A Recipe for Success: A Data-Driven Exploration of Factors Influencing Restaurant Check-Ins on Yelp

Seminar Paper / Term Paper

Submitted to

Dr. Keyvan Dehmamy

Goethe University Frankfurt

Frankfurt am Main

by

Illia Tykhyi

Ginnheimer Landstraße 42

Frankfurt am Main 60487

s6508253@stud.uni-frankfurt.de

Program: M. Sc. Data Science and Marketing Analytics

Contents

List of Abbreviations

ZCTA……..……..……ZIP Code Tabulation Area

VIF…….Variance Inflation Factor

TDL……...……..…. Top Decile Lift

OCR……..…….…….. Online Customer Review

SMOTE…….……Synthetic Minority Over-sampling Technique

k-NN…….……k-Nearest Neighbors

SVMs……...…….… Support Vector Machines

ML……...……Machine Learning

ADASYN…….……...Adaptive Synthetic Sampling

NN……..……Neural Networks

List of Tables

List of Figures

Introduction

This study investigates the multifaceted factors influencing customer attendance at restaurants in Philadelphia, Pennsylvania, utilizing a rich dataset provided by Yelp, a leading platform connecting consumers with local businesses. Recognizing that attendance is a crucial driver of revenue and profitability, we aim to provide actionable insights for restaurant owners and investors seeking to optimize their business performance. Our analysis leverages Yelp's extensive data repository, encompassing business attributes, images, customer reviews, star ratings, and crucially, check-in frequencies, which serve as a direct proxy for customer attendance.

Our central hypothesis proposes that check-in frequency is influenced by a complex interplay of factors, including:

Internal Restaurant Attributes: Price range, number of reviews and average star rating, number of photos, availability of specific amenities - Happy Hour, delivery, takeout, etc.

External Competitive Factors: business proximity, landmarks nearby

Demographic Factors: General population density in the restaurant's zip code. Age group concentrations (youth, adults, elders, children) in the restaurant's zip code.

We hypothesize that restaurants with lower price ranges, attractive amenities like happy hour, strong online reputations (higher review count and average star rating), appealing visuals (more photos), fewer nearby competitors, higher number of nearby landmarks and locations in densely populated areas with a favorable age group composition will exhibit higher check-in frequencies.

To begin, we review existing research to identify knowledge gaps and assess the suitability of the chosen Yelp variables for our analysis. We then introduce the Yelp platform to provide context

for the specific internal data points used. We then introduce the Yelp platform to provide a broader understanding of the specific variables used.

Following this, we delve into the details of the dataset construction. This includes examining both internal Yelp data (e.g., restaurant attributes, reviews) and external data sources (e.g., demographics) acquired for Philadelphia restaurants. We describe the data preparation process, which involved addressing inconsistencies by marking missing information and removing entries with no reviews. Additionally, all data points were converted to a consistent format (typically a 0-1 scale) to facilitate analysis. We then tackled challenges like high correlations between demographic variables and imbalanced check-in data. After, the data was split into training and evaluation sets for machine learning model development.

The next step involves presenting the machine learning process itself. This includes the specific algorithms employed and the methods used to evaluate their performance. Finally, after identifying the most effective model, we leverage the results from a logit regression analysis, alongside the findings from the neural network, to generate recommendations for businesses.

Literature Review

The influence of population density, age demographics, online reputation, business proximity, price range, number of photos on profile and various physical attributes on restaurant attendance in Philadelphia is a complex but fascinating phenomenon. In this literature review we will explore research supporting our hypotheses regarding these factors and how they relate to restaurant check-in frequency.

Research on urbanization economies suggests that larger, denser cities tend to benefit from increased firm productivity (Collier, Jones, and Spijkerman 2018). This finding provides support for our hypothesis that restaurants in areas with higher general populations will see higher check-in frequencies on Yelp. The economic benefits of densely populated areas seem to translate into greater activity and patronage for businesses, including restaurants.

The influence of different age groups on restaurant attendance is well-documented. Young adults (18-24) often prioritize casual dining and social experiences (Quinn et al., 2019). This suggests that restaurants located in areas with a higher concentration of youth might see higher check-in frequencies. The role of social media in restaurant decisions further reinforces this hypothesis, as younger generations frequently utilize platforms like Yelp for restaurant discovery and reviews (Bilgihan et al., 2016; Ukpabi & Karjaluoto, 2016).

Adults (25-64) hold a larger share of working positions and, consequently, have greater disposable income (Statista, n.d.). This makes them a key target demographic for restaurants. Our hypothesis proposes that restaurants in areas with higher adult concentrations would see higher check-in frequencies. This hypothesis finds indirect support from research on the impact of location decision on small enterprise performance (Lumbwe et al. 2018). Their findings highlight the importance of proximity to customers for business success, suggesting that

restaurants located near high concentrations of adults could benefit from increased foot traffic and patronage.

In contrast to younger demographics, elderly consumers (65+) might exhibit lower restaurant attendance, potentially due to their preference for specific cuisine types and lower mobility (Guido et al., 2022). Additionally, their lower engagement with technology (Rybaczewska & Sparks, 2022) might lead to less online exploration of new restaurants.

Family-oriented dining is another significant factor. Our hypothesis predicts that restaurants in areas with a higher concentration of children will experience higher check-in frequencies. This is supported by research that suggests accessibility of the business location is crucial for attracting families (Mkwanazi and Mbohwa, 2016b). Restaurants strategically located in family-friendly areas are likely to benefit from higher check-in frequencies.

The importance of online reputation is undeniable. A survey by Restaurant Insights found that a significant portion of US restaurant-goers (68%) visit restaurants based solely on positive online reviews (Restaurant Insights, 2017). This reinforces our hypothesis that restaurants with a higher review count will see higher check-in frequencies due to increased online visibility and brand awareness.

A restaurant's average star rating acts as a signal of its quality, influencing consumer decisions (de Langhe, Fernbach, & Lichtenstein, 2014). A higher average star rating could translate into higher check-in frequencies, suggesting that customer satisfaction drives repeated visits.

Numerous studies highlight the strong correlation between positive reviews and increased sales, booking intentions, and improved profitability for restaurants (Xie et al., 2016; Yang & Mai, 2010; Wang et al., 2021). This phenomenon is driven by the nature of restaurants as experience goods, where quality cannot be readily assessed before consumption (Tsao et al., 2015). Online Customer Reviews (OCRs) act as a valuable source of information for customers, reducing perceived risks and influencing their decisions (Filieri, 2016; Manes & Tchetchik, 2018).

The study of OCRs in the restaurant context is further complicated by the interplay of multiple cues, including brand strength, that customers use to evaluate a restaurant's quality. Existing research reveals a significant moderating effect of brand on the relationship between OCRs and restaurant performance (Ho-Dac et al., 2013; Wang et al., 2021). While positive reviews often contribute to increased success for lesser-known restaurants, strong brands may diminish the influence of positive OCRs (Wang et al., 2021). This can be attributed to the strong signals that established brands transmit, reducing customer reliance on OCRs for information (Akdeniz et al., 2013).

Visual appeal is also essential. The number of photos associated with a restaurant on Yelp suggests visual appeal and potential for user engagement (Pieters, Wedel, & Batra, 2010; Nardini, LeBoeuf, & Lutz, 2013). Our hypothesis proposes that restaurants with higher photo counts might experience greater online engagement, leading to higher check-in frequencies.

Studies like (Kwok and Yu 2013) highlight the increasing influence of visual cues, such as photos, on customer engagement and online reputation. Restaurants should consider maximizing their visual presence on platforms like Yelp.

Price is a crucial factor in consumer decision-making, often acting as a cue for perceived value and quality (Lichtenstein & Burton, 1989; Rao & Monroe, 1989). While higher prices may suggest higher quality and exclusivity for some, others prioritize affordability. Our hypothesis suggests that restaurants with a lower price range might see higher check-in frequencies, as they are more accessible and affordable for a wider range of customers.

The distance of restaurants to tourist attractions and landmarks can significantly influence their success. Research suggests that tourists often prioritize convenience when making dining choices (Tussyadiah, 2012), making restaurants near popular destinations appealing due to increased foot traffic. Furthermore, being located near a well-known landmark can enhance a restaurant's visibility and brand awareness, potentially drawing attention from tourists seeking unique experiences or photo opportunities. However, proximity to tourist areas can also bring challenges, including higher operating costs and increased competition.

Beyond broad location factors, recent research has delved deeper into the nuanced impact of proximity on restaurant success. The concept of proximity suggests that a restaurant's physical location relative to other businesses can have a significant impact on its visibility, customer traffic, success and in sequence - attendance. Several theoretical frameworks and empirical studies provide insights into the complicated nature of this impact.

Spatial Interaction Theory: Kivell and Shaw (1980) proposed the Spatial Interaction Theory, which posits that the influence of distance can be offset by a store's attractiveness. This principle suggests that proximity to other successful restaurants, whether they offer similar cuisines or distinct experiences, can influence a restaurant's own success. Successful restaurants located nearby may attract a larger customer pool, enhancing the visibility and accessibility of surrounding establishments, including the restaurant under consideration.

Agglomeration Effects: The principle of minimum differentiation, as proposed by Brown (1989) and Hotelling (1990), highlights the tendency of similar businesses to cluster together. This clustering phenomenon, known as agglomeration, can lead to both positive and negative consequences for individual businesses. On the one hand, clustering can increase customer traffic due to the concentration of similar offerings within a specific area. On the other hand, increased competition within a cluster can also lead to a decline in the profitability of individual restaurants.

Neighborhood Influence: (Hu et al. 2014) explored the influence of geographical neighbors on rating predictions using a Yelp dataset. Their findings support the notion that proximity to other businesses, particularly successful ones, can affect a restaurant's online ratings, which significantly impacts its perceived success. Customers often rely on online reviews to make informed dining decisions, and positive reviews from nearby restaurants can indirectly benefit a restaurant's online reputation.

