Goethe-Universität | 60629 Frankfurt am Main
Fachbereich Wirtschaftswissenschaften | Prüfungsamt
(Stand: 04.06.2018)

GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Ehrenwörtliche Erklärung/ Statutory Declaration

An das Ende von schriftlichen Arbeiten (bspw. Seminararbeiten, Bachelor- oder Masterarbeiten) ist eine ehrenwörtliche Erklärung zu setzen. Sie hat folgenden Wortlaut:

Ehrenwörtliche Erklärung

"Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst habe. Wörtlich übernommene Sätze oder Satzteile sind als Zitat belegt, andere Anlehnungen, hinsichtlich Aussage und Umfang, unter Quellenangabe kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen und ist nicht veröffentlicht. Sie wurde nicht, auch nicht auszugsweise, für eine andere Prüfungs- oder Studienleistung verwendet."

Ort, Datum: Frankfurt Am Main, 14/07/2023 Unterschrift: [illegible] Manuel

Bei schriftlichen Arbeiten, die zusätzlich in elektronischer Form einzureichen sind, ist jedoch folgender Wortlaut zu verwenden:

Ehrenwörtliche Erklärung

"Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst habe. Wörtlich übernommene Sätze oder Satzteile sind als Zitat belegt, andere Anlehnungen, hinsichtlich Aussage und Umfang, unter Quellenangabe kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen und ist nicht veröffentlicht. Sie wurde nicht, auch nicht auszugsweise, für eine andere Prüfungs- oder Studienleistung verwendet. Zudem versichere ich, dass die von mir abgegebenen schriftlichen (gebundenen) Versionen der vorliegenden Arbeit mit der abgegebenen elektronischen Version auf einem Datenträger inhaltlich übereinstimmen."

Ort, Datum: Frankfurt Am Main, 14/07/2023 Unterschrift: [illegible] Manuel

---

A statutory declaration is to be included at the end of every written work (e.g. Seminar paper, Bachelor's or Master's thesis).
The translation is as follows:

Statutory Declaration

"I herewith declare that I have composed the present thesis myself and without use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The thesis in the same or similar form has not been submitted to any examination body and has not been published. This thesis was not yet, even in part, used in another examination or as a course performance."

Place, Date: Frankfurt Am Main, 14/07/2023 Signature: [illegible] Manuel

In case of written work, which is also to submit on a data carrier, the translation is as follows:

Statutory Declaration

"I herewith declare that I have composed the present thesis myself and without use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The thesis in the same or similar form has not been submitted to any examination body and has not been published. This thesis was not yet, even in part, used in another examination or as a course performance. Furthermore I declare that the submitted written (bound) copies of the present thesis and the version submitted on a data carrier are consistent with each other in contents."

Place, Date: Frankfurt Am Main, 14/07/2023 Signature: [illegible] Manuel

AACSB
ACCREDITED

Unlocking Insights: Machine Learning Analysis of Yelp Data for Check-In Predictions

Seminar Paper / Term Paper

Submitted to Prof. Dr. Dehmamy

Faculty of Economics and Business, Goethe University Frankfurt, Frankfurt am Main

Manuel Scionti - scionti.manuel@hotmail.it - Erasmus Program

Abstract

This research paper explores and predicts customer check-in behavior using the Yelp Open Dataset. Leveraging machine learning models and fine-tuning a neural network, we achieve accurate check-in predictions. Valuable marketing insights are extracted, including peak check-in hours, influential factors, and location-specific trends. These insights enable businesses to optimize operations, enhance customer engagement, and attract more check-ins. By leveraging data analytics, businesses gain a competitive edge and effectively meet customer needs.

Contents

List of Figures

List of Tables

1 Introduction

This report describes a detailed investigation using the Yelp Open Dataset to understand and predict customer check-in behaviors. We used various machine learning techniques and a neural network to create an accurate model for predicting check-ins. These insights are valuable for businesses looking to attract more customers. First, we collected and prepared the Yelp Open Dataset, which contains information about businesses and customer feedback. We selected relevant features like business characteristics, review-related data, and user demographics to analyze check-in behaviors. By transforming the data into a binary classification problem, we could categorize businesses as having high or low check-in frequencies. We then tested different machine learning models, such as logistic regression, decision trees, random forests, support vector machines, naive Bayes, gradient boosting, and ensemble methods, to find the most effective one. We evaluated their performance using metrics like accuracy, precision, recall, and F1-score. Once we identified the best model, we fine-tuned a neural network using techniques like grid search, cross-validation, and adjusting hyperparameters. In addition to accurate predictions, our study uncovered useful marketing insights. Analyzing check-in patterns helped identify peak hours for customer visits, allowing businesses to optimize their operations and staff accordingly. We also discovered factors associated with high check-in frequencies, such as specific business attributes and positive reviews, which businesses can focus on to improve customer engagement. Furthermore, we examined check-in behavior across different locations, providing businesses with information to tailor their marketing strategies to regional preferences. The report is organized as follows: Section 2 reviews relevant literature on consumer behavior, user-generated media, and data science. Section 3 outlines the key performance indicators used in our study. Section 4 provides details about the algorithms, hyperparameter tuning, experimental setup, and evaluation criteria. Section 5 presents the study's results, and Section 6 discusses concluding remarks and limitations.

2 Literature Review

2.1 The Theory of Consumer Behaviour

Consumer behaviour in the restaurant choice context, as per the marketing literature, is moulded by a variety of intrinsic and extrinsic elements that shape perceptions and decision-making processes. Recognizing these motives can provide restaurants with a competitive advantage, as customer satisfaction is a core element of consumer decision-making and many restaurant value strategies. Research has repeatedly shown that aspects like food quality, service quality, cost, overall enjoyment, and the restaurant ambiance significantly influence customer satisfaction (Duarte Alonso et al.). Customer satisfaction, seen as the overall contentment level with a product or service experience, is closely tied to customer loyalty, which in turn leads to higher sales and profitability for the restaurant (Parsa et al., 2012; Saad Andaleeb Conway, 2006). Loyalty is expressed through behavioural, attitudinal, and composite aspects, all guiding the customer to favour one company over others, repurchase their products or services, and speak positively about the experience to others (Bowen Chen, 2001). In the context of specific restaurant attributes, the overall environment and design seem to have less of an impact on customer satisfaction compared to factors like staff responsiveness, price-value ratio, and food quality (Parsa, H. G., Gregory, A., Self, J. T., & Dutta, K. (2012)). However, the prioritization of these factors can vary based on the type and price range of the restaurant. External influences like weather conditions and the influence of other individuals also play a pivotal role in moulding consumer behaviour. For instance, rainy weather is associated with larger order sizes, while higher temperatures decrease them, with both sunny and rainy weather contributing to increased daily sales more than cloudy weather (Tian, X., Cao, S., Song, Y. (2021). Furthermore, negative restaurant experiences, such as extended waiting times, can lead to adverse consumer behaviours like reduced visits and a longer return time (De Vries et al., 2018). In general, improved customer satisfaction not only encourages return intention and positive word-of-mouth but also solidifies customer loyalty, enhancing the restaurant's reputation and leading to increased revenue (Kim et al., 2009). Therefore, restaurant managers should make efforts to cultivate customer loyalty in potential customers and secure a faithful customer base for lasting business success.

