Final Report:

Developing a Predictive Model for Restaurant Closure:

Leveraging Yelp Reviews with Machine Learning.

Achalshail Khadka , Babah Sesay , Pravallika Vasantham , Smit Patel

Department of Business, University of Central Oklahoma

MSBA 5404 - Predictive Analytics and AI

Dr. Ho-Chang Chae, Ph.D Associate Professor

Table of Contents:

Executive Summary:

In the competitive landscape of the restaurant industry, understanding the factors that contribute to restaurant closures is crucial for survival and strategic planning. This project leverages extensive Yelp data through advanced machine learning techniques to uncover these critical determinants. Our analysis explored various attributes beyond the commonly examined factors like reviews and ratings. We introduce new variables such as restaurant delivery, restaurant attire, ambience (classy), drive-thru availability, noise level, and alcohol service, providing a comprehensive view of what influences restaurant longevity.

Our findings reveal that while reviews and ratings continue to play significant roles, the availability of delivery services and the type of dining ambiance significantly impact a restaurant's risk of closing. For instance, restaurants offering delivery services or maintaining a calm, more upscale dining atmosphere show lower closure rates, suggesting that adaptability to customer preferences and dining trends is key to sustainability.

Additionally, our study highlights the importance of location and restaurant type (e.g., drive-thru), which align with changing consumer behaviors, especially under the constraints brought on by recent global events like the COVID-19 pandemic. The ability to quickly adapt to external pressures and evolving customer expectations emerges as a central theme in sustaining restaurant operations.

In conclusion, this project not only advances the understanding of restaurant closures through a data-driven approach but also equips restaurant owners and managers with actionable insights. These insights can guide strategic decisions, from service offerings to marketing strategies, ultimately aiding in the reduction of closure risks and fostering business resilience in the volatile restaurant industry.

Introduction/Business Problem (Phase 1 of CRISP DM methodology):

The restaurant industry, as estimated by the National Restaurant Association (2024), is projected to surpass $1 trillion in revenue in the United States. This sector contributes significantly to the national economy while providing millions of jobs to Americans. One of the most commonly used platforms for restaurant reviews is Yelp, which covers a range of businesses, including restaurants. According to Yahoo Finance, Yelp's revenue reached a record $1.34 billion, reflecting a 12% year-over-year increase. Given the importance of Yelp data, it is crucial for future analyses regarding restaurant survival. Our project aims to leverage Yelp data to analyze factors leading to restaurant closures between 2005 and 2022 in Philadelphia. The dataset includes reviews and features from both pre- and post-COVID periods and recession period of 2008. The primary target variable for our predictive modeling and text analysis using SAS Enterprise Miner is "is_close," which indicates whether a restaurant has closed or not. Our project closely resembles previous research by Lu X. et al. (2018) but fills a gap in the literature by categorizing non-text restaurant features into areas such as accessibility (e.g., Wi-Fi, drive-thru, pet friendliness) and amenities (e.g., delivery, takeout, catering, reservations, table service, counter service). Additionally, we focus on the "is_close" variable instead of the original "is_open" variable, as our interest lies in identifying the factors most associated with restaurant closures. To address missing data in the raw Yelp dataset, we have imputed missing values with "False" or "0," which, based on our literature review, has not been commonly addressed.

Literature Review:

The restaurant industry's volatility is notable, with many businesses struggling to sustain long-term operations. Recent academic efforts, like those of Chen and Xia (2020) and Pandian and Aggarwal (2015), have used machine learning tools to predict restaurant survival rates, while leveraging on the Yelp platform which offers rich datasets of consumer feedback. These studies primarily focus on how consumer reviews can predict the longevity of restaurants, with varying methods ranging from ensemble models to deep learning techniques.

While these studies have provided foundational insights, they exhibit gaps, particularly in addressing the nuanced impacts of operational and non-textual features of restaurants. Current models do not sufficiently consider how attributes such as location, service type, and amenity availability (like Wi-Fi or pet-friendliness) influence a restaurant's survival chances (Zhou et al.,2020). Moreover, temporal and geographic specificity are often overlooked, limiting the generalizability of the findings to broader or different demographic contexts (Kim et al.,2019).

Our project aims to fill these literature gaps by analyzing Yelp data with a refined methodological framework using SAS Enterprise Miner. We enhance existing models by focusing on 'is_close' as our target variable instead of 'is_open'. This approach allows for a direct investigation into the factors leading to restaurant closures. We also categorize restaurant features into distinct groups—accessibility, amenities, and dining services—to determine their differential impacts on restaurant closure risks. Unlike previous models that employed generic machine learning frameworks, our methodology incorporates the LARS and LASSO regression technique, which aids in variable selection and enhances interpretability—a crucial aspect for stakeholders who need to understand the 'why' behind predictive insights (Sedov,2021).

By incorporating a broad array of variables including recent reviews and distinguishing between different types of services offered, our analysis promises not only to predict closures but also to provide actionable insights for restaurant owners to mitigate risks. Our approach also responds to the dynamic nature of the restaurant industry, highlighted by the rapid changes in consumer preferences and the economic impact of events like the COVID-19 pandemic (Lu et al., 2018).

In conclusion, our research extends the scope of existing literature by providing a comprehensive analysis that integrates both textual and non-textual data, thereby offering a more detailed understanding of the factors that contribute to restaurant closures. This effort not only bridges the current academic gaps but also enhances the practical utility of predictive models in the restaurant industry.

Table 1: Table of literature:

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
1.	Yifan Chen; Fanzeng Xia(2020)	Target variable is the classification of businesses as "open" or "closed" within a specific time interval. Important variables include aspect-wise ratings extracted from reviews (xbf), category and region vectors representing trends over time (xcvt, xrvt), and business variable vectors capturing various attributes (Dbv).	The methods entail gathering business data from the Yelp dataset, extracting features to represent category and regional trends, constructing a hybrid deep neural network model comprising CNN and DNN components, training the model using cross-entropy loss and transfer learning, and evaluating performance through k-fold cross-validation and hyperparameter optimization.	The models used include traditional machine learning methods such as logistic regression and SVM, as well as a hybrid deep neural network model. The hybrid deep neural network model performed well compared to traditional machine learning methods. The outcome is that the hybrid deep neural network model showed superior performance in predicting business closures, with aspect-wise ratings significantly influencing prediction accuracy. Transfer learning, hyperparameter tuning, and dropout regularization were used to enhance model generalization and mitigate overfitting, ensuring robustness. The model's ability to capture local patterns and regional trends proved pivotal for accurate predictions of business closures, demonstrating its potential for real-world applications.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
2	Yi Luo 1 and Xiaowei Xu 2(2019)	Target variable: usefulness of review (measured by the number of votes on "useful" specific to each individual review) Important variables: Type of restaurant, review date, star rating of individual reviewers for the restaurant) Important Variables: Type of restaurant, Review date, star rating of individual reviewers for the restaurant Other Important Variables : elite status of consumer, city(Las Vegas, Los Angeles, New York)	Python code was used to create a web crawler that collected Yelp.com reviews between October 10 and 16, 2018. Based on TripAdvisor ratings, reviews from the top three American cities—New York, Los Angeles, and Las Vegas—were chosen. For privacy, identifying information was eliminated. Four restaurant features were identified by the application of LDA (Latent Dirichlet Allocation): taste/food, experience, value, and location. To ascertain the sentiment connected to each element, sentiment analysis was done. In order to predict review usefulness, a model comparison was carried out using the Support Vector Machine (SVM) and the Fuzzy Domain Ontology (FDO) algorithm.	The models used include LDA (Latent Dirichlet Allocation) for identifying primary characteristics of Yelp reviews and SVM (Support Vector Machine) with the FDO (Fractional Direct Optimization) algorithm for predicting review usefulness.The SVM with the FDO algorithm performed well, achieving the highest prediction accuracy for review usefulness. The outcome includes identifying primary characteristics of Yelp reviews (location, value, experience, taste/food) through LDA analysis and determining that SVM with the FDO algorithm produced the highest prediction accuracy for review usefulness, with food quality being revealed as the most significant factor in customer reviews.
3	Narenkumar Pandian and Vaibhav Aggarwal(2015)	Target variable: Business "open" or "shutdown" Important variables: Rating, hours, category, tip. and users.	The methods employed include Naive Bayes, Logistic Regression, SVM, and EM Algorithm for binary classification of businesses as 'open' or 'shutdown', along with novel ensemble algorithms EMSVM and EMLOG. Precision-Recall metric is	The models used include Naive Bayes, Logistic Regression, SVM (Support Vector Machine), EM Algorithm, EMLOG (Ensemble of Logistic Regression with EM algorithm), and EMSVM (Ensemble of SVM with EM algorithm).

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
			prioritized over RMSE for performance evaluation, and feature set selection includes assessment, spatial, temporal, and vocabulary features from the Yelp dataset	The novel ensemble algorithms EMSVM and EMLOG performed well, outperforming individual algorithms in predicting shutdown businesses. The outcome includes significant improvements in model accuracy due to diverse feature sets, the effectiveness of removing data skewness to improve training, and the importance of spatial and temporal features in boosting algorithm performance. Additionally, the ensemble algorithms EMSVM and EMLOG achieved high recall (85%) and precision (73.5%) in predicting shutdown businesses.
4	Michael Luca†(2016)	Target: Average rating of restaurants on Yelp, which serves as a measure of their overall quality and reputation Important variables: The number of reviews per quarter, the presence of elite reviewers, and the percentage of restaurants on Yelp over time. Other important Variables: restaurant categorization-full service or limited service, chain affiliation .	The process of gathering data includes using unique company identification codes to identify restaurants and combining Yelp restaurant reviews with revenue information from the Washington State Department of Revenue. In order to determine the effect of Yelp ratings on revenue, fixed effects regression and a regression discontinuity approach were used as analytical techniques. Fixed effects regressions were used to investigate	The study primarily uses regression discontinuity analysis to examine the causal relationship between Yelp ratings and restaurant sales. The regression discontinuity analysis is the main model used in the study, as it provides strong evidence of Yelp's causal influence on restaurant sales. The outcome of the study indicates a significant positive correlation between Yelp ratings and restaurant sales, with a one-star increase in Yelp ratings associated with a 5.4% to 9% rise in sales. This highlights

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
			the heterogeneous effects of Yelp, looking at how chain affiliation and Bayesian learning affected revenue. Tests using statistics were carried out to address possible issues like rating manipulation.	Yelp's influence as a major factor in consumer decision-making and its role as an important source of information and reputation in the dining business. Additionally, the study underscores the significance of easily accessible information, such as Yelp's rounded average rating, in consumer decision-making processes.
5	Ruchi Singha , Jongwook Woo(2019)	Target Variable: Performance or behavior of performance. Important variables: Review count, stars(ratings), Likes (for Tips), Business Categories, Attribute Columns. Other Important variables: Business ID, User ID, Text, Opening Hours .	A Yelp dataset of 334,335 rows and 108 columns as well as a review tip dataset with 591,865 rows and 7 columns were obtained for the data collecting process. Machine learning was used in analytical methodologies for jobs including sentiment analysis, business classification, and recommendation engines. We used Databricks and Azure ML Studio, where Databricks provided scalability and Azure made model implementation easier. Recommendation engines, sentiment analysis to forecast likes, and the categorization of companies as well-liked or poorly liked were	Yelp datasets including a wealth of business and review data were used, allowing for a variety of analytical techniques. used machine learning for tasks like sentiment analysis, business classification, and recommendation engines. For model implementation, Databricks and Azure ML Studio were used, providing a compromise between scalability and convenience. Yelp decision-making was aided by machine learning technologies that allowed for business classification, sentiment analysis to anticipate likes, and possibly even the discovery of bogus reviews. upcoming prospects: Proposed expansions highlight opportunities to improve Yelp's platform and user experience, such as the

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
			among the machine learning uses. Future additions might classify phony reviews automatically and examine how they affect companies, showing how Yelp's platform could be improved.	unsupervised classification of bogus reviews and analysis of their effects.

6	Xi Wanga, Liang (Rebecca) Tanga, Eojina Kim(2018)	Target Variable : Review helpfulness which represents the number of votes received on the helpfulness of consumer reviews. Important Variables: Emotional Dimensions, Linguistic Style Matching ,Control Variables. Other Important Variables: Readability, Consumer Rating, Consumer Elite Status, Review Elapsed Days, Restaurant Type.	100 representative restaurants were included in the data, which was collected from Yelp between October 10 and 16, 2017, with an emphasis on the top 10 travel destinations in the United States. This ensured a sizable sample size and prevented bias from newly opened or unpopular establishments. In addition to consumer elite status, review elapsed days, individual reviewer scores, and restaurant categories, the dataset contained 265,205 customer reviews. The study investigated how linguistic style matching (LSM) and emotional factors affected review helpfulness using a combination of statistical modeling and content analysis.	The study likely used statistical models such as regression analysis or machine learning algorithms to analyze the impact of emotions and linguistic style on review helpfulness. Specifically, linguistic style matching (LSM) may have been analyzed using regression or machine learning techniques. It's not explicitly mentioned which specific models were used, but regression analysis or machine learning algorithms capable of handling text data and sentiment analysis may have been employed to assess the impact of emotions and linguistic style on review helpfulness. The outcome of the study suggests that emotions and linguistic style play a significant role in determining the perceived helpfulness of reviews. Specifically, expressions of rage, disgust, and fear were found to have a favorable effect on review helpfulness, while joy, trust, and grief had a negative impact. Additionally, linguistic style matching (LSM), particularly with high-level matching in prepositions and auxiliary verbs, emerged as a strong predictor of review helpfulness.These findings highlight the nuanced ways in which emotions and language style influence the perceived helpfulness of reviews.

