pdf

Luca, M. (2016). Reviews, reputation, and revenue: The case of yelp.com. St. Louis: Federal Reserve Bank of St Louis. Retrieved from

http://vortex3.uco.edu/login?url=https://search-proquest com.vortex3.uco.edu/docview/1698415032accountid=14516

Luo, X., Homburg, C. (2007). Neglected Outcomes of Customer Satisfaction. Journal of Marketing, 71(2), 133-149.

Mao, Z. (2006). Investigation of the relationship between firm -wise financial factors and firm performance in the hospitality industry (Order No. 3244001). Available from ProQuest Central; ProQuest Dissertations & Theses Global. (304963288). Retrieved from http://vortex3.uco.edu/login?url=https://search-proquest-com.vortex3.uco.edu/docview/304963288?accountid=14516

McDonnell, S., 2020. What Percentage of Sales Are from Drive Through Windows at Fast Food Restaurants? [online] Small Business - Chron.com. Available at: <https://smallbusiness.chron.com/percentage-sales-drive-through-windows-fast-food-restaurants-75713.html>

McLynn, K. (2018, August 22). Total US Restaurant Count Stands At 660,755 in Spring 2018. Retrieved March 21, 2020, from https://www.npd.com/wps/portal/npd/us/news/press-releases/2018/total-us-restaurant-count-at-660755-in-spring-2018-a-one-percent-drop-from-last-year-reports-npd/

Namkung, Y. and Jang, S. (2008), "Are highly satisfied restaurant customers really different? A quality perception perspective", International Journal of Contemporary Hospitality Management, Vol. 20 No. 2, pp. 142-155.

National Restaurant Association. (2020). 2020 National Statistics: Restaurant Industry Facts at a Glance. Retrieved from https://restaurant.org/research/restaurant-statistics/restaurant-industry-facts-at-a-glance

Parikh, A., Behnke, C., Vorvoreanu, M., Almanza, B., & Nelson, D. (2014). Motives for reading and articulating user-generated restaurant reviews on yelp.com. Journal of Hospitality and Tourism Technology, 5(2), 160-176. doi:http://dx.doi.org/10.1108/JHTT-04-2013-0011

Ozdemir, V.E. and Hewett, K. (2010) 'The effect of collectivism on the importance of relationship quality and service quality for behavioral intentions: A cross-national and cross-contextual analysis', Journal of International Marketing, 18:1, pp.41-62.

Parsa, H. G., Self, J. T., Njite, D., & King, T. (2005). Why restaurants fail. Cornell Hotel and Restaurant Administration Quarterly, 46(3), 304-322. Retrieved from http://vortex3.uco.edu/login?url=https://search-proquest-com.vortex3.uco.edu/docview/209705978?accountid=14516

Parsa, H. G., van der Rest, J.,I., Smith, S. R., Parsa, R. A., & Bujisic, M. (2015). Why restaurants fail? part IV: The relationship between restaurant failures and demographic factors. Cornell Hospitality Quarterly, 56(1), 80. Retrieved from http://vortex3.uco.edu/login?url=https://search-proquest com.vortex3.uco.edu/docview/1644489819?accountid=14516

Severson, K., & Yaffe-Bellany, D. (2020). Independent Restaurants Brace for the Unknown. Nytimes.com. Retrieved 16 July 2020, from https://www.nytimes.com/2020/03/20/dining/local-restaurants-coronavirus.html.

Shellenberger, P. (2017), "Predicting Yelp Food Establishment Ratings Based on Business Attributes". Honors Theses and Capstones. 374. Retrieved from https://scholars.unh.edu/honors/374

Snow, D., (2018). Predicting Global Restaurant Facility Closures. The Alan Turing Institute; New York University (NYU) - Finance and Risk Engineering Department; University of Auckland. https://ssrn.com/abstract=3420490 or http://dx.doi.org/10.2139/ssrn.3420490

Tripathi, G., & Dave, K. (2016). Assessing The Impact Of Restaurant Service Quality Dimensions On Customer Satisfaction And Behavioral Intentions. Journal of Services Research, 16(1), 13-39. Retrieved from http://vortex3.uco.edu/login?url=https://search-proquest-com.vortex3.uco.edu/docview/1890207041?accountid

Wrulich, M., Hirschberg, C., Rajko, A., & Schumacher, T. (2020). McKinsey. McKinsey&Company. Retrieved from https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-changing-market-for-food-delivery#.