2.2 Digital Footprint and User-Generated Content

The significant impact of online platforms and user-generated content on business performance is increasingly recognized, especially in the hospitality sector. Both the quantity and quality of online mentions can significantly affect a restaurant's performance and survival chances (Li et al., 2022; Pantelidis, 2010). Anderson and Magruder (2012) support this, showing that even a marginal half-star improvement on online review databases can lead to a 19% increase in restaurant revenues. Interestingly, the evolution from traditional Word-of-Mouth (WOM) to electronic WOM (e-WOM) has altered consumer behaviour and business impact dynamics (Huete-Alcocer, N. (2017)). Nowadays, consumers often consider online reviews, particularly on platforms like Yelp, as a critical component of their decision-making process (Lim Van Der Heide, 2015; Pan et al., 2018). Luca (2016) and Zhang et al. (2010) suggest that substantial volumes of consumer reviews have a positive correlation with a restaurant's online popularity and revenue growth. For example, a one-star rise in a restaurant's rating on Yelp is associated with a 5-9% increase in sales (Luca, 2016). The influence of online reviews extends beyond just a numerical rating. User-generated content, including detailed reviews and photographs, enhances the impact of these reviews (Ceylan et al., 2021). Bakhshi et al. (2014) observed that Yelp reviews considered ”funny” and ”useful” were typically more critical and received lower ratings, while ”cool” reviews were more positive and had higher scores. This underscores the idea that the review tone, whether positive or negative, can have significant implications for a restaurant (Saad Andaleeb Conway, 2006). In addition, platforms like Yelp operate not only as review sites but also as digital communities. Parikh et al. (2015) highlighted that the adoption of Yelp is primarily driven by a sense of community involvement and trust. Yelp's community-centric model encourages the exchange of information and fosters a strong sense of trust, with 97% of users finding reviews on the platform trustworthy (Yelp Investors Report, 2023). This perception of reliability, along with factors like the number of online reviews and the user's friends count, often influences Yelp-users' behaviour (Lim and Van Der Heide, 2015). To summarize, the interplay between online presence, user-generated content, and the digital community fostered by platforms like Yelp greatly influences consumer behaviour and business performance in the restaurant industry. These digital platforms provide both a numerical and descriptive perspective to gauge customer satisfaction and make informed decisions. Hence, it's crucial for businesses to maintain a positive online reputation, considering the significant commercial impact these factors can have.

2.3 The Role of Data Science and Machine Learning in Analyzing User Behaviour

The intricacy of consumer behaviour calls for substantial data and robust tools to accurately model it. In this respect, data science methods are particularly valuable because of their ability to handle big data and make accurate predictions about user behaviour, such as personal preferences and the act of clicking the ”check-in” button on Yelp (Provost & Fawcett, 2013). Data science methods have been successfully applied in consumer behaviour theory, such as detecting significant correlations between customer satisfaction and loyalty (Szymanski and Henard, 2001). Further, recognizing that user-generated content (UGC) contributes significantly to the data used in consumer behaviour research emphasizes the need for effective sentiment analysis. Xu, Wu, and Wang (2015) have developed a successful model for such analysis, using a Binarized Naive Bayes model to predict user sentiment based on Yelp reviews. While check-in and review actions are distinct on Yelp, a correlation may exist, warranting further exploration. Machine learning techniques are pivotal in creating models that predict user behaviour. These techniques can be trained on various types of data, including user interactions, demographics, and previous behaviour. Models such as Random Forests or

Gradient Boosting Machines can handle multiple data sources and model complex, non-linear interactions (Breiman, L. (2001); Friedman, J. H. (2001)). Though our current research doesn't employ Social Network Analysis (SNA) due to a greater focus on physical restaurant attributes over user social attributes, it's worth acknowledging the potential usefulness of SNA in mapping relationships between Yelp users. Scott (2017) advocates the utility of SNA in understanding social behaviour and identifying influential users or closely connected communities that could affect 'check-in' behaviour. Time series analysis may be useful in examining the temporal aspect of user behaviour, identifying trends and seasonality in 'check-ins', which could be used to predict future trends (Box, G. E., & Jenkins, G. M. (1970)).

3 Dataset

The objective of this research is to pinpoint what factors contribute to customer satisfaction. Given the inherent difficulty in quantifying customer satisfaction, this study employs Yelp check-ins as a representative measure. Yelp's check-in feature allows customers to record or share their restaurant visits with other users on the platform. It's generally assumed that a customer's decision to check-in is tied to a positive dining experience, either because they wish to remember the restaurant for future visits or to share their experience with friends(Li et al., 2022; Pantelidis, 2010). Therefore, understanding the motivations behind these check-ins is of significant importance.

3.1 Yelp

Yelp has risen to prominence in recent years as an online review and rating platform. It offers a digital forum where customers can share experiences, assign ratings, and pen reviews about a wide array of businesses, including eateries. Users on Yelp can grade businesses from one to five stars and supply comprehensive written evaluations. The variety and richness of data available on Yelp yield precious insights into customer experiences and viewpoints (Lim Van Der Heide, 2015; Pan et al., 2018). It encapsulates details like the average rating of a business, individual ratings and reviews, comments concerning specific facets of a business (for example, service, atmosphere, food quality), and other contextual information (such as the date of visit or type of event). Additionally, Yelp features an option for users to upload photos, offering a visual dimension that can augment understanding of a business. The data sourced from Yelp presents several benefits for both researchers and businesses. Primarily, it supplies a vast amount of user-generated content, encapsulating real-time feedback and opinions. This data can prove especially useful in examining customer satisfaction, pinpointing patterns, and deriving insights into what influences customer preferences. Additionally, Yelp's popularity and extensive usage make the data indicative of a diverse customer demographic. Researchers can utilize Yelp's data for data analysis and application of machine learning methods to extract pertinent information. By employing sophisticated algorithms, it's possible to analyze large datasets, recognize trends, and reveal concealed patterns that can guide business strategies. Moreover, Yelp's provision of historical data enables longitudinal studies and the analysis of trends over time. Nevertheless, it's crucial to acknowledge that Yelp data could contain inherent biases. Users opting to leave reviews might not represent the whole customer base, and extreme ratings or subjective opinions could distort the overall business perception (Ceylan et al., 2021). Thus, researchers need to exercise caution in data interpretation and consider potential biases when formulating conclusions.

3.1.1 Yelp Open Dataset

The Yelp Open Dataset, publicly released by Yelp, serves as a treasure trove of data from one of the top online review platforms. Crafted to aid research and analysis, it affords access to a sizeable, diverse collection of business-related data and user reviews. Industries represented in the dataset are varied, spanning from eateries to retail outlets and service providers. A distinctive attribute of the Yelp Open Dataset is its wide-ranging coverage. It incorporates businesses from numerous cities and countries, thereby enabling researchers to delve into regional disparities, cultural impacts, and industry-specific dynamics. The Yelp Open Dataset is provided in a user-friendly format, typically as structured JSON files or SQL tables, thereby facilitating easy analysis using various tools and programming languages. Comprehensive documentation and guidance help users navigate the dataset and comprehend its structure. The Yelp Open Dataset can be utilized to undertake a broad spectrum of studies and analyses. Researchers can delve into subjects such as sentiment analysis, customer behavior modeling, recommendation systems, and business performance assessment, among others. By applying data analytics and machine learning techniques to this comprehensive dataset, researchers can extract valuable insights into consumer preferences, identify factors impacting business success, and formulate strategies to boost customer satisfaction. The specific variables included in the Yelp Open Dataset may differ across versions and releases. However, some common variables typically included in the dataset that have been also used for our analysis are:

- Business Details: This encompasses information about businesses listed on Yelp, such as their unique identifiers, names, addresses, contact details, operation hours, categories (like restaurants, retail, health-care), and geographic coordinates.
- Reviews: User-generated reviews are included in the dataset, featuring the review text, reviewer's username or identifier, review date, and the star rating assigned by the reviewer.
- Ratings: This variable symbolizes the cumulative rating given to a business based on all received reviews. It provides a comprehensive measure of customer satisfaction and serves as a vital metric for evaluating business performance.
- User Information: The dataset may incorporate information about review-submitting users, like user-names, profile details, and other user-specific attributes that help portray the reviewers.
- Review Metadata: Beyond the review text and ratings, the dataset may also include metadata related to each review, including the number of votes the review has received (useful, funny, cool), the number of views, and other engagement-related metrics.
- Business Attributes: These variables encapsulate specific characteristics or features tied to a business, such as Wi-Fi availability, credit card acceptance, outdoor seating, or parking facilities. These attributes offer insights into the amenities and services offered by the businesses.