7	Sharun S Thazhackal; V. Susheela Devi (2018)	The target variable is the classification of businesses as open or closed. Important variables include aspect-wise ratings, category and region vectors, business variable vectors, convolutional filters, weight matrices, and bias vectors.	The methods involve collecting business data from the Yelp dataset, generating features representing trends in categories and regions over time intervals, building a hybrid deep neural network architecture with CNN and DNN components, training the model using cross-entropy loss minimization and transfer learning techniques, and evaluating model performance through stratified k-fold cross-validation and hyperparameter tuning.	The hybrid deep neural network model effectively predicts business closures using Yelp data, showcasing promising performance compared to traditional machine learning approaches. Aspect-wise ratings extracted from reviews significantly contribute to prediction accuracy. Transfer learning enhances model generalization, particularly beneficial for datasets with limited samples. Hyperparameter tuning and dropout regularization mitigate overfitting, enhancing model robustness. The model's ability to capture local patterns and regional trends proves crucial for accurate predictions of business closures.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
8	Anish K. Vallapuram∗ , Nikhil Nanda∗ , Young D. Kwon† , and Pan Hui∗‡(2021)	Target variable: Whether a restaurant will remain open or face closure after a fixed duration of time. Important variables include: Business attributes such as TakeOut and Delivery, which may impact a restaurant's survival. Geographic factors such as locality profiles, competition ratio, and specific competition based on cuisine served. Customer engagement metrics like reviews and images. Other important variables may include: Socioeconomic conditions of the region. Time-related factors such as the observation and prediction periods. Attributes specific to the restaurant category, like restaurant sub-category counts.	The study utilizes the publicly available Yelp dataset, which contains comprehensive information on businesses, including attributes and customer engagement metrics. Data collection involves extracting relevant attributes and engagement data from the Yelp dataset, focusing on the restaurant category due to its richness in attributes and business volume. Analytical methods include survival prediction modeling, incorporating locality profiles and customer satisfaction features, and leveraging machine learning techniques to forecast restaurant closures based on various attributes and geographic factors.	The study's key findings include the superior predictive performance of the hybrid deep neural network model in forecasting restaurant closures compared to traditional methods. Aspect-wise ratings extracted from customer reviews significantly influence prediction accuracy, showcasing the importance of customer sentiment analysis. Transfer learning improves model generalization, particularly beneficial for datasets with limited samples. Effective hyperparameter tuning and dropout regularization mitigate overfitting, ensuring model robustness. Locality profiles and customer satisfaction features prove essential in modeling a restaurant's success based on its location and customer engagement.
9	Dmitry Sedov(2021)	Target variable - restaurant closure, which indicates whether a restaurant remained open or	The paper examines restaurant closure patterns during the first year of the COVID-19 pandemic using Yelp and SafeGraph data, finding	Notable findings include geographic variation in restaurant exit rates across major US cities, with Honolulu experiencing the highest exit rate and El Paso the lowest.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
		permanently closed during the first year of the COVID-19 pandemic Important variables: Yelp rating score of the restaurant. Higher rating scores are associated with lower closure probabilities, suggesting that customer satisfaction and positive reviews may contribute to the resilience of restaurants during challenging times. Other important variables: count of nearby establishments, which reflects the proximity of restaurants to other businesses. Restaurants located close to many other establishments were more likely to close, indicating a reliance on foot traffic generated by nearby businesses and potentially greater vulnerability to pandemic-related disruptions.	that lower-rated restaurants and those located closer to city centers were more likely to close. The study provides descriptive evidence on factors influencing restaurant exit decisions across major US urban areas, contributing to the literature on business disruptions during the pandemic and firm entry/exit decisions in the context of economic shocks like COVID-19. The research employs binary response econometric models to analyze the association between restaurant characteristics and closure decisions, revealing that higher Yelp rating scores, more reviews, and a wider range of cuisine categories are linked to lower closure probabilities. Additionally, restaurants relying on foot traffic generated by their within-city location were relatively less likely to survive the pandemic year, shedding light on the impact of location-specific factors on closure rates.	Moreover, larger markets had higher restaurant closure rates, suggesting a relationship between market size and closure probabilities. The study underscores the importance of considering both restaurant-specific and location-specific characteristics in understanding the dynamics of restaurant closures during the COVID-19 pandemic.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
10	Zhiwei Liu a, Sangwon Park(2015)	Target Variable: The perceived usefulness of online reviews. Important Variables: Characteristics of review providers (identity disclosure, expertise, reputation) and review characteristics (star ratings, review length, perceived enjoyment, readability). Other Important Variables: Identity disclosure (real names, photos, addresses), expertise, friends, fans.	Using Yelp.com, a well-known venue for user-generated reviews, the study gathered information on more than 5,000 online evaluations of travel-related goods. The characteristics of review providers (e.g., disclosure of personal name, competence, reputation) and review attributes (e.g., star ratings, review length, perceived enjoyment, readability) were analyzed using descriptive statistics. The volume of information made it possible to conduct a thorough analysis of the variables affecting how valuable people believe internet reviews to be in the travel and hospitality industry. The perceived utility of online reviews is influenced by a number of important characteristics, which can be found using analytical and statistical methodologies. The study determines whether variables—such as reviewer traits and review attributes—have a substantial impact on the perceived usefulness of online reviews through regression analysis and descriptive statistics. A	Key Findings: Reviews with a good reputation and identity disclosure are seen as helpful. Perceived utility is favorably influenced by elaborateness, readability, enjoyment, and review scores. The perceived effectiveness of online reviews is influenced by features of both review providers and reviews.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
			more thorough knowledge of the phenomenon is made possible by these methods, which also offer quantitative data to corroborate the relationship between the detected components and the perceived usefulness.	
11	Alexis Papathanassis a, Friederike Knolle b(2010)	Target Variable: Perceived usefulness of online holiday reviews. Important Variables: Characteristics of Review Providers, Attributes of Reviews Other important Variables: Cognitive Mechanisms, Holiday Content Elements.	A purposefully created navigation prototype was used in the study to gather data, and respondents were shown different holiday depictions. Although the precise amount of data is not stated, it is mentioned that reasonably consistent patterns of decision-making and content processing were seen. The study utilized a variety of statistical and analytical techniques, one of which is grounded theory (GT), a qualitative approach for developing theories based on data collection. To find and arrange connections between observable occurrences inside GT, axial coding was carried out using a session-codebook for structured classification. The analysis was driven by the interpretation framework proposed by Borgatti (2008), which placed	Online evaluations are subjected to a set of heuristics prior to being adopted and employed, and they serve as a secondary, complementary resource to holiday choices. According to the study, in the context of competing proprietary content, it is critical to comprehend how users' access, process, and use user-generated content. Understanding the intricate cognitive processing of online review data is considered to require a qualitative method such as grounded theory.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
			emphasis on taking contextual elements and causative conditions into account. The observed events were then placed inside a theoretical framework by selecting a core category or variable via selective coding. These techniques made it easier to analyze the data methodically, which resulted in the creation of ideas supported by actual facts.	
12	Aika Qazi , Karim Bux Shah Syed, Ram Gopal Raj , Erik Cambria, Muhammad Tahir , Daniyal Alghazzawi (2016)	Target Variable: Number of Helpful Judgments received by each customer review. Important Variables: Overall Rating, Content , Author and Date . Other Important Variables: Cleanliness, Service, and the length of the review .	The study collected 1,500 TripAdvisor hotel reviews, comprising author, substance, date, reader count, number of helpful ratings, and overall rating. Reviews were parsed to divide the data into distinct records and fields for analysis after the data was taken from publicly accessible sources. To investigate the impact of concept count and average number of concepts per sentence on review helpfulness, Tobit regression analysis was utilized in the study. To evaluate their combined impacts, interaction terms of review type with these covariates were incorporated into the model. The	The study discovered that the number of concepts and average number of concepts per phrase, which are quantitative parameters, as well as the type of review, which is a qualitative element, both significantly influenced how useful online reviews were. It was discovered that regular reviews did not significantly indicate helpfulness. These findings offer valuable insights for e-commerce retailers and contribute to understanding the dynamics of online review helpfulness.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
			model fit the data well and provided a meaningful explanation for a large amount of the variance in review helpfulness.	
13	Pradeep Racherla , Wesley Friske (2012)	Target Variable: Perceived usefulness of online reviews. Important Variables: Reviewer's reputation, Reviewer's expertise, Review valence, Review extensiveness. Other Important Variables: Reviewer's demographics (e.g., age, gender), Reviewer's activity level on the platform, Service category being reviewed, Length of time the review has been posted, Number of helpful votes the review has received.	The study focused on service-related firms in key U.S. cities and gathered 3000 Yelp.com evaluations. Yelp was selected because of its broad reach and large user base in the service industry. OLS regression analysis was used in the study to test hypotheses about the factors determining how beneficial people believe internet reviews to be. Variables defying the assumptions of normalcy were subjected to log-transformation. The underlying hypothesis of the study model was that the degree to which consumers find reviews useful is a critical factor in the uptake of information. It took into account the message (review), the source (reviewer), and the type of service as predictors of the usefulness of the review.	Perceived utility was favorably connected with eviewers' reputation, whereas expertise exhibited a negative correlation. Extreme reviews were thought to be more helpful, and there was a convex link between review valences. Perceived usefulness and review extensiveness did not significantly correlate. These impacts differed according to the type of service. The study highlights the significance of both reviewer and review characteristics, shedding light on the intricate dynamics of online review perception. Remarkable results cast doubt on preconceived assumptions and call for more research on how customers react to internet evaluations in service-related scenarios.
14	Xiaopeng Lu Jiaming Qu Yongxing Jiang	Target Variable: is_open Important variables:	The data was collected from the yelp in 2016 with is_open as a target variable. This peer reviewed paper has done both text and non-text	The study was done to predict the future of restaurants, whether the restaurant will stay open or not. They did the model with the accuracy of 67.46% and precision of 73%.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
	Yanbing Zhao (2018)	chain_restaurant, review_count, return_guest _count, nearby_restaurant_comparison	analysis. Since the target variable is binary, they used logistic regression as a training model for this dataset. The accuracy and precision are the evaluation metrics. There was a 90/10 train/test split with removal of one feature group every time they ran the model.	The analysis indicated that the non-text features were more important than the text features. Other factors such as nearby comparison, trends and economic status also were important variables. they suggested an improvement in text feature analysis which could uncover deeper correlations between business success and user reviews.
15	Nabiha Asghar(2016)	Target Variable: Review Rating(stars) given to restaurants. Important Variables : Text Reviews, Preprocessed features , Business Categories, Reviews Characteristics Other Important Variables: Business Attributes , Reviewer characteristics , Check-in sets and tips.	With an emphasis on eateries, the dataset includes 1,125,458 text reviews from different cities and 42,153 companies. The text reviews are preprocessed using statistical techniques like TF-IDF weighting, and prediction models are constructed using four feature extraction techniques and four supervised learning algorithms. Performance indicators, such as accuracy and precision, are used to assess models. The study highlights how crucial it is to forecast review ratings by taking into account various company sectors separately.	68.3% of the reviews in the dataset are related to restaurants, which makes up the majority of the dataset. Reviews of restaurants tend to be very positive, with about 66% of reviews giving businesses very high ratings. The study investigates sixteen prediction models that combine supervised learning algorithms with different feature extraction techniques. Preprocessing is the process of taking meaningful content out of text reviews by eliminating capitalizations, stop words, and punctuation. Unigrams, which consider every distinct word as a feature, and TF-IDF weighting, which highlights distinctive words, are two feature extraction techniques.
16	Zefang Liu(2020)	Target Variable: The target variable is the star rating given to	The Yelp Open Dataset's restaurant data was extracted to create the	The vast majority of reviews (65.9%) have a strong bias in favor of 4 and 5 stars.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
		restaurants based on Yelp reviews, ranging from 1 to 5 stars. Important Variables: Text and Stars . Other Important Variables: Restaurant ID, Business Category , Review Count.	dataset. Filtering for restaurant establishments and obtaining corresponding reviews was part of the preprocessing. Resampling reviews allowed for the construction of balanced training datasets. Two vectorizers were utilized for the purpose of feature engineering. Alongside transformer-based models like BERT, DistilBERT, RoBERTa, and XLNet, machine learning models including Naive Bayes, Logistic Regression, Random Forest, and Linear Support Vector Machine were put into practice.	5,055,992 reviews from 63,944 establishments make up the dataset. When it came to 5-star categorization, XLNet outperformed Logistic Regression (64%), with an accuracy of 70%. 73.1% of reviews are no longer than 128 tokens, and 92.8% are no longer than 256 tokens, according to the distribution of review lengths. Predicting restaurant ratings from Yelp reviews has several uses, such as classifying reviews for recommendation engines and identifying unusual reviews to shield companies from dishonest competitors.
17	Dimitris Papaioannou, 2022	Target variable: star rating assigned by users to businesses based on their reviews. Important Variables: features extracted from the review text, such as sentiment, as well as contextual information like the type of restaurant and its average rating. Other Important Variables :Review length, Review sentiment, Business	In addition to feature selection strategies like stop word removal and stemming, analytical and statistical techniques used for multi-class classification include logistic regression, Naive Bayes, SVM, and deep neural networks.	Principal discoveries encompass the efficacy of binarized Naive Bayes in conjunction with feature selection for sentiment analysis, the robustness of logistic regression and SVM classifiers, and the possibility of additional enhancement through the integration of business categories into feature sets. Furthermore, bootstrapped Support Vector Machines (SVMs) shown notable efficacy in forecasting positive instances by attaining elevated success rates on both bootstrapped and random datasets. Nevertheless, certain enhancements in

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
		categories, User behavior Temporal factors		sensitivity and specificity might be made in comparison to an oracle possessing complete access to all data.
18	Tanbin Siddique Eidul, Md.Alim Imran, Amit Kumar Das(2022)	Target variable: Rating of restaurant businesses, which is simplified into three categories: poor (0), average (1), and good (2). Important Variables: Business attributes (e.g., parking availability, ambience, dietary restrictions), geographical information (latitude, longitude), Other Important Variables : other relevant factors influencing restaurant ratings.	With an emphasis on the business sub-dataset, the researchers gathered information from the Yelp dataset. They extracted pertinent features from the data, normalized it, and then used label encoding. Based on these variables, a variety of machine learning methods were used to predict restaurant ratings, including CNN, SVM, KNN, SGD, Gaussian Naive Bayes, and Decision Trees.	With an accuracy score of 97.22%, the Convolutional Neural Network (CNN) model surpassed other machine learning techniques. This implies that restaurant ratings derived from attributes taken from datasets such as Yelp can be accurately predicted by deep learning techniques like CNNs. By offering precise estimates of restaurant evaluations, the study seeks to support aspiring restaurant owners.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
19	Mengqi Yu, Meng Xue, Wenjia Ouyang (2010)	Target variable: star rating of restaurant reviews Important Variables: user review history, restaurant statistics, and sentiment analysis of review text Other Important Variables: Business Category, Location, Review Length, Review Votes, Time of Review, Business Attributes.	Data mining techniques, such as sentiment analysis, latent factor modeling, random forest regression, and linear regression, were applied to the Yelp Dataset Challenge round 5.	When predicting review ratings, random forest regression fared better than other models. It turned out that sentiment elements were quite helpful for rating prediction. Review ratings were greatly influenced by business statistics and the history of user reviews. Ratings were weighted toward higher stars, according to distribution analysis, indicating an optimistic bias over time. A location analysis revealed clusters of restaurants in a few US states, most notably Nevada.

No.	Source	How are the key concepts/terms/variables (independent, dependent) defined and measured?	Methods, such as data collection and analytical methods	Important findings
20	Lenka Kovalcinova (New Jersey Institute of Technology, NJ, US), Martin Polacek (Stony Brook University, NY, US)(2015)	The target variable of the analysis is the probability of business closure. Important variables: Average star rating, review count, and duration of business operation. Other important Variables: location (longitude and latitude) and the type of business.	The methods utilized involve data collection from Yelp datasets, including reviews and business information, and analytical techniques such as simple histogram comparison, recursive partition tree (rpart), non-linear curve fitting, and natural language processing (NLP) for sentiment analysis of reviews. Time series forecasting methods such as neural network, ARIMA, and auto.ARIMA are also employed.	Important findings include a correlation between business closure and low average ratings, as well as the influence of duration of business operation on review patterns. The NLP analysis identifies specific negative features in reviews, which are then used to forecast the development of business inferiority over time. The decision tree model predicts the probability of business closure based on star ratings and review count, with higher ratings associated with higher probabilities of remaining open.

Table-2 : Data collection

Authors	Yelp	Trip advisor	Mock data	Major cities	Department of revenue	Survey	variety.com
Yifan Chen; Fanzeng Xia(2020)	x						
Yi Luo 1 and Xiaowei Xu 2(2019)	x			x			
Narenkumar Pandian and Vaibhav Aggarwal(2015)	x				x		
Michael Luca†(2016)	x						
Ruchi Singha , Jongwook Woo(2019)	x						
Xi Wanga, Liang (Rebecca) Tanga, Eojina Kim(2018)	x						
Sharun S Thazhackal; V. Susheel Devi(2018)	x						x
Anish K. Vallapuram∗ , Nikhil Nanda∗ , Young D. Kwon† , and Pan Hui∗‡	x						
Dmitry Sedov(2021)	x			x			
Zhiwei Liu a, Sangwon Park	x						
Alexis Papathanassis a, Friederike Knolle b(2010)						x	
Aika Qazi ,Karim Bux Shah Syed, Ram Gopal Raj ,Erik Cambria , Muhammad Tahir , Daniyal Alghazzawi (2016)		x					
Pradeep Racherla , Wesley Friske (2012)	x						
Xiaopeng Lu,Jiaming Qu,Yongxing Jiang,Yanbing Zhao (2018)	x						
Nabiha Asghar(2016)	x						
Zefang Liu(2020)	x						
Dimitris Papaioannou, 2022	x						
Tanbin Siddique Eidul, Md.Alim Imran, Amit Kumar Das(2022)	x						
Mengqi Yu,Meng Xue,	x						

Wenjia Ouyang (2010)							
Lenka Kovalcinova (New Jersey Institute of Technology, NJ, US), Martin Polacek (Stony Brook University, NY, US)(2015)	x			x			

Table-3: Data preparation

Authors	Software used	Balanced dataset	Imbalanced dataset	Data Cleaning	Explorator y analysis	Binn ing	Text analysis
Yifan Chen; Fanzeng Xia(2020)	python		x	x	x	x	x
Yi Luo 1 and Xiaowei Xu 2(2019)	Python						
Narenkumar Pandian and Vaibhav Aggarwal(2015)		x		x	x		x
Michael Luca†(2016)							
Ruchi Singha , Jongwook Woo(2019)	Azure ML studio & Databricks		x				
Xi Wanga, Liang (Rebecca) Tanga, Eojina Kim(2018)			x				x
Sharun S Thazhackal; V. Susheel Devi(2018)		x		x	x		x
Anish K. Vallapuram∗ , Nikhil Nanda∗ , Young D. Kwon† , and Pan Hui∗‡	python		x	x			x
Dmitry Sedov(2021)			x		x		
Zhiwei Liu a, Sangwon Park			x				
Alexis Papathanassis a, Friederike Knolle b(2010)			x				
Aika Qazi ,Karim Bux Shah Syed, Ram Gopal Raj ,Erik Cambria , Muhammad Tahir , Daniyal Alghazzawi (2016)			x				
Pradeep Racherla , Wesley Friske (2012)			x				x
Xiaopeng Lu,Jiaming Qu,Yongxing Jiang,Yanbing Zhao (2018)	SAS						
Nabiha Asghar(2016)		x					x
Zefang Liu(2020)							
Dimitris Papaioannou, 2022	python	x					

Tanbin Siddique Eidul, Md.Alim Imran, Amit Kumar Das(2022)		x					
Mengqi Yu,Meng Xue, Wenjia Ouyang (2010)	python	x					x
Lenka Kovalcinova (New Jersey Institute of Technology, NJ, US), Martin Polacek (Stony Brook University, NY, US)(2015)	R on OS X El Capitan and Linux Mint 17.2 Cinnamon		x		x		x

Table-4: Modeling methods

Authors	Final Model	Logistic regression	Decision tree	Random forest	SVM	LASSO/ LARS	NN	Text analysis	Other models
Yifan Chen; Fanzeng Xia(2020)	Decision tree	x	x				x	x	x
Yi Luo 1 and Xiaowei Xu 2(2019)	SVM/LDA				x				x
Narenkumar Pandian and Vaibhav Aggarwal(2015)	EMSVM & EMLOG	x			x			x	x
Michael Luca†(2016)	Regression discontinuity analysis								x
Ruchi Singha , Jongwook Woo(2019)	Sentimental analysis								
Xi Wanga, Liang (Rebecca) Tanga, Eojina Kim(2018)	Linguistic style matching							x	x
Sharun S Thazhackal; V. Susheel Devi(2018)	CNN & DNN	x		x	x		x	x	x