Location-Based Recommendations: Recent studies, such as those by Wang et al. (2016) and Eravci et al. (2016), have developed solutions to recommend optimal locations for expanding restaurant businesses, highlighting the critical role of location in attracting customers and driving business success. These studies underscore the importance of considering proximity to customer clusters, existing competitors, and key amenities when evaluating potential restaurant locations.

Finally, we consider the influence of attributes on restaurant attendance. A restaurant's atmosphere plays a crucial role in shaping customer emotions and behavior (Ryu & Jang, 2007, 2008). Research highlights the power of atmospherics, emphasizing the need for a well-considered ambiance that enhances the dining experience. While the impact of atmosphere can be both positive and negative, restaurants must carefully consider design and layout to create an experience that aligns with their target market.This includes factors like visual appeal, sound and music, comfort and functionality, and specific amenities such as televisions, outdoor seating, Wi-Fi, and other features.

Convenience and accessibility are important factors in customer decision-making. Studies (Klassen et al., 2005; Woo et al., 2012) have shown that easy access, short walking distances, and parking availability can significantly influence revisit intentions. This underscores the importance of a restaurant's location and accessibility in relation to potential customers.

Yelp introduce

Yelp, the online review platform founded in 2004 in San Francisco has become a dominant force in the crowded landscape of crowdsourced review platforms. With a market capitalization of $2.5 billion (Companies Market Cap, 2024), Yelp's influence on consumer behavior, particularly within the restaurant industry, is undeniable. The platform's impact extends beyond simple star ratings, shaping consumer decisions, influencing restaurant reputations, and informing strategic marketing initiatives.

For consumers, yelp acts as a digital town square where real-life dining experiences and opinions converge. It empowers individuals to explore Philadelphia's dynamic restaurant scene, discover hidden culinary gems, and make informed dining decisions based on the collective wisdom of other Yelpers. Yelp's user-centric design encourages to write detailed reviews, allowing consumers to explore nuances of a restaurant's ambiance, cuisine, service quality, and overall value. This depth of information differentiates Yelp from other platforms, making its reputation as a trusted source for discerning diners.

Yelp's interface allows users to easily filter restaurants based on location, cuisine type, price range, and user ratings. The platform's integration of user-generated photos provides potential diners with visual insights into a restaurant's atmosphere and offerings.

For Philadelphia restaurant owners, Yelp represents both a challenge and an opportunity. Managing online reviews is crucial for maintaining a positive reputation, attracting new customers, and fostering brand loyalty. Recognizing this, Yelp provides businesses with tools to

manage their online presence and valuable insights from customer feedback. Restaurants can directly engage with customers by responding to reviews, addressing concerns, and demonstrating their commitment to customer satisfaction. Yelp also offers analytics dashboards to track review data, providing insights into customer sentiment, highlighting areas for improvement, and enabling data-driven business decisions.

This study leverages Yelp's rich dataset to meticulously analyze the factors influencing customer attendance at Philadelphia restaurants. We move beyond star ratings, utilizing check-in frequency as a direct measure of customer visits. Our analysis use internal restaurant attributes (price range, star rating, number of reviews, photo count) and link with factors like competition and demographic trends. By uncovering these intricate relationships, we aim to provide Philadelphia restaurateurs with actionable insights to optimize their offerings, attract customers, and thrive in a competitive market.

Overview over the Yelp data

This study utilizes a subset of Yelp's large dataset, focusing on restaurant information, user reviews, and check-in data. While the original dataset covers over eight million reviews, 200,000 businesses, and two million users across ten metropolitan areas, our analysis specifically focuses on restaurants in Philadelphia, Pennsylvania, between January 1st, 2018, and December 31st, 2019. This refined subset allows for a focused examination of restaurant attendance patterns within a specific geographic location and timeframe. To ensure data consistency and capture potential trends, the dataset was sub-settled to include only businesses categorized as restaurants and located within Philadelphia city limits. Data spanning two years (2018-2019) was included to capture potential trends and variations in restaurant attendance over time. Data was aggregated at the yearly level, meaning the check-in data for each restaurant represents the total number of check-ins within each calendar year (2018 and 2019). To understand the factors influencing restaurant check-in frequency, we employ a combination of existing variables from the Yelp dataset, supplemented with data from external sources, and construct new metrics. These include
Yearly Check-in Frequency (2018-2019): our dependent variable, representing the total number of check-ins per restaurant per year.
Internal Restaurant Attributes:
Price range: derived from the 'attributes.pricerange2' variable, this indicates the average cost of dining.
Availability of amenities: features like happy hour, delivery, takeout, etc., are extracted from the 'attributes' column.
Number of reviews: this variable represents the volume of user reviews for each restaurant in a given year.
Average star rating: this captures the overall customer sentiment towards each restaurant based on reviews.
Number of photos: reflects a restaurant's visual appeal and user engagement on Yelp.
Number of attributes: this reflects the diversity of features and services offered by a restaurant, potentially indicating a wider range of customer preferences it caters to.
External Competitive Factors:

Business proximity: number of competitor restaurants, calculated using geospatial data, helps understand the competitive landscape. This metric considers the category of a restaurant (as indicated by the column 'category') and calculates the overlap of categories within the 5km radius.
Nearby Landmarks: number of landmarks within a 1 km radius, calculated using geospatial data, helps understand the potential for increased foot traffic and tourist appeal associated with the restaurant's location. Landmarks often attract tourists and locals alike, potentially leading to higher restaurant attendance.
Demographic Factors: Age group concentrations (youth, adults, elders, children) within the restaurant's zip code: These variables, sourced externally and merged with the Yelp data, provide insights into the demographic makeup of the restaurant's surrounding area. Population data was acquired from the 2018 American Community Survey Single-Year Estimates, available through the U.S. Census Bureau. The population data was merged with the Yelp dataset via longitude and latitude matching the geometry of *ZCTA*5 zip codes with the position of a restaurant.

This combination of variables allows for a multi-faceted analysis of restaurant attendance in Philadelphia, considering both internal restaurant characteristics and external influences.

Data clean and Methodology

Data Cleaning and Standardization

	n	mean	sd	median	min	max	skew	kurtosis	se
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
A psych: 37 × 9									
check_in_count	5133	33.97	66.70	13.00	1.00	1522.00	8.39	132.02	0.93
log_check_in_count	5133	2.55	1.45	2.56	0.00	7.33	0.01	-0.66	0.02
average_stars	5133	3.64	0.93	3.83	1.00	5.00	-0.93	0.54	0.01
n_photo	5133	6.61	12.46	2.00	0.00	157.00	4.77	32.99	0.17
review_count	5133	25.01	40.72	13.00	1.00	744.00	6.18	63.21	0.57
scaled_capped_business_proximity	5133	0.34	0.98	0.00	-0.46	4.85	2.16	5.64	0.01
landmarks_nearby	5133	13.43	7.25	13.00	0.00	32.00	0.25	-0.32	0.10
scaled_capped_surrounding_children	5133	0.03	0.67	0.00	-1.08	2.28	0.98	0.83	0.01
scaled_capped_surrounding_youth	5133	0.21	0.92	0.00	-1.25	2.13	0.76	-0.64	0.01
scaled_capped_surrounding_adults	5133	0.00	0.76	0.00	-2.09	1.89	0.36	-0.41	0.01
scaled_capped_surrounding_elders	5133	0.06	0.65	0.00	-1.52	1.19	0.31	-0.97	0.01
attribute_count	5133	10.94	4.46	11.00	0.00	24.00	0.12	-0.39	0.06
restaurantspricerange2	5133	1.66	0.60	2.00	1.00	4.00	0.52	0.57	0.01
alcohol	5133	0.65	0.91	0.00	0.00	2.00	0.74	-1.38	0.01
wifi	5133	1.02	1.00	2.00	0.00	2.00	-0.04	-1.99	0.01
isromantic	5133	0.03	0.16	0.00	0.00	1.00	5.99	33.90	0.00
isintimate	5133	0.04	0.20	0.00	0.00	1.00	4.63	19.48	0.00
istouristy	5133	0.01	0.12	0.00	0.00	1.00	8.38	68.31	0.00
ishipster	5133	0.06	0.23	0.00	0.00	1.00	3.77	12.22	0.00
isdivey	5133	0.03	0.18	0.00	0.00	1.00	5.29	25.94	0.00
isclassy	5133	0.27	0.44	0.00	0.00	1.00	1.04	-0.91	0.01
istrendy	5133	0.10	0.30	0.00	0.00	1.00	2.61	4.82	0.00
isupscale	5133	0.01	0.11	0.00	0.00	1.00	9.24	83.48	0.00
iscasual	5133	0.55	0.50	1.00	0.00	1.00	-0.20	-1.96	0.01
parking_garage	5133	0.08	0.28	0.00	0.00	1.00	2.99	6.94	0.00
parking_street	5133	0.70	0.46	1.00	0.00	1.00	-0.86	-1.25	0.01
parking_validated	5133	0.03	0.16	0.00	0.00	1.00	5.78	31.42	0.00
parking_lot	5133	0.20	0.40	0.00	0.00	1.00	1.49	0.21	0.01
parking_valet	5133	0.02	0.14	0.00	0.00	1.00	6.64	42.10	0.00
bikeparking	5133	0.77	0.42	1.00	0.00	1.00	-1.32	-0.27	0.01
outdoorseating	5133	0.42	0.49	0.00	0.00	1.00	0.32	-1.90	0.01
restaurantstableservice	5133	0.51	0.50	1.00	0.00	1.00	-0.03	-2.00	0.01
byob	5133	0.28	0.45	0.00	0.00	1.00	0.99	-1.03	0.01
businessacceptscreditcards	5133	0.89	0.31	1.00	0.00	1.00	-2.58	4.64	0.00
hastv	5133	0.74	0.44	1.00	0.00	1.00	-1.09	-0.82	0.01
restaurantsreservations	5133	0.34	0.47	0.00	0.00	1.00	0.68	-1.54	0.01
restaurantsdelivery	5133	0.77	0.42	1.00	0.00	1.00	-1.26	-0.42	0.01

Table 1: Summary statistics of the input data

The final dataset contains information on a large number of restaurants, providing a wealth of insights into their characteristics and offerings. For example, 70% of restaurant have street parking, suggesting it is the most common option, when the second – “Parking lot” offered by only 20%. Interestingly, a surprisingly small percentage of restaurants (around 2%) offer valet parking, suggesting this is a premium service reserved for a select few. Similarly, a significant majority of restaurants (89%) accept credit cards, highlighting the prevalence of this payment method. However, less than half of the restaurants (44%) offer free Wi-Fi, with only 44 restaurants offering paid Wi-Fi. This suggests that free Wi-Fi is increasingly becoming a

standard expectation for diners, while paid Wi-Fi remains a niche offering. Additionally, around 70% of restaurants have a TV, what indicates informative trends.

