Store Location Influence on Consumer Spending Behavior

Likitha Sivananjegowda
22265298
School of Computing
Dublin City University
Dublin, Ireland
Likitha.shivananjegowda2@mail.dcu.ie

Seliz Suresh Koshy
22265479
School of Computing
Dublin City University
Dublin, Ireland
seliz.koshy2@mail.dcu.ie

Emmanuel Chakkalakkal Baby
22261866
School of Computing
Dublin City University
Dublin, Ireland
emmanuel.chakkalakkalbaby2@mail.dcu.ie

Abstract

Consumer spending behaviour is the process of selecting, purchasing or utilisation of commodities or services for their essential needs. Understanding their purchase behaviour and tailoring products that meet user expectations improves a business in the long run. Features like location, reviews, ratings, etc., help predict customer spending behaviour and help managers design their marketing programs.

To analyse this, we propose a novel model that considers factors influencing customer spending behaviour. We used a restaurant delivery app dataset, Swiggy, obtained from Kaggle and analysed data to identify the key features. Additionally, this paper will review existing works, the data mining methodology used, and the evaluation results.

Dataset: Swiggy Dataset from Kaggle

Gitlab Repository: Gitlab Repository containing code,report and video presentation

Research Question: How does store location influence consumer spending behaviour?

I. INTRODUCTION

With more than 100k eateries listed on Swiggy, one of the top food delivery services in India, We generated this dataset so that others may use their abilities to better understand this sector. It is also the largest online food ordering and delivery service in India, and it ranks first on the list of startups in India that have become unicorns. Founded in Bangalore in 2014, the firm has since spread to more than 100 Indian locations. Swiggy launched speedy pick-up and drop-off meal delivery services to make people's lives easier. It provides a single point of contact to make requests from a wide selection of restaurants as well as a complete meal ordering and delivery system that links foodies and local restaurants.

II. RELATED WORK

To further explore the research question, we researched similar papers worked on the topic.

The paper [1] presents a machine learning-based approach to predict user purchase behaviour. The paper highlights the importance of predicting user behaviour in e-commerce applications and proposes a framework based on machine learning algorithms to predict user purchase behaviour. Xiang et al. discuss using machine learning algorithms such as decision tree, random forest XGBoost, LightGBM, and combined models of the latter two for predicting user behaviour. The authors conclude that the proposed approach of the combined model of LightGBM & XGBoost has a much better F1 score than that of the single decision tree and random forest algorithms.

Consequently [2] examines how different menu price formats affect restaurant consumer behaviour. The study found that how prices are presented on a menu can influence customers' perception of value and affect their willingness to spend money. Specifically, menus that present prices in a clear and easy-to-read format and techniques like minimising the use of dollar signs can lead to higher sales and customer satisfaction. The study also highlights the importance of understanding how customers decide what to order and how prices can influence those decisions. The paper [3] focuses on developing a method to identify sentences in consumer product reviews that express how the product is used. The authors use natural language processing techniques to analyse a large dataset of product reviews and identify sentence patterns indicative of priceless expressions. The results show that their approach effectively detects usage expressions, which can help improve the understanding of how products are used and for generating product usage information for consumers. The authors conclude that their method can be applied to other domains to extract useful information from text, such as in medical and technical manuals.[4] The paper examines consumer shopping and spending behaviour across different retail formats, including traditional brick-and-mortar stores, online stores, and hybrid models that combine online and offline channels. [5] The study is based on a comprehensive analysis of existing literature and original data collected from surveys and transactional data from a large US-based retailer. The authors explore various factors influencing consumer behaviour, including convenience, price, product selection, and brand loyalty.

In a similar manner, [6] uses a combination of LSTM and RF by constructing a chronological sequence of consumer-commodity behaviour and other characteristics. This model inputs the constructed consumer product time series data into an LSTM-RF model and selects the best features for the model based on importance level ranking, which produces better classification results. Next the paper [7] explores the increasingly driven by convenience, with many consumers choosing to shop online or through hybrid models that offer flexible options such as buying online and pick up in store (BOPIS). The authors note that although traditional brick-and-mortar stores continue to play an essential role in consumer shopping behaviour, particularly for specific product categories and demographic groups, retailers that can offer a seamless omnichannel experience across multiple formats are likely to be the most successful in the current retail landscape.

