Restaurant Startup Analysis

Aryan Shrestha
Department of Computer Science
(of College of Arts and Science)
Georgia State University
Atlanta, USA
ashrestha6@student.gsu.edu

***Abstract*—An increase in cutting-edge tools and machine learning for data analysis has facilitated the establishment of numerous firms a lot simpler and easier than before. Various perspectives concerning the population, culture, way of life, and consumer desires, and ratings have allowed us to make wise selections in the past putting a business idea into action. I examine one such aspect in this project: food and their suppliers. If what is provided does not attain the appropriate target market, the business is likely to fail or not be long-term sustainable. In this project I try to answer various such questions by analyzing multiple restaurant's data for a particular location. The goal of this project is to study the various factors that are directly responsible for the success of a newly start-up restaurant in a particular area which will help the owners to better understand their product and gain desired customers.**

Keywords—data mining, restaurant, startup

I. INTRODUCTION

The most recognizable representation of a lively and thriving civilization are restaurants. Most of them represent the unique food taste of the culture in a particular location. It serves as a center of unity for those that visit and spend quality time with family and friends or those who host crucial business gatherings. In order to set up a restaurant, it's crucial to think about the following: 1) Determining the target customers 2) The uniqueness offered by the restaurant 3) What effects do the products have on the area where the restaurant is built. The dataset I'll be using is provided by Zomato Media Pvt. Ltd which is a leading Indian restaurant aggregator and food delivery start-up that provides information on menus and user-reviews for different restaurants across different cities in India and operates a delivery service for its client restaurants. I'll be only using the data of the restaurants located in the city of Bangalore. Various types of restaurants from all over the world are found in Bangalore with nuanced cuisine. The food industry is always at a rise here and the number of restaurants is increasing day by day. I am planning to use advanced data mining techniques considering the factors which influence a potential business's success restaurant such as the menu, prices, and other factors based on the demographic requirements and expectations.

II. RELATED WORK

There is some very interesting research and analysis going on in examining the restaurant's data so as to provide the best products and services to the customers. According to [1], seasonal and predictable elements define restaurants together with unique consumer demand, which create a challenge for restaurants to operate efficiently. Another such work is accessible in [2] in which how customers view the consumer values in dining out experiences is assessed, and the results are compared with other studies on service quality and meal experience [3]. This research aimed to comprehend the perspectives of executives at significant U.S. restaurant chains with relation to the method, cause, and presenting menus with healthier alternatives. This data analysis has been very helpful in understanding the features best required for a startup restaurant.

III. DATASET

A. Data Source

The dataset used is provided by Zomato Media Pvt. Ltd which is a leading Indian restaurant aggregator and food delivery start-up that provides information on menus and user-reviews for different restaurants across different cities in India and operates a delivery service for its client restaurants. The dataset is available to everyone on the Kaggle platform. It is a collection of records that shows hundreds of restaurants across the city of Bangalore. Each row contains seventeen features each pertaining to different details of the restaurant such as its location, details of the items offered, pricing, the ratings given by consumers etc. It contains around 51717 rows with 17 columns.

Here, the target variable will be 'rate' and the aggregated ratings of the restaurant are predicted given the scenarios and thereby determine whether a restaurant with the given features is likely to be successful or not.

B. Variables in our dataset

- url: The restaurant's website url
- address: Address of the restaurant
- name: Name of the restaurant
- online_orders: It specifies if the restaurant takes online orders or not.
- book_table: Indicates whether or not the restaurant offers table reservations.
- rate: The restaurant's rating out of 5
- votes: Number of votes received by the restaurant on Zomato
- phone: the restaurant's phone number
- location: the neighborhood in which the restaurant is located.
- rest_type: specifies the type of restaurant.
- disk_liked: Indicates which dishes were popular among customers in that restaurant.
- Cuisines: Cuisines available at the restaurant
- approx_cost (for two people): Estimated cost of food in that restaurant for two people.

- Reviews_list: Reviews given by users
- menu_item: The restaurant's menu
- listed_in(type): Specifies the type of service provided by a restaurant
- listed_in(city): The restaurant is on the city's list.