The dataset underwent several cleaning steps to ensure data quality and consistency. Firstly, if an attribute wasn't explicitly mentioned in the Yelp review, it was marked as "NA" (missing) for that restaurant. This helps us understand which attributes are commonly reported and which are not. Secondly, observations with a review count of 0 were dropped from the analysis. This is crucial since a review count of 0 leads to a misleading 0 average star rating, which significantly distorts the overall analysis.

To standardize the data, all attributes were adjusted to a 0-1 state. This makes it easier to compare and analyze different attributes using statistical methods. The exception to this rule was the 'alcohol' attribute, which was manually categorized as "Non_alcohol", "Beer_and_wine", or "Full_bar", representing a range of alcohol offerings. Similarly, the "wifi" attribute was categorized as "None", "Paid", or "Free".

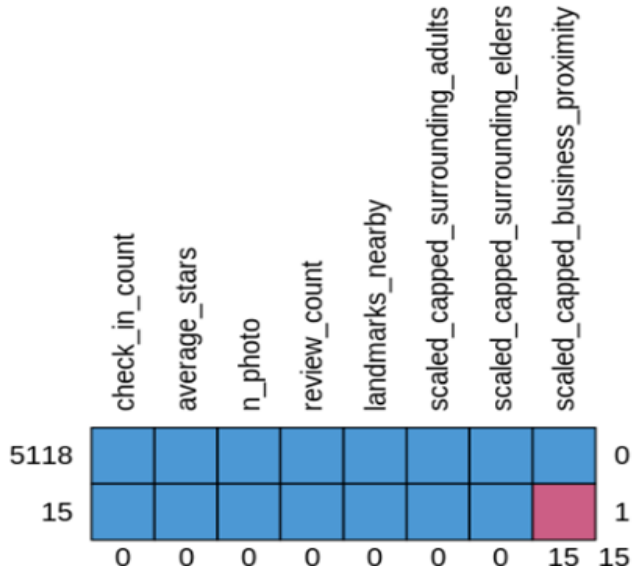

Figure 1: Pattern of missingness

Finally, 15 observations had missing values for the attribute 'scaled_capped_business_proximity'. These were imputed using the Mice method, which ensures the imputation process considers the correlations between different attributes. The rule of thumb was complied – to fill in no more than 5 percent of the values in a dataset. (in this case, 5).

This cleaned and standardized dataset forms the foundation for a more robust and meaningful analysis of the restaurant data, allowing for a deeper understanding of customer preferences, restaurant offerings, and the overall dynamics of the restaurant industry.

Exploratory Data Analysis

To examine multicollinearity in our data we firstly use Variance Inflation Factor (VIF) and tolerance - measures used to assess multicollinearity in regression models. VIF indicates how much the variance of a regression coefficient is inflated due to the linear relationship with other predictor variables, with higher values signifying greater multicollinearity. Tolerance, the reciprocal of VIF, represents the proportion of a variable's variance that is not explained by its correlation with other predictors, with lower values indicating a higher risk of multicollinearity. Our analysis reveal potential multicollinearity between the demographic variables 'scaled_capped_surrounding_adults' and 'scaled_capped_surrounding_elders'. With a tolerance and VIF - 0.13, 7.82 and 0.12, 8.14 respectively. While these values exceed some commonly used thresholds for multicollinearity (e.g., tolerance < 0.25 and VIF > 4), they do not reach the more conservative threshold of tolerance < 0.1 and VIF > 10. However, it is crucial to acknowledge that these are not absolute cutoffs for serious multicollinearity (O'Brien, 2006).

O'Brien (2006) cautions against relying solely on these rules of thumb, emphasizing that VIF values need to be considered in the context of other factors influencing the stability of regression coefficients, such as the overall model fit (R2), sample size, and the variance of the dependent variable.

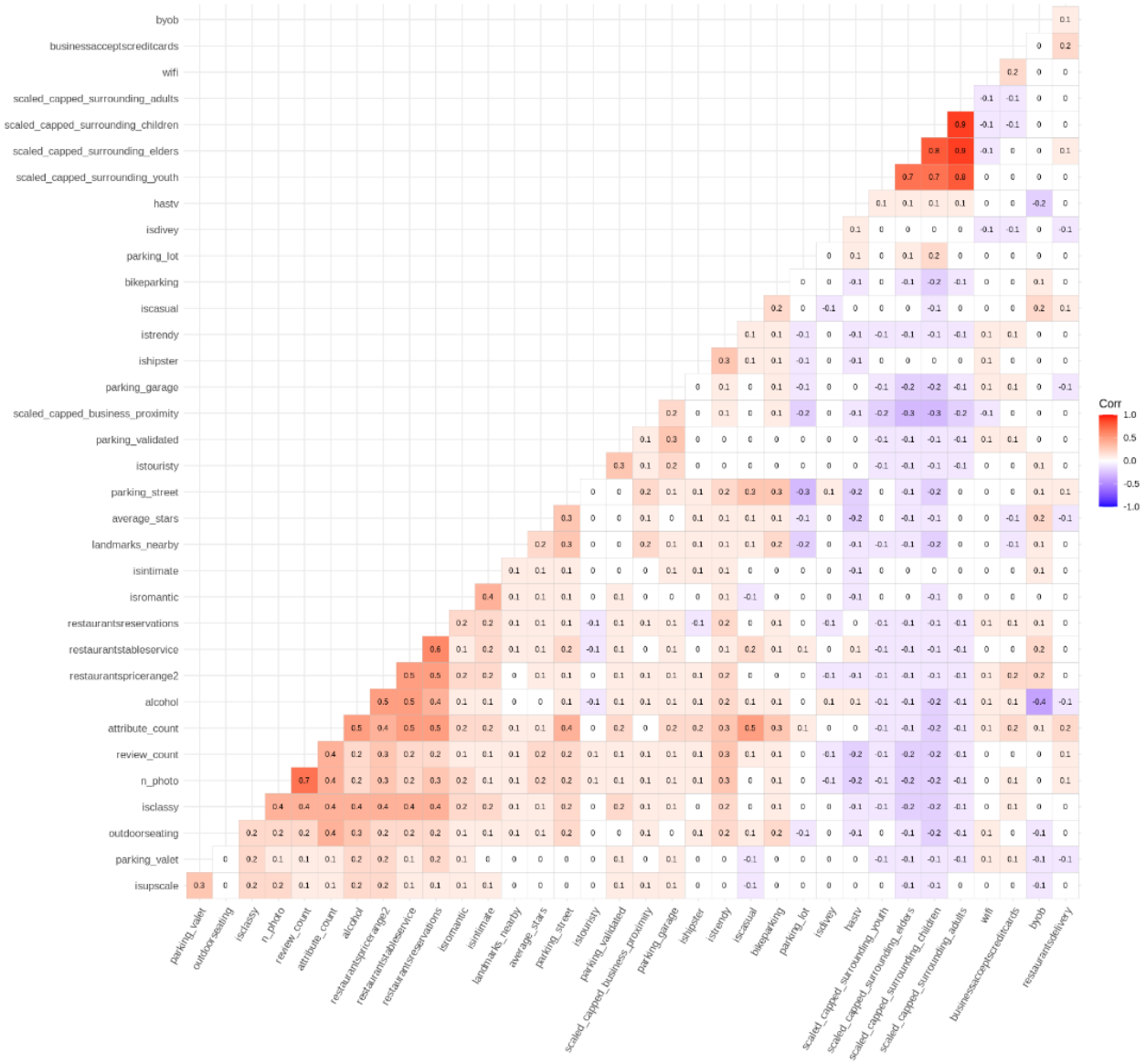

Figure 2: Correlation between variables

By analyzing the correlation plot we can make informed decisions about which variables to include in our model, striking a balance between minimizing multicollinearity and maximizing the predictive power and explanatory value of the model. Plot reveals interesting relationships within the restaurant data. Notably, there's a strong correlation between the demographic variables of Children, Youth, Adults, and Elders. This is unsurprising, given that these groups represent sequential stages of life. However, this high correlation can lead to issues in statistical modeling, potentially causing multicollinearity issues.

Further, analysis of the data reveals that Yelp users between the ages of 18-34 represent only 27% of all Yelp users, based on data from (Yelp, 2024. This implies that by retaining the Adults and Elders variables in our model, we represent over 73% of the Yelp user base. This is

significant because it emphasizes the importance of these demographic groups for understanding the restaurant landscape and customer behavior.