Overall, the paper provides valuable insights into the complex factors influencing consumer shopping and spending behaviour across different retail formats and offers practical recommendations for retailers looking to stay competitive in a rapidly evolving marketplace.

III. Proposed method

To highlight how we implemented the model we have used the KDD process model to work on this project. In the following sections we give a description of the tasks involved in each phase.

1. Data Collection

For our initial data selection we have gone through multiple dataset that revolve around hospitality services or retail store data and we selected the Swiggy dataset as it had the features we were looking for. This dataset is of approximately 45.66 MB size and consists of around 142k million records. This dataset contains 10 different features of each restaurant listed on the website of Swiggy as well as additional columns such as weekly_avg_salary and no_years_open that we

have derived from the dataset. For training the data we dropped columns like id,link,name as they do not relate to our research question. We noticed that the columns city,cuisine,rating and cost are the key features that will help our predictive model.

2. Data Cleaning

To make sure the original data was properly prepared for modelling, numerous data cleaning and preparation activities were conducted. These comprised:

- Initially we have used the raw data to check for class imbalance which provides a useful way to quickly visualise the distribution of classes within each target variable in a dataset.
- The sum of null values in the dataset and checked the percentage of missing values in each column. Here it was found that some columns such as 'name','rating','rating_count','cost','cuisine','address' have very minute null values like 0.06% or 0.09% so we remove them and for columns like lic_no having 0.15% of null values we imputed those rows with the mode values. Overall, our data loss came up to a 0.1% during the cleaning process.
- Explored on the dataset which had very minute percentage ie, 0.9 and 0.6 of missing values so just we just dropped the columns
- With the null values in the lic_no column with the mode (i.e., the most frequent value) based on each licence number and filters out rows where certain columns have null values using boolean indexing. It removes rows where the name, rating, rating_count, cost, cuisine, address, and lic_no columns have null values. This is done using the - data[column].isnull() syntax, which returns the rows where the column is not null.
- We found that only 0.1% of the data was lost during the cleaning process. This suggests that the cleaning process was relatively successful in retaining most of the original data.
- Replacing the remaining 0.0 values in the 'rating' column with the mean value of the column, rounded to 2 decimal places. This step appears to be a way to handle missing or invalid values in the 'rating' column.
- We then perform data cleaning on the 'cost' column, by converting the string values to float values and removing the '₹' symbol.

Features	About	DType
city	The city where the restaurant is located	object
rating	Rating of the Restaurant	int64
rating_count	Number of People given the Rating	object
cost	Cost of eating in that restaurant	object
cuisine	Cuisines that restaurant serves	object
no_years_open	Derived from the lic_no column by using the 4th and 5th character contains the last 2 digits and since	int64

	fssai was established 2008 in India only those records are selected.	
weekly_avg_salary	Derived the weekly avg salary for each state by combining a dataset from moneymint which contains avg salary by month for each state.	int64

Table 1: Features used in the dataset

3. Modelling

We employed three machine learning models XGBoost Regressor, Random Forest Regressor, and Decision Tree Regressor, to accomplish this task using Python. These models were chosen because of their ability to handle non-linear relationships, robustness against overfitting, and ease of implementation. For XGBoost Regressor, we used the library 'xgboost,' and for Random Forest Regressor and Decision Tree, we used another library, 'scikit-learn [8].' By comparing the performance of these models, we aimed to identify the most effective approach for predicting restaurant ratings based on the given features, ultimately helping restaurant owners and stakeholders make informed decisions to enhance their services and customer satisfaction. This study aims to forecast restaurant ratings using three distinct machine-learning techniques: Random Forest Regressor, XGBoost Regressor, and Decision Tree Regressor. The dataset includes information on restaurant locations, types of cuisine, average costs, and customer ratings. The project is broken down into six primary steps: loading the dataset, pre-processing the data, setting model parameters, training the models, evaluating their performance, and making predictions for new restaurants.