The Yelp Open Dataset also contains extra variables and auxiliary data depending on the specific release. These can be user-uploaded photographs, check-in data, business hours on different days, pricing information, and more.

3.2 Physical attributes

The business attributes in the Yelp Open Dataset offer an in-depth understanding of each business's unique traits and features. These attributes give a glimpse into the services, facilities, and other elements that might impact customer experiences. Most of these variables were encoded as dummies or binary. Though the precise assortment of business attributes can change across different versions of the dataset, the following are some frequently seen attributes:

Tab. 1: Summary statistics and scales of the physical attributes variable subset

Statistic	N	Mean	St. Dev.	Min	Median	Max
ch_in	50,529	0.018	0.133	0	0	1
business_price	50,529	1.616	0.602	1	2	4
business_open	50,529	0.786	0.410	0	1	1
business_park	50,529	0.397	0.489	0	0	1
wifi_dummy	50,529	0.441	0.497	0	0	1
tv_dummy	50,529	0.561	0.496	0	1	1
bikeparking_dummy	50,529	0.639	0.480	0	1	1
goodforgroups_dummy	50,529	0.840	0.366	0	1	1
outdoorseating_dummy	50,529	0.348	0.476	0	0	1
creditcardpayment_dummy	50,529	0.987	0.112	0	1	1
noise_level	50,529	1.906	0.548	1	2	3
alcohol_dummy	50,529	0.612	0.487	0	1	1

1. Wi-Fi: This attribute indicates whether a business offers Wi-Fi connectivity to its customers. It provides insights into the availability of internet access, which can be important for customers who require connectivity during their visits.

2. Accepts Credit Cards: This attribute indicates whether a business accepts credit card payments. It helps customers determine the payment methods available and can be a crucial factor for those who prefer using cards instead of cash.

3. Outdoor Seating: This attribute indicates whether a business provides outdoor seating options. It is particularly relevant for restaurants and cafes, as it informs customers about the availability of outdoor spaces to dine or relax.

4. Takes Reservations: This attribute denotes whether a business accepts reservations. It can be useful for customers who prefer to plan their visits in advance or avoid potential wait times.

5. Waiter Service: This attribute indicates whether a business offers waiter service. It distinguishes establishments that provide table service from those where customers typically order and collect their food or items themselves.

6. Delivery: This attribute indicates whether a business offers delivery services. It is especially pertinent for restaurants and other food-related businesses, as it signifies the convenience of having items delivered to a customer's specified location.

7. Good for Kids: This attribute suggests whether a business is suitable for children. It is particularly relevant for family-oriented establishments, such as family-friendly restaurants or entertainment venues.

8. Good for Groups: This attribute suggests whether a business is suitable for larger groups of people. It provides insights into the availability of space, seating arrangements, and amenities that can accommodate gatherings or events.

9. Parking: This attribute indicates whether a business provides parking facilities. It informs customers about the availability and convenience of parking options near the establishment.

3.3 Users generated content

The user-generated content variables in the Yelp Open Dataset provide valuable insights into the reviews and feedback submitted by users on the platform. These variables offer a closer look at the opinions, sentiments, and experiences shared by customers. While the exact variables may vary across different releases of the dataset, here are some commonly encountered user-generated content variables:

Tab. 2: Summary statistics and scales of the user characteristics variable subset

Statistic	N	Mean	St. Dev.	Min	Median	Max
ch_in	50,529	0.018	0.133	0	0	1
cum_n_tips	50,529	7.293	9.059	1	4	67
cum_max_friends	50,529	283.612	386.464	1	134	3,382
cum_max_u_elite	50,529	2.699	3.091	0	1	9
cum_max_us_fans	50,529	14.635	31.363	0	5	305
cum_max_us_tip	50,529	65.362	150.592	1	9	1,216
avg_tips_stars	50,529	3.426	0.870	1.000	3.500	5.000
cum_n_review	50,529	41.130	44.215	3	24	251
avg_sentiment_score_review	50,529	0.210	0.128	−0.243	0.245	0.551
sum_elite_status	50,529	8.168	9.937	0	4	58
max_friends_count	50,529	573.844	789.607	0	328	6,873
male_to_female_tips_ratio	50,529	1.338	0.887	0.200	1.000	6.000
sum_fans	50,529	228.614	477.618	0	91	5,136
avg_review_stars	50,529	3.436	0.863	1.000	3.609	5.000
male_to_female_review_ratio	50,529	1.322	0.810	0.250	1.062	5.000
n_photo	50,529	1.944	3.633	0	0	24

1. Review Text: This variable contains the written content of the reviews submitted by users. It provides detailed descriptions of their experiences, including feedback on various aspects such as food quality, service, ambiance, pricing, and more. Analyzing the review text allows for sentiment analysis, topic modeling, and understanding specific mentions or sentiments related to a business.

2. Star Rating: The star rating variable represents the rating given by users to a business. It is typically a numerical value ranging from one to five stars, reflecting the user's overall satisfaction level with their experience. The star rating provides a quick summary of customer sentiment towards a business.

3. Review Date: This variable indicates the date when a review was posted by a user. It enables temporal analysis, allowing researchers to observe trends, changes, or seasonality in customer feedback over time.

4. User ID: This variable represents a unique identifier assigned to each user. It allows for tracking and analyzing the activities and contributions of individual users, including their reviewing behavior and engagement patterns on the platform.

5. Votes: The votes variable captures the number of votes that a review has received from other users. These votes are typically categorized as ”useful,” ”funny,” or ”cool,” allowing users to express their agreement, appreciation, or amusement with a particular review. Analyzing vote patterns can reveal popular or impactful reviews within the dataset.

6. Review Metadata: Additional metadata associated with each review may be included in the dataset. This can include information such as the number of views a review has received, the number of comments or replies it has generated, or other engagement-related metrics. These metadata provide insights into the visibility and interactions surrounding a review.

Researchers can leverage these user-generated content variables to gain a deeper understanding of customer experiences, sentiments, and preferences. By analyzing the review text, sentiment analysis techniques can be applied to identify positive or negative sentiments, extract topics or themes, or identify specific keywords or phrases that are commonly mentioned in reviews. The star rating and other metadata provide summary measures of satisfaction and engagement. Analyzing user activities and voting patterns can help identify influential or popular reviews within the dataset. These user-generated content variables, along with other variables in the Yelp Open Dataset, enable researchers to explore customer feedback, evaluate business performance, build recommendation systems, and gain valuable insights into consumer behavior and sentiments towards businesses listed on Yelp.