IV. DATA PREPROCESSING

Data has been essential to businesses like banking, insurance, and retail, telecom and e-commerce for the business development and expansion. It's crucial to both collect the data and clean the data properly prior to utilizing it. Data redundancies and discrepancies or irrelevance yields subpar outcomes. The quality of the data is crucial if businesses want to use it to optimize their operations and boost profitability. Poor data could completely mislead and influence how the model interprets the data, resulting in inaccurate assumptions. It is important to assess the dataset's quality and make sure that extra noise is removed for efficient data analysis and machine learning models. to examine the connections between distinct variables. Before modeling our predictor, it is crucial to take into account the raw data's most pertinent variables. However, there are some variables that have little to no effect on the prediction model such as url, phone etc. These least likely used features are not used as they provide very less insights.

A. Dropping unwanted columns

Even though there are 17 features, I will focus only on the most crucial ones and eliminate the other columns. The dropped columns are url, address, phone, location, dish_liked, reviews_list and menu_item. Now, there are only 10 features.

B. Dropping duplicate rows

After checking the dataset, it is found that there are 124 duplicate rows in the dataset. These rows adversely affect the result and should be removed. After removing the repeated rows, the number of rows present are 51593. This will help reduce the complexity in data with respect to processing time.

C. Cleaning individual rows

Each row contains a lot of redundant data that needs to be taken care of. Data may have a lot of structural errors and can exist in different forms but have the same meaning. These errors will cause unwanted insights and misinterpretations. So, these errors should be dealt with. Firstly, the redundant data from the name column is removed. All the punctuations, numbers, special characters etc. are removed and only the alphabets are retained in this step. Secondly, the unique characters are checked for the ratings column and there are nan, new and - values instead of the usual ratings. Also, there are string values containing /5 character sets. All this insignificant data is removed, and the values of the ratings column is converted into numeric values. Thirdly, the cost columns have string values containing commas. These values are removed and similar to the above column, we convert it into numeric values.

D. Handling missing values

There are always some missing values in any dataset. Null values are the values that occur due to the value at the given cell being missing or deliberately chosen to not to be mentioned. Among the fundamental checks, it is important to look whether the data has any null values. And also to understand if the null value conveys some information or it was absent due to some man-made error like missing entries or it actually represents unavailability of data. All the features are checked for missing values.

Some features like ratings have 10003 null values. This means a significant number of records in the dataset do not have a legitimate value. It is unwise to fill these values or guess an entry for these null values. Similarly, other features such as rest_type, cuisines and cost have 227, 45 and 344 missing values respectively. So, I removed all the corresponding rows which contain null values. After dropping the rows with null values, the number of rows in the dataset is reduced to 41190.

E. Scaling and Label Encoding

When the data is supplied or more precisely, required to be presented in a numerical format rather than a category one, machine learning algorithms will perform more accurately. This is explained by the fact that a computer only comprehends numbers in the first two digits. Any model training that includes categorical predictors starts with label encoding. In order to encode a dataset, some number labels are assigned to categorical values. This crucial data preprocessing step is necessary since models like linear regression and others only require numerical data. The predictors in the dataset were label encoded to adhere to convention and aid in training the model better. Scaling is important for models like PCA and linear regression, so the data is scaled accordingly for these models. However, it is not necessary for other models like decision tree, random forest and so on.

V. DATA EXPLORATION

The percentage of restaurants that offer table booking and the restaurants that take online orders is shown in Figure 1 and Figure 2 respectively. Based on this, we can deduce that only 15.2% of the restaurants have table booking option while 65.7% have online order option.