1. Dataset loading: The necessary pandas and NumPy libraries are imported, and the Swiggy dataset is read from a CSV file ('swiggy_final_versionV2.csv'), which contains restaurant data used for rating predictions.
2. Data pre-processing: Unnecessary columns are removed, and the 'cuisine' and 'city' columns are converted into categorical data types. The dataset is then split into training and testing sets, dividing the features (X) from the target variable (ratings, y). The 'cuisine' and 'city' columns are one-hot encoded to create a numerical representation compatible with machine learning algorithms.
3. Parameter definition: The code sets up a dictionary of parameters for each of the three models (Random Forest Regressor, XGBoost Regressor, and Decision Tree Regressor) to manage their behaviour during training.
4. Model training: Each model is initialized with the specified parameters and trained using the restaurant features and ratings from the training set.
5. Model evaluation: The trained models are employed to predict ratings for the test set. Performance metrics like Mean Absolute Error, Root Mean Squared Error, and R-squared are then calculated to gauge each model's accuracy in predicting restaurant ratings.
6. Rating prediction: Use the trained models to estimate a new restaurant's rating based on its features. A new Data Frame containing the restaurant's features is created, and the 'cuisine' and 'city' columns are one-hot

encoded and aligned with the training data columns. The models then predict the rating, and the results are displayed.

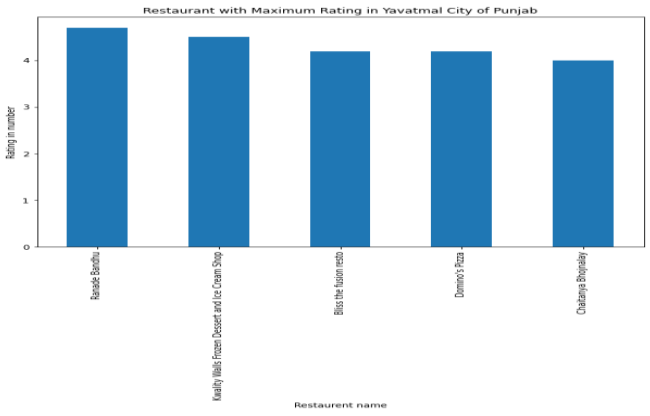
According to the evaluation results, the Decision Tree Regressor outshines the Random Forest Regressor and XGBoost Regressor in forecasting restaurant ratings. The Decision Tree Regressor has the lowest MAE and RMSE values and the highest R-squared value, indicating that it offers the most precise predictions among the three models.

4. Visualization

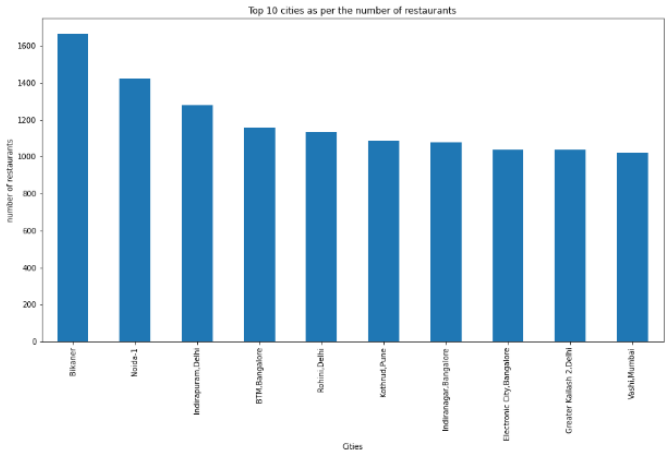
With more than 100k eateries listed on Swiggy, one of the top food delivery services in India, I generated this dataset so that others may use their abilities to better understand this sector.

By responding to the following questions, the interface aids in making the outcomes clear.

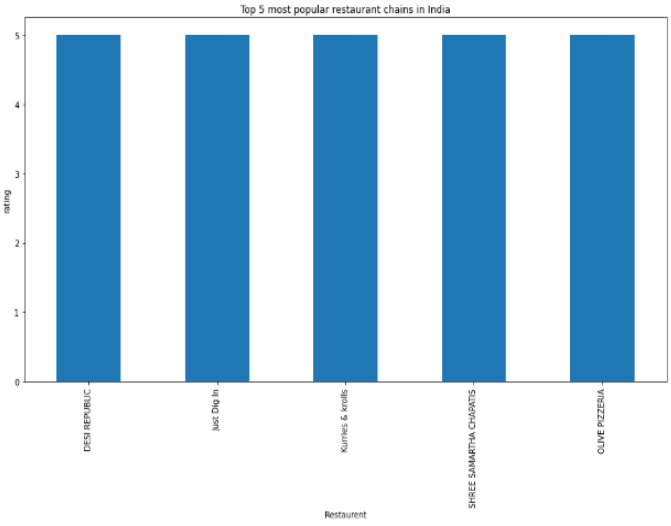
Restaurant with Maximum Rating in particular?