3.4 External Data

Tab. 3: Summary statistics and scales of the external variable subset

Statistic	N	Mean	St. Dev.	Min	Median	Max
ch_in	50,529	0.018	0.133	0	0	1
business_lat	50,529	41.120	0.021	41.050	41.117	41.202
business_long	50,529	−81.565	0.059	−81.655	−81.569	−81.465
PRCP	50,529	37.210	72.292	0.000	7.286	574.167
SNOW	50,529	4.494	18.731	0.000	0.000	193.000
SNWD	50,529	18.107	49.924	0.000	0.000	285.750
TMAX	50,529	172.344	113.947	−128.000	200.000	338.667
TMIN	50,529	71.591	103.597	−229.000	81.667	241.500
TOBS	50,529	110.223	105.009	−167.000	128.000	275.000
TOBS_1	50,529	110.614	104.207	−161.000	128.000	275.000
TOBS_2	50,529	111.113	103.415	−161.000	128.000	275.000
TOBS_3	50,529	111.397	102.839	−161.000	128.000	275.000
TOBS_4	50,529	111.687	102.410	−161.000	128.000	275.000
weekend	50,529	0.425	0.494	0	0	1

Beyond the factors directly related to the consumer (termed as social attributes) and those pertaining to the restaurant itself (referred to as physical characteristics), external elements can also significantly affect consumer patterns. For instance, varying climatic conditions have been shown to exert a considerable influence on business performance and sales (Palka, 2017; Tian et al., 2021). According to Tian et al. (2021), overcast weather has a lesser impact on revenue compared to either sunny or rainy conditions. Based on these findings, we felt it essential to incorporate weather-related external data into our model, which we sourced from the National Climatic Data Center (NOAA). As displayed in Table 3, we included indicators for snow (SNOW SNWD), rainfall (PRCP), and temperature (TMAX, TMIN, TOBS). Moreover, the restaurant's geographical location and the specific day of the week may sway a consumer's choice to patronize a particular restaurant. For instance, leisure activities such as dining out are typically more prevalent on weekends. Thus, we integrated a 'weekend' attribute into our model to assess the influence of weekends on check-in behavior. Additionally, we incorporated the restaurant's longitude and latitude data, obtained from the Yelp dataset. This data not only played a critical role in the model but also enabled us to acquire weather information from NOAA, as it allowed identification of the nearest weather station based on the restaurant's coordinates.

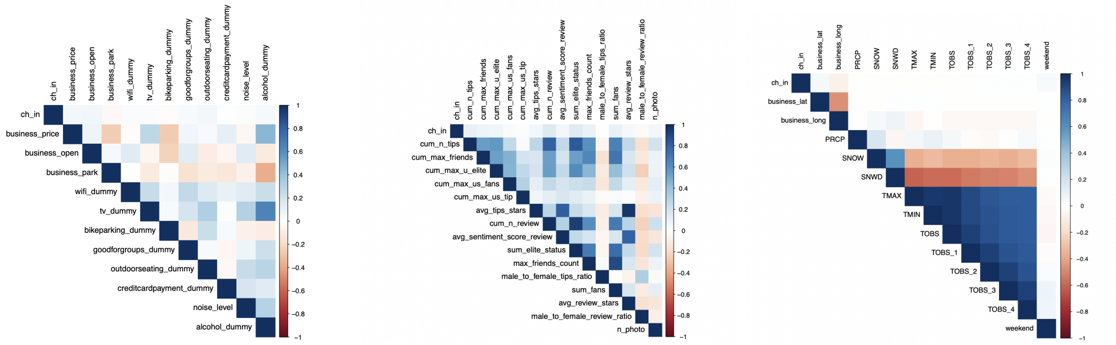
4 Methodology

This section will show the main phases and methodologies of the analysis.

4.1 Handling missing data

Handling missing data is a crucial step in every machine learning analysis since some algorithms can't cope with them. Simply ignoring the missing values would lead to algorithm failure. There exist different techniques to handle missing data and fill them with meaningful values. The chosen methodology was the *predictive mean matching technique* using the R package 'mice'. Predictive mean matching is employed as it enhances the robustness of inferences drawn from the missing data, as demonstrated by Morris et al. (2014). This method offers greater precision compared to alternative approaches such as replacing missing values with the median or mean. Furthermore, further analysis should be conducted to detect potential outliers. Boxplots and summary statistics were employed to visually inspect each variable individually and identify any outliers. However, no outliers are observed in the respective variables, thus no specific outlier treatment is applied.

4.2 Variable selection



(a) Physical Correlation

(b) Social Correlation

(c) External Correlation

Fig. 1: Dataset Correlation Plots

Since the final dataset presents a high number of variables, was key to select a subset of the most relevant predictors. Variable selection is more a heuristic process rather than an exact science. In order to choose the best variables for our analysis, we decide to use two different methodologies. The first one was to fit a logistic regression on our data and inspect the coefficients. More specifically, by looking at each p-value is possible to detect all the predictors that have the strongest influence on our target variable 'check-in'. The second method was to use a Stepwise Variable Selection method based on the Akaike Information Criterion (AIC), which gave us a subset of variables. In order to choose the best and final subset of variables to work with, we decided to create a combination of variables given by both methods, plus some variables that we thought might be important from a business point of view. The final subset of variable is shown in Table 4

Tab. 4: Final subset of variables used for the analysis

Statistic	N	Mean	St. Dev.	Min	Median	Max
ch_in	50,529	0.018	0.133	0	0	1
business_open	50,529	0.786	0.410	0	1	1
business_park	50,529	0.397	0.489	0	0	1
wifi_dummy	50,529	0.441	0.497	0	0	1
goodforgroups_dummy	50,529	0.840	0.366	0	1	1
outdoorseating_dummy	50,529	0.348	0.476	0	0	1
creditcardpayment_dummy	50,529	0.987	0.112	0	1	1
n_photo	50,529	1.944	3.633	0	0	24
cum_n_tips	50,529	7.293	9.059	1	4	67
avg_tips_stars	50,529	3.426	0.870	1.000	3.500	5.000
cum_n_review	50,529	41.130	44.215	3	24	251
male_to_female_tips_ratio	50,529	1.338	0.887	0.200	1.000	6.000
business_lat	50,529	41.120	0.021	41.050	41.117	41.202
business_long	50,529	−81.565	0.059	−81.655	−81.569	−81.465
TOBS_3	50,529	111.397	102.839	−161.000	128.000	275.000
TOBS_4	50,529	111.687	102.410	−161.000	128.000	275.000
weekend	50,529	0.425	0.494	0	0	1

4.3 Oversampling - SMOTE

During the analysis it was found that the target variable 'check-in' suffered from a high class imbalance with a proportion of 90%-10%. Specifically, the 'no check-in' level was the majority class and the 'yes check-in' level was the minority class. To solve this problem, we decided to use oversampling techniques. In this case, the method chosen was SMOTE. By applying SMOTE to our dataset we manage to rebalance the classes to almost 50%-50%. There is much evidence in the literature that oversampling techniques can improve the performance of machine learning(Chawla et al. 2002).

4.4 Model Selection

Tab. 5: Model Descriptions

Model	**Description**
Artificial Neural Networks (ANN)	Neural networks, based on the way humans think, emulate the information processing, storage, and learning of the human brain (Abraham, 2005).
Support Vector Machines (SVM)	Supervised learning models that analyze data and recognize patterns, primarily used for classification and regression (Badea, 2014).
Decision Trees (DTs)	Tree-shaped structures that represent sets of decisions, suitable for generating rules in classification or regression tasks (Jun Lee & Siau, 2001).
Random Forests (RF)	Combine multiple weak learners (small DTs) into a single classification decision (Breiman, 2001).
Gradient Boosting (GBM)	Forward learning ensemble method that sequentially fits DTs to improve predictive power (Friedman, 2002).
Naïve Bayes (NB)	Probabilistic classifier based on Bayes' theorem, assuming independence of features (Rish, 2001).
Logistic Regression (LR)	Regression method for binary classification, employing a sigmoid function (Hosmer et al., 2013).

