Analysing Success Factors and Predicting Ratings for Upcoming Restaurants

[1]Abhijith S., [2]Animesh Khare, [3]Aniruddha Krishna Jha

[1, 2, 3]Department of Computer Science Engineering

PES University, Bangalore, India

[1]abhijiths.bang@gmail.com, [2]animesh1.khare1@gmail.com, [3]aniruddhakj100@gmail.com

***Abstract*—Predictive models are often used by investors to decide whether a budding business would be profitable in their domain. The success of a new restaurant can be similarly predicted based on past data relating to the location and services offered. We develop a predictor model to find out the popularity and success a new restaurant can generate with an extensive study. This study aims to offer huge insights on which factors determine the success of a new restaurant.**

***Keywords*—Restaurant, Factors of Success, Predictor**

I. INTRODUCTION

Greek philosopher Heraclitus once famously said that "change is the only constant in life." This statement holds true in regards to our lifestyle and eating habits as well. With the boom of delivery services and takeouts in India, businesses like Zomato, Ola Eats, Swiggy have popped off. E-commerce as a whole has become the backbone for a lot of industries and the restaurants and even small shops feast off of it. Bengaluru, being an IT hub in India, invites a lot of diverse cultures to its abode. Thus, it has become a key destination for foodies, with all the variety available. Everyday new restaurants and dhabas pop-up in the city, thereby matching the increasing demand with increasing supply. However, the established restaurant chains find it much easier than their newly established counterparts in terms of competition. There are a lot of factors that contribute to this, the location, the kind of style a restaurant goes with, the popularity associated with the chain, the initial ratings etc. In this study, we will be talking about a few of those factors.

II. LITERATURE REVIEW

Prior studies relating to this domain involve determining factors that affect the rating of a restaurant, how restaurant types and location affects the profit margin from a business standpoint, what are the most popular and most favored restaurant chains in particular cities and building recommender systems for suggesting good and better restaurants adjusted according to a user's preference. A lot of studies have been done in this domain and regard, but most of them actually aim to proceed towards some sort of recommender system for the customers. In this regard, we look at a few studies carried out in different regions and draw insights from them

A. ARIZONA - RATINGS

A 2015 study by Sahoo & Saunil[1] looks into finding the ideal conditions and practices needed by up and coming restaurants to minimise competition and maximise initial gains. The purpose of this research was to determine the factors that affect the rating of a restaurant, to find out what the successful businesses in the restaurant industry do differently that leads to higher customer satisfaction, translating to higher ratings and revenues. The data available on Yelp for the state of Arizona, USA was utilised to gather insights on the factors that lead to positive customer sentiments as reflected by their reviews and to train a model that can predict whether a restaurant will be successful. primarily conducted 2 different analyses.

A multiple linear regression analysis was carried out to find the impact of different types of

independent variables on the rating, the dependent variable, with the assumption that the data had negligible multicollinearity, possessed multivariate normality and was homoscedastic.

Furthermore they performed a Time Series Analysis on text-based user reviews to see whether the general sentiment, which is identified by an R shiny app through the analysis of patterns in reviews, varied depending on the time of the year. The key findings of these two analyses were that Fine Dining restaurants with a relaxed ambience have the highest chance of being successful and that festive seasons garner lower sentiments due to the increase in prices.

B. KOREA

Jeong-Hee, Myoung Soo, Kyu-Suk in 2018[2] have suggested and developed research hypotheses for restaurant industries in South Korea. These hypotheses aim to find out moderating effects of restaurant types and locations, expected outcome being that the research would help start-up restaurant managers get useful insights to better manage their businesses. A research model was built and evaluated on IBM's SPSS 18.0 through multiple regressions.

The key findings state that for Korean style restaurants, ops management and customer relations management were positively associated with financial performance. Also, financial performance was negatively affected by external environments which included government regulations, labor shortages, high taxes, credit card fees, and high initial investments, etc. Management of employees turns out to be a crucial factor for success.

Customer Relationship Management (CRM) is one of the main positive factors affecting business performance. In addition, ops management strongly affects business performance.

C. BANGALORE & CHENNAI

A 2020 research[3] carried out in a similar domain but different context provided new intuitions and insights, The motive behind this research was to find out the most popular restaurants in major cities, Bangalore and Chennai in this case, determine the popular cuisines savoured by people and use data science to analyze datasets and thus assisting people in finding out where they can find their preferred cuisine, and also providing them with the maximum number of choices possible to serve their purpose. Here the authors have made their analysis on Zomato, a popular online food delivery app. An effort is made by the authors to analyze Zomato's food delivery process.