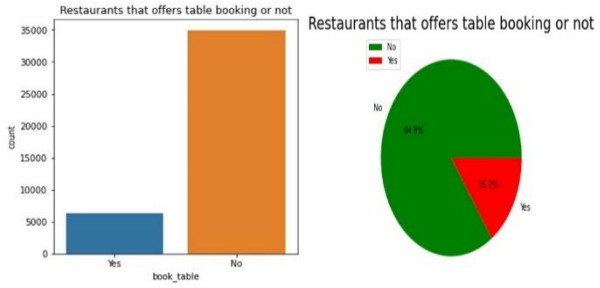

Figure 1: Restaurants that offer table booking or not

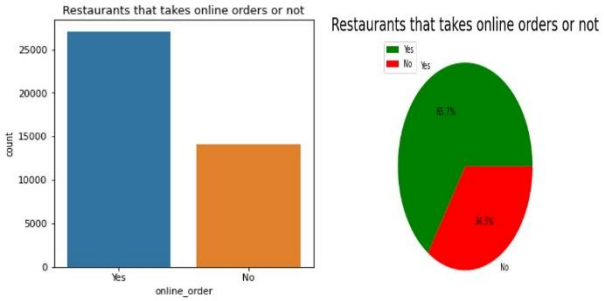

Figure 2: Restaurant that takes online order or not

Figure *3* shows the most common ratings of the restaurants which is our target variable. The distribution is slightly left skewed and the majority of the restaurants have a rating of 3.9. Many restaurants have ratings between 3.6 and 3.9 and very few restaurants have low ratings. The correlation between ratings and average cost for two table is shown in Figure *4* and the most expensive restaurants have ratings between 3 and 4.5.

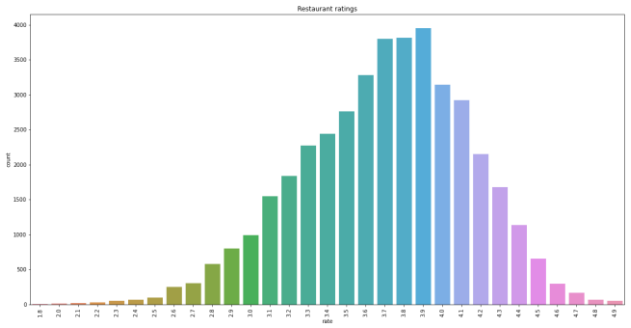

Figure 3: Most common ratings

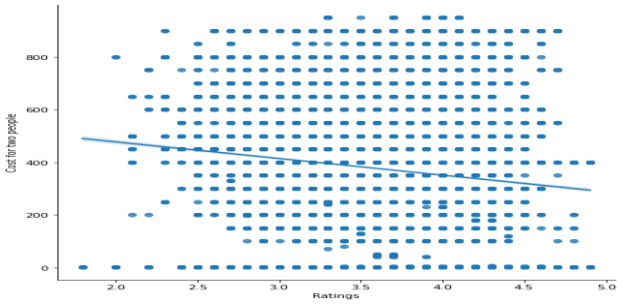

Figure 4: Ratings vs average cost for two people

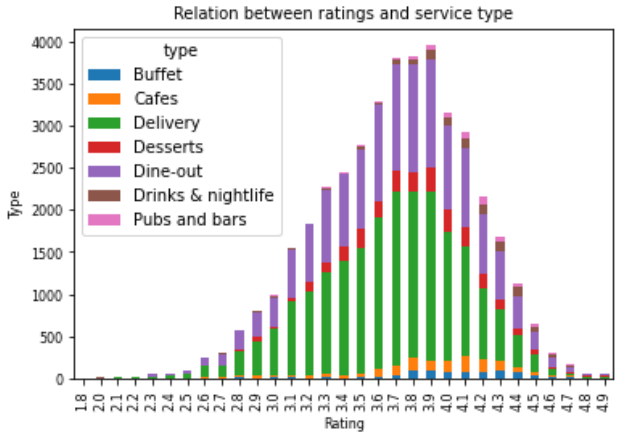

Figure 5: Ratings vs service types

From Figure *5*, the type of restaurants that have high ratings are mostly delivery and dine-out places. The relation between average cost of two people and restaurants can be seen in Figure *6*. The cafes are the most expensive type of restaurants in Bangalore followed by delivery places. Figure *7* shows the best restaurants chain based on popularity. Café coffee day and Onesta are the most popular with over 80 outlets throughout the city. Smally's restro café is the least popular with just over 50 chains throughout the city.