APPENDIX

Yelp Dataset Code

```
#Import necessary libraries and packages
import pandas as pd
import numpy as np
from pandas.io.json import json_normalize
import matplotlib.pyplot as plt
import seaborn as sns

#Import primary CSV data file
business_original = pd.read_csv('business.csv')

#First view of data
business_original.head(15)

#Determine the actual span of our data
business_original.shape

#Unique cities in dataset
business_original['city'].nunique()

#Assess the data types for ease of analysis
business_original.info()

#Check for duplicate records in the dataset
business_original.duplicated().sum()

#Check for Null values
business_original.isnull().sum()

#check what percentage of hours is null
(business_original.hours.isnull().sum()/len(business_original))*100

#hours is not really useful and has a bad format so we drop it
business_original.drop('hours', axis=1, inplace=True)

#updated span/size of dataset
business_original.shape

#Analyze the state column to determine which states will be of use
business_original['state'].value_counts()

#Visulaize the distribution above
ax = business_original['state'].value_counts()
```

```
ax.plot.bar(figsize = (16,4), title="Count of Business Records for each State")

#Graph new order of states
filt = ['AZ','NV','NC','OH','PA']
state_filt= business_original['state'].isin(filt)
graph=business_original[state_filt]

ax_1 = graph['state'].value_counts()
ax_1.plot.bar(figsize = (16,4), title="Count of Business Records for each State")

#Hence, filter needs only relevant states
filt1 = ['AZ','NV','NC','OH','PA']
state_filt1= business_original['state'].isin(filt1)
business = business_original[state_filt1]
business.head()

business['state'].value_counts()

#How many records do we have left to work with?
business.shape

#Begin exploration of categories
#Check for null values
business['categories'].isnull().sum()

#Replace null values
business["categories"].fillna("",inplace=True)

#Reset index and drop unneccessry columns
business=business.reset_index().drop(columns=['Unnamed: 0','index'])

#Filter out only records that fall into important categories
targets = ['Restaurants', 'Fast
Food','Shopping','Beauty','Spa','Nightlife','Auto', 'Arts','Entertainment','Active Life']
business=business[business.categories.str.contains('|'.join(targets))]

#What do we have left?
business.shape

#CREATE FUNCTION TO SINGLE OUT AREA OF PRIMARY INTEREST FOR ANALYSIS
def Restaurant(x):
    if ('restaurants' in x.lower()) or ('fast food' in x.lower()) or ('restaurant' in x.lower()):
      return 1
    else:
      return 0
```

```
business["Restaurant"] = business["categories"].apply(Restaurant)
business[["categories","Restaurant"]].head(10)

business["Restaurant"].sum()
```

Extracting Attributes

```
#Expand attributes columns by splitting and create dummy variables
business["attributes"]=business["attributes"].str.replace("{","")
business["attributes"]=business["attributes"].str.replace("}","")
business["attributes"]=business["attributes"].str.replace("'","")
business["attributes"]=business["attributes"].str.replace("","")
business["attributes"]=business["attributes"].astype(str)
pd.set_option('display.max_columns', 50)
business.head()

#Create Parking variable
def Parking(x):
    if ('valet: True' in x) or ('garage: True' in x) or ('lot: True' in x):
        return 1
    else:
        return 0

business['Parking']=business['attributes'].apply(Parking)

#Create Kid_friendly variable
def Kid_friendly(x):
    if 'GoodForKids: True' in x:
        return 1
    else:
        return 0

business['Kid_friendly']=business['attributes'].apply(Kid_friendly)

#Create Reservations variable
def Reservations(x):
    if 'RestaurantsReservations: True' in x:
        return 1
    else:
        return 0

business['Reservations'] = business['attributes'].apply(Reservations)

#Create Price range variable
def Price_Range(x):
```

```
    if 'RestaurantsPriceRange2: 1' in x:
        return 1
    elif 'RestaurantsPriceRange2: 2' in x:
        return 2
    elif 'RestaurantsPriceRange2: 3' in x:
        return 3
    else:
        return 4

business['Price_Range'] = business['attributes'].apply(Price_Range)

#Create creditcard variable
def Credit_card(x):
    if "BusinessAcceptsCreditCards: True" in x:
        return 1
    else:
        return 0

business['Credit_card'] = business['attributes'].apply(Credit_card)

#Create wheelchair access variable
def wheelchair_access(x):
    if 'WheelchairAccessible: True' in x:
        return 1
    else:
        return 0

business['wheelchair_access'] = business['attributes'].apply(wheelchair_access)

#Create breakfast variable
def good_for_breakfast (x):
    if 'breakfast: True' in x:
        return 1
    else:
        return 0

business['good_for_breakfast'] = business['attributes'].apply(good_for_breakfast)

#Create lunch variable
def good_for_lunch (x):
    if 'lunch: True' in x:
        return 1
    else:
        return 0

business['good_for_lunch'] = business['attributes'].apply(good_for_lunch)
```

```
#Create dinner variable
def good_for_dinner (x):
    if 'dinner: True' in x:
        return 1
    else:
        return 0

business['good_for_dinner'] = business['attributes'].apply(good_for_dinner)

#Create alcohol variable
def alcohol (x):
    if ('Alcohol: ufull_bar' in x) or ('Alcohol: ubeer_and_wine' in x):
        return 1
    else:
        return 0

business['alcohol'] = business['attributes'].apply(alcohol)

#Create happyhour variable
def happyhour (x):
    if 'HappyHour: True' in x :
        return 1
    else:
        return 0

business['happyhour'] = business['attributes'].apply(happyhour)

#Create wifi variable
def wifi (x):
    if ('WiFi: ufree' in x) or ('WiFi: free' in x) or ('WiFi: yes' in x) or ('WiFi:
uyes' in x) or ('WiFi: True' in x) or ('WiFi: uTrue' in x):
        return 1
    else:
        return 0

business['wifi'] = business['attributes'].apply(wifi)

#Create table service variable
def table_service (x):
    if 'RestaurantsTableService: True' in x :
        return 1
    else:
        return 0

business['table_service'] = business['attributes'].apply(table_service)
```

```
#Create Entertainment
def Entertainment (x):
    if ('HasTV: True' in x) or ('dj: True' in x) or ('background_music: True' in x) or ('jukebox:
True' in x) or ('live: True' in x) or ('video: True' in x) or ('karaoke: True' in x):
        return 1
    else:
        return 0

business['Entertainment'] = business['attributes'].apply(Entertainment)

#Create takeout variable
def takeout (x):
    if 'RestaurantsTakeOut: True' in x :
        return 1
    else:
        return 0

business['Takeout'] = business['attributes'].apply(takeout)

#Create Noise_Level variable

def Noise_Level(x):
    if ('NoiseLevel: uquiet' in x) or ('NoiseLevel: quiet' in x):
        return 1
    elif ('NoiseLevel: uaverage' in x) or ('NoiseLevel: average' in x):
        return 2
    elif ('NoiseLevel: uloud' in x) or ('NoiseLevel: loud' in x):
        return 3
    else:
        return 4

business['Noise_Level'] = business['attributes'].apply(Noise_Level)

#Create Reservations variable

def Reservations (x):
    if 'RestaurantsReservations: True' in x :
        return 1
    else:
        return 0

business['Reservations'] = business['attributes'].apply(Reservations)

#Create Delivery variable
```

```
def Delivery (x):
    if 'RestaurantsDelivery: True' in x :
        return 1
    else:
        return 0

business['Delivery'] = business['attributes'].apply(Delivery)
```