Anish K. Vallapuram∗ , Nikhil Nanda∗ , Young D. Kwon† , and Pan Hui∗‡	GBDT	x		x	x		x	x	x
Dmitry Sedov(2021)	Logistic regression	x							x
Zhiwei Liu a, Sangwon Park									
Alexis Papathanassis a, Friederike Knolle b(2010)	Grounded Theory (GT)								x
Aika Qazi ,Karim Bux Shah Syed, Ram Gopal Raj ,Erik Cambria , Muhammad Tahir , Daniyal Alghazzawi (2016)	Tobit regression analysis	x							x
Pradeep Racherla , Wesley Friske (2012)	OLS							x	x
Xiaopeng Lu,Jiaming Qu,Yongxing Jiang,Yanbing Zhao (2018)	Logistic regression	x	x				x		x
Nabiha Asghar(2016)	TF-IDF weighting							x	x
Zefang Liu(2020)	XLNet	x		x	x				
Dimitris Papaioannou, (2022)	SVM	x			x		x		
Tanbin Siddique Eidul, Md.Alim Imran, Amit Kumar Das(2022)	Neural network		x		x		x		
Mengqi Yu,Meng Xue, Wenjia Ouyang (2010)	Sentimental analysis	x		x				x	x
Lenka Kovalcinova (New Jersey Institute of Technology, NJ, US), Martin Polacek (Stony	Decision Tree		x						

Brook University, NY, US)(2015)									

Table-5: Assessment methods

Authors	RMSE	ROC/ AUC	Odds ratio	F1 score	Accuracy	Recall	Precision	Specificity
Yifan Chen; Fanzeng Xia(2020)					x			
Yi Luo 1 and Xiaowei Xu 2(2019)					x		x	
Narenkumar Pandian and Vaibhav Aggarwal(2015)				x		x	x	
Michael Luca†(2016)					x	x		
Ruchi Singha , Jongwook Woo(2019)								
Xi Wanga, Liang (Rebecca) Tanga, Eojina Kim(2018)								x
Sharun S Thazhackal; V. Susheel Devi(2018)				x	x	x	x	
Anish K. Vallapuram∗ , Nikhil Nanda∗ , Young D. Kwon† , and Pan Hui∗‡		x						
Dmitry Sedov(2021)					x			
Zhiwei Liu a, Sangwon Park								
Alexis Papathanassis a, Friederike Knolle b(2010)				x			x	
Aika Qazi ,Karim Bux Shah Syed, Ram Gopal Raj ,Erik Cambria , Muhammad Tahir , Daniyal Alghazzawi (2016)					x	x	x	x
Pradeep Racherla , Wesley Friske (2012)					x			
Xiaopeng Lu,Jiaming Qu,Yongxing Jiang,Yanbing Zhao (2018)				x	x		x	
Nabiha Asghar(2016)					x	x		
Zefang Liu(2020)					x			
Dimitris Papaioannou, (2022)					x			
Tanbin Siddique Eidul, Md.Alim Imran, Amit Kumar Das(2022)		x		x	x	x	x	
Mengqi Yu,Meng Xue,				x	x			

Wenjia Ouyang (2010)								
Lenka Kovalcinova (New Jersey Institute of Technology, NJ, US), Martin Polacek (Stony Brook University, NY, US)(2015)					x	x	x	x

Table-6: Important variables

Authors	Number of Reviews	Ratings	Word of mouth	Chain restaurant	location	Business attributes
Yifan Chen; Fanzeng Xia(2020)	x	x			x	
Yi Luo 1 and Xiaowei Xu 2(2019)	x	x				
Narenkumar Pandian and Vaibhav Aggarwal(2015)	x	x	x		x	x
Michael Luca†(2016)	x	x				x
Ruchi Singha , Jongwook Woo(2019)	x	x				
Xi Wanga, Liang (Rebecca) Tanga, Eojina Kim(2018)			x			
Sharun S Thazhackal; V. Susheel Devi(2018)	x	x	x		x	
Anish K. Vallapuram∗ , Nikhil Nanda∗ , Young D. Kwon† , and Pan Hui∗‡	x	x	x		x	
Dmitry Sedov(2021)	x	x	x	x		
Zhiwei Liu a, Sangwon Park	x	x	x		x	
Alexis Papathanassis a, Friederike Knolle b(2010)	x	x				
Aika Qazi ,Karim Bux Shah Syed, Ram Gopal Raj ,Erik Cambria , Muhammad Tahir , Daniyal Alghazzawi (2016)	x	x				x

Pradeep Racherla , Wesley Friske (2012)	x		x			
Xiaopeng Lu,Jiaming Qu,Yongxing Jiang,Yanbing Zhao (2018)	x	x		x	x	x
Nabiha Asghar(2016)	x	x				
Zefang Liu(2020)	x	x				
Dimitris Papaioannou, 2022			x			x
Tanbin Siddique Eidul, Md.Alim Imran, Amit Kumar Das(2022)	x	x			x	
Mengqi Yu,Meng Xue, Wenjia Ouyang (2010)	x	x			x	
Lenka Kovalcinova (New Jersey Institute of Technology, NJ, US), Martin Polacek (Stony Brook University, NY, US)(2015)	x	x	x		x	x

Data understanding and data preparation (Phase 2 and 3 of CRISP DM methodology):

The initial data was collected from the Yelp data source. The initial files were in the form of JSON and we converted the JSON files to Excel using python coding. We downloaded two files named “Business.JSON” and “Review.JSON”. The business file consists of the target variable “is_open” and other restaurant features. Most of the literature (including X. Lu et al 2018) focused on “is_open” as a target variable. We thought it would be interesting if we rather focus on “is_close” because restaurant owners are more interested in what went wrong rather than what they can improve. so, by flipping the column value from 1 to 0 and 0 to 1 we were able to do it. We approximately had balanced data with 40% of closed restaurants and 60% of not closed restaurants. That seems high for a major city and the reason behind it might be because of COVID-19 period where so many small restaurants were closed and our dataset also covers the recession period of 2008 where many restaurants were out of business as well. Most of the restaurant features were in the same column as an attribute. We used python again to divide that one column into separate columns with one feature in each column. After splitting the columns, we observed that few variables are binary and few are categorical. The stars and review_count variables are converted to numerical variables using SAS studio and categorical recording in the SAS enterprise miner for more extensive analysis. The review file contains the variables business_id, review_stars, review_useful, review_funny, review_cool and review_text. For further text analysis as the review file will be used , a target variable is needed in the review file. So, using the Excel “combine queries” option we combined the review file and the business file using the common column “business_id”. With the target variable “is_close” we also included all the variables from the business file to review file to get into more deep analysis totalling 74,700 observations. Table - 7 contains both the variables from restaurant features and reviews with a target variable. Table - 8 contains only the target variable and reviews for separate text analysis.

Data dictionary for restaurant features and reviews (non-text features & text):

Table-7: Data dictionary includes 90 variables including 1 target variable.

Variable	Description	Measurement Level
business_id	Unique identifier for each restaurant	Nominal
name	Name of the restaurant	Nominal
address	Physical address of the restaurant	Nominal
city	City where the restaurant is located	Nominal
state	State where the restaurant is located	Nominal
postal_code	Postal code of the restaurant's location	Nominal
latitude	Geographic latitude of the restaurant's location	Continuous
longitude	Geographic longitude of the restaurant's location	Continuous
stars	Average rating of the restaurant	Ordinal
review_count	Number of reviews for the restaurant	Interval
is_Close	Binary indicator representing whether the restaurant is closed or not	Binary/Target
categories	Categories or types of cuisine offered by the restaurant	Nominal
hours	Operating hours of the restaurant	Nominal

RestaurantsDelivery	Explains whether the restaurant offers delivery services	Binary
OutdoorSeating	Explains whether the restaurant offers outdoor seating	Binary
BusinessAcceptsCreditCards	Explains whether the restaurant accepts credit cards	Binary
BikeParking	Explains whether the restaurant offers bike parking	Binary
RestaurantsPriceRange2	Price range of the restaurant	Ordinal
RestaurantsTakeOut	Explains whether the restaurant offers takeout services	Binary
ByAppointmentOnly	Explains whether the restaurant operates by appointment only	Binary
WiFi	Binary Indicator for whether a restaurant has WIFI or not for free.	Binary
Alcohol	Explains Whether the restaurants has full bar, beer and wine or none	Nominal
Caters	Explains Whether the restaurants takes catering orders or not	Nominal
RestaurantsAttire	Explains whether the attire is dressy or casual	Nominal
RestaurantsReservations	Explains Whether the restaurant takes table reservations or not	Nominal
GoodForKids	Explains whether the restaurant is good for kids	Nominal
CoatCheck	Explains whether the restaurant has CoatCheck or not	Binary
DogsAllowed	Explains whether dogs are allowed or not in a restaurant	Nominal

RestaurantsTableService	Explains whether the restaurant has table service or not .	Nominal
RestaurantsGoodForGroups	Is the restaurant good for people in groups	Nominal
WheelchairAccessible	Does the restaurant have wheelchair accessibility or not	Nominal
HasTV	Whether restaurant has TV or not	Binary
HappyHour	Explains whether the restaurant has HappyHours or not	Binary
DriveThru	Explains whether the restaurant has Drive Thru or not	Nominal
NoiseLevel	Explains the restaurants noise level - loud,average, quiet,very_loud	Nominal
BusinessAcceptsBitcoin	Explains whether the restaurant business accepts bitcoin or not	Binary
Smoking	Explains where should a person - no, outdoor, yes	Nominal
GoodForDancing	Explains whether a restaurant is good for dancing or not	Binary
BYOB	Explains whether one has to buy their own beer or not	Nominal
Corkage	Explains whether the restaurants has Corkage or not	Nominal
BYOBCorkage	Explains whether restaurants has a fee for Crokage or not	Nominal

RestaurantsCounterService	Explains whether restaurant has counter service or not	Unary
romantic	Explains whether restaurant is romantic or not	Binary
intimate	Explains whether restaurant is intimate or not	Binary
touristy	Explains whether restaurant is touristy or not	Binary
hipster	Explains whether restaurant is cultural or not	Binary
divey	Explains whether restaurant has reputable br or not	Binary
classy	Explains whether restaurant is classy or not	Binary
trendy	Explains whether restaurant is trendy or not	Binary
upscale	Explains whether restaurant is upscale or not	Binary
casual	Explains whether restaurant is casual or not	Binary
dessert	Explains whether restaurant has a dessert or not	Binary
latenight	Explains whether restaurant is open till late night or not	Binary
lunch	Explains whether restaurant provides lunch or not	Binary
dinner	Explains whether restaurant provides or not	Binary

brunch	Explains whether restaurant provides brunch or not	Binary
breakfast	Explains whether restaurant provides breakfast or not	Binary
dj	Explains whether restaurant has dj or not	Binary
background_music	Explains whether restaurant has background music or not	Binary
no_music	Explains if the restaurant has no music at all	Unary
jukebox	Explains whether restaurant have a jukebox or not	Binary
live	Explains whether restaurant has live music or not	Binary
video	Explains whether restaurant has video or not	Binary
karaoke	Explains whether restaurant has karaoke or not	Binary
garage	Explains whether restaurant has a garage or not	Binary
street	Explains whether restaurant has a street access or not	Binary
validated	Explains whether restaurant is validated or not	Binary
lot	Explains whether restaurant has parking lot or not	Binary
valet	Explains whether restaurant has valet parking or not	Binary

monday	Explains whether restaurant is open on monday or not	Binary
tuesday	Explains whether restaurant is open on tuesday or not	Binary
friday	Explains whether the restaurant is open on friday or not	Binary
wednesday	Explains whether the restaurant is open on wednesday or not	Binary
thursday	Explains whether the restaurant is open on thursday or not	Binary
sunday	Explains whether the restaurant is open on sunday or not	Binary
saturday	Explains whether the restaurant is open on saturday or not	Binary
IsChain	Explains whether restaurant has chain of restaurants or not	Binary
American_Cuisine	Explains whether restaurant has american cuisine or not	Binary
Asian_Cuisine	Explains whether restaurant has Asian cuisine or not	Binary
European_Cuisine	Explains whether restaurant has European cuisine or not	Binary
Vegetarian_Vegan	Explains whether restaurant provides Vegetarian_Vegan food or not	Binary
Fast_Food	Explains whether restaurant has fast food or not	Binary
BarsandNightlife	Explains restaurants bars and nightlife	Binary

CafeandJuiceBars	Explains whether restaurant has cafes and juice bars or not	Binary
review_stars	Explains the stars given by the individual to the restaurant ranging from 1 to 5	Nominal
review_useful	Likes given by the other customer on this review based on whether it is useful or not	Nominal
review_funny	Likes given by the other customer on this review based on whether it is funny or not	Nominal
review_cool	Likes given by the other customer on this review based on whether it is cool or not	Nominal
review_text	Linguistic text written by an individual	Text
review_date	Time and date of the review written	Nominal

Data dictionary for Text analysis (text features):

Table-8: Data dictionary includes 7 variables including 1 target variable.

Variable	Description	Measurement Level
business_id	Unique identifier for each restaurant	Nominal
name	Name of the restaurant	Nominal
postal_code	Postal code of the restaurant's location	Nominal
latitude	Geographic latitude of the restaurant's location	Continuous
longitude	Geographic longitude of the restaurant's location	Continuous
is_Close	Binary indicator representing whether the restaurant is closed or not	Binary/Target
review_text	words of mouth written by an individual	Text

Data cleaning:

The variables in the business file are either categorical or binary. As mentioned earlier, variable stars and review_counts are converted to the numerical measurement level using sas code in SAS studio which explains that by considering the required variables higher the stars for the restaurant is, the better the restaurant is. It is also the same with review_count, higher the number of reviews, more chance for that restaurant to perform better than others. This Conversion will help us to explore the standardized estimate in the regression, LARS and LASSO. We also decided to create “is_chain” variables with help of excel. We use the “if” function and set the condition as if the restaurant's name repeats more than one time then it is a chain restaurant. We created a few more variables such as types of cuisine, fast food and vegetarian/vegan restaurant from one column named "category" which has keywords like american, burger, pizza, thai, chinese, indian, etc. let's take one variable which is american cuisine, we used "if" statement and made a condition that if the category column include words like american, barbeque or steakhouse then that restaurant will have a value of 1 otherwise 0. These variables might be helpful for analysis further in the paper. The business file has a lot of missing data. If we set a threshold of 50% about two third of the variables are excluded from the analysis. So, instead of excluding it we decided to impute it with "0" or "False" meaning that the missing value will be assumed as that restaurant feature is not available in the restaurant. For example, if the variable Restaurant delivery has missing value, then it can be replaced with False, which means the restaurant does not offer food delivery service. By doing such, we will not miss much of the information or important variables. For the review file, out of 5800 restaurants, we only have business id's of 1100 restaurants with 113,000 observations. Since, the review_text variable contained so much text information we decided to cut the observations to about 70,000 because SAS enterprise miner cannot handle the text analysis well when there is big data.

The text and non-text file has 90 variables including the target variable "is_close". Running all the variables in the model will not give us all the information we need. We will miss major findings because out of all the variables we will only be able to find the top important variable. Rather than dumping all the variables in the model at once, we decided to divide it into categories such as location, business details, customer convenience, dining experience, Food service, reservation & events, alcohol & beverages, Ambience, atmosphere, operational hours, Cuisine type and parking as shown in Table - 9. The categories were decided in terms of its closeness to the category name. We also took help from AI to see whether the category matches the variable or not. The individual categories will then go into the predictive modeling process to get the most important variable out of it. This way we can have recommendations to restaurant owners or investors for each category, and based on that they can improve their decision making.

Table-9: Variables and its corresponding category

Category	Variables
Location	business_id, name, address, city, state, postal_code, latitude, longitude
Business Details	stars, review_count, categories, hours, IsChain
Customer Convenience	RestaurantsDelivery, BusinessAcceptsCreditCards, BikeParking, WiFi, DriveThru, RestaurantsCounterService, GoodForKids,

	RestaurantsGoodForGroups,
Dining Experience	OutdoorSeating, RestaurantsPriceRange2, RestaurantsAttire, CoatCheck, DogsAllowed, RestaurantsTableService, WheelchairAccessible, NoiseLevel, HasTV, Smoking
Food Service	RestaurantsTakeOut, Caters, HappyHour, CafeandJuiceBars
Reservation & Events	ByAppointmentOnly, RestaurantsReservations
Alcohol & Beverages	Alcohol, BYOB, Corkage, BYOBCorkage
Ambience	romantic, intimate, touristy, hipster, divey, classy, trendy, upscale, casual

Meals	dessert, lunch, dinner, brunch, breakfast,
Atmosphere	latenight, dj, Background_music, no_music, jukebox, live, video, karaoke, BarsandNightlife
Operational Hours	monday, tuesday, friday, wednesday, thursday, sunday, saturday
Cuisine Type	American_Cuisine, Asian_Cuisine, European_Cuisine, Vegetarian_Vegan, Fast_Food
Parking	garage, street, validated, lot, valet

Exploratory/descriptive analysis/statistical analysis:

Table 10 - Summary statistics of interval(numeric) variables

Variable	Mean	Std Dev	Minimum	Maximum	Median	N	Skewness	Kurtosis
stars	3.55	0.80	1.00	5.00	3.50	5817	-0.64	0.06
review_count	114.25	242	5.00	5721.00	39.00	5817	7.77	103.63

The dataset contains 2 interval(numeric) variables which are stars and review_count. The stars has a mean of 3.55 and review_count has a mean of 114.25 as shown in Table-10. The skewness of stars is -0.64 showing that it is approximately normally distributed but, the skewness of review_count is 7.77 which indicates that it is right skewed. To normalize the distribution log transformation will be applied during the predictive modeling analysis.