The correlation plot also highlights a strong relationship between review_count and photo_count (correlation coefficient of 0.7). This suggests that restaurants with a high number of reviews tend to have a higher number of photos. Several reasons could explain this correlation: Active users: Restaurants with more reviews are likely frequented by more active Yelp users who are more inclined to upload photos. Attractiveness: Restaurants with higher photo counts might be perceived as more visually appealing and, consequently, attract more reviewers. Despite the correlation, retaining both review_count and photo_count in the model is important for next reasons: While both variables capture user engagement, they offer unique insights. Review_count reflects the overall volume of user feedback, while photo_count provides a measure of visual appeal. Additionally, the combined effect of these variables helps us better understand the user experience and behavior on Yelp.

Data Preprocessing for Machine Learning

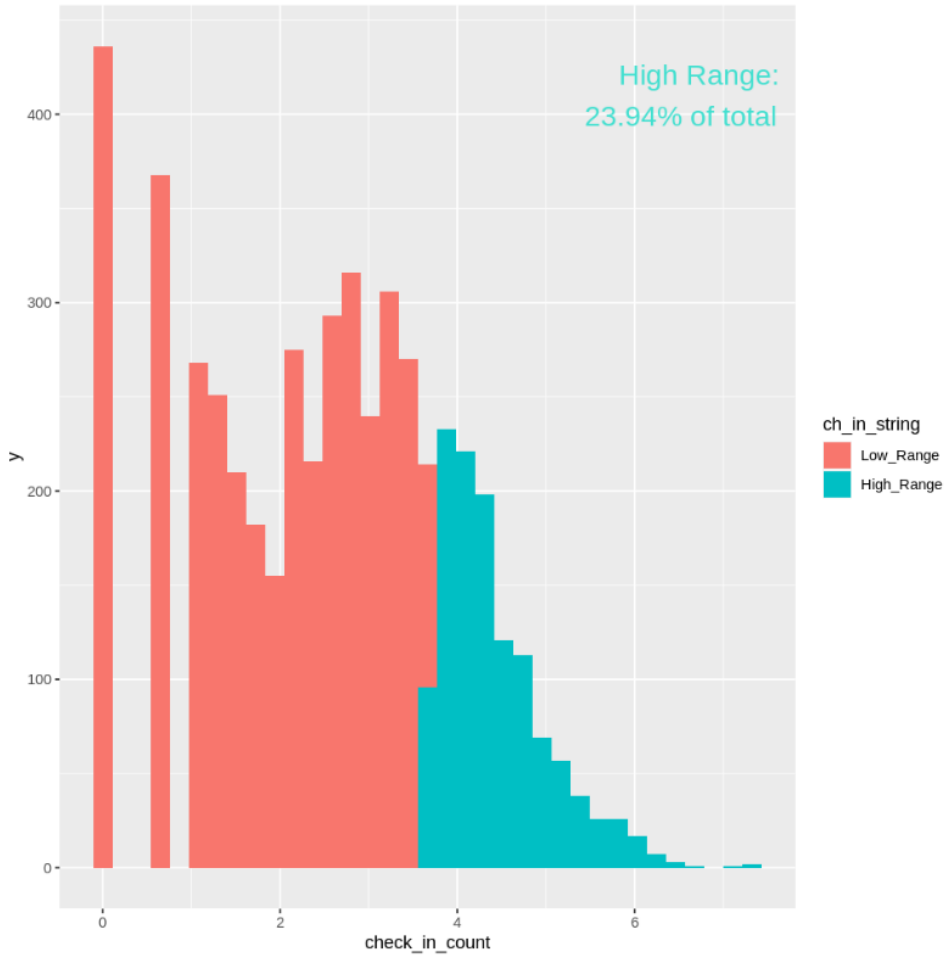

Figure 3: Histogram of binned, grouped data

Before diving into machine learning, we took several steps to prepare the data for optimal performance. One crucial step is binning, which involves grouping continuous data into discrete categories or bins. Implementing Equal-Width Binning, method where the data range is divided into bins of equal size, we handle the large range of check_in_count into two classes: "High Range" and "Low Range". This resulted in a "Low Range" bin containing 2922 restaurants, while the "High Range" bin contained 927 observations. The "High Range" representing minority bin contained the top 24% of values, representing a significant portion of the data. While 24:76 ratio is not specifically characterize high imbalance we decided to use an usual method such as Synthetic Minority Over Sampling Technique (SMOTE) to solve the problem (Hamid, Yusoff, & Mohamed, 2022).

SMOTE is a powerful technique used to balance datasets where one class (in this case, the "High Range" bin) is significantly smaller than the other. SMOTE generates synthetic data points that are similar to existing data points in the minority class, effectively increasing the size of the minority class without altering the distribution of the original data. After applying SMOTE, the "High Range" bin grew to 2781 restaurants, while the "Low Range" bin remained at 2922. We implemented SMOTE using a default value of K (nearest neighbors) of 5 and did not explore other SMOTE variations, such as Borderline-SMOTE or ADASYN.

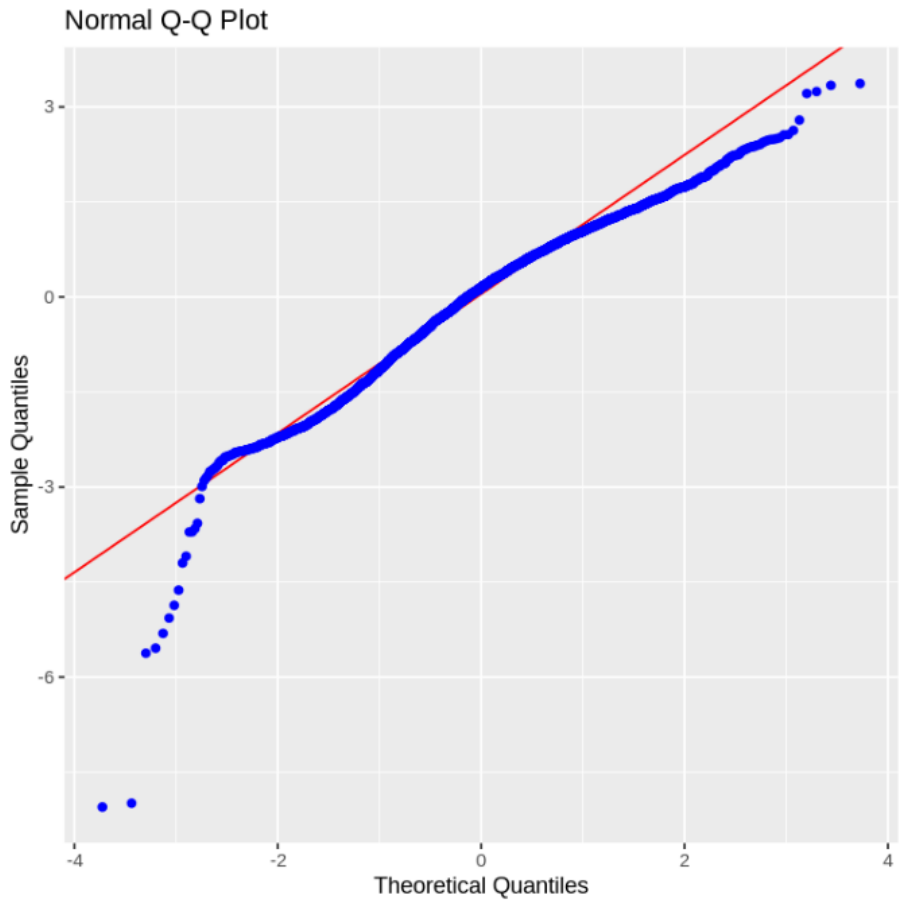

Figure 4: Quantile–quantile probability plot

The analysis of our data revealed significant deviation from the standard assumptions of normality and homoscedasticity, highlighting the importance of addressing these issues for a more reliable analysis. The Anderson-Darling test for normality and Breusch-Pagan test for heteroskedasticity strongly rejected the null hypotheses of normality and homoscedasticity, respectively. These findings confirmed that the residuals were not normally distributed and that the error term's variance varied across observations. Furthermore, normal plot clearly shows us negative skew. To ensure all attributes have similar scales, we applied Min-Max Scaling, a normalization technique that scales all attributes to the range of 0-1. This helps machine learning algorithms converge faster and perform more accurately by minimizing the impact of attributes with drastically different scales.

Finally, we split the data into two sets: Test and Evaluation. The Test set, representing 75% of the total data, was used to train our machine learning models. The Evaluation set, representing the remaining 25% of the data, was reserved for evaluating the performance of the trained models on unseen data, providing an unbiased estimate of model accuracy.

By carefully preparing the data through these steps, we ensured that our machine learning models receive the optimal information and perform more accurately on the chosen task

Models evaluation and Tuning

We ran different machine learning models to test their performance regarding our data. Each ML model has specific properties in terms of how it represents data and how it learns from that data. For the performance we look at four different measures that overall show how well the models solve the task. We will briefly overview five common ML models: Logit, Naive Bayes, k-Nearest Neighbors (k-NN), Support Vector Machines (SVMs), Decision Trees, and Neural Networks. Afterwards we will evaluate their overall performance regarding the task.

The *Logit* model is a statistical method for modeling binary outcome variables. It specifies the conditional mean of a discrete outcome variable as a logistic function of covariates (Agresti, 2002; Cramer, 2003). This model assumes that the probability of an observation being in a particular category is expressed as a nonlinear function of a set of predictors. A key advantage of the Logit model is its ability to handle non-linear relationships and predict probabilities. However, it can be difficult to interpret the model coefficients directly, especially when there are omitted variables (Breen et al., 2013). The *Naive Bayes* classifier is a probabilistic model based on Bayes' theorem (Domingos and Pazzani, 1996). It is used for classification tasks, where each instance is described by a conjunction of attribute values, and the goal is to predict the probability of that instance belonging to a specific class. Naive Bayes makes the simplifying assumption that attribute values are conditionally independent given the class label. This simplifies the learning process but if the independence assumption is violated it can lead to inaccurate predictions. However, it often works remarkably well in many text classification problems despite this assumption (Lang, 1995; Joachims, 1996).