Several methods were utilized in predicting check-in behavior. The focus is on widely recognized and popular models commonly employed for binary classification tasks, including churn prediction, as discussed by Badea (2014) and Shaaban et al. (2012). Specifically, the methods employed are Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DTs), Random Forests (RF), Gradient Boosting (GBM), Naïve Bayes (NB), and Logistic Regression (LR). These methods exhibit variations in terms of performance and execution speed, both of which are critical considerations in determining the optimal model. Evaluating their individual strengths and weaknesses enables the selection of the most suitable model for the task at hand (Table 5).

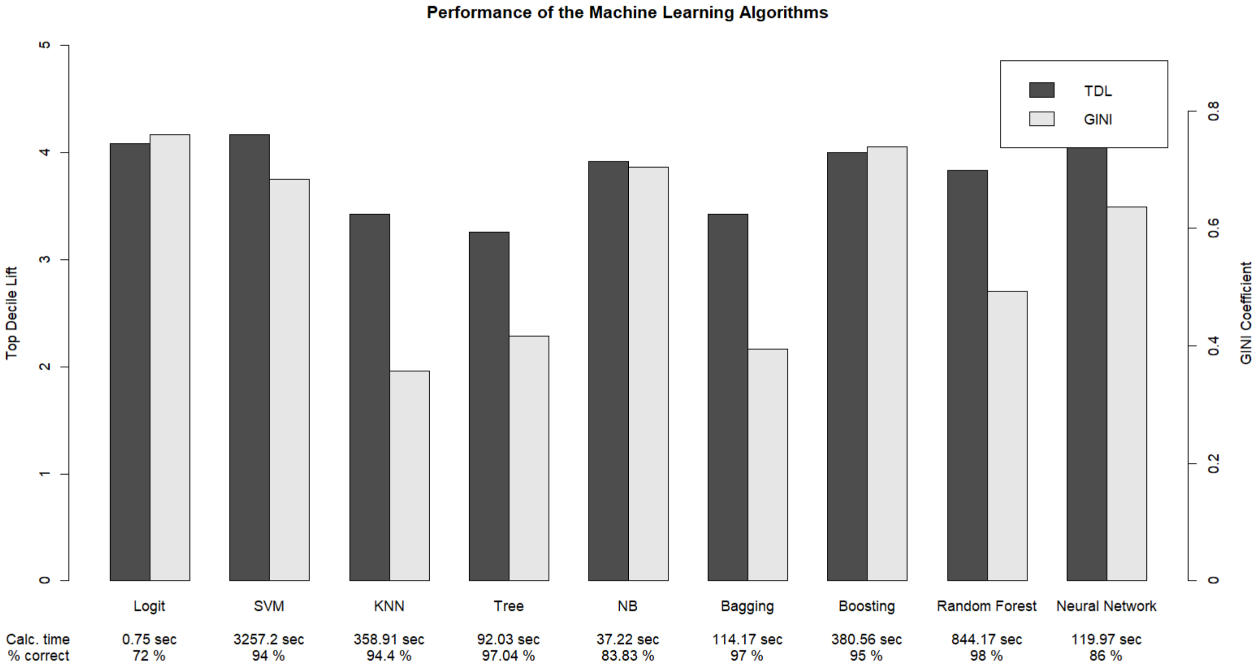


Fig. 2: Models performance comparison

Figure 2 gives an overview and comparison of the different ML models used. In order to proceed with the comparison, it is important to define the different metrics we are considering. The first is the so-called 'Gini coefficient'. The Gini coefficient is a measure of inequality that assesses the predictive power of a classification model. In the context of binary classification, it represents the ability of the model to correctly classify instances according to their predicted probabilities. A Gini coefficient value of 0 indicates random predictions, while a value of 1 indicates perfect separation between the positive and negative classes. A higher Gini coefficient indicates a more accurate model with better discriminative power.

The second metric is 'accuracy'. It is a widely used metric that measures the overall correctness of the predictions made by a classification model. It represents the ratio of correct predictions (both true positives and true negatives) to the total number of predictions. While accuracy is intuitive and easy to interpret, it may not provide a complete picture of model performance, especially in unbalanced datasets where classes are unevenly distributed.

The third is 'Top Decile Lift', which measures the performance of a model in identifying the top 10% of instances with the highest predicted probabilities. It quantifies how much better the model is at identifying positive instances compared to a random selection. A higher lift value indicates a better ability to identify the most important cases. This metric is particularly relevant when the focus is on capturing a specific segment of the population with a higher likelihood of the target variable. When comparing these metrics, it is important to consider the specific context and objectives of your classification problem. In fact, the Gini coefficient provides a comprehensive measure of predictive and discriminatory power. It is useful if you want to prioritise accurate classification, or if the positive and negative classes need to be well separated. When discrimination is critical, the Gini coefficient may be an appropriate metric to consider. Accuracy is a general metric that reflects the overall correctness of the predictions. It is suitable when the classes are balanced and you are interested in the overall performance of the model without focusing on any particular segment. However, accuracy may not be reliable when dealing with unbalanced datasets where the class distribution is skewed. Top Decile Lift is particularly relevant when identifying the best performing instances is critical to your business objective. It helps to assess the model's ability to prioritise cases with a higher probability

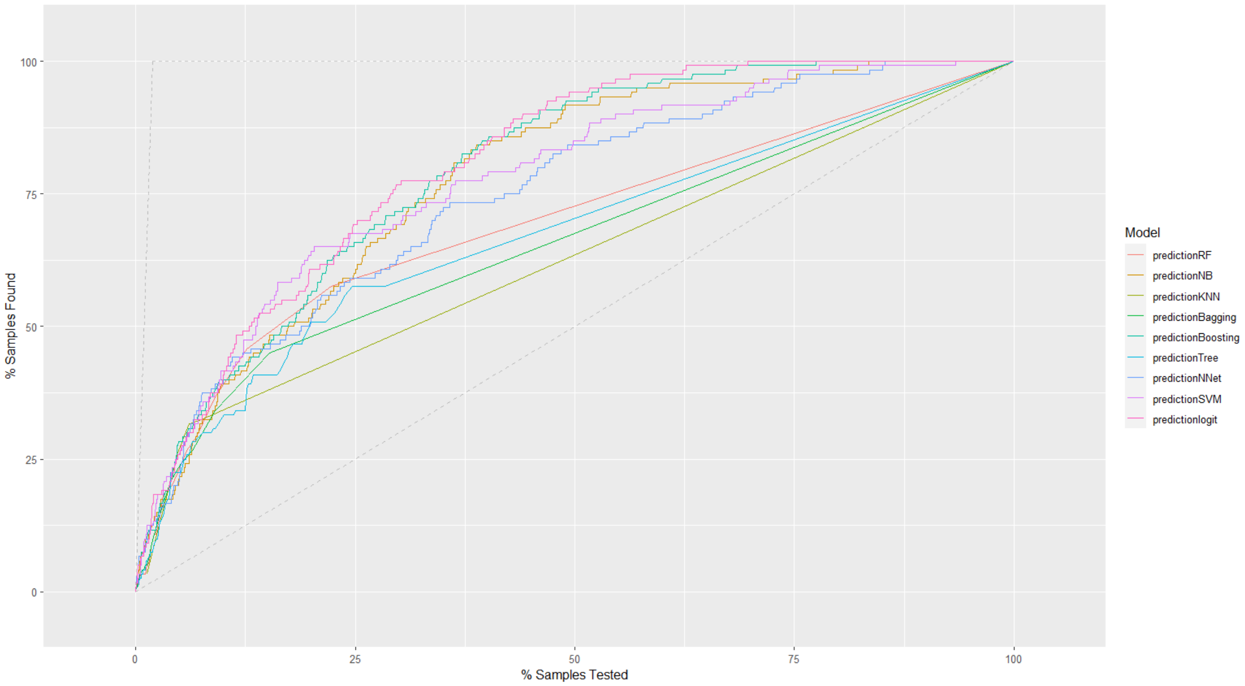
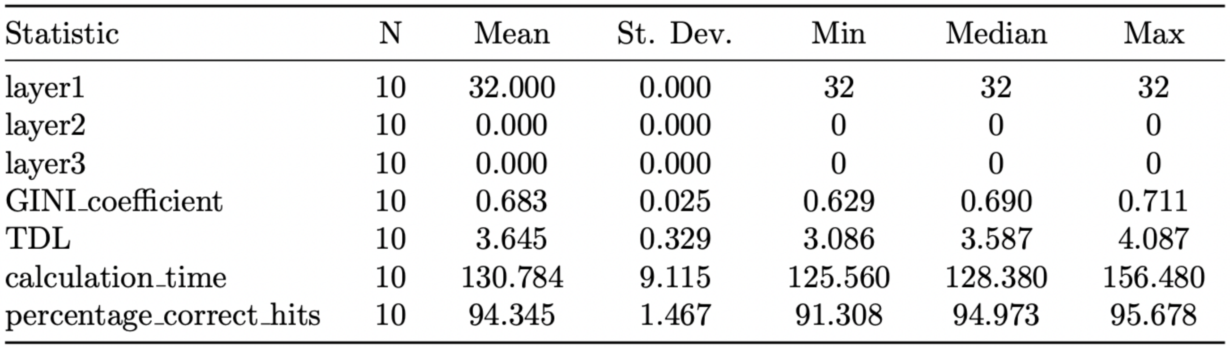