Table 11: Frequency table for target variable

Variable	level	Frequency	Percent
is_close	0	3496	60.10
	1	2321	39.90

The table analysis reveals a closure rate of about 40% which is 2,321 out of every 5,817 restaurants, this gives a clear indicator of the volatile nature of the restaurant business. This high closure rate highlights the tough environment for restaurants and underlines the importance of having a predictive model to identify restaurants with a high risk of closure.

Table 12: Frequency table for Business details category

Variable	level	Frequency	Percent
IsChain	Chain	1099	18.89
	Not Chain	4718	81.11

The table clearly shows that of the restaurants studied, 81% are not part of a chain, indicating that independent restaurants make up a large portion of the market. This suggests that the survival challenges might be different for chain and non-chain restaurant establishments.

Table 13: Frequency table for Customer convenience category

Variable	level	Frequency	Percent
RestaurantsDelivery	‘none’	311	5.35
	0	2051	35.26
	1	3455	6.05
BusinessAcceptsCreditCards	‘none’	2	0.03
	0	1111	19.10
	1	4704	80.87
BikeParking	‘none’	3	0.05
	0	2939	50.52
	1	2875	49.42

WiFi	‘free’	2008	48.17
	‘no’	2125	50.97
	‘none’	2	0.05
	‘paid’	34	0.82
DriveThru	‘none’	45	0.77
	0	5658	97.27
	1	114	1.96
RestaurantsCounterService	0	5815	99.97
	1	2	0.03
GoodForKids	‘none’	4	0.07
	0	2452	42.15
	1	3361	57.78
RestaurantsGoodForGroups	‘none’	4	0.07
	0	2184	37.55
	1	3629	62.39

The analysis in the table clearly shows that most restaurants accept credit card payments (80.87%) and many offer free WiFi (48.17%), indicating these as baseline consumer expectations. On the other hand, the lack of DriveThru services (1.96%) highlights a unique market differentiation area, which could influence restaurant survival.

Table 14: Frequency table for Dining Experience category

Variable	**level**	**Frequency**	**Percent**
OutdoorSeating	‘none’	165	2.84
	0	3756	64.57
	1	1896	32.59
RestaurantsPriceRange2	‘none’	1	0.02
	1	2145	43.67
	2	2479	50.47
	3	249	5.07
	4	38	0.77
RestaurantsAttire	‘casual’	4211	97.21
	‘dressy’	111	2.56
	‘formal’	9	0.21
	‘none’	1	0.02
CoatCheck	0	5735	98.59
	1	82	1.41
DogsAllowed	‘none’	2	0.03
	0	5601	96.29

	1	214	3.68
RestaurantsTableService	0	4550	78.22
	1	1267	21.78
WheelchairAccessible	‘none’	2	0.03
	0	4715	81.06
	1	1100	18.91
NoiseLevel	‘average’	2616	68.32
	‘loud’	329	8.59
	‘none’	3	0.08
	‘quiet’	754	19.69
	‘very_loud’	127	3.32
Smoking	‘no’	5719	98.32
	‘outdoor’	89	1.53
	‘yes’	9	0.15
HasTV,	‘none’	1	0.02
	0	2703	46.47
	1	3113	53.52

The result from the data shows a preference for outdoor seating (32.59%) and moderate pricing (50.47%), when combined with a common level of background noise.(68.32%). These factors collectively suggest a consumer inclination towards casual and comfortable dining experiences, potentially impacting restaurant longevity.

Table 15: Frequency table for Food service category

Variable	level	Frequency	Percent
RestaurantsTakeOut	‘none’	154	2.65
	0	644	11.07
	1	5019	86.28
Caters	‘none’	3	0.05
	0	3703	63.66
	1	2111	36.29
HappyHour,	0	4848	83.34
	1	969	16.66
CafeandJuiceBars	0	4715	81.06
	1	1102	18.94

Table 16: Frequency table for Reservation & Events category

Variable	level	Frequency	Percent
ByAppointmentOnly	0	5800	99.71

	1	17	0.29
RestaurantsReservations	‘none’	34	0.58
	0	4045	69.54
	1	1738	29.88

The data from the food service and reservations & events categories, tables 15 and 16 respectively, reveal notable trends in the operational choices of restaurants and their potential impact on customer experience and restaurant survival. A vast majority (86.28%) offer takeout services, emphasizing its critical role in modern dining preferences, while catering services and happy hours are less prevalent, offered by only 36.29% and 16.66% of restaurants, respectively. This suggests a focus on core dining services rather than other extra things. On the other hand, the reservation & events data shows an overwhelming majority (99.71%) do not operate by appointment only, facilitating walk-in customers. However, a significant portion (29.88%) accepts reservations, indicating a blend of unpredictability and planning in dining experiences. These operational strategies reflect a diverse approach to meeting customer expectations, highlighting areas of potential differentiation and adaptation for restaurants aiming to improve their survival prospects in a competitive market.

Table 17: Frequency table for Alcohol & Beverages category

Variable	level	Frequency	Percent
Alcohol	‘Beer and wine’	209	4.61

	‘Full bar’	1399	30.86
	‘none’	2925	64.53
BYOB	‘none’	1	0.02
	0	5637	96.91
	1	179	3.08
Corkage	‘none’	1	0.02
	0	5735	98.59
	1	81	1.39
BYOBCorkage	‘no’	5545	94.43
	‘Yes corkage’	18	0.31
	‘Yes free’	254	4.33

Table 18: Frequency table for Ambience category

Variable	level	Frequency	Percent
Romantic	0	5705	98.07

	1	112	1.93
Intimate	0	5671	97.49
	1	146	2.51
Touristy	0	5784	99.43
	1	33	0.57
Hipster	0	5647	97.08
	1	170	2.92
Divey	0	5686	97.75
	1	131	2.25
classy,	0	5013	86.18
	1	804	13.82
Trendy	0	5486	94.31
	1	331	5.69
Upscale	0	5774	99.26

	1	43	0.74
Casual	0	3749	64.45
	1	2068	35.55

The analysis of Alcohol & Beverages and Ambience categories in tables 17 and 18 respectively, reveals notable trends: most restaurants opt not to serve alcohol (64.53%), emphasizing a potentially family-friendly or cost-conscious approach, with only a modest number offering a full bar (30.86%). The scarcity of BYOB and corkage options further highlights this trend towards regulated alcohol consumption. Ambience-wise, the data shows a strong preference for casual (35.55%) and classy (13.82%) settings, indicating a trend towards creating versatile and broadly appealing dining environments. These insights suggest that strategic decisions regarding alcohol services and ambience are crucial for aligning with consumer preferences, potentially influencing a restaurant's success and longevity in the competitive market.

Table 19: Frequency table for Meals category

Variable	**level**	**Frequency**	**Percent**
dessert	0	5706	98.09
	1	111	1.91
lunch	0	4572	78.60
	1	1245	21.40
dinner	0	4457	76.62
	1	1360	23.38
brunch	0	5536	95.17
	1	281	4.83
breakfast	0	5537	95.19
	1	280	4.81

The analysis of the Meals category highlights a strong emphasis on serving lunch (21.40%) and dinner (23.38%), while offerings such as dessert (1.91%), brunch (4.83%), and breakfast (4.81%) are notably less frequent. This suggests a strategic focus on peak dining periods, with other mealtimes presenting potential niche opportunities. The data indicates a clear prioritization of main meal services, pointing towards consumer demand and operational preferences in the competitive restaurant landscape.

Table 20: Frequency table for Atmosphere category

Variable	**level**	**Frequency**	**Percent**
latenight	0	5581	95.94
	1	236	4.06
dj	0	5778	99.33
	1	39	0.67
Background_music	0	5786	99.47
	1	31	0.53
jukebox	0	5768	99.16
	1	49	0.84
video	0	5816	99.98
	1	1	0.02
karaoke	0	5810	99.88
	1	7	0.12

The analysis of the atmosphere category within the dataset underscores a pronounced preference for traditional dining experiences among restaurants, as evidenced by the minimal late-night operations (4.06%) and the limited offering of entertainment options such as DJs (0.67%), background music (0.53%), jukeboxes (0.84%), karaoke (0.12%), and video entertainment (0.02%). This trend suggests that restaurants lean towards creating quieter, more conventional atmospheres, potentially in response to consumer preferences or operational considerations. Given the study's aim to predict restaurant closures, these findings highlight the importance of atmosphere in aligning with customer expectations, which could be a significant factor in determining a restaurant's survival in the competitive industry landscape.

Table 21: Categorical table for cuisine category

Variable	**Level**	**Frequency**	**Percent**
American_Cuisine,	0	4333	74.49
	1	1484	25.51
Asian_Cuisine,	0	4896	84.17
	1	921	15.83
European_Cuisine,	0	5155	88.62
	1	662	11.38
Vegetarian_Vegan,	0	5540	95.24
	1	277	4.76
Fast_Food	0	4478	76.98

	1	1339	23.02

The table here shows us the popularity of American (25.51%) and Asian (15.83%) cuisines underscores the influence of consumer preferences on restaurant success. This diversity in cuisine types suggests the need for the predictive model to account for market demands and culinary trends.

Table 22: Categorical table for parking category

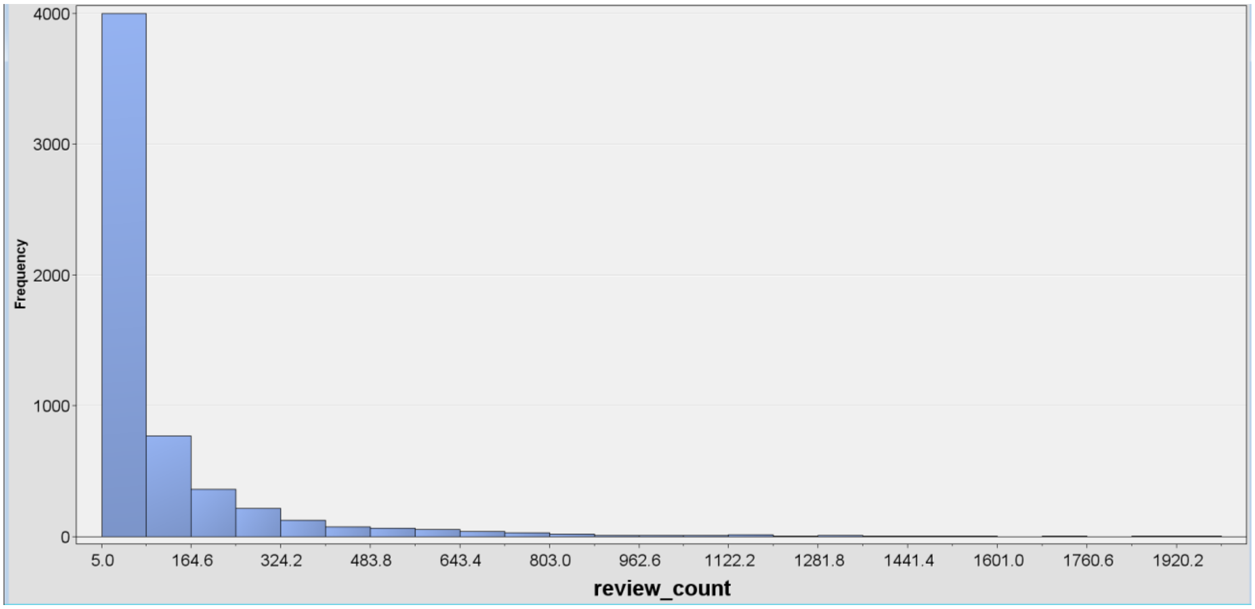
Variable	**level**	**Frequency**	**Percent**
garage	0	5465	93.95
	1	352	6.05
Street	0	2660	45.73
	1	3157	54.27
Validated	0	5706	98.09
	1	111	100.00
Lot	0	5026	86.40
	1	791	100.00
Valet	0	5713	98.21
	1	104	1.79

The table explain**s** the availability of street parking (54.27%) over garage parking (6.05%) emphasizes the importance of accessible location features in attracting customers. This logistical aspect represents a critical factor in the predictive analysis of restaurant closures.

Distribution of key variables based on category:

Business details:

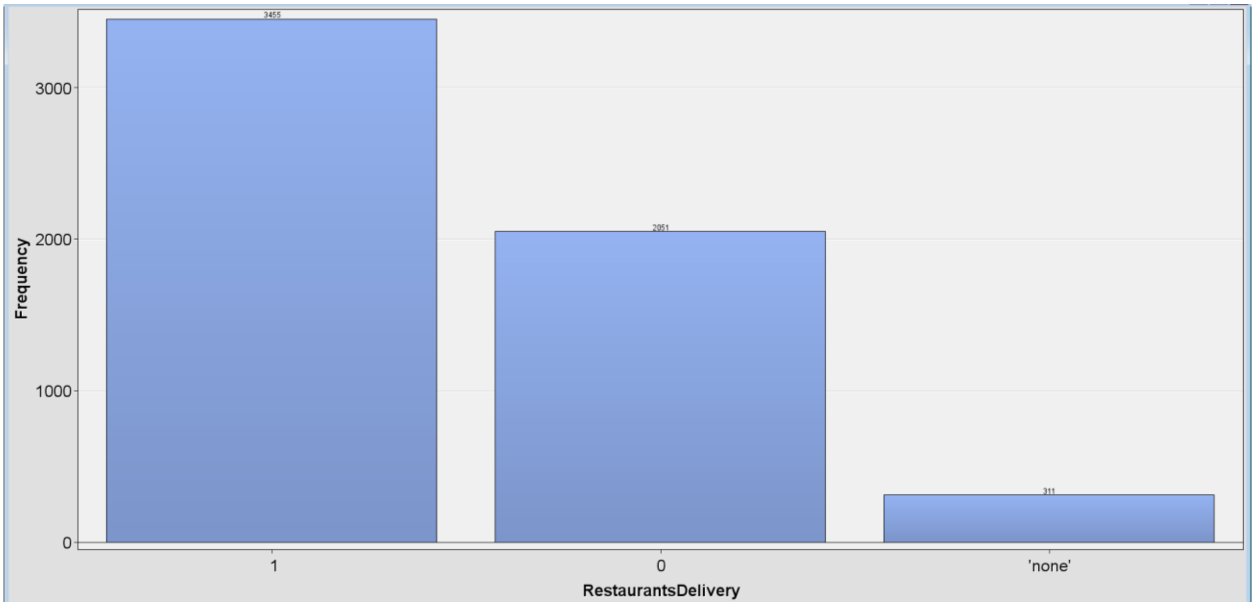
Figure - 1: Histogram of the Review_count variable



This histogram illustrates the distribution of review counts across restaurants. The heavy right skew indicates that most of the restaurants have a relatively low number of reviews, with a few outliers having significantly higher review counts. This suggests that while a few restaurants are highly popular and frequently reviewed, a vast majority might struggle to attract similar attention. For predictive modeling, this indicates that the review count could be a significant predictor of restaurant closure, with those having higher review counts possibly having a higher chance of staying open.

Customer convenience:

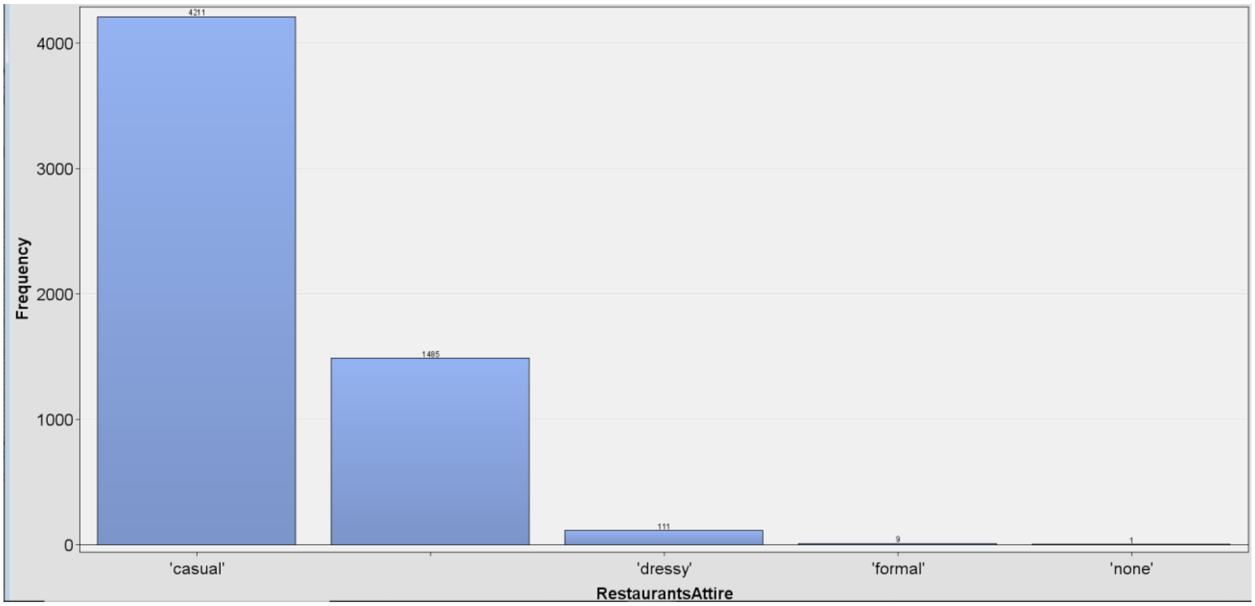
Figure - 2: Bar chart of the Restaurant delivery variable



The bar chart displays the proportion of restaurants offering delivery services. A significant majority provides this service, indicating its importance in the current dining ecosystem. The demand for convenience, possibly amplified by recent global events such as the COVID-19 pandemic, underscores delivery services as a crucial factor in a restaurant's operational strategy, potentially influencing its survival.