The *k-Nearest Neighbors (k-NN)* algorithm is an instance-based learning method, which means that it stores the training examples and uses them directly to make predictions for new data points (Cover and Hart, 1967). k-NN classifies a new datapoint based on the class labels of the k nearest neighbors to the new point in the training dataset. It is a simple and easy-to-implement algorithm but can be computationally expensive for large datasets. Moreover, it is sensitive to the curse of dimensionality, meaning that its performance can degrade significantly when there are many irrelevant features (Moore and Lee, 1994). *Support Vector Machines (SVMs)* are a set of supervised learning methods used for both classification and regression (Cortes and Vapnik, 1995). For classification, SVMs aim to find a hyperplane that maximally separates different classes of data points in a high-dimensional feature space. The hyperplane is defined by a set of support vectors, which are the data points closest to the decision boundary. SVMs are known for their good generalization capabilities and have been successfully applied to many problems such as text classification and image recognition. However, SVMs can be computationally expensive to train, especially for large datasets. *Decision Trees* are a type of supervised learning model that uses a tree-like structure to represent decisions and their possible consequences (Breiman et al., 1984). Each node in the tree represents a test on a specific attribute of the data, and each branch represents a possible outcome of the test. The algorithm learns a decision tree by iteratively splitting the data based on the attribute that yields the most information gain. Decision trees are relatively easy to interpret, but they are sensitive to overfitting and require careful pruning to avoid this problem (Mingers, 1989). *Bagging (Bootstrap Aggregating)* is an ensemble learning technique that combines multiple decision trees to improve prediction accuracy and reduce variance (Breiman, 1996). It works by creating multiple training sets with replacement (bootstrapping) from the original dataset, then training a separate decision tree on each of these sets. The predictions for a new datapoint are then combined through averaging or majority voting. Bagging helps to reduce variance by averaging over the predictions of multiple models, making it more robust to noisy data or outliers. *Random Forests* are a powerful ensemble learning method that extends bagging by introducing randomness in feature selection as well (Breiman, 2001). Each decision tree in a random forest is trained on a random subset of the features. This helps to reduce correlation between trees and improves the model's generalization ability. Random Forests are often preferred over single decision trees due to their ability to handle high-dimensional datasets, reduce overfitting, and offer robust prediction accuracy. *Boosting* is an ensemble learning technique that sequentially combines weak learners to create a strong learner. It iteratively weights the training data and focuses on the instances that are misclassified by the previous models (Freund and Schapire, 1997). Boosting methods like AdaBoost and Gradient Boosting are known for achieving high accuracy, especially in the case of imbalanced datasets or complex decision boundaries. The idea is to generate a model that progressively improves performance by focusing on the most difficult instances (Breiman, 1998). *Neural Networks* are a class of ML models inspired by the structure of the human brain (McCulloch and Pitts, 1943). They consist of interconnected nodes, called neurons, organized in layers. Each neuron receives input from the previous layer and calculates a weighted sum of these inputs, followed by a non-linear activation function. Neural networks have a high representational power and have proven to be highly effective in tasks such as image recognition, natural language processing and speech recognition. However, training deep neural networks (networks with many layers) can be computationally expensive and prone to vanishing or

exploding gradients. Recent advances in deep learning, however, have addressed these issues and have led to significant performance gains (LeCun et al., 2015).

The performance of the presented machine learning algorithms is assessed using several metrics: the Gini coefficient, Top Decile Lift (TDL), accuracy rate, and runtime. The Gini coefficient is a statistical measure of inequality. In the context of machine learning, it's primarily used to assess the performance of classification models, particularly in the context of decision trees and rule-based models (Breiman et al., 1984; Quinlan, 1986). It measures the model's ability to differentiate between positive and negative instances, essentially indicating how well the model "separates" the classes. A Gini coefficient of 0 indicates a model that makes no distinction between classes, a coefficient of 1 suggests perfect separation, while 0.5 represents a model that performs no better than a random guess. Higher Gini coefficients generally indicate better model performance in separating classes. TDL is another metric used in classification models, particularly for marketing applications (Campbell et al., 2000; Tong and Koller, 2001). It measures the lift, or improvement, achieved by the model in targeting the top 10% of the population (in terms of predicted probability of a positive outcome). It is calculated by comparing the response rate of the top decile to the average response rate across the entire population. A TDL of 1 indicates no lift, meaning the model provides no improvement over random selection. A TDL greater than 1 indicates the model effectively targets the top decile, leading to a higher response rate in that group. Higher TDL values suggest better model performance for targeted marketing efforts. The accuracy rate, also known as overall accuracy, is a straightforward measure of how well a model performs in classification tasks (Duda and Hart, 1973). It's simply the proportion of correctly classified instances out of the total number of instances. An accuracy rate of 1 indicates a model that correctly classifies all instances, while a rate of 0 suggests the model makes no correct predictions. However, accuracy can be misleading in cases of imbalanced datasets where one class dominates. For example, a model predicting the occurrence of a rare event might have a high accuracy by simply classifying all instances as the non-rare event(He and Garcia, 2009). The next measure, also known as computational time or runtime, is crucial for practical applications. It quantifies the time required for the model to learn and make predictions (Cortes and Vapnik, 1995). The complexity of a model and the size of the dataset strongly influence runtime. Balancing performance with efficiency is often a key consideration in choosing an appropriate algorithm.