Top 10 cities as per the number of restaurants listed?



Top 5 most popular restaurant chains in India?



5. Evaluation/Results:

This project used three evaluation measures: Mean Absolute Error, Root Mean Square and R-squared

over the pre-processed data of 148242 records, and Table 1 illustrates the results obtained.

Models	Mean Absolute Error	Root Mean Square	R-squared
Random Forest Regressor	1.103	1.154	0.015
XGBoost Regressor	0.783	1.186	-0.039
Decision Tree Regressor	0.212	0.425	0.866

Table 1**:** Results of each model on the dataset

While investigating the outcomes the team noticed that Decision Tree Classifier seem to perform better in terms of the score compared to the other models. To evaluate the rating on each model we provided the same input for each model i.e, city : 'Abohar', rating_count : 1.0, cost : 200.0, 'cuisine: 'Beverages,Pizzas', weekly_avg_salary: 4529.75, no_years_open: 1. The predicted rating is depicted in Table 2 with Decision Tree Regressor having closer value to the actual rating.

Model Used	Actual Rating	Predicted Rating
Random Forest Regressor	1.61	2.550259556345501
XGBoost Regressor	1.61	1.5029217
Decision Tree Regressor	1.61	1.6428380192028247

Table 2: Outcomes obtained from each model

Our expected outcome was XGBoost Regressor having more accuracy compared to the other models and this could possibly be due to the method of encoding done on the categorical columns.

IV. Conclusions and future work

To summarise, we have selected the Swiggy dataset to identify the factors that influence customer spending behaviour and derived additional features to better predict our research question. We have then applied various pre-processing steps to prepare the model for prediction. The models used are Decision Tree Regressor, XGBoost Regressor, Random Forest Regressor and from our observations we noted that Decision Tree Regressor proved to predict closer to the actual rating compared to the other models.

As part of future work this dataset can be further improved by using the combined LSTM-rf model as a similar type of dataset would work better on this model as seen in the related works.

V.REFERENCES

[1].Xiang Zhai, Peng Shi, Liang Xu, Yalong Wang, and Xi Chen. 2020. Prediction model of User Purchase Behaviour based on machine learning. *2020 IEEE International Conference on Mechatronics and Automation (ICMA)* (2020). DOI:http://dx.doi.org/10.1109/icma49215.2020.9233677

[2]. Sybil S. Yang, Sheryl E. Kimes, and Mauro M. Sessarego. 2009. Menu price presentation influences consumer purchase behavior in restaurants. *International Journal of Hospitality Management* 28, 1 (2009), 157–160. DOI: http://dx.doi.org/10.1016/j.ijhm.2008.06.012

[3]. Lahiri, S., Vydiswaran, V.V. and Mihalcea, R., 2017, November. Identifying usage expression sentences in consumer product reviews. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 394-403)

[4] Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings of the 20th international conference on Computational Linguistics (COLING '04). Association for Computational Linguistics, USA, 841–es. https://doi.org/10.3115/1220355.1220476

[5].Jan Valendin, Thomas Reutterer, Michael Platzer, Klaudius Kalcher, Customer base analysis with recurrent neural networks, International Journal of Research in Marketing, Volume 39, Issue 4,2022, Pages 988-1018, ISSN 0167-8116, https://doi.org/10.1016/j.ijresmar.2022.02.007.

[6]. W. Hu and Y. Shi, "Prediction of online consumers' buying behaviour based on LSTM-RF model," 2020 5th International Conference on Communication, Image and Signal Processing (CCISP), Chengdu, China, 2020, pp. 224-228, doi: 10.1109/CCISP51026.2020.9273501.

[7]. Xiao, Y., Wang, Y., Mao, H. and Xiao, Z., 2016, December. Predicting restaurant consumption level through social media footprints. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 3328-3338).

[8]. http://scikit-learn.org/