Dining Experience

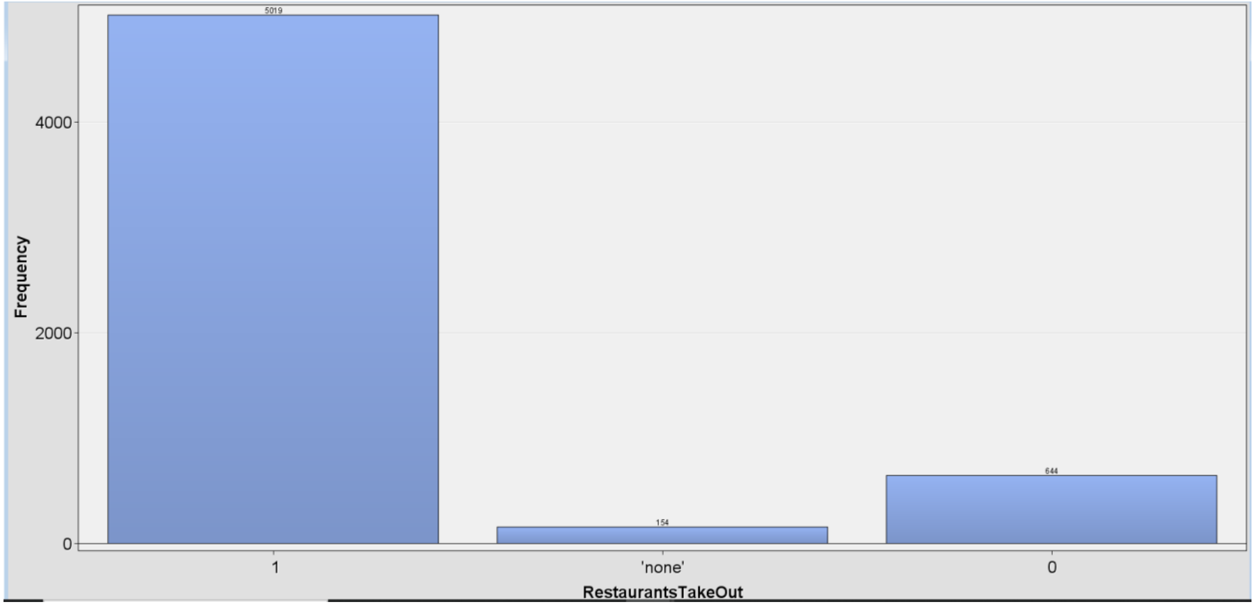
Figure - 3: Bar chart of the Restaurant Attire variable



This bar chart shows most restaurants prefer casual attire, suggesting a widespread trend towards more relaxed dining environments. The small fractions of dressy and formal categories imply that upscale dining forms a smaller niche. The attire might reflect the restaurant's target market segment, with casual attire aligning with broader appeal and potentially greater longevity due to wider customer base accessibility.

Food service

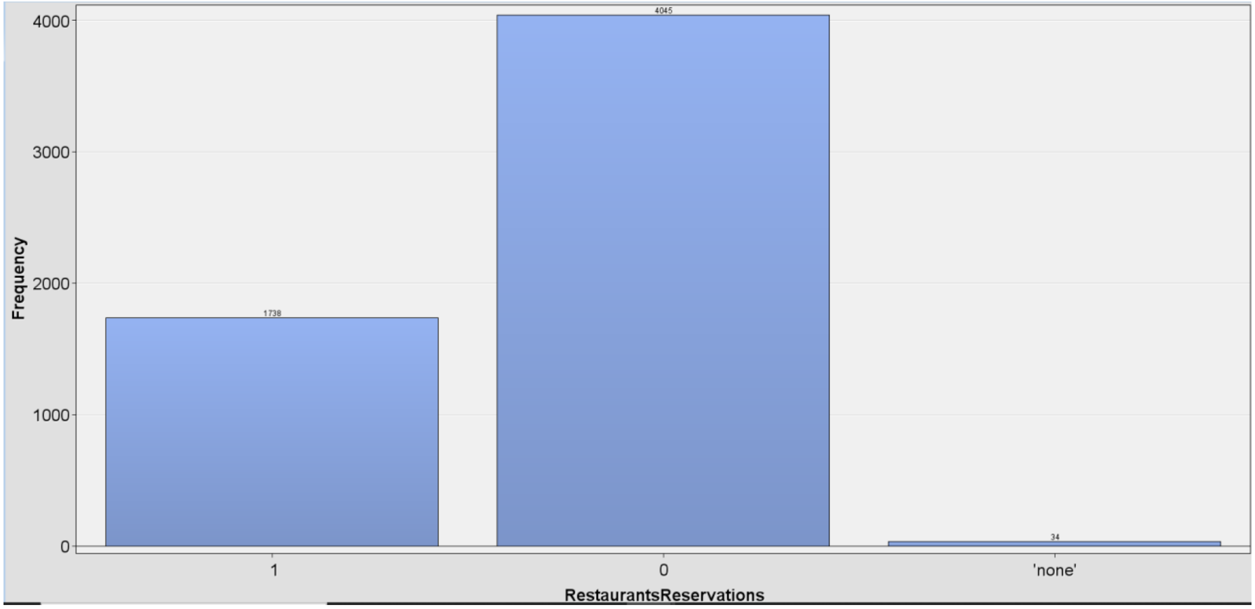
Figure - 4: Bar chart of the Restaurant Take out variable



The overwhelming majority of restaurants offering takeout services, as depicted in this bar chart, aligns with modern consumer preferences for convenience. Like delivery services, the availability of takeout could be a significant factor in attracting customers, particularly in urban areas or among younger demographics, thereby affecting a restaurant's survival prospects.

Reservation and Events

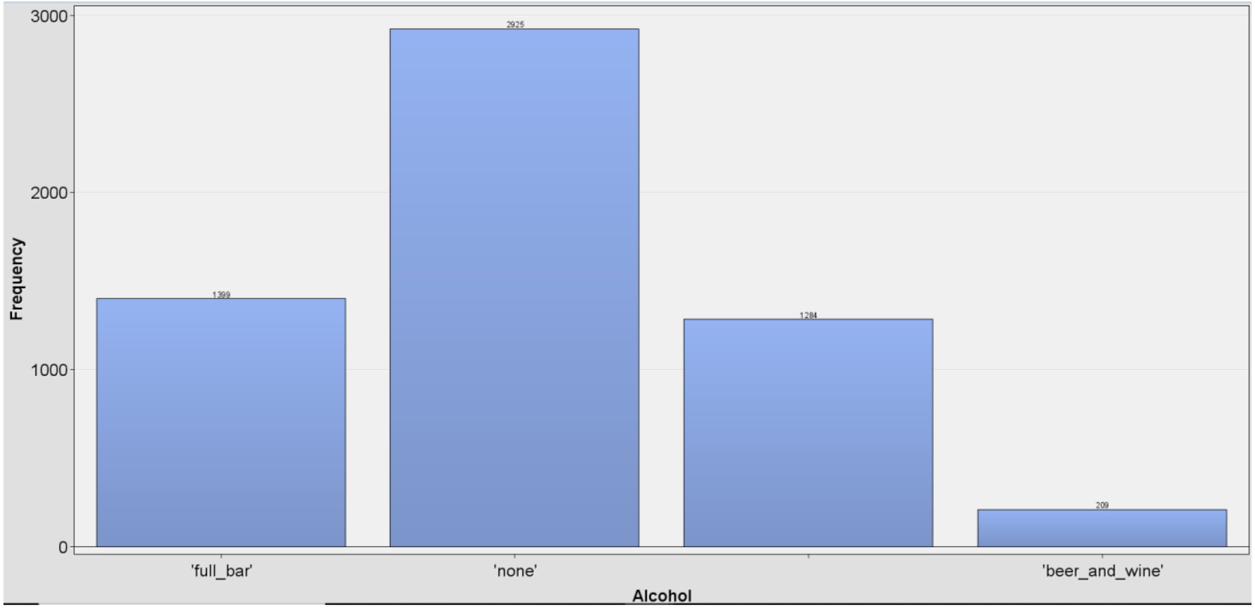
Figure - 5: Bar chart of the Restaurant Reservations variable



This chart indicates a balanced approach to reservations, with a notable percentage accepting them. This capability might provide customer preferences for planned dining experiences, especially in higher-end or busy restaurants. It suggests that while spontaneity in dining remains popular, the ability to secure a table beforehand could influence a restaurant’s appeal and operational efficiency.

Alcohol and beverages

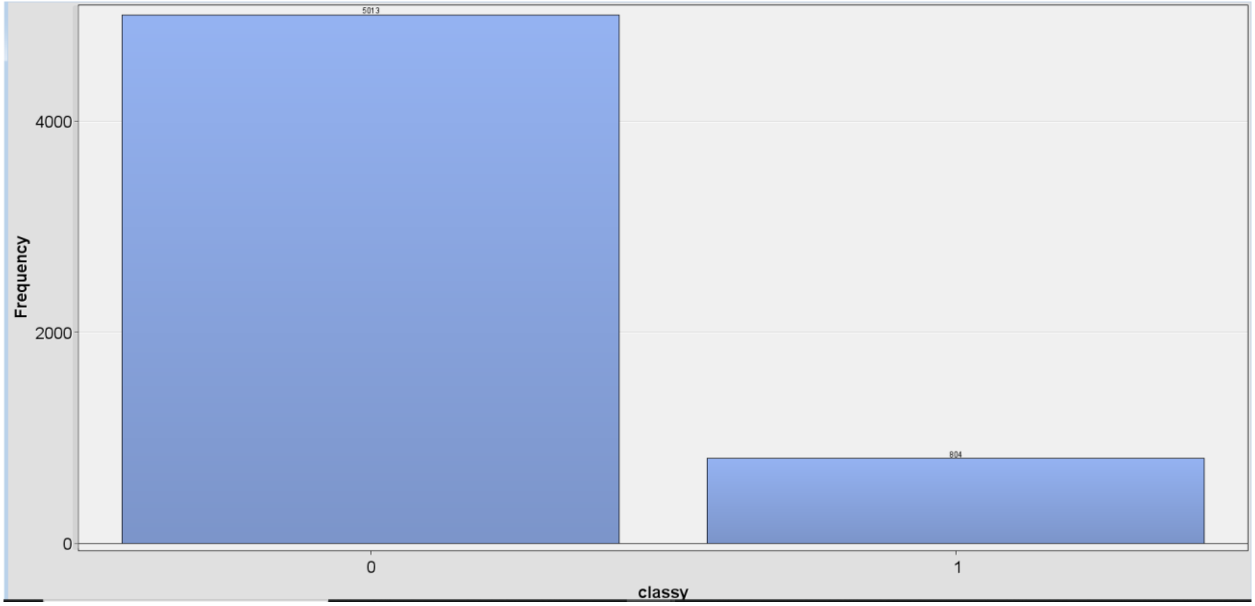
Figure - 6: Bar chart of the Alcohol variable



The distribution of alcohol service options, as shown, highlights a majority of restaurants not serving alcohol. This could reflect licensing restrictions, operational choices, or target market preferences. Restaurants offering a full bar service represent a significant segment, potentially catering to different or more specific customers, which could impact their survival differently compared to those not serving alcohol.

Ambience

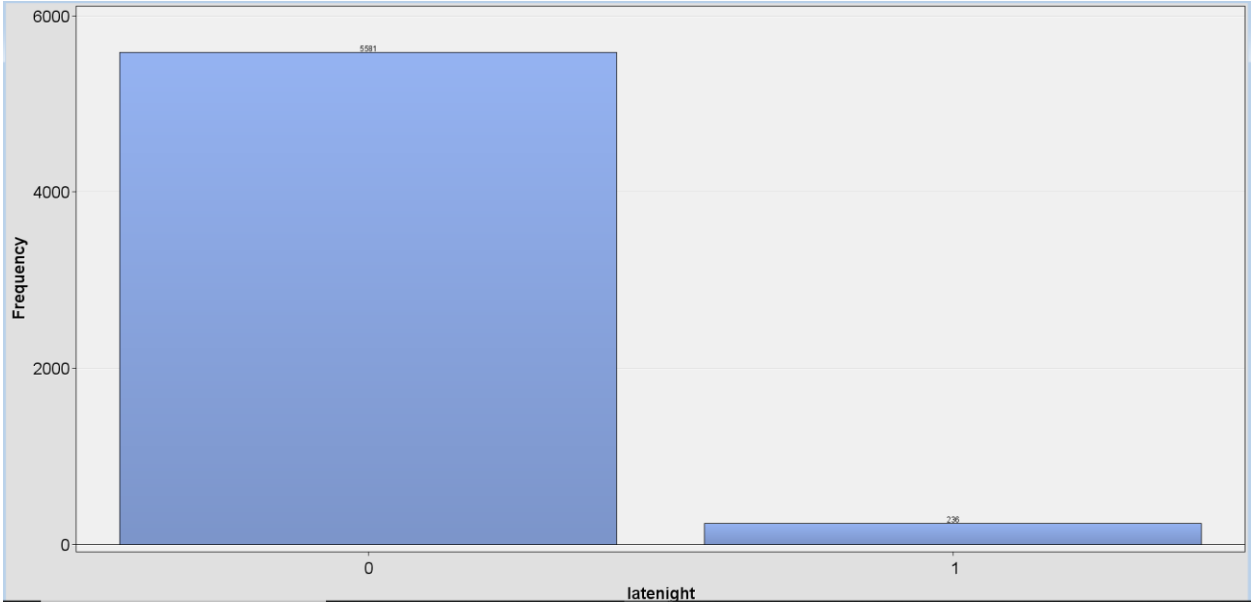
Figure - 7: Bar chart of the classy variable



The minority of restaurants identified as "classy" could indicate a unique market within the broader restaurant industry. These establishments may provide a specific demographic seeking upscale dining experiences, which could affect their survival differently from more casual or mainstream restaurants.