Extracting Categories

```
#Create FastFood variable
def FastFood (x):
    if 'Fast Food' in x :
        return 1
    else:
        return 0

business['FastFood'] = business['categories'].apply(FastFood)

#Create Ethnicity variable
def ethnicity (x):
    if ('american' in x.lower()) or ('burgers' in x.lower()):
        return 'American'
    elif 'chinese' in x.lower():
        return 'Chinese'
    elif ('mexican' in x.lower()) or ("tex-mex"in x.lower()):
        return 'Mexican'
    elif 'italian' in x.lower():
        return 'Italian'
    elif ('japanese' in x.lower()) or ('sushi' in x.lower()):
        return 'Japanese'
   # elif 'thai' in x.lower():
      #  return 'Thai'
   # elif 'indian' in x.lower():
      #  return 'Indian'
   # elif 'korean' in x.lower():
      #  return 'Korean'
    else:
        return 'other'

business['Ethnicity'] = business['categories'].apply(ethnicity)

#Remove foreign symbols from name to allow for counting chains
business["name"]=business["name"].str.replace(' ','')
business["name"]=business["name"].str.replace("'","")
```

```
business["name"]=business["name"].str.replace(',',"")
business["name"]=business["name"].str.replace('.',"")

business["name"]=business["name"].astype(str)
business["name"]=business["name"].str.lower()

#Select only restaurants for data analysis before chain is counted
Rest_filt= business["Restaurant"]==1
Restaurant=business[Rest_filt]
Restaurant.head(10)

#Create chain counts column by counting occurence of names
Restaurant['Chain_Counts'] = Restaurant.groupby(['name'])['name'].transform('count')

#Declare chain if chain counts is 4 or more.
def Chain (x):
    if x >= 4 :
        return 1
    else:
        return 0

#Create Is_Chain column
Restaurant['Is_Chain'] = Restaurant['Chain_Counts'].apply(Chain)

#Drop longitude and latitude since they are not needed
Restaurant.drop(columns=['longitude','latitude'], inplace=True)

#Confirm shape of DF
Restaurant.shape

#Check for number of Open restaurants
Restaurant['is_open'].sum()

#Check for number of Closed restaurants
len(Restaurant['is_open'])-(Restaurant['is_open'].sum())

#Check again for null values
Restaurant.isnull().sum()

#Make pie chart to show distribution of open and closed businesses'

# Pie chart
labels = ["Open", 'Closed']
sizes = [23867, 11438]
#colors
colors = ['Lime','Red']
```

```
fig1, ax1 = plt.subplots(figsize=(10,5))
fig1.subplots_adjust(0.3,0,1,1)
patches, texts, autotexts = ax1.pie(sizes, colors = colors, labels=labels, autopct='%1.1f%%', start
angle=90)
for text in texts:
    text.set_color('black')
    text.set_size(12)
for autotext in autotexts:
    autotext.set_color('black')
    autotext.set_size(14)

# Equal aspect ratio ensures that pie is drawn as a circle
ax1.axis('equal')
plt.tight_layout()
plt.show()

Restaurant.state.value_counts()

Restaurant.postal_code.value_counts() #Reject

#Check for ethnicity distribution
#Looks very skewed so it may not be used. There are 600 levels. This does not seem feasible for
analysis within this time frame.
Restaurant.Ethnicity.value_counts()
Restaurant.head()
```