Fig. 3: Models comparison according to their ROC curves

of the target variable. If capturing the most important cases or focusing on a specific segment is important, Top Decile Lift can provide valuable insights.

In addition to these metrics, we also decide to look at the models' ROC curve and their area under the curve AUC. Figure 3 gives us a complete overview of the performance of the difference models. More specifically, the Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the performance of a binary classification model at different discrimination thresholds. It is created by plotting the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis as the threshold for classifying positive instances is varied. In this way, we can see how different models perform at different thresholds. As our computing power was limited, we also took into account the time it took to train the model. These unusual metrics can be very significant in a business context. Since training performance and training time might represent a cost to the business. As far as our models are concerned, almost all of them show good accuracy. The SVM would have been an excellent candidate as the final model, but as mentioned above, the model takes too long to train, probably due to the high dimensionality of the dataset. For this reason, our final choice was the neural network, which generally shows high values in all metrics and relatively short training times. Specifically, as shown in Figure X, the neural network has an accuracy of 86%, a Gini coefficient of 0.65 and a very high TDL of around 4.

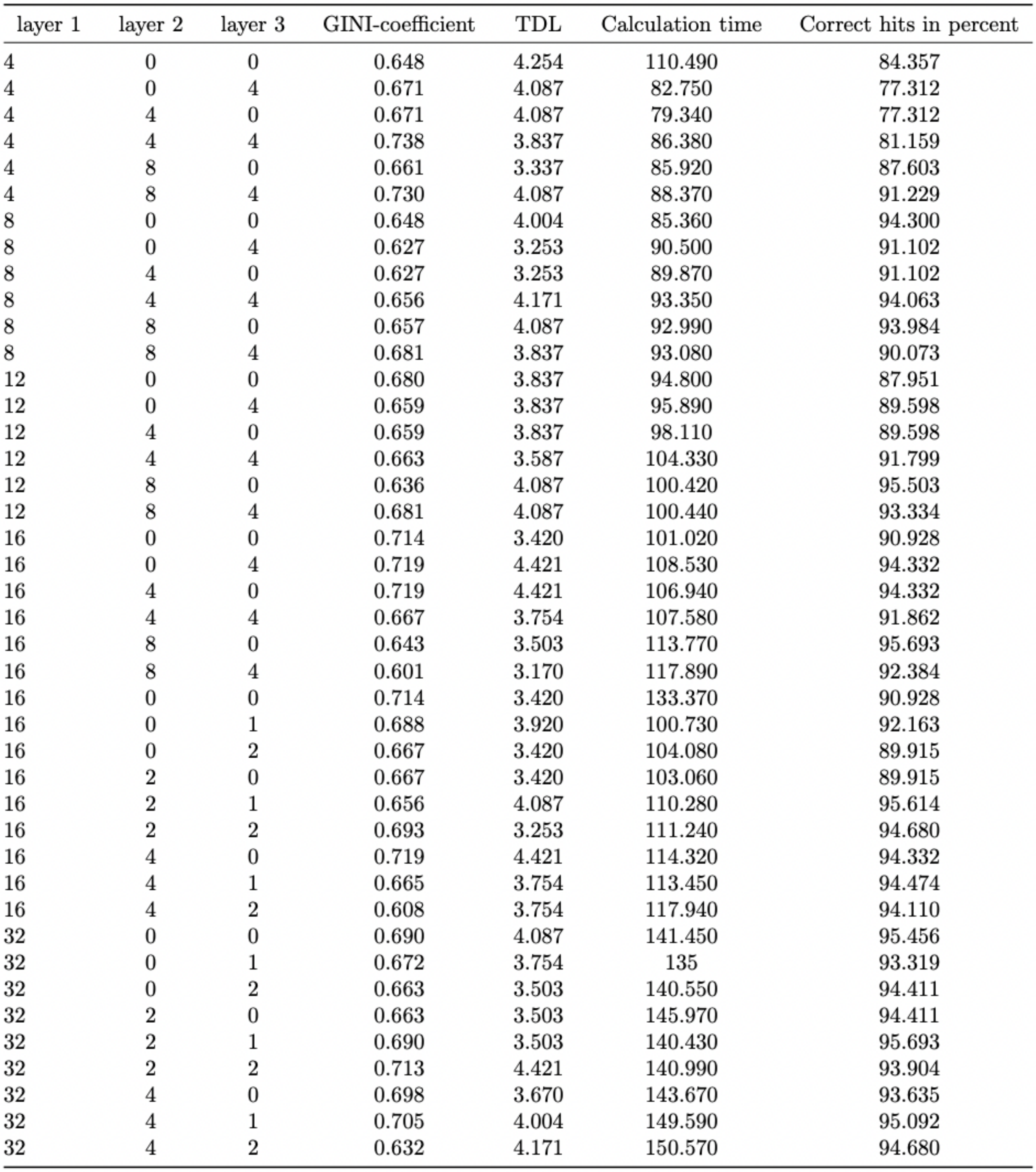
4.5 Hyperparameters tuning



Statistic	N	Mean	St. Dev.	Min	Median	Max
layer1	10	32.000	0.000	32	32	32
layer2	10	0.000	0.000	0	0	0
layer3	10	0.000	0.000	0	0	0
GINI_coefficient	10	0.683	0.025	0.629	0.690	0.711
TDL	10	3.645	0.329	3.086	3.587	4.087
calculation_time	10	130.784	9.115	125.560	128.380	156.480
percentage_correct_hits	10	94.345	1.467	91.308	94.973	95.678