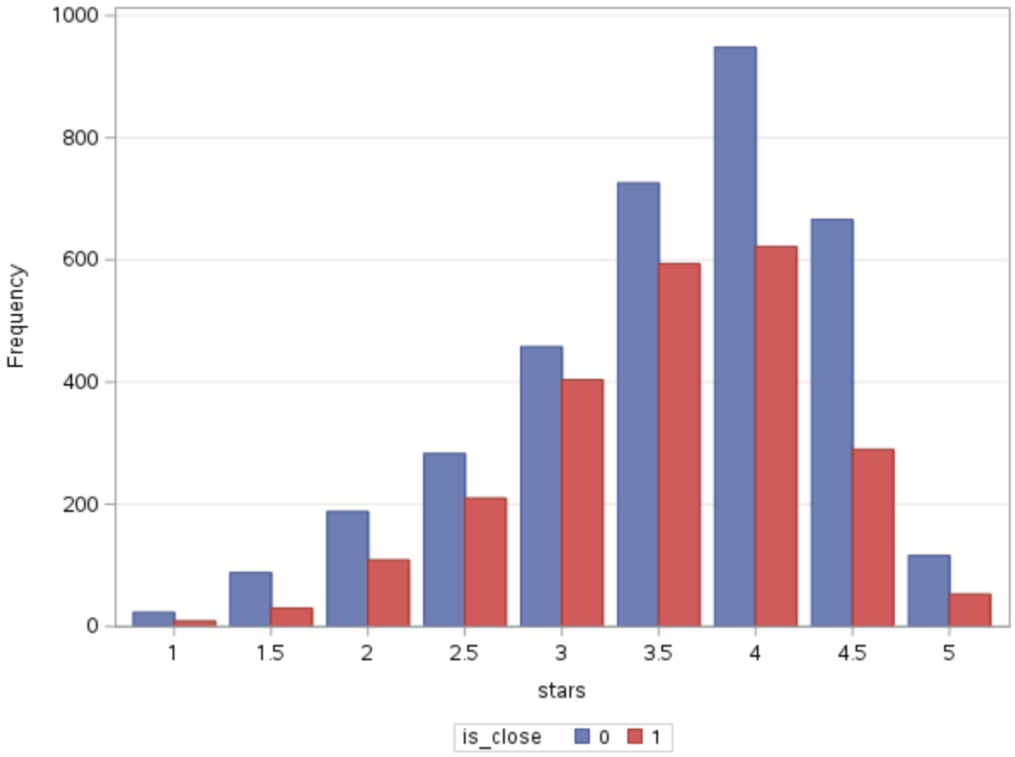
Atmosphere

Figure - 8: Bar chart of the late-night variable



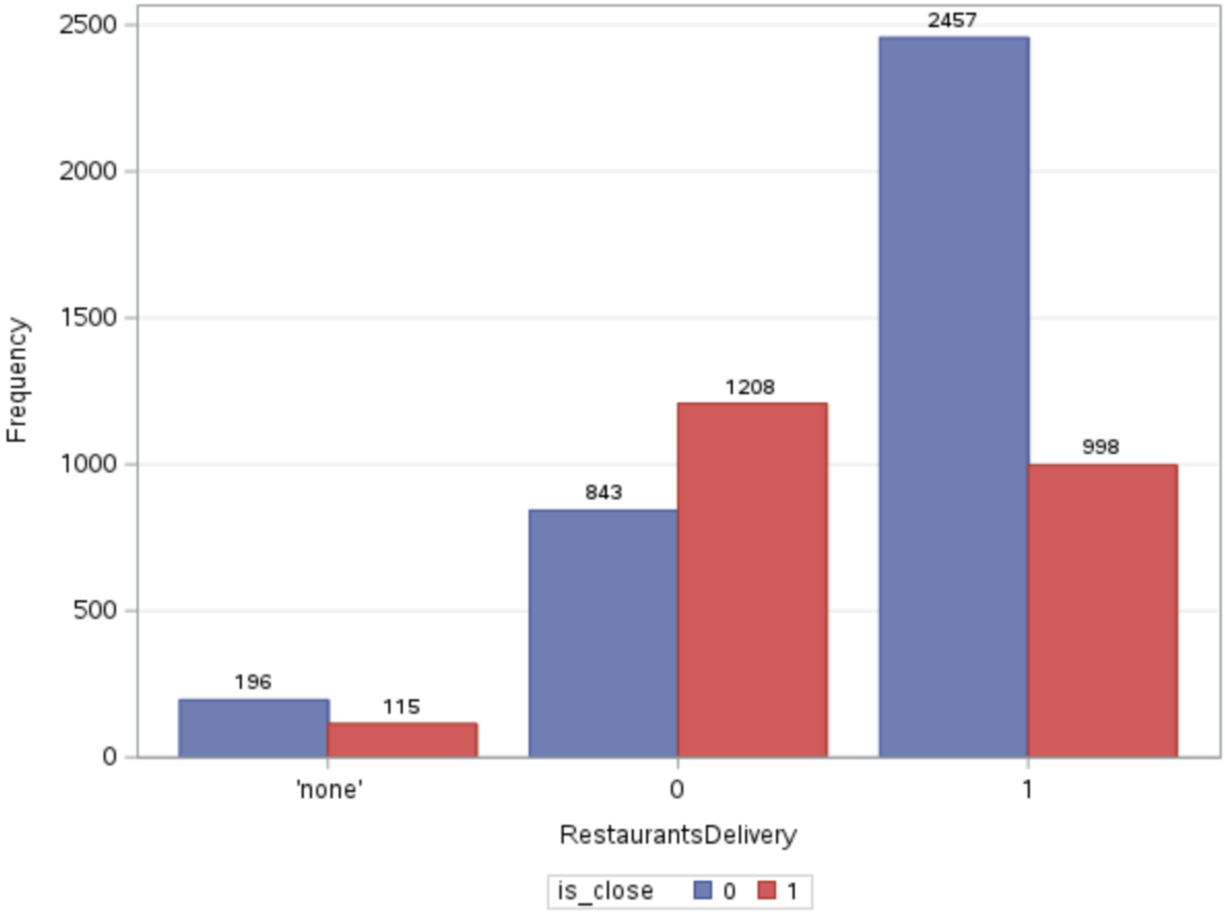
A small percentage of restaurants operate late at night, suggesting to meet the needs of the market. The decision to offer late-night services could be influenced by location, target demographic, or operational capabilities. This service could attract a specific customer segment, potentially impacting the restaurant's survival by catering to after-hours demand.

Figure - 9: Bar chart of stars by target variable



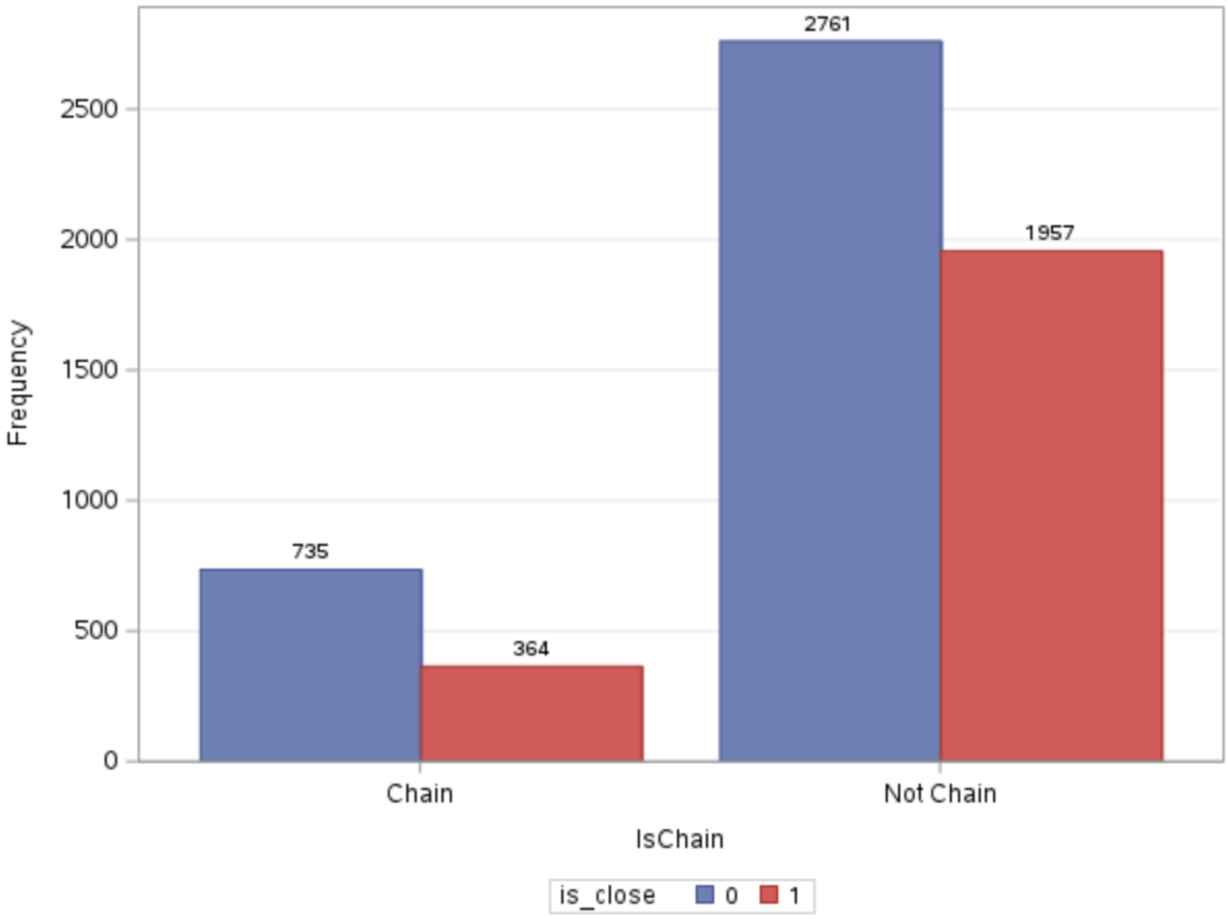
The chart above clearly shows the distribution of restaurant star ratings in relation to closure status reveals a bimodal pattern, with peaks around ratings of 3.5 and 4.5 stars. Notably, restaurants with lower ratings (1-2.5 stars) exhibit a higher frequency of closures, while those rated higher (3-5 stars) predominantly remain open. This suggests that higher star ratings, indicative of greater customer satisfaction and perceived quality, play a crucial role in a restaurant's likelihood to stay in business. Hence, maintaining high standards in service and food quality is essential for enhancing customer ratings and decreasing the risk of closure.

Figure - 10: Bar chart of Restaurant delivery by target variable



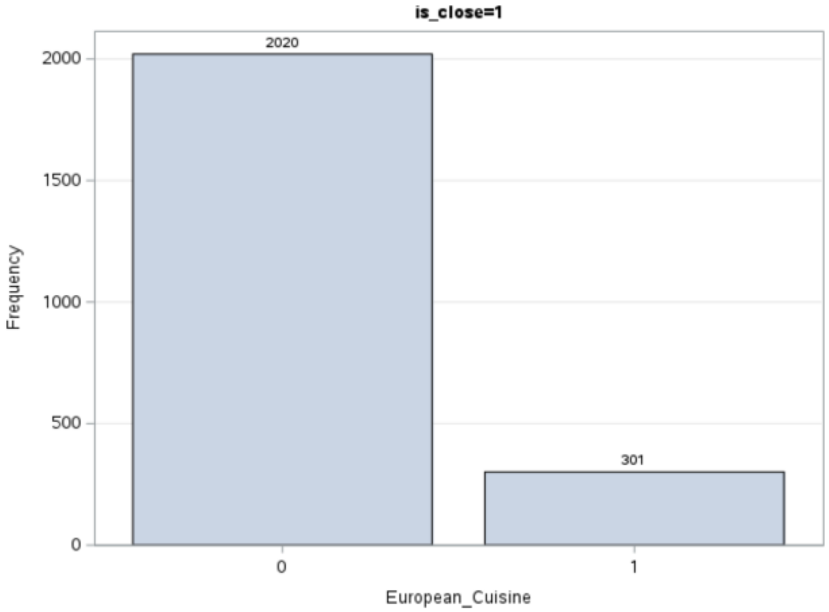
The restaurant delivery chart above underscores the importance of delivery services in restaurant operations, showing that restaurants offering delivery tend to have a lower frequency of closure. This trend highlights the role of delivery services in increasing accessibility and convenience, which can attract more customers and boost revenue—a vital factor for survival. The analysis suggests that integrating delivery services could be a strategic approach to improve a restaurant's resilience against closure, particularly important in urban environments or during situations that restrict in-person dining, such as pandemics.

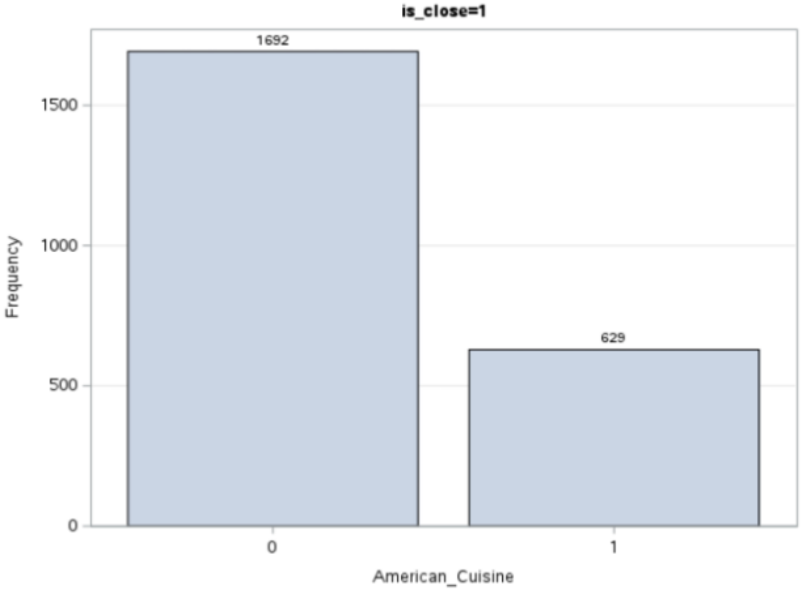
Figure - 11: Bara chart of Is_chain by target variable

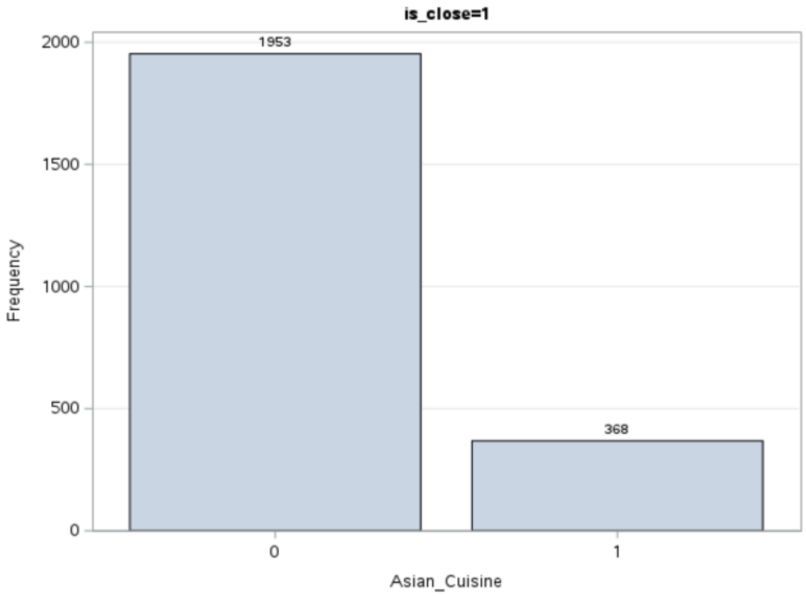


The chart examines the impact of chain affiliation on restaurant closure rates. It illustrates that chain restaurants experience significantly lower closure rates compared to independent establishments. Chains often benefit from established brand recognition, centralized marketing strategies, and robust financial support, which can help mitigate economic challenges and sustain operations. The data indicates that being part of a restaurant chain can be a protective factor against closure. For independent restaurants, forming alliances or considering franchise opportunities could be effective strategies to harness some benefits typically enjoyed by chain establishments.

Figure - 12: Bar charts of European, American and Asian cuisine by target variable = 1