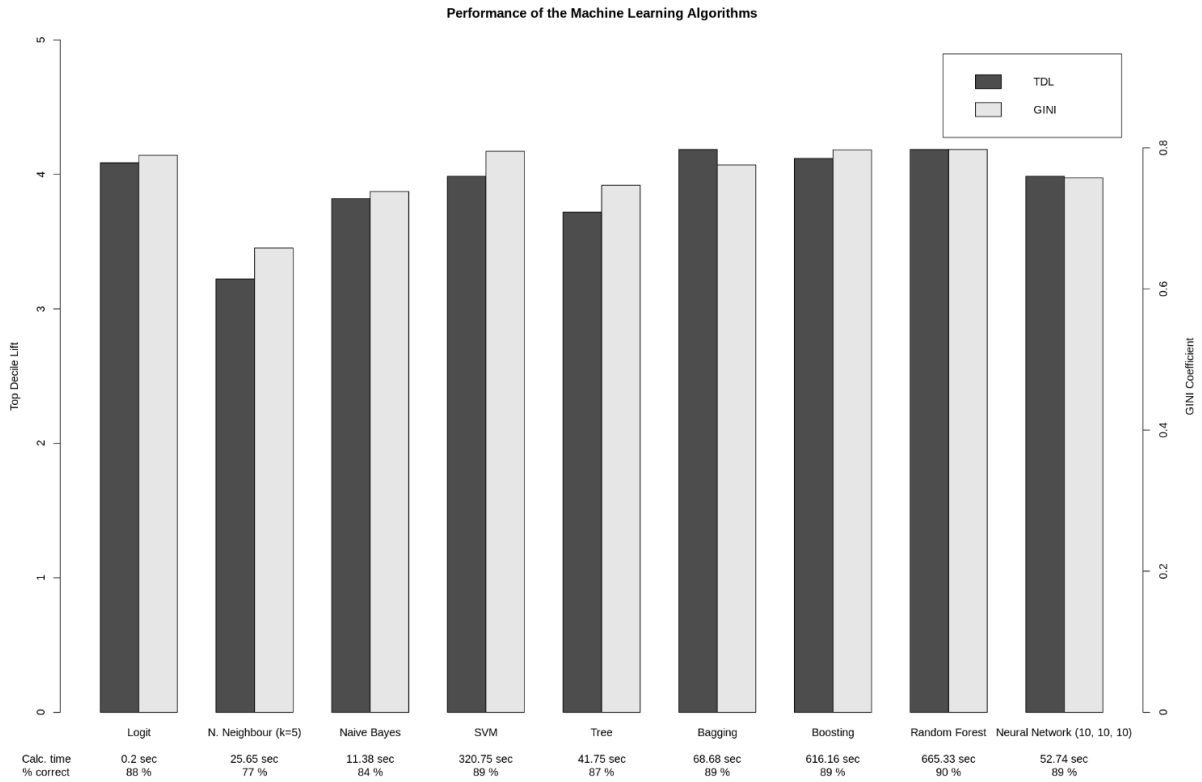

Figure 5: Performance of different machine learning algorithms

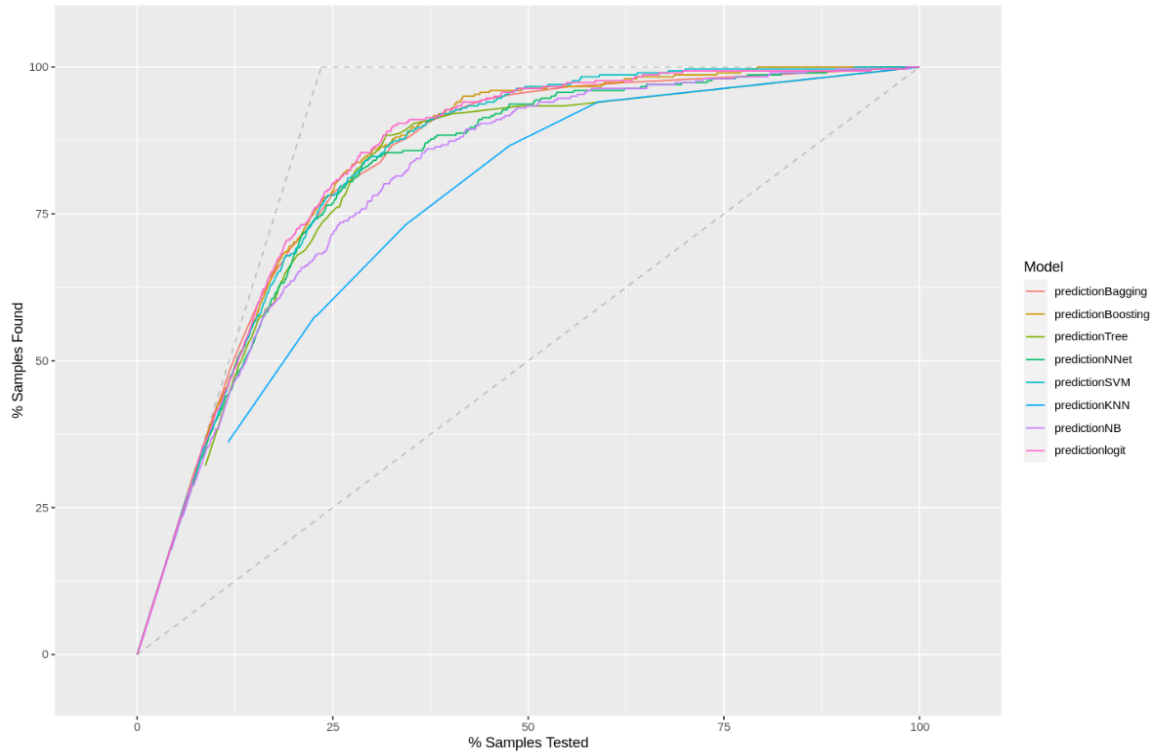

Figure 6: Lift curves for different machine learning algorithms.

We compared the performance of several machine learning models and found that K-Nearest Neighbors, Decision Trees and Naive Bayes models underperformed. While other models achieved good and consistent results, SVM, Boosting, and Random Forest were computationally expensive for our dataset. As a result, we focused on Neural Networks for further optimization.

We experimented with various Neural Network architectures, changing the number of layers and hidden units. The Neural Network with architecture (10, 10, 10) performed best, achieving a Target Dependent Lift (TDL) of 3.99, a Gini coefficient of 0.76, and a hit rate of approximately 89%. Based on these findings, we recommend using the Neural Network (10, 10, 10) for further calculations.

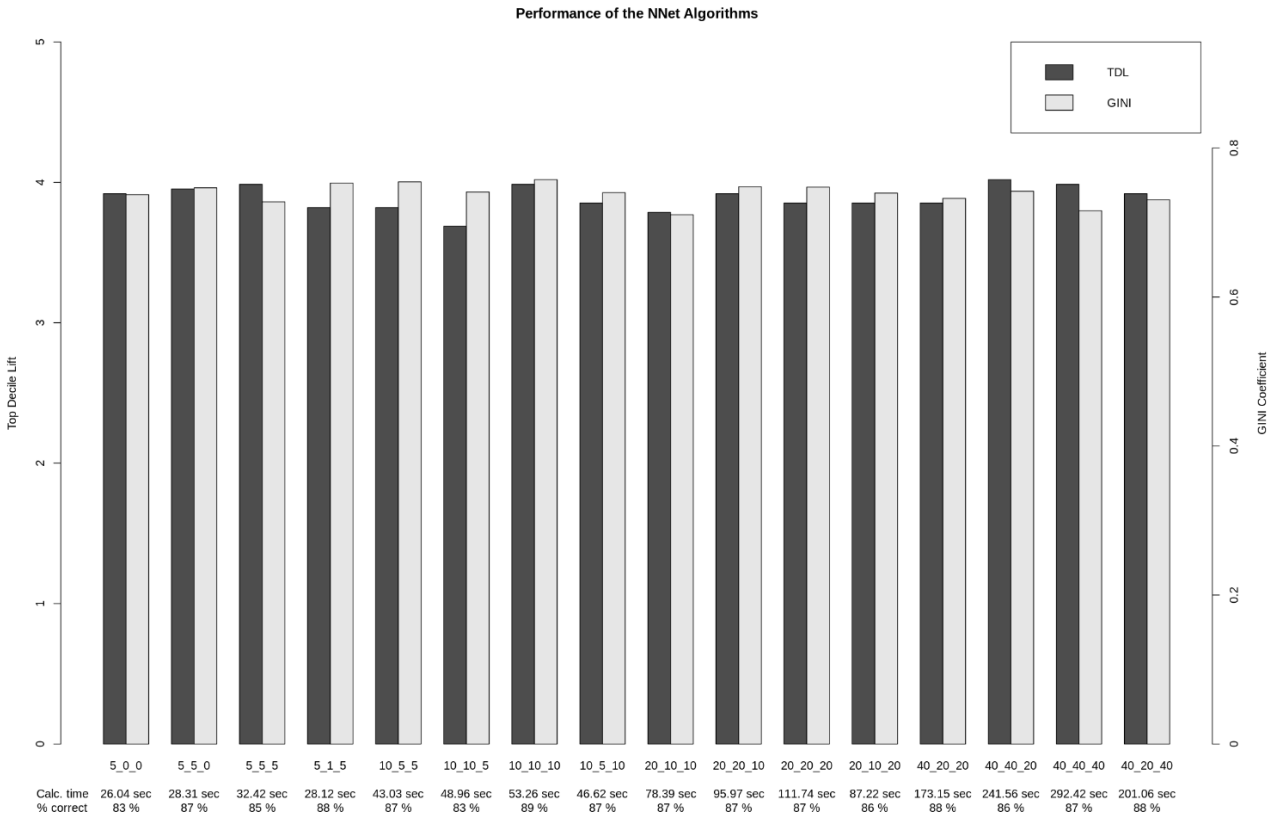

Figure 7: Performance of NNet Algorithms

Worth mentioning, we tested the performance of k-Nearest Neighbors (k-NN) models with different values for k: 7, 5, 3, and 1. Our results showed that the model with k = 5 performed

best overall. This finding highlights the importance of careful hyperparameter tuning, as the optimal k value can significantly influence the model's performance.

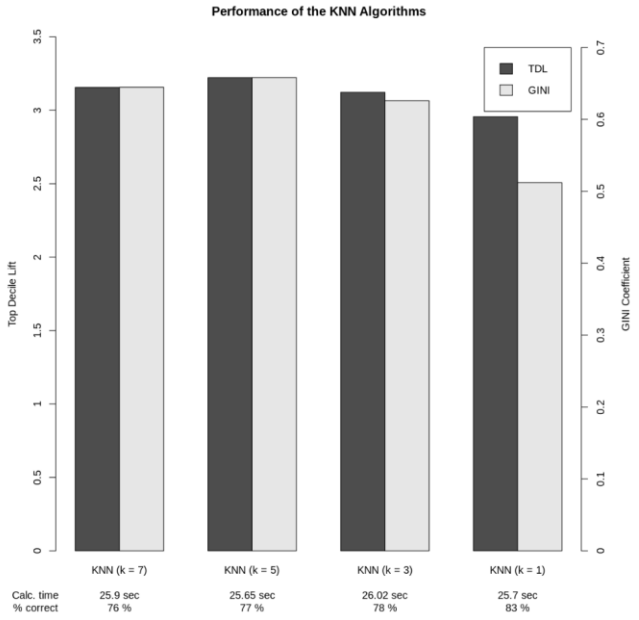

Figure 8: Performance of the KNN Algorithms

Results and business recommendations

Machine Learning Model Findings

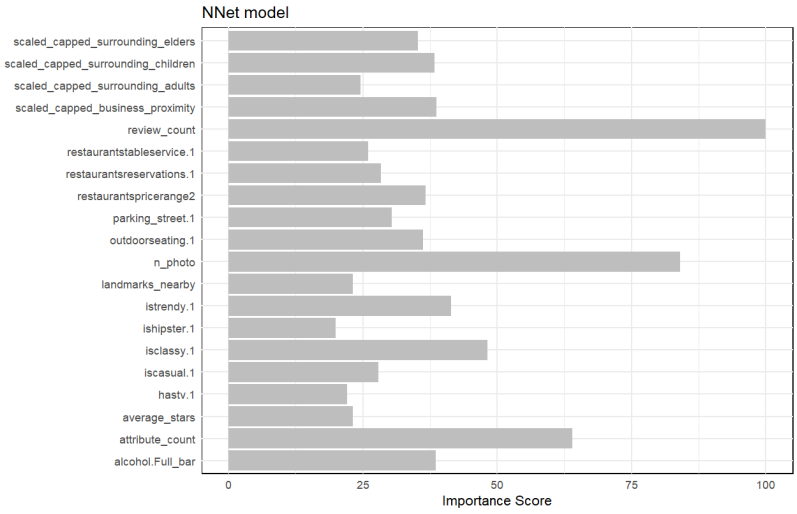

Figure 9: Importance score of 20 most important variables

Figure 9 highlights the 20 most influential variables for the chosen NN machine learning algorithm. Variables like review count, photo count and attribute showed the strongest predictive power.

While three models (NN, Bagging, LOGIT) had comparative performance, we decided to use results of the LOGIT regression model to provide specific business recommendations.

Business Recommendations Using Logistic Regression

Figure 7 presents the results of the Logit regression model used for deriving business recommendations. We categorized the results based on segments described in Chapter 4 (Overview over the yelp data).

Based on the analysis of the Yelp data and the logit regression model results, several key recommendations come out for restaurant owners and investors seeking to improve customer attendance in Philadelphia. These recommendations are aligned with our initial hypotheses regarding the influence of internal restaurant attributes, external competitive factors, and demographic factors. Most significant variables showed varied effects, differing in both direction (positive/negative) and strength (magnitude).