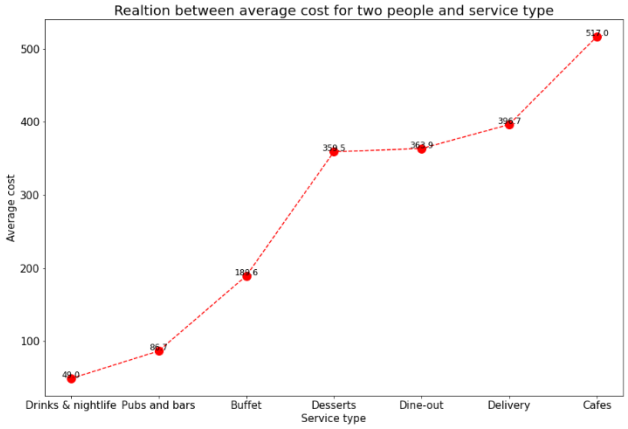

Figure 6: Cost vs types

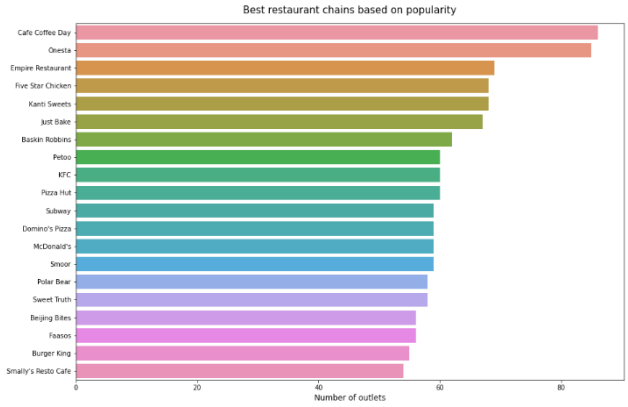

Figure 7: Most popular restaurants

VI. MODELS

A. *Linear Regression*

Linear regression models predict a continuous target when there is a linear relationship between the target and one or more predictors [4]. A supervised machine learning method called linear regression models the association between a response variable and one or more explanatory variables. The linear predictor functions are used to model each and every linear relationship. The connection between the response variable and the independent variables is assumed to be linear in a linear regression model, in other words. It is primarily employed in jobs involving prediction and forecasting. Details on the various ways that linear regression is implemented can be found in the references.

There is no linear trend among the features as seen in Figure 8. With this model, I got an accuracy of 0.46 which means that this model is not the best fit for this dataset.

B. *Random Forest Regression*

Decision tree algorithms are efficient in eliminating columns that don't add value in predicting the output. In some cases, we are even able to see how a prediction was derived by backtracking the tree. However, this algorithm doesn't perform individually when the trees are huge and hard to interpret. Such models are often referred to as weak models. The model performance is improvised by taking an average of several such decision trees derived from the subsets of the training data. This approach is called random forest classification [4].

Up until this point, all our analyses had merely highlighted the non-linearity of the data, so it was critical to utilize a non-linear model to make predictions. Decision trees are frequently used to describe nonlinear relationships between training observations and provide the necessary response predictions. Decision trees can end up overfitting the data with a very complex structure. They tend to perform quite well on training data but usually fail in accuracy and efficiency during validation. Ideally, our goal should be to minimize error due to bias and variance. This is where Random forests come to the rescue. A random forest is a collection of decision trees whose results are aggregated into one final outcome. Random Forests reduce variance by training on different samples of the data by choosing subsets of decision trees and also leverage the possibility of choosing subsets of features to train a set of sub trees in comparison to others in the entire model. This random choice of features helps in reducing correlation between different base trees, hence, limiting the error due to bias or variance. This very reason in accordance with the non-linearity of data was the biggest motivation to choose Random Forest regressor model for the original dataset and then perform the predictions to evaluate the results. This model gave an r2-score accuracy of 0.90 which is quite impressive.