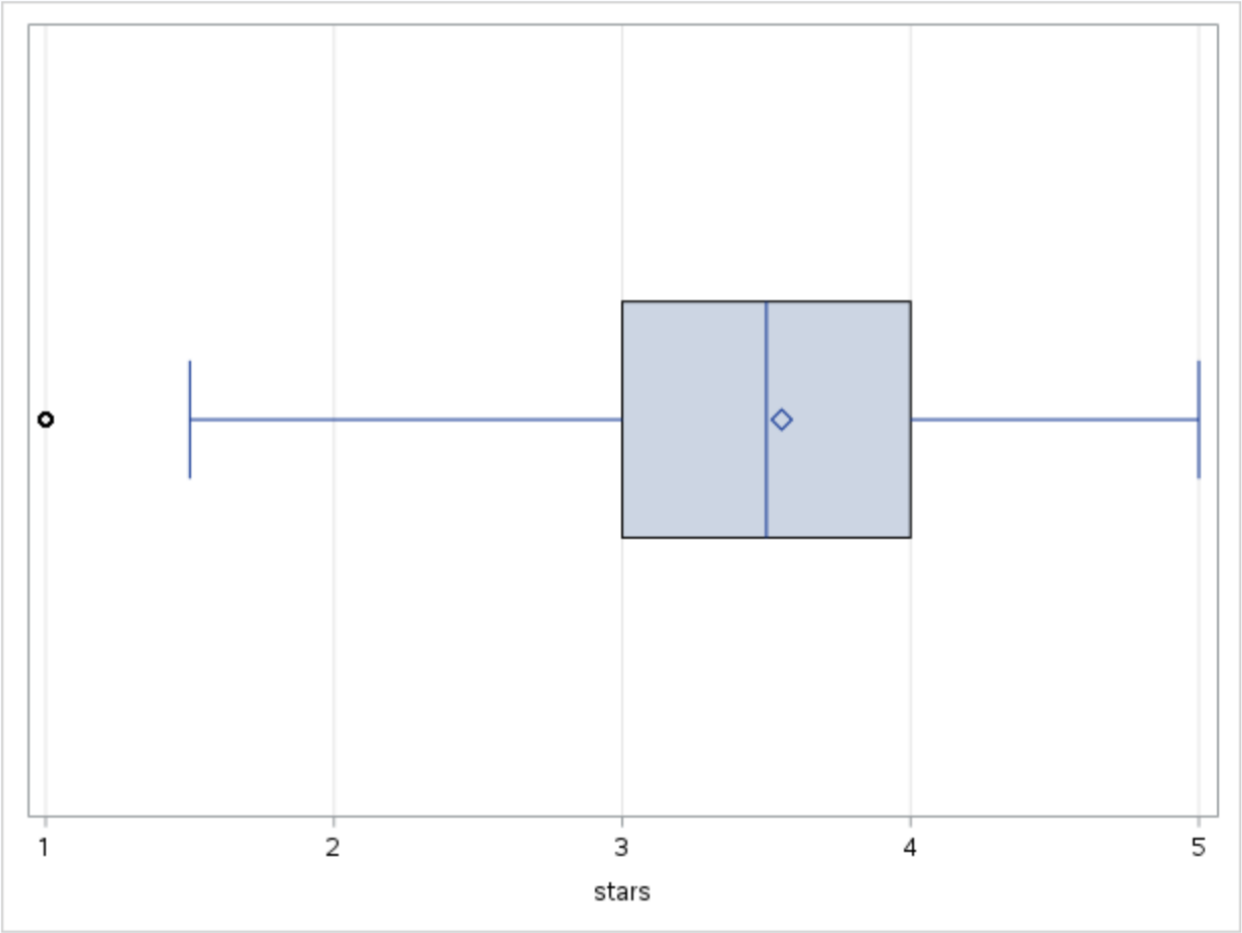






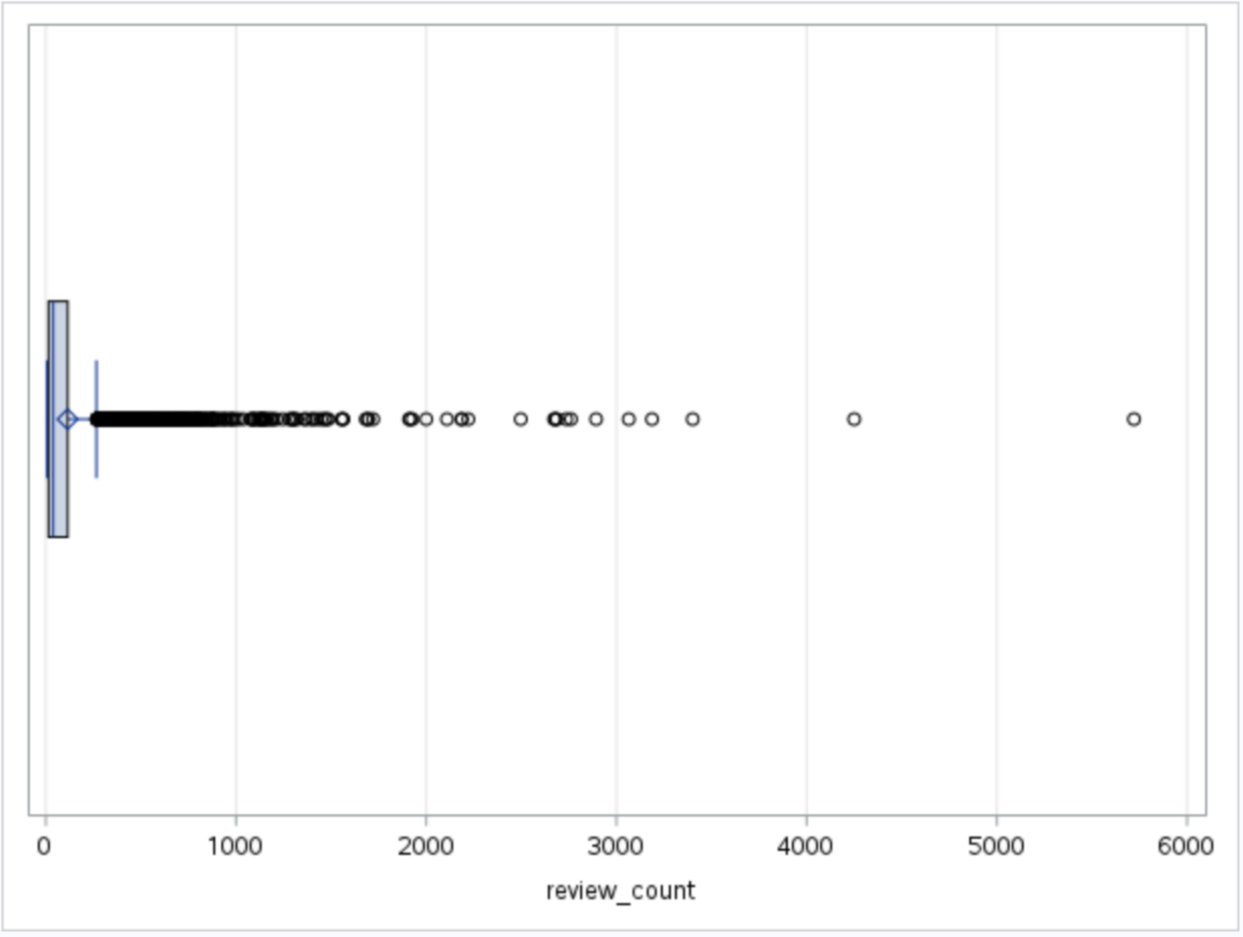
The provided graphs illustrate the proportion of closed restaurants in the year 2005 to 2022 that offered European, American, and Asian cuisines. Specifically, the graphs focus on restaurants that have ceased operations (is_close=1), and the classification is based on whether these closed establishments offered the specified cuisines (1 for yes and 0 for no). For European cuisine, 15% of the closed restaurants offered European dishes, while 85% did not. In the case of American cuisine, 37% of the closed restaurants served American food, compared to 63% that did not. Lastly, for Asian cuisine, 19% of the closed restaurants featured Asian menus, whereas 81% of the closures did not involve Asian cuisine. These figures reflect the distribution of cuisine offerings among restaurants that were unable to sustain operations during the period analyzed.

Figure - 13: Box plot of the stars variable



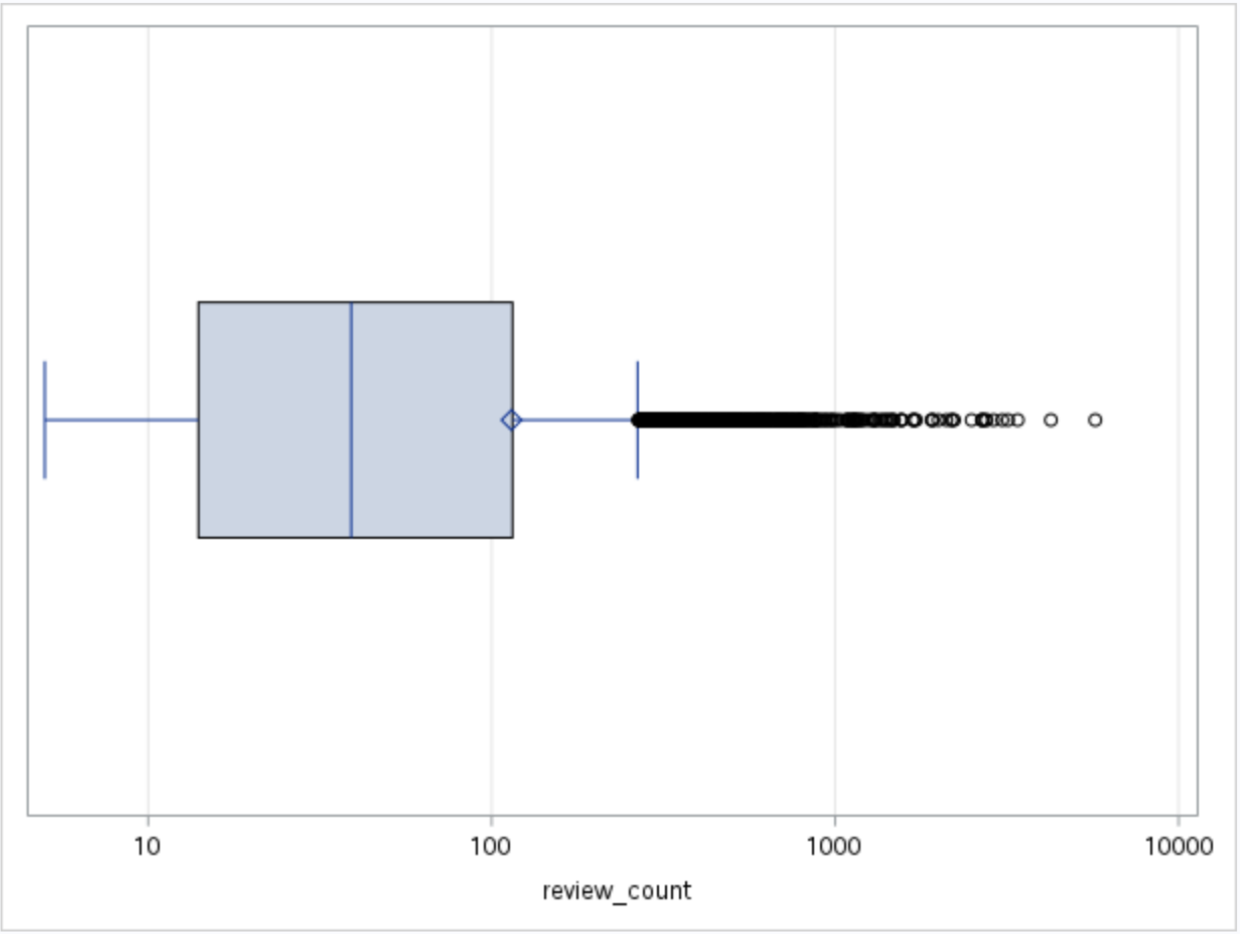
The provided box plot visualizes the distribution of star ratings for restaurants. The median rating appears to be in between 3.5 to 4-star range, with the box representing the middle 50% of the data clustered tightly around this median, indicating a majority of the restaurants have ratings in this range. The presence of outliers on the lower end suggests a small number of restaurants have ratings significantly below the median. This concentration around the median suggests a competitive market where many restaurants perform at a similar level in terms of customer-perceived quality. For predictive analysis, the proximity of the ratings to the median could suggest that even slight variations in star ratings might influence a restaurant's likelihood of survival, with those rated below the median potentially at higher risk of closure.

Figure - 14: Box plot of the review_count



Since it was difficult to see the box plot (figure-14) due to so many outliers, we have decided to use a logarithmic scale which shows clear graph of the box plot and outliers (figure-15)

Figure - 15: Logarithmic scale Box plot of the review_count.



The box plot for the review count variable displays a log-transformed scale, which is used to manage the wide variance in review counts and handle the skewness evident from the presence of outliers. The median value is situated toward the lower end of the scale, indicating that a typical restaurant has a relatively moderate number of reviews. However, the distribution's long right tail, with many outliers, points to a few restaurants with a significantly higher count of reviews, suggesting these are much more frequently reviewed or possibly more popular. For predictive modeling, the difference in review counts could be an indicator of restaurant visibility and popularity, which may correlate with their survival chances. The high number of reviews for certain restaurants might also be associated with longevity and success in the restaurant industry.

Correlation analysis:

Table - 23: Correlation matrix of interval variables

Pearson Correlation Coefficients, N = 5817		
	stars	**review_count**
stars	1.00000	0.14694
review_count	0.14694	1.00000

The correlation coefficient between "stars" and "review_count" is 0.14694 (Table - 23), indicating a weak positive correlation. This suggests that there is a slight tendency for businesses with higher star ratings to have more reviews, but the relationship is not strong.

Chi-square test for the key variables:

Table – 24: Chi-square test for Restaurant delivery and Target variable

Statistic	**DF**	**Value**	**Prob**
Chi-Square	2	484.5979	<.0001
Likelihood Ratio Chi-Square	2	483.5354	<.0001
Mantel-Haenszel Chi-Square	1	278.3764	<.0001

Phi Coefficient		0.2886	
Contingency Coefficient		0.2773	
Cramer's V		0.2886	

The table above summarizes the results of a chi-square test that evaluates the relationship between a restaurant's delivery service and the variable of interest (is_close) that indicates whether the restaurant is closed or not closed. The standard chi-square test and chi-square likelihood ratio test both produce very high values (484.5979 and 483.5354, respectively) with two degrees of freedom, with p-values as low as 0.0001 indicating a statistically significant association. It shows that there is a very close relationship between a restaurant's delivery capacity and its operational status.

Table – 25: Chi-square test for Restaurant Attire and Target variable

Statistic	DF	Value	Prob
Chi-Square	3	6.2794	0.0988
Likelihood Ratio Chi-Square	3	6.7810	0.0792
Mantel-Haenszel Chi-Square	1	0.4857	0.4858
Phi Coefficient		0.0381	
Contingency Coefficient		0.0380	
Cramer's V		0.0381	

The table presents the results of various chi-square tests that examine the relationship between restaurant attire and a target variable (restaurant closure). The traditional chi-square test and the likelihood ratio chi-square test, each with 3 degrees of freedom, indicate a weak association between the variables (p-values of 0.0988 and 0.0792, respectively). However, these results do not reach traditional levels of statistical significance ($p < 0.05$).

Table – 26: Chi-square test for Alcohol and Target variable

Statistic	DF	Value	Prob
Chi-Square	2	10.5457	0.0051
Likelihood Ratio Chi-Square	2	10.5204	0.0052
Mantel-Haenszel Chi-Square	1	3.0896	0.0788
Phi Coefficient		0.0482	
Contingency Coefficient		0.0482	
Cramer's V		0.0482	

The table above shows the results of a chi-square test examining the relationship between alcohol variable and the dependent variable (is_close). The standard Chi-Square test, with 2 degrees of freedom, shows a value of 10.5457 and a statistically significant p-value of 0.0051, suggesting a strong association between alcohol consumption and the target variable.

Table – 27: Chi-square test for Restaurant Reservations and Target variable

Statistic	**DF**	**Value**	**Prob**
Chi-Square	2	89.1473	<.0001
Likelihood Ratio Chi-Square	2	88.2296	<.0001
Mantel-Haenszel Chi-Square	1	66.1432	<.0001
Phi Coefficient		0.1238	
Contingency Coefficient		0.1229	
Cramer's V		0.1238	

The table above shows the results of various chi-square tests examining the relationship between restaurant reservations (a categorical independent variable) and a target variable (is_close). The p-value is highly significant (less than 0.0001), all tests provide strong evidence to reject the null hypothesis, suggesting that there is a significant relationship between the two variables. The chi-square statistic and chi-square likelihood ratio both showed similar values (89.1473 and 88.2296, respectively) with 2 degrees of freedom, indicating the robustness of the results.

Table – 28: Chi-square test for classy and Target variable

Statistic	DF	Value	Prob
Chi-Square	1	140.5187	<.0001
Likelihood Ratio Chi-Square	1	151.4292	<.0001
Mantel-Haenszel Chi-Square	1	140.4946	<.0001
Phi Coefficient		-0.1554	
Contingency Coefficient		0.1536	
Cramer's V		-0.1554	

The table above shows the results of a chi-square test examining the relationship between the variables 'is_close' and 'classy'. A chi-square statistic of 140.5187 with degrees of freedom and a p-value less than 0.0001 indicates a statistically significant relationship between the two variables.

One sample T-test:

Table – 29: One sample T-test for Review_count

Tests for Normality				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.325829	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	181.0438	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	912.7077	**Pr > A-Sq**	<0.0050

N	Mean	Std Dev	Std Err	Minimum	Maximum
5817	114.3	242.0	3.1730	5.0000	5721.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
114.3	108.0	120.5	242.0	237.7	246.5

DF	t Value	Pr > |t|
5816	36.01	<.0001

The one-sample t-test conducted on the variable 'review_count' aimed to determine if the mean review count of 114.3 significantly differs from a hypothesized population mean. Based on the results, the t-statistic of 36.01 with 5816 degrees of freedom resulted in an extremely low p-value (<0.0001), indicating strong evidence against the null hypothesis of no difference. This suggests that the mean review count of 114.3 is significantly different from the hypothesized population mean. Additionally, prior to conducting the t-test, normality assumptions were assessed using Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests, all of which showed statistically significant departures from normality (p < 0.01).

Figure - 16: Variable worth plot from stat explore node

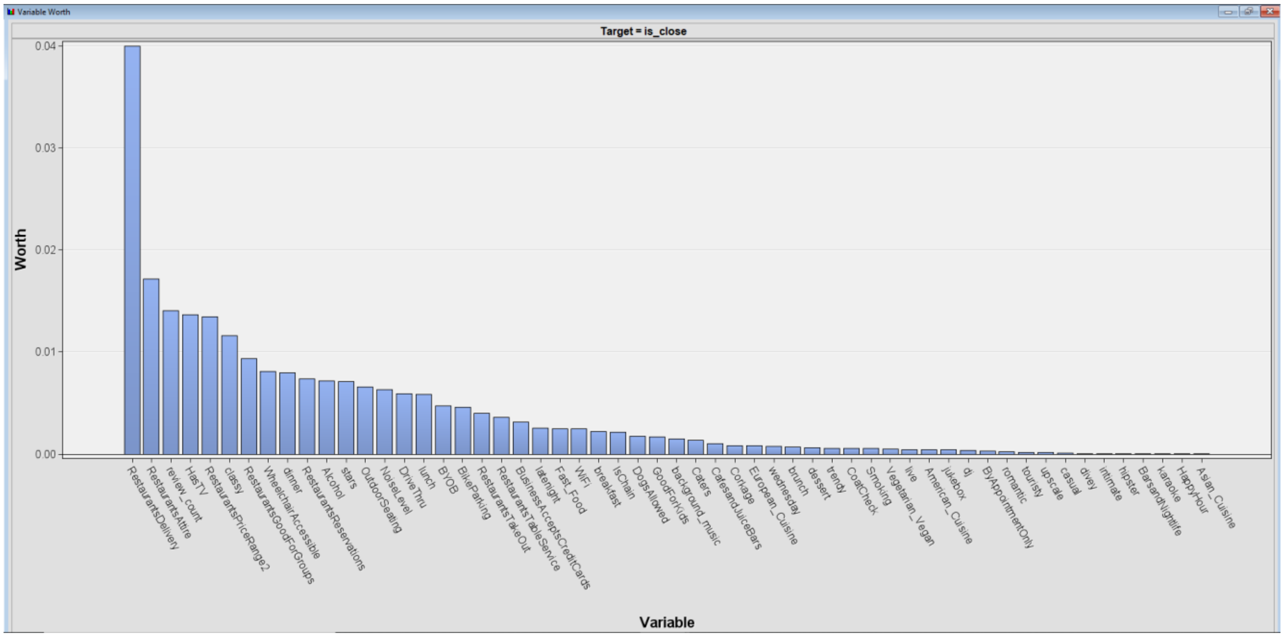


Figure - 16 shows the variable worth plot without any data preparation. It shows that the restaurant delivery shows the highest worth among all the variables followed by the restaurant attire, review counts of the restaurant, if the restaurant has T.V. or not and restaurant price range from 1 to 5.

Modeling –

We used SAS enterprise miner to develop the predictive model. The non-text and text file was combined to test with different types of models to choose the best performing out of it. Since our data contains only two interval variables other variables are categorical, there might be a chance that one model will work better than the other. The models which are most used in machine learning are logistic regression, decision tree, neural network and partial least square (PLS). There are some other models which are not that common are Principal component analysis (PCA), least angle regression (LARS), Least absolute shrinkage and selection operator (LASSO), Adaptive LASSO and support machine vector (SVM). For our predictive modeling, we will be testing all the models to see which model turns out to be a champion model.

Logistic regression – The logistic regression estimates the odds ratio i.e., the odds of the positive outcome for an input variable with respect to a target variable.It excels in managing both categorical and numerical data and is particular utility in scenarios that require probability computation .make decisions, or group the data.

Decision tree – The decision tree has the ability to handle missing data without imputation of the missing values. It makes a tree-like structure of the variables starting with the most important to least important. The end part of the tree is called leaves because there is no further division of the variables after that. Decision tree is renowned for its ease of interpretation and its ability to provide valuable insights.