Fig. 4: Neural Network best hyperparameter configuration

Once the best model was identified, model optimisation steps were performed. Hyperparameter tuning or fine-tuning is a fairly common technique in machine learning. It is the process of finding the optimal values for the hyperparameters of a machine learning model. Hyperparameters are parameters that are set before the model is trained and are not learned from the data. They control the behaviour and performance of the model, such as the learning rate, the regularisation strength or the number of hidden units in a neural network. The purpose of tuning is to find the combination of hyperparameter values that maximises the performance or minimises the error of the model for a given task or data set. Their choice can greatly affect the predictive ability and generalisation performance of the model. A common approach is grid search. Grid search involves specifying a grid of hyperparameter values to explore. The grid is defined by selecting specific values or a range of values for each configuration. The model is then trained and evaluated for each combination in the grid. The performance of the model is typically measured by a performance metric, such as accuracy or mean squared error, and the combination of hyperparameters that gives the best performance is selected as the optimal set. Grid search exhaustively searches all possible combinations in the grid, which can be computationally expensive when the search space is large. To alleviate this, Random Search can be used as an alternative. In Random Search, instead of systematically exploring all combinations, random combinations of hyperparameters are sampled from the defined search space. This allows for a more efficient search, especially when some of the parameters are less influential than others. In the case of our model, various combinations were tried out using the Grid Search technique. Specifically, we tried to find the right combination between the number of layers the model should have and the number of neurons within the single layer. As shown in Figure 4, The best configuration proved to be the one with only one layer and 32 neurons. Achieving an accuracy of 96%.

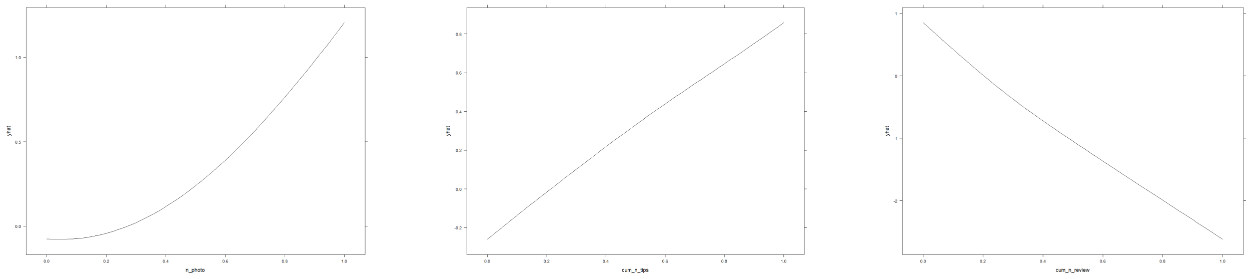


layer 1	layer 2	layer 3	GINI-coefficient	TDL	Calculation time	Correct hits in percent
4	0	0	0.648	4.254	110.490	84.357
4	0	4	0.671	4.087	82.750	77.312
4	4	0	0.671	4.087	79.340	77.312
4	4	4	0.738	3.837	86.380	81.159
4	8	0	0.661	3.337	85.920	87.603
4	8	4	0.730	4.087	88.370	91.229
8	0	0	0.648	4.004	85.360	94.300
8	0	4	0.627	3.253	90.500	91.102
8	4	0	0.627	3.253	89.870	91.102
8	4	4	0.656	4.171	93.350	94.063
8	8	0	0.657	4.087	92.990	93.984
8	8	4	0.681	3.837	93.080	90.073
12	0	0	0.680	3.837	94.800	87.951
12	0	4	0.659	3.837	95.890	89.598
12	4	0	0.659	3.837	98.110	89.598
12	4	4	0.663	3.587	104.330	91.799
12	8	0	0.636	4.087	100.420	95.503
12	8	4	0.681	4.087	100.440	93.334
16	0	0	0.714	3.420	101.020	90.928
16	0	4	0.719	4.421	108.530	94.332
16	4	0	0.719	4.421	106.940	94.332
16	4	4	0.667	3.754	107.580	91.862
16	8	0	0.643	3.503	113.770	95.693
16	8	4	0.601	3.170	117.890	92.384
16	0	0	0.714	3.420	133.370	90.928
16	0	1	0.688	3.920	100.730	92.163
16	0	2	0.667	3.420	104.080	89.915
16	2	0	0.667	3.420	103.060	89.915
16	2	1	0.656	4.087	110.280	95.614
16	2	2	0.693	3.253	111.240	94.680
16	4	0	0.719	4.421	114.320	94.332
16	4	1	0.665	3.754	113.450	94.474
16	4	2	0.608	3.754	117.940	94.110
32	0	0	0.690	4.087	141.450	95.456
32	0	1	0.672	3.754	135	93.319
32	0	2	0.663	3.503	140.550	94.411
32	2	0	0.663	3.503	145.970	94.411
32	2	1	0.690	3.503	140.430	95.693
32	2	2	0.713	4.421	140.990	93.904
32	4	0	0.698	3.670	143.670	93.635
32	4	1	0.705	4.004	149.590	95.092
32	4	2	0.632	4.171	150.570	94.680

Fig. 5: Overview of the Grid Search used for the fine-tuning of the Neural Network

4.6 Variable importance and partial dependencies

Since, however, the ultimate goal of this analysis is to give insights to the business, it was only right to focus and devote space to the explainability of the model, especially in the presence of a neural network, which is often described as a real 'black box'. Trying to overcome this problem, this section proposes an in-depth study of the importance and influence of the individual variables within the model. We go on to analyse the impact of the variables globally and then, in order to also be able to provide precise indications from a business point of view, we analyse the so-called dependency plots, whereby the various predictors are taken individually and are analysed as a function of the target variable. In this way we can understand their individual influence and also the direction - positive or negative - they have in relation to the target variable. Since the ultimate goal of this analysis is to provide insights to the business, it was necessary to focus on and dedicate attention to the explainability of the model, especially when dealing with a neural network, often described as a true 'black box'. To address this issue, this section proposes a comprehensive study of the importance and influence of individual variables within the model. By analyzing the impact of variables globally and then, in order to provide specific business insights, examining the so-called dependency plots, where each predictor is analyzed individually in relation to the target variable. This allows us to understand their individual influence and the direction - positive or negative - they have in relation to the target variable. As seen in Figure 7, the most important variables globally for the model are the cumulative number of reviews, the cumulative number of tips, and the number of photos, reinforcing and making it even clearer how the aspect of user-generated content (UGC) is of primary importance for predicting check-ins. Other variables such as 'weekend' or 'business_park' follow in importance.