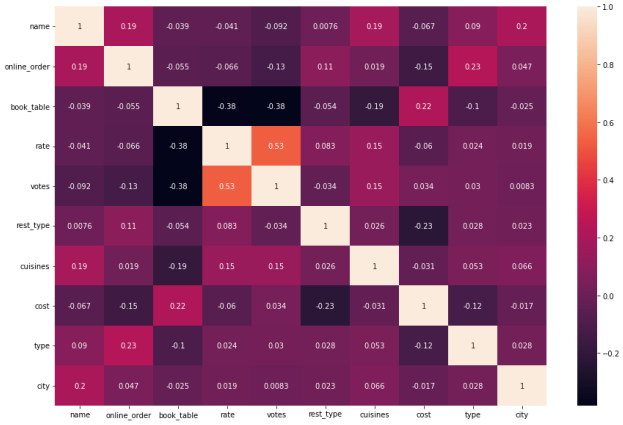

Figure 8: Correlation graph

C. *Decision Tree Regression*

A decision tree-based model builds a set of rules from the training data to be able to predict the outcome. For the sake of understanding, this algorithm is compared to trees formed through decisions. The model contains branches that represent the rules that lead to the path of the outcome, that is, the leaf. Each prediction path leads to a leaf that contains multiple values. The same principle is applied to classification-type problems as well. For regression-type problems, the final prediction is usually the average of all of the values contained in the leaf it falls under [4]. This model gave an accuracy of 0.87 which is the lowest after linear regression.

D. *Gradient Boosting and XGB*

Gradient boosted trees are also a type of ensemble learning. They are based on the method called boosting, which involves training a model one after another based up on the outputs from the previous models. In gradient boosted trees, we calculate the error from the previous model, also known as residuals. Now we define another model that is trained on this residual. The resulting model is the sum of previous model and the model trained on residuals. This process is repeated until convergence. Even though gradient boosted trees out perform random forest models, they are computationally expensive because they are built sequentially. A specific implementation called XGBoost is used to overcome this issue [4].

While using gradient boosting and XGB, I got an accuracy of 0.91 and 0.93 respectively. XGB performs well on the dataset than gradient boosting.

E. *PCA*

Principal component analysis, or PCA, is a statistical technique to convert high dimensional data to low dimensional data by selecting the most important features that capture maximum information about the dataset. The features are selected on the basis of variance that they cause in the output. The feature that causes highest variance is the first principal component. The feature that is responsible for second highest variance is considered the second principal component, and so on. It is important to mention that principal components do not have any correlation with each other. It is imperative to mention that a feature set must be normalized before applying PCA [5].

PCA was employed in this study to condense the dataset into a smaller collection of representative variables that together account for the majority of the variability in the original dataset. Here, the goal was to determine whether one characteristic in particular outperformed others in the dataset in describing the data. In such a circumstance, we can use the feature that almost precisely characterizes the data (almost because, often, one feature does not completely describe the outcome in real-world scenarios) and avoid the aspects that are least important to the outcome.