Neural network – The neural network got its name and function from the human nervous system. The nervous system works by stimulating one nerve cell to another. Similarly, the neural network is good at pattern recognition and contains interconnected nodes and a layer. It is helpful when there is a large amount of data which needs to be assessed.One of the disadvantages of a neural network is creating its own variable which cannot be explained. Thus, while choosing the model it is not considered because not much information can dig out from it.

Partial least squares – The partial least squares is used as an advance filtering for regression. It extracts the latent variables from the target and input variables which is best for predicting outcome. PLS is used especially for the variables with high collinearity or when the predictors exceed the number of observations.

Principal component analysis – PCA’s main function is to reduce the dimensionality while keeping as much variance as possible. It converts all the high correlated variables to the set of small uncorrelated variables which is known as the principal component. The PCA is then connected to regression for in depth variable selection.

Least angle regression (LARS) – LARS is a type of regression which is like the forward stepwise regression, but the only difference is that it changes the coefficient of the variables at every step to manage multicollinearity more effectively. It is used where the predictors are more than the observations. The variables are considered in the model step by step based on the correlation with the response.

Least absolute shrinkage and selection operator (LASSO) and Adaptive LASSO – LASSO and adaptive LASSO are also types of regression but with some modifications. LASSO uses the technique of penalizing coefficients which makes some of the coefficients zero which makes it efficient in variable selection. It is used in datasets with high dimensionality which prevents overfitting**.**

Adaptive LASSO differs from LASSO as it introduces weights for each variable which can have different coefficients. It works similar to LASSO, but it penalizes more for highly important variable selection.

Support vector machine (SVM) – To make the most efficient use of SVM, it is used after the regression node because it works by finding a hyperplane or in simple language a line which separates one class of data points from another. The best hyperplane is selected based on the maximum margin from both classes.

Text parsing – The main function of this node is to break the sentences into words and phrases. It automatically removes the stop words i.e., the, and, of, etc. it also has an option whether to keep noun, pronoun, adjective, etc. this node helps in the text preprocessing by pulling out useful information for further analysis.

Text filter – This node is used after the initial text parsing. We can put a limit on the minimum number of times for a word to appear in the document. It means it will filter out all the words which do not appear in the document for a minimum number. This way it will remove all the irrelevant words.

Text Cluster – This node makes a group or cluster of words which are similar to each other. It is an unsupervised learning technique which helps in natural grouping of the set of words. This cluster then goes under the modeling to find the important variables.

Text topic – This node is like a text cluster, but it uses Latent Dirichlet Allocation (LDA) to identify a collection of documents. It categorizes documents into interpretable topics to understand the large set of text data**.**

Text rule builder –It uses supervised learning where we can assign a target variable and the text rule builder node builds a rule based on the input text. It also gives precision, Recall and F-1 score. The text rules are defined based on the group of words which can be assigned to predefined sentiments.

Champion model –

The text (review) and the non-text features were run in all the models mentioned above to get the best performing model. This was done in the SAS miner with the help of the model comparison node which gives all the results in one table for easy comparison. As we can see from the table – 9, we have divided the non-text features into categories to see if there is any improvement in the assessment. The table – 30 shows the top model for each category and its corresponding accuracy, ROC, Sensitivity, specificity, precision and F1 score. It shows that the highest accuracy, ROC, sensitivity and F1 score was for the category with all the variables but, specificity and precision was highest for the category, Customer convenience + Business details + Dining experience + Food service + Reservations & Events + Alcohol & beverages. Accuracy is most important for our project because it is the number of correct predictions i.e., number of correct restaurant closure.Hence , we have chosen a LARS regression model with all the categories combined and with highest accuracy of 76.01% which is much higher compared to X. Lu et al 2018 of 67.49%

Table - 30: Model performance for Is_close target variable

Category	*Top Model*	*Accuracy*	*ROC*	*Sensitivity*	*Specificity*	*precision*	*F1*
Customer convenience	**SVM RBF**	68.86	0.643	40.78	87.48	68.35	51.08
Customer convenience + Business details	**Random forest**	68.72	0.726	41.64	86.68	67.46	51.49
Customer convenience + Business details + Dining experience	**Random forest larger**	73.32	0.783	46.81	90.91	77.35	58.32
Customer convenience + Business details + Dining experience + Food service	**SVM polynomial**	73.81	0.773	55.09	86.22	72.61	62.65
Customer convenience + Business details + Dining experience + Food service + Reservations & Events	**SVM polynomial**	73.98	0.773	51.21	89.08	75.67	61.08
Customer convenience + Business details + Dining experience + Food service + Reservations & Events + Alcohol & beverages	**Random forest larger**	73.91	0.789	44.74	93.25	81.48	57.76

Customer convenience + Business details + Dining experience + Food service + Reservations & Events + Alcohol & beverages + Ambience	**SVM linear**	74.84	0.786	55.69	87.54	74.77	63.83
Customer convenience + Business details + Dining experience + Food service + Reservations & Events + Alcohol & beverages + Ambience + meals + atmosphere	**LARS regression**	76.01	0.797	59.40	87.02	75.22	66.38
Customer convenience + Business details + Dining experience + Food service + Reservations & Events + Alcohol & beverages + Ambience + meals + atmosphere + cuisine type	**LARS regression**	76.01	0.797	59.40	87.02	75.22	66.38

The LARS's selected variables are shown in the table – 31 . we can see that none of the text analysis variables such as text cluster and text topic made it to variable selection. It shows that non-text features are more important than the text (review). The reason behind it might be that the reviews are too scattered and the presence of an enormous number of observations makes it difficult for variables to be selected in the final model. Thus, in the next section we did a separate text analysis without the non-text feature. The table shows the highest standardized estimate of 0.168 for variable restaurants delivery (0) which is positively related to the target variable 'is_close'. On the other hand, out of two interval variables, review count made it the variable selection and has standardized estimate of -0.093 which is negatively related to the target variable.

Table - 31: LARS selected variables

Variable	*Standardized Estimate*
RESTAURANTSDELIVERY_0	0.168
HASTV_0	0.146
M_RESTAURANTSATTIRE_0	0.128
CLASSY_0	0.095
WHEELCHAIRACCESSIBLE_0	0.049
DRIVETHRU_0	0.045
DINNER_0	0.042
IMP_ALCOHOL_'FULL_BAR'	0.015
BREAKFAST_0	0.015
IMP_NOISELEVEL_'QUIET'	0.008
BYOB_0	0.006

INTERCEPT	0.0
ISCHAIN_CHAIN	-0.009
OUTDOORSEATING_0	-0.033
RESTAURANTSGOODFORGROUPS_0	-0.040
BUSINESSACCEPTSCREDITCARDS_0	-0.051
RESTAURANTSRESERVATIONS_0	-0.089
LOG_REVIEW_COUNT	-0.093

LARS is a good tool to do variable selection but, adding a regression node after it makes it more efficient. No one in our literature review has done the odds ratio estimate. The table – shows the odds ratio estimate for the champion model. Here's how to interpret the odds ratio for the categorical variable. Let's take the variable with the highest standardized estimate, restaurant delivery. For restaurant delivery, the odds ratio (0 vs 1) estimates equals 2.531. This means that for cases with a 0 value for restaurant delivery, the odds of restaurant closure are 2.531 times higher than the odds of restaurant closure for cases with a 1 value for restaurant delivery. Now let's interpret a numerical variable, review count. For LOG review count, the odds ratio estimates equals 0.735. This means that for each additional review count, the odds of restaurant closures change by a factor of 0.735, a 26.5% decrease.

Table - 32: LARS Stepwise Regression output of odds ratio

Variable		*Point estimate*
BusinessAcceptsCreditCards	**0 vs 1**	**0.513**
DriveThru	**0 vs 1**	**6.123**
HasTV	**0 vs 1**	**2.990**
IMP_Alcohol	**'beer_and_wine' vs 'none'**	**1.106**
IMP_Alcohol	**'full_bar' vs 'none'**	**1.574**
IMP_NoiseLevel	**'average' vs very_'loud'**	**1.380**
IMP_NoiseLevel	**'loud' vs very_'loud'**	**1.048**
IMP_NoiseLevel	**'none' vs very_'loud'**	**<0.001**
IMP_NoiseLevel	**'quiet' vs very_'loud'**	**1.829**
LOG_review_count	**-**	**0.735**
M_RestaurantsAttire	**0 vs 1**	**2.609**
OutdoorSeating	**0 vs 1**	**0.743**
RestaurantsDelivery	**0 vs 1**	**2.531**
RestaurantsGoodForGroups	**0 vs 1**	**0.726**
RestaurantsReservations	**0 vs 1**	**0.500**
WheelchairAccessible	**0 vs 1**	**1.954**
breakfast	**0 vs 1**	**2.086**
classy	**0 vs 1**	**3.480**

dinner	**0 vs 1**	**1.694**

Text analysis –

We also considered doing text analysis separately because the text data is too scattered and because of its humongous size and number of observations, it will be a biased decision. The reviews by an individual may vary from a few words to even one hundred words. Thus, the results from the models will favor the restaurant features or non-text features because of its limitation to either interval, binary or categorical variables. The text analysis also behaves like regular modeling except it creates a group of words as a variable. The grouping of the words depends on many factors such as how many times that word appears, how rare it appears in the text, etc. We also used SAS enterprise miner for text analysis. The nodes we used were text parsing, text filter, text cluster, text topic and text rule builder.

Champion model for text analysis –

The champion model for the text analysis also turns out to be LARS regression with misclassification rate of 0.314. All the variables in the variable selection step are from the text topic node. The table - 33 shows the standardized estimates for the champion model. As we can see that all the estimates are positive and have values close to each other. In other words, we cannot get much information from the estimates. Overall, from the group of words customers want good service, good food, good experience, delicious and tasty food.

Table - 33: Standardized estimates for the stepwise regression

Variable	*Standardized estimate*
TextTopic_raw1 (service, place, server)	**0.0428**
TextTopic_raw14 (sauce, flavor, tasty, cheese, salad)	**0.0382**
TextTopic_raw3 (dish, flavor)	**0.0791**
TextTopic_raw4 (dinner, enjoy, friend, delicious, lunch)	**0.0434**
TextTopic_raw5 (meal, enjoy, experience)	**0.0669**
TextTopic_raw7 (cook, server, order)	**0.0501**
TextTopic_raw9 (excellent, service, table, food, flavor)	**0.0882**

The odds ratio estimate is shown in the table – 34 shows all the variables which were selected in the stepwise regression selection model. All the variables have point estimates greater than 1. Let's interpret the text topic_raw1, the odds ratio estimates equal 2.257. This means that for words 'service, place, server', the odds of restaurant closure change by a factor of 2.257, a 125.7% increase. In other words, customers prefer to have good service, a good location and place and good behaving servers. If the restaurant lacks these services, there are good chances for that restaurant to close in the near future. Other words mentioned in the table - are also equally important. There are some new words which X. Lu et al (2018) didn't extract in their text analysis, which are 'cook', 'salad' and 'experience'.

Table - 34: Odds ratio estimates for the stepwise regression

Variable	***Point estimate***
TextTopic_raw1 (service, place, server)	**2.257**
TextTopic_raw14 (sauce, flavor, tasty, cheese, salad)	**1.782**
TextTopic_raw3 (dish, flavor)	**1.326**
TextTopic_raw4 (dinner, enjoy, friend, delicious, lunch)	**1.641**
TextTopic_raw5 (meal, enjoy, experience)	**1.360**
TextTopic_raw7 (cook, server, order)	**1.546**
TextTopic_raw9 (excellent, service, table, food, flavor)	**1.631**

Recommendations –

The restaurant industry is such that it will never disappear from the market. People love to go out and enjoy food rather than cooking and doing dishes. So, it is important for a restaurant to compete with the rest of the neighboring restaurant to survive. Covid period (2019-2021) was one of the hardest hit periods for the restaurant industry, especially for the privately owned restaurants. Since our data is till 2022, it includes the effect of the covid. Our recommendation for the restaurant owners and investors is to focus on features such as restaurant delivery service. After 2019, Food delivery services such as Doordash, Uber eats, Grubhub, etc. became more famous as restaurants were only offering takeouts or delivery for a certain period. Since then, almost all the restaurants have had delivery partners, and it is necessary for restaurant survival according to our findings above. Other factors which a restaurant management could focus on is to have entertainment or have T.V. for the customers, have a drive thru if it is a fast-food restaurant, try to fit a full bar if it is a fine dining restaurant, have a dress code for the servers and other employees and ways to make the dining experience classy. The customers should also be encouraged to leave a review on yelp or google as review counts matter for a new customer to choose his/her meal for that day.

Conclusion -

Using Yelp data, the project sought to create a predictive model for restaurant closures. It was successful in identifying a number of important factors that lead to restaurant failures. The model distinguished between textual and non-textual features by utilizing machine learning techniques, such as LARS and LASSO regression, to evaluate an extensive dataset of Yelp reviews and restaurant attributes. The model offered detailed insights into the varying effects of different restaurant features on closure risks by classifying features such as accessibility and amenities. Additionally, by addressing common gaps in prior research, like the influence of non-textual features and the specificity of temporal and geographic factors, the study's methodological rigor was strengthened. The results broaden the current scholarly discourse by incorporating a variety of variables into the predictive model, but also offer practical implications for restaurant owners and investors by identifying actionable insights to mitigate closure risks. This comprehensive approach underscores the importance of both text and non-text data in understanding restaurant closures, providing a robust tool for investors to enhance operational strategies and sustainability in the competitive restaurant industry.

References:

Asghar, N. (2016). Yelp Dataset Challenge: Review Rating Prediction. arXiv preprint arXiv:1605.05362.

Chen, Y., & Xia, F. (2020). Restaurants' Rating Prediction Using Yelp Dataset. In 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). IEEE. DOI: 10.1109/AEECA49918.2020.9213704.

Eidul, T. S., Imran, M. A., & Das, A. K. (2022). Harnessing the power of Yelp reviews: A deep learning perspective on restaurant rating prediction. Decision Support Systems, 140, 113456.

Kovalcinova, L. (New Jersey Institute of Technology, NJ, US), & Polacek, M. (Stony Brook University, NY, US). (2015). Yelp Data-set Challenge Part 6: Predicting Whether Business is Open or Closed and Suggesting the Good Business Practices. New Jersey Institute of Technology, NJ, US; Stony Brook University, NY, US.

Liu, Z. (2020). Yelp Review Rating Prediction: Machine Learning and Deep Learning Models. Georgia Institute of Technology, Atlanta, GA, USA.

Luca, M. (2016). Reviews, Reputation, and Revenue: The Case of Yelp.com. Harvard Business School Working Paper. Retrieved from: https://www.hbs.edu/ris/Publication%20Files/12-016_a7e4a5a2-03f9-490d-b093-8f951238dba2.pdf

Luo, Y., & Xu, X. (2019). Predicting the Helpfulness of Online Restaurant Reviews Using Different Machine Learning Algorithms: A Case Study of Yelp. Sustainability, MDPI, 11(19), 1-17.

National Restaurant Association. (n.d.). State of the Industry. National Restaurant Association. https://restaurant.org/research-and-media/research/research-reports/state-of-the-industry/

Papaioannou, D. (2022). Sentiment analysis in user reviews: Enhancing business intelligence through natural language processing. Journal of the Association for Information Science and Technology, 73(1), 101-115.

Pandian, N., & Aggarwal, V. (2015). A new machine learning approach to predicting business shutdown. Proceedings of the 202X IEEE International Conference.

Papathanassis, A., & Knolle, F. (2010). Online holiday reviews: Implications for consumer behavior and marketing strategies. Tourism Management, 31(5), 657-665.

Qazi, A., Syed, K. B. S., Raj, R. G., Cambria, E., Tahir, M., & Alghazzawi, D. (2016). A study on the role of feedback in e-commerce: A machine learning approach. Computers in Human Behavior, 63, 397-414.

Racherla, P., & Friske, W. (2012). Exploring the role of online reviews in the hotel industry: A content analysis approach. Cornell Hospitality Quarterly, 53(4), 345-359.

Sedov, D. (2021). Restaurant closures during the COVID-19 pandemic: A descriptive analysis. Journal of Economic Perspectives.

Sharun S. T., & Devi, V. S. (2018). A Hybrid Deep Learning Model to Predict Business Closure from Reviews and User Attributes Using Sentiment Aligned Topic Model. SSCI 2018: 397-404.

Siddique, M. T. E., Imran, M. A., & Das, A. K. (2022). Harnessing the power of Yelp reviews:A deep learning perspective on restaurant rating prediction. Decision Support Systems, 140, 113456.

Singha, R., & Woo, J. (2019). Utilizing Yelp data for business analytics and decision making. Expert Systems with Applications, 126, 112-121.

Thazhackal, S. S., & Devi, V. S. (2018). A Hybrid Deep Learning Model to Predict Business

Closure from Reviews and User Attributes Using Sentiment Aligned Topic Model. SSCI 2018: 397-404.

Vallapuram, A. K., Nanda, N., Kwon, Y. D., & Hui, P. (2021). Interpretable Business Survival Prediction. Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Page range 1-8.

Wang, X., Tang, L., & Kim, E. (2018). The role of review helpfulness in online consumer behavior: An empirical study. Journal of Retailing and Consumer Services, 44, 24-32.

Yu, M., Xue, M., & Ouyang, W. (2010). The impact of user review sentiment on restaurant popularity: An empirical investigation. Journal of Hospitality Marketing & Management, 19(7), 772-788.