Our findings strongly suggest that casual dining is a driving force in Philadelphia. Restaurants with a casual ambiance are significantly more likely to attract customers and generate higher check-in frequencies. This confirms our hypothesis that restaurants with a casual ambiance will exhibit higher check-in frequencies, aligning with research that shows young adults often prioritize casual dining and social experiences (Quinn et al., 2019). Creating a welcoming, relaxed, and comfortable atmosphere is crucial for success. While romantic, intimate, and upscale ambiances may appeal to a niche market, they are not significant drivers of high check-in frequency in Philadelphia. The negative association between "isromantic.1" and "isintimate.1" and check-in frequency could be related to a perceived higher price point associated with these ambiances, potentially discouraging some diners. They might also attract a smaller, more specific clientele, making them less likely to achieve high check-in frequencies. Instead, consider focusing on creating a welcoming and inclusive atmosphere, potentially incorporating elements of casual, trendy, hipster or even "divey" vibes. This supports our hypothesis that a casual atmosphere will be a key driver of success. The model reveals a strong positive association between the number of photos on a restaurant's profile and higher check-in frequency. This underscores the importance of maximizing visual appeal, as our hypothesis stated, and aligns with research that highlights the influence of visual cues on customer engagement (Kwok & Yu, 2013). Credit card acceptance shows a strong level of significance in our model, indicating a clear benefit for businesses to consider offering this payment method. The negative magnitude of the restaurantsreservations.1 variable might indicate that customers might perceive a longer waiting time for a table at restaurants that require reservations, which could make them less attractive for quick meals or last-minute dining decisions. Other amenities were unsignificant.

Encouraging customers to leave with review photos that showcase the restaurant's ambiance, food, and overall experience might be essential to maximize check-in frequency. Serving alcohol, even if limited to beer and wine, is a significant factor driving higher check-in frequency, confirming our hypothesis. Restaurants that offer alcohol service are more likely to attract customers, particularly those seeking a more social dining experience. Review counts and star rating are strongly associated with higher check-in frequency, but they have less influence on check-in compared to other significant variables. This aligns with research that suggests positive reviews can contribute to increased success for lesser-known restaurants. However, strong brands may diminish the influence of positive reviews (Wang et al., 2021). Restaurants should actively manage their online presence, responding to both positive and negative reviews to increase review count and to build trust, foster customer engagement. This aligns with research emphasizing the importance of actively managing online reviews (Sparks & Browning, 2010; Xie et al., 2016). The model indicates that higher price ranges are associated with lower check-in frequencies, supporting our hypothesis regarding affordability. While offering high-quality ingredients and experiences is essential, striking a balance between quality and affordability can be a winning strategy. This aligns with research on price cues and perceived value (Lichtenstein & Burton, 1989; Rao & Monroe, 1989). The interpretation of the wifi.Paid variable may be deceptive. While it suggests that paid Wi-Fi is not a significant factor, this could be due to the relatively low prevalence of paid Wi-Fi offerings. It's more likely that restaurants in Philadelphia are moving towards free Wi-Fi as a standard amenity. The model suggests that offering parking lots can be a contributing factor to higher check-in frequency. Restaurants might consider offering parking lots, especially if located in areas where street parking is limited. However, providing valet parking does not appear to significantly influence check-in frequency, potentially due to its higher cost.

While the model did not find a strong association between business proximity and high check-in frequency, it is still important to understand the competitive landscape within a 5km radius. The concept of spatial interaction theory suggests that proximity to other successful restaurants can influence a restaurant's own success (Kivell and Shaw, 1980). Even if it's not a major factor in this model, analyzing the types of restaurants in the area can reveal opportunities for differentiation and market niches to target. While proximity to landmarks was not a significant predictor in this model, besides our model, it remains a potential factor for attracting tourists and locals alike, supporting our initial hypothesis. Restaurants located near landmarks often benefit from increased foot traffic and brand awareness. Consider strategically utilizing proximity to landmarks to market your restaurant and attract customers seeking unique experiences. This aligns with research that suggests tourists often prioritize convenience when making dining choices (Tussyadiah, 2012).

The model strongly suggests that restaurants located in areas with a higher concentration of adults are more likely to have higher check-in frequencies, supporting our hypothesis about adult demographics. This suggests that adults are a key target market for Philadelphia restaurants. Tailoring marketing efforts and offerings to appeal to adults can be a strategic approach to increasing attendance. This aligns with research that indicates adults hold a larger share of

working positions and have greater disposable income (Statista, n.d.). The model reveals a negative association between high concentrations of children and elders and high check-in frequency. While this doesn't mean avoiding these demographics entirely, it does highlight the need to carefully consider the needs and preferences of families and older adults.

While this study provides valuable insights into variables driving check-in behavior, some crucial factors require further investigation for definitive recommendations. Studying restaurant categories beyond the initial comparison could reveal more profitable options. Additionally, focusing solely on Yelp users might not capture the entire customer base. Examining other data sources could provide a more complete picture of customer behavior and preferences. Finally, while check-in behavior is an important metric, it's just one aspect of customer behavior. Further research could explore additional factors influencing customer decisions and overall business success.

```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -5.226527   0.462659 -11.297  < 2e-16 ***
average_stars                    0.216489   0.077763   2.784 0.005370 **
n_photo                          0.039985   0.008754   4.568 4.93e-06 ***
review_count                     0.119082   0.004417  26.960  < 2e-16 ***
scaled_capped_business_proximity -0.031480   0.056846  -0.554 0.579724
landmarks_nearby                 0.008547   0.008101   1.055 0.291378
scaled_capped_surrounding_children -0.941202   0.233457  -4.032 5.54e-05 ***
scaled_capped_surrounding_youth   0.150458   0.097160   1.549 0.121488
scaled_capped_surrounding_adults  1.553644   0.296557   5.239 1.62e-07 ***
scaled_capped_surrounding_elders -1.627013   0.297948  -5.461 4.74e-08 ***
attribute_count                  -0.004942   0.023711  -0.208 0.834893
restaurantspricerange2           -0.517816   0.129897  -3.986 6.71e-05 ***
alcohol.Beer_and_wine             1.017319   0.208582   4.877 1.08e-06 ***
alcohol.Full_bar                  0.768877   0.195168   3.940 8.16e-05 ***
wifi.Free                         0.058642   0.105675   0.555 0.578944
wifi.Paid                         0.526416   0.670089   0.786 0.432107
isromantic.1                     -2.210573   0.437642  -5.051 4.39e-07 ***
isintimate.1                     -0.664237   0.265213  -2.505 0.012261 *
istouristy.1                      1.340946   0.394155   3.402 0.000669 ***
ishipster.1                       0.783086   0.194085   4.035 5.47e-05 ***
isdivey.1                         0.834697   0.280277   2.978 0.002900 **
isclassy.1                        0.275518   0.133065   2.071 0.038401 *
istrendy.1                        0.752616   0.163379   4.607 4.09e-06 ***
isupscale.1                      -0.234871   0.510036  -0.460 0.645159
iscasual.1                        0.641655   0.138306   4.639 3.49e-06 ***
parking_garage.1                  0.089515   0.187911   0.476 0.633811
parking_street.1                  0.173755   0.163713   1.061 0.288537
parking_validated.1              -0.241132   0.315124  -0.765 0.444154
parking_lot.1                     0.381273   0.163078   2.338 0.019388 *
parking_valet.1                  -0.488337   0.392857  -1.243 0.213853
bikeparking.1                     0.209297   0.154119   1.358 0.174458
outdoorseating.1                 -0.049637   0.116261  -0.427 0.669417
restaurantstableservice.1        -0.063672   0.152886  -0.416 0.677069
byob.1                            0.073472   0.158969   0.462 0.643953
businessacceptscreditcards.1      0.829843   0.209443   3.962 7.43e-05 ***
hastv.1                          -0.070620   0.123640  -0.571 0.567881
restaurantsreservations.1        -0.529490   0.146763  -3.608 0.000309 ***
restaurantsdelivery.1            -0.289953   0.130756  -2.218 0.026589 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7902.6  on 5702  degrees of freedom
Residual deviance: 3016.7  on 5665  degrees of freedom
AIC: 3092.7

Number of Fisher Scoring iterations: 7
```

Table 2: Logit regression coefficients

Conclusion and Future Work

This work aimed to identify factors within a Yelp dataset that contribute to a restaurant's success. Customer check-in behavior was used as a proxy metric for success. At first, this study sheds light on the critical role of different restaurant factors and review elements played in shaping consumer behavior and ultimately, driving revenue. Secondly, a brief overview of Yelp and the type of data was given. Then, we described the data preparation process, which involved addressing inconsistencies within Yelp dataset and challenges like high or imbalanced check-in data. Then, machine learning algorithms were evaluated, with the Neural Network model demonstrating the best performance in predicting check-in behavior. Finally, logistic regression was used to generate business recommendations, focusing on the most impactful coefficients.

Our findings align with existing literature regarding the influence of certain variables on consumer behavior. However, the importance of these variables varies. For example, star rating and review count have lower impact compared to adults concentration variable and restaurant attributes. This suggests that even before opening, restaurants can take steps to increase their chances of success by optimizing these factors.

While this study explored some key attributes, others remain uninvestigated or require further exploration. Here are some potential areas for future research:
Analyze additional restaurant attributes and their influence on customer behavior.
Explore alternative dependent variables beyond check-in behavior, potentially including revenue or customer satisfaction metrics.
Include data on weather to gain a more comprehensive understanding of an influence of uncontrollable factors.
By delving deeper into these areas, we can develop a more holistic understanding of the factors driving restaurant success and provide even more valuable insights for business owners

Bibliography

Agresti, A. 2002. Categorical Data Analysis. New York: Wiley.