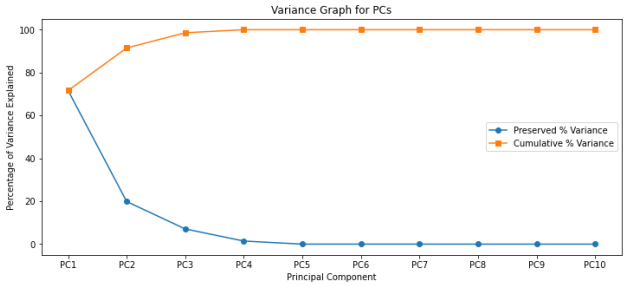

Figure 9: PCA Variance Graph

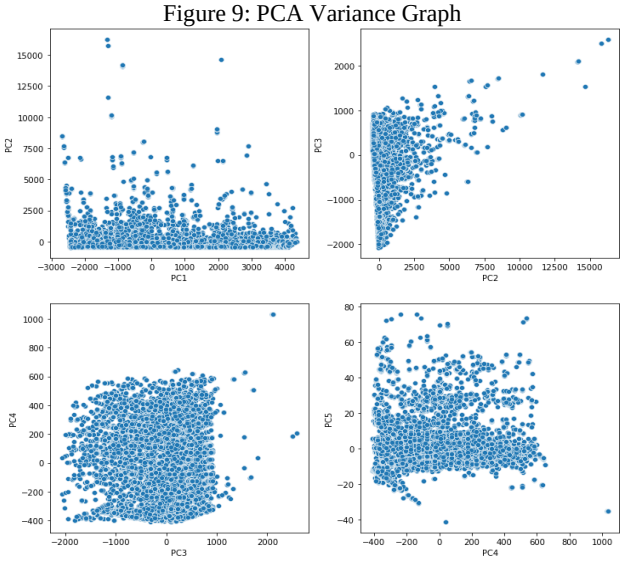

Figure 10: PCA Observations

From Figure 9, we can analyze that the first PCA was able to explain 72% of the variance in the data. This means some of the individual features are highly correlated. *Figure 10* shows the observations spread in 2-d plane using four different PCs. However, just using PCA is not enough. So, I applied Random Forest Regressor on all 10 principal components if there is any improvement in predicting the outcome. By applying this technique, I got an r2-score accuracy of 0.99 which is the highest out of all the models.

All the models and their r2-score can be seen in *Table 1* below.

Table 1

Models	R2 Score
Linear Regression	0.46
Random Forest	0.90
Decision Tree	0.87
Gradient Boosting	0.91
XGB	0.93
Random Forest on PC	0.99

VII. K-FOLD CROSS VALIDATION

Datasets are typically split into two halves, the training set and the test set, in machine learning and data mining. This method enables us to assess the model's performance on unobserved data by training it on a portion of the available samples and then testing it on a completely different subset from the same sample space. Different training and test subgroup combinations yield different outcomes and hence produce varying levels of accuracy. This is the precise problem that arises from bias or variance in the data selections used for training and testing. Cross validation is the most honorable method of dataset validation. When a dataset is divided into many folds and train and test evaluations are iteratively applied to each fold, cross validation is used to evaluate and validate the dataset. One such method is k-fold cross validation, which divides the dataset into k folds, keeps one subset for testing, and conducts training on the remaining k-1 subsets. Up until every subset has had an opportunity to become a test set at least once, this process is repeated. The performance of the random forest regressor was assessed using K-fold CV. The model's average accuracy was 0.97 after using 12-fold CV. The cross-validation on random forest regressor on PCs is shown in *Figure 11*.

Figure 11: Cross-Validation

VIII. CONCLUSION

Based on the small collection of features that were selected for the model to predict the success of a startup restaurant, Random Forest on Principal Component Analysis is the best appropriate model among all the techniques used for the prediction of the target variable. As long as the set of attributes stays the same, the model's architecture can be extended for use across other datasets from various cities throughout India. The outcomes of this model will help potential restaurant owners and other interested parties make well-informed decisions before starting a chain of restaurants in or around Bangalore. The goal of future study will be to provide the groundwork for a scalable model that can generalize parameters to successfully operate across many cities and produce positive outcomes with larger data sets.

IX. REFERENCES

[1] Mhlanga, O. (2018),” Factors impacting restaurant efficiency: a data envelopment analysis”, Tourism Review, Vol. 73 No. 1, pp. 82-93.

[2] Consumer values among restaurant customers, International Journal of Hospitality Management, Volume 26, Issue 3,2007,Pages 603-622,ISSN 0278-4319, https://doi.org/10.1016/j.ijhm.2006.05 004.

[3] Karen Glanz, Ken Resnicow, Jennifer Seymour, Kathy Hoy, Hayden Stewart, Mark Lyons, Jeanne Goldberg, How Major Restaurant Chains Plan Their Menus: The Role of Profit, Demand, and Health, American Journal of Preventive Medicine, Volume 32, Issue 5, 2007, Pages 383-388, ISSN 0749-3797.

[4] *Regression Algorithms*. (n.d.). Retrieved from https://developer.ibm.com/tutorials/learn-regression-algorithms-using-python-and-scikit-learn/

[5] *Principal Component Analysis*. (n.d.). Retrieved from https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/