(a) Partial Dependency Plot - Number of Photos

(b) Partial Dependency Plot - Cumulative Number of Tips

(c) Partial Dependency Plot - Cumulative Number of Reviews

Fig. 6: Partial dependency plots of cumulative number of reviews, tips, and number of photos

However, it is important to note that Figure 6 provides a measure of the overall relevance of the variable but does not provide any information regarding the positive or negative direction of the variable. To obtain this information, it is necessary to look at the Partial Dependency Plots. Quite interesting is the behavior of the variable regarding the cumulative number of reviews, which as regards its partial dependency plot, seems to have - for some reason that should be investigated with further research - a negative trend. This means that as reviews increase, the likelihood of a check-in appears to decrease. While as regards the cumulative number of tips and photographs, the general trend is positive

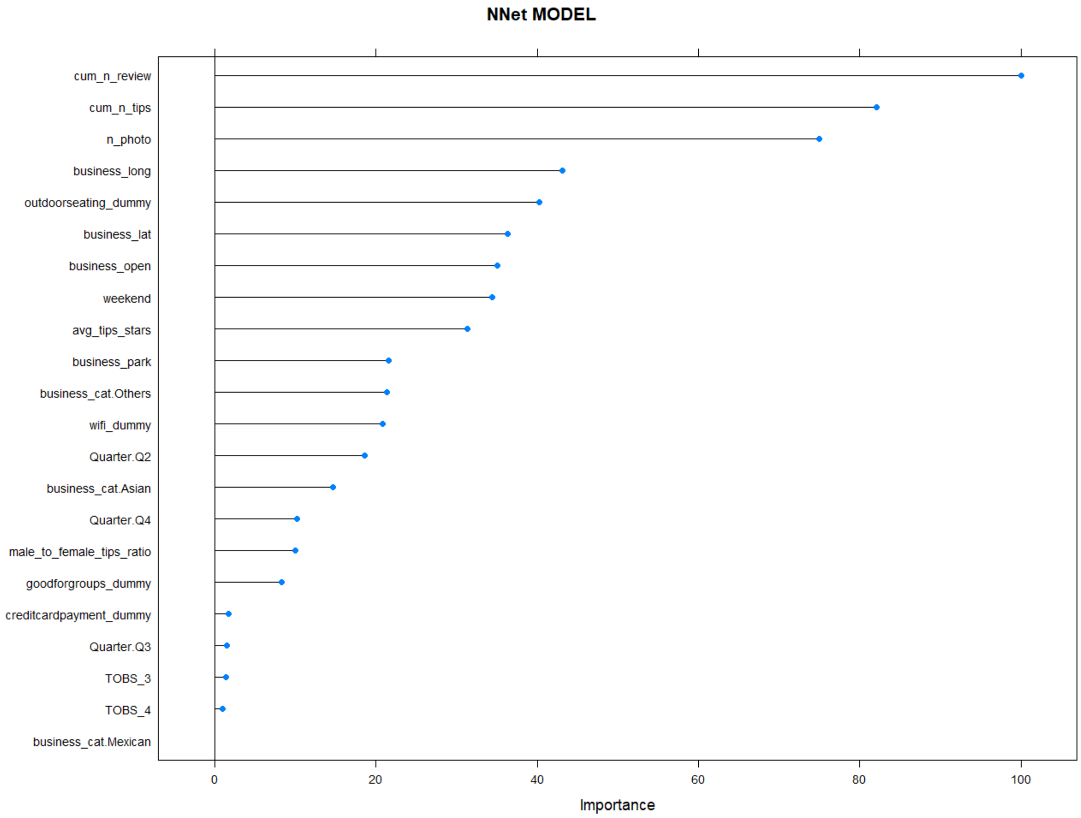


Fig. 7: Variables importance

5 Conclusion

In this study, we conducted a comprehensive analysis using various machine learning models to predict customer check-in behaviors. The target variable, check-in, was rebalanced using the SMOTE oversampling technique to address the issue of class imbalance. Among the evaluated models, the neural network stood out as the best performer, achieving an accuracy of 96% and demonstrating superior predictive capabilities compared to other models such as logistic regression, decision trees, random forests, support vector machines, naive Bayes, and gradient boosting. The neural network configuration with a single layer and 32 neurons proved to be optimal in this case. The utilization of SMOTE oversampling played a crucial role in improving the model's performance. By generating synthetic instances of the minority class, SMOTE effectively balanced the dataset and mitigated the impact of class imbalance. This led to enhanced accuracy and predictive power in identifying check-in behaviors. Additionally, we conducted an analysis of variable importance to identify the factors influencing check-in patterns. The cumulative number of reviews, cumulative number of tips, and number of photos emerged as the most significant variables, emphasizing the critical role of user-generated content (UGC) in predicting check-ins. Other variables such as 'weekend' and 'business_park' also demonstrated notable importance.

References

[1] Anderson, M., & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563), 957-989.

[2] Bakhshi, S., Kanuparthy, P., & Shamma, D. A. (2014). If it is funny, it is mean: Understanding social perceptions of yelp online reviews. In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work* (pp. 46-52).

[3] Bowen, J. T., & Chen, S. L. (2001). The relationship between customer loyalty and customer satisfaction. *International Journal of Contemporary Hospitality Management*, 13(5), 213-217.

[4] De Vries, J., Roy, D., & De Koster, R. (2018). Worth the wait? How restaurant waiting time influences customer behavior and revenue. *Journal of Operations Management*, 63, 59-78.

[5] Duarte Alonso, A., O'Neill, M., Liu, Y., & O'Shea, M. (2013). Factors Driving Consumer Restaurant Choice: An Exploratory Study From the Southeastern United States. *Journal of Hospitality Marketing and Management*, 22(5), 547-567.

[6] Li, H., Chen, Q. X., Liang, S., & Yang, J. J. (2022). The power of internet exposure: influence of online news coverage on restaurant survival. *International Journal of Contemporary Hospitality Management*, 34(4), 1399-1422.

[7] Li, H., Meng, F., Jeong, M., & Zhang, Z. (2020). To follow others or be yourself? Social influence in online restaurant reviews. *International Journal of Contemporary Hospitality Management*, 32(3), 1067-1087.

[8] Lim, Y. S., & Van Der Heide, B. (2015). Evaluating the wisdom of strangers: The perceived credibility of online consumer reviews on Yelp. *Journal of Computer-Mediated Communication*, 20(1), 67-82.

[9] Liu, P., & Tse, E. C. Y. (2018). Exploring factors on customers' restaurant choice: an analysis of restaurant attributes. *British Food Journal*, 120(10), 2289-2303.

[10] Mahesh, B. (2018). Machine Learning Algorithms-A Review. *International Journal of Science and Research.*

[11] Palka, A. (2017). Consumer preferences on impulse ice cream. *Towaroznawcze Problemy Jakości*, 51(2).

[12] Pan, X., Hou, L., Liu, K., & Niu, H. (2018). Do reviews from friends and the crowd affect online consumer posting behaviour differently? *Electronic Commerce Research and Applications*, 29, 102-112.

[13] Pantelidis, I. S. (2010). Electronic meal experience: A content analysis of online restaurant comments. *Cornell Hospitality Quarterly*, 51(4), 483-491.

[14] Parikh, A. A., Behnke, C., Nelson, D., Vorvoreanu, M., & Almanza, B. (2015). A Qualitative Assessment of Yelp.Com Users' Motivations to Submit and Read Restaurant Reviews. *Journal of Culinary Science and Technology*, 13(1), 1-18.

[15] Parsa, H. G., Gregory, A., Self, J. T., & Dutta, K. (2012). Consumer Behaviour in Restaurants: Assessing the Importance of Restaurant Attributes in Consumer Patronage and Willingness to Pay. *Journal of Services Research*, 12(2).

[16] Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51-59.

[17] Saad Andaleeb, S., & Conway, C. (2006). Customer satisfaction in the restaurant industry: An examination of the transaction-specific model. *Journal of Services Marketing*, 20(1), 3-11.

[18] Sagiroglu, S., & Sinanc, D. (2013). Big Data: A Review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42-47).

[19] Tian, X., Cao, S., & Song, Y. (2021). The impact of weather on consumer behavior and retail performance: Evidence from a convenience store chain in China. *Journal of Retailing and Consumer Services*, 62.

[20] Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, 29(4), 694-700.

[21] Yelp. (2023). Yelp - Company - Fast Facts. Retrieved July 13, 2023, from: https://www.yelp-press.com/company/fast-facts/default.aspx

6 Appendix

The analysis code is available here:
https://github.com/mnlscn/yelp-project/blob/main/full-code-appendix.txt