Akdeniz, B., Calantone, R. J., & Voorhees, C. M. (2013). Effectiveness of marketing cues on consumer perceptions of quality: The moderating roles of brand reputation and third-party information. *Psychology and Marketing*, *30*(1), 76–89.

Albayrak, A. (2014). Müşterilerin restoran seçimlerini etkileyen faktörler: İstanbul örneği [Factors that influence of customers' restaurant selection: The case of Istanbul]. Anatolia: Turizm Araştırmaları Dergisi, 15(2), 213–226.

Bart de Langhe, Philip M. Fernbach, Donald R. Lichtenstein, Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings, *Journal of Consumer Research*, Volume 42, Issue 6, April 2016, Pages 817–833, https://doi.org/10.1093/jcr/ucv047

Bilgihan, L., Barreda, K. R., Okumus, K., & Nusair, F. (2016). Consumer perception of knowledge-sharing in travel-related online social networks. Tourism Management, 52, 287-296.

Breen, R., A. Holm, and K. B. Karlson. 2013. "Total, direct, and indirect effects in logit and probit models". Sociological Methods & Research. 42: 164–191

Breiman, L. (1998). Arcing Classifiers. The Annals of Statistics, 26(3), 801–824. http://www.jstor.org/stable/120055

Breiman, L. 1996. "Bagging predictors". Machine learning. 24(2): 123–140.

Breiman, L. 2001. "Random forests". Machine learning. 45(1): 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and regression trees. Belmont, CA: Wadsworth International Group.

Brown, S. (1989). Retail Location Theory: The Legacy of Harold Hotelling. *Journal of Retailing*, *65*(Winter), 450–470.

Collier, P., Jones, P., & Spijkerman, D. (2018). Cities as Engines of Growth: Evidence from a New Global Sample of Cities. University of Oxford.

Companies Market Cap. (2024). Yelp (YELP) - Market capitalization. Retrieved from https://companiesmarketcap.com/yelp/marketcap/

Cortes, C. and V. Vapnik. 1995. "Support-vector networks". Machine learning. 20(3): 273–297

Cramer, J. S. 2003. Logit Models from Economics and Other Fields. Cambridge, UK: Cambridge Univ. Press.

Domingos, P. and M. Pazzani. 1996. “Beyond independence: Conditions for the optimality of the simple Bayesian classifier”. In: ICML. 105–112

Eravci, B., Bulut, N., Etemoglu, C., & Ferhatosmanoğlu, H. (2016). Location recommendations for new businesses using check-in data. In *Proceedings of the 16th IEEE International Conference on Data Mining Workshop (ICDMW)* (pp. 1110–1117). IEEE.

Filieri, R. (2016). What makes an online consumer review trustworthy? *Annals of Tourism Research*, *58*, 46–64.

Freund, Y. and R. E. Schapire. 1997. “A decision-theoretic generalization of on-line learning and an application to boosting”. Journal of Computer and System Sciences. 55(1): 119–139.

Guido, G., Ugolini, M. M., & Sestino, A. (2022). Active ageing of elderly consumers: insights and opportunities for future business strategies. In SN Business & Economics.

H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239

Hamid, M. H. A., Yusoff, M., & Mohamed, A. (2022). Survey on highly imbalanced multi-class data. International Journal of Advanced Computer Science and Applications, 13(6), 413-428.

Ho-Dac, N. N., Carson, S. J., & Moore, W. L. (2013). The effects of positive and negative online customer reviews: Do brand strength and category maturity matter? *Journal of Marketing*, *77*(6), 37–53.

Hotelling, H. (1990). Stability in Competition. In *The Collected Economics Articles of Harold Hotelling* (pp. 50–63). Springer.

Hu, L., Sun, A., & Liu, Y. (2014). Your neighbors affect your ratings: On geographical neighborhood influence to rating prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 345–354). ACM.

Joachims, T. 1996. “A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization”. Technical Report. Carnegie Mellon University.

Jones, P., Collier, P., & Spijkerman, D. (2020). Cities as Engines of Growth: Evidence from a New Global Sample of Cities. Journal of Applied Business and Economics, 22(2). https://doi.org/10.33423/jabe.v22i2.2808

Kivell, P., & Shaw, G. (1980). The Study of Retail Location. In *Retail Geography* (Edited by J. A. Dawson). Croom Helm.

Klassen, J. K., Trybus, E., & Kumar, A. (2005). Planning food services for a campus setting. *International Journal of Hospitality Management*, *24*(4), 579-609.

Kwok, L., & Yu, B. (2013). Spreading social media messages on Facebook: An analysis of restaurant business-to-consumer communications. *Cornell Hospitality Quarterly*, *54*(1), 84–94.

Lang, K. 1995. “Newsweeder: Learning to filter netnews”. In: Proceed-ings of the 12th International Conference on Machine Learning (ICML). 331–339.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. “Gradient-based learning applied to document recognition”. Proceedings of the IEEE. 86(11): 2278–2324

Lichtenstein, D. R., & Burton, S. (1989). The relationship between perceived and objective price-quality. Journal of Marketing Research, 26, 429–443.

Lumbwe, A. K., Nwobodo-Anyadigwu, E., & Mbohwa, C. (2018). The impart of Location decision on Small, Micro, and Medium Enterprises’performance in Johannesburg. Proceedings of the International Conference on Industrial and Operations management, Paris, France, July 26-27, pp. 1205-1216.

McCulloch, W. and W. Pitts. 1943. “A logical calculus of the ideas immanent in nervous activity”. Bulletin of Mathematical Biophysics. 7: 115–133

Mingers, J. 1989. “An empirical comparison of pruning methods for decision-tree induction”. Machine Learning. 4(2): 227–243.

Moore, A. W., & Lee, M. S. (1994). Efficient algorithms for minimizing cross validation error. In Proceedings of the 11th International Conference on Machine Learning (pp. 3745). San Mateo, CA: Morgan Kaufmann.

Nardini, G., LeBoeuf, R. A., & Lutz, R. J. (2013). When a picture is worth less than a thousand words. Association for Consumer Research 2013 North American Conference, Chicago (USA), (October 4, 2013.)

O'Brien, R. M. (2006). A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 40(5), 673–690.

Pieters, R., Wedel, M., & Batra, R. (2010). The stopping power of advertising: Measures and effects of visual complexity. Journal of Marketing, 74, 48–60.

Quinn, J.M., Wood, W., & Duke (2019). Habits Across the Lifespan 1 Running head: HABITS ACROSS THE LIFESPAN Habits Across the Lifespan.

Rao, A. R., & Monroe, K. B. (1989). The effect of price, brand name, and store name on buyers’ perceptions of product quality: An integrative review. Journal of Marketing Research, 26, 351–357.

Rybaczewska, M., & Sparks, L. (2021). Ageing consumers and e-commerce activities. International Journal of Retail & Distribution Management, 49(5), 472–491. https://doi.org/10.1108/IJRDM-04-2020-0125

Ryu, K., & Jang, S. (2007). The effect of environmental perceptions on behavioral intentions through emotions: the case of upscale restaurants. *Journal of Hospitality & Tourism Research, 31*(1), 56-72.

Ryu, K., & Jang, S. (2008). DINESCAPE: A scale for customers’ perception of dining environments. *Journal of Foodservice Business Research, 11*(1), 2-22.

Statista. (2023). U.S. mean disposable household income by age 2022. https://www.statista.com/statistics/825883/us-mean-disposable-household-income-by-generation/

T. Cover and P. Hart, "Nearest neighbor pattern classification," in IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, January 1967, doi: 10.1109/TIT.1967.1053964.

TouchBistro. (2017). How Diners Choose Restaurants. Restaurant Insights Report.

Tsao, W. C., Hsieh, M. T., Shih, L. W., & Lin, T. M. Y. Y. (2015). Compliance with eWOM: The influence of hotel reviews on booking intention from the perspective of customer conformity. *International Journal of Hospitality Management*, *46*, 99–111.

Tussyadiah, I. (2012). Tourist motivations and travel behavior: A review of recent research. Tourism Management, 33(1), 1–14.

Ukpabi, W. H., & Karjaluoto, H. (2016). Consumer acceptance of mobile internet in travel and tourism industry. International Journal of Information Management, 36(3), 469-479.

Wang, F., Chen, L., & Pan, W. (2016). Where to place your next restaurant?: Optimal restaurant placement via leveraging user-generated reviews. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 2371–2376). ACM.

Wang, M. and Zhou, X.: GEOGRAPHY MATTERS IN ONLINE HOTEL REVIEWS, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLI-B2, 573–576, https://doi.org/10.5194/isprs-archives-XLI-B2-573-2016, 2016

Wang, Y., Kim, J., & Kim, J. (2021). The financial impact of online customer reviews in the restaurant industry: A moderating effect of brand equity. *International Journal of Hospitality Management*, *95*, 102895.

Woo, C. S., Kim, W. G., & Lee, H. (2012). The influence of the quality of the physical environment, food, and service on restaurant image, customer perceived value, customer satisfaction, and behavioral intentions. *International Journal of Contemporary Hospitality*, *24*(2), 200-223.

Xie, K. L., Zhang, Z., Zhang, Z., Singh, A., & Lee, S. K. (2016). Effects of managerial response on customer eWOM and hotel performance: Evidence from TripAdvisor. *International Journal of Contemporary Hospitality Management*, *28*(9), 2013–2034.

Yang, S. B., Hlee, S., Lee, J., & Koo, C. (2017). An empirical examination of online restaurant reviews on Yelp.com: A dual coding theory perspective. *International Journal of Contemporary Hospitality Management*, *29*(2), 817–839.

Yelp. (2024, July). Fast facts. [Press release]. yelp-press.com