Python and its packages were used to complete and implement their research model namely Numpy, Pandas, Matplotlib, Scikit-learn and Seaborn. They have also implemented various algorithms like Linear regression, Logistic regression, Decision tree regression and Random forest regression to enhance their research. For example, they have used Linear regression to predict the number of restaurants based on the city of choice and so on.

Some of the conclusions of their research suggested that the most popular cuisines of Bangalore were North Indian,Chinese and South Indian dishes. Then, the most popular restaurants in Chennai were found to be Shri Krishna Bhavan, Hotel Pandian, and Sultan Biryani. Also, more importantly it was found that the accuracy of Bangalore dataset was 84% whereas Chennai's dataset was approximately 99% and that orders placed online was 62% and offline order was 29%. Thus from this research, the most popular cuisines, restaurants in Bangalore & Chennai were inferred and also a key finding suggested that online orders are by far outnumbering offline orders, which suggests that online food orders are becoming more of a commonplace in our society.

III. METHODOLOGY

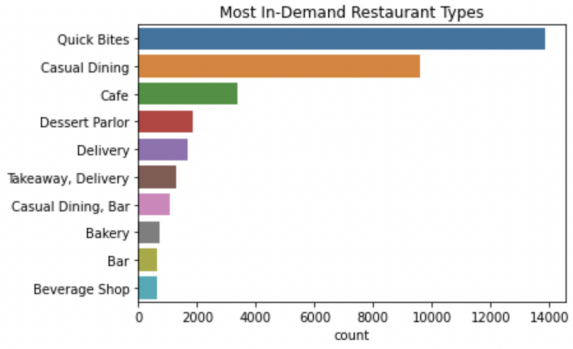
One of the most common approaches when analyzing user reviews in the current world is by looking for patterns in the textual reviews that users post. However, it is our belief that those studies lead to biased results due to their assumption that users post both good and bad reviews with equal dedication and effort [4] . Just on general observation, it can be noticed that people tend to be more vocal about the negatives of any experience than about the positives on any reviewing platform. The exact opposite is observed however when those platforms are linked to, or are themselves, a social media based platform

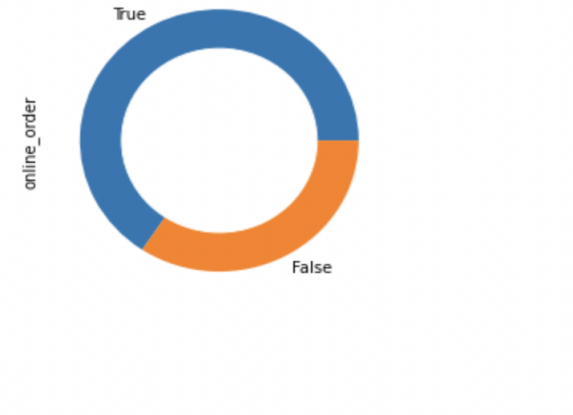
as people then tend to put an extra effort into painting a more positive picture of their lives. But again, these businesses might get subjected to review bombing, where even bots are created just to give out negative reviews. To ensure such a bias does not creep into our study, we will be focusing on facilities provided by the restaurants, their prices, their location and certain other variables that can be verified as facts to see their impact on the average numerical rating. We will also be more focused towards how the numbers and numerical aspects direct the ratings of restaurants or restaurant chains.

A. PREPROCESSING

The dataset that we will be working on for this study and analysis can be found at kaggle here: https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants.

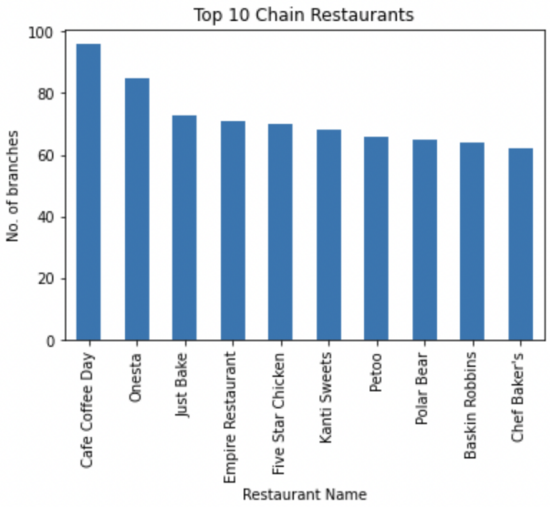
After a closer inspection on the data, we found that it is quite dirty. We have dropped the NaN values from the dataset to work with cleaner values, but before we take a grasp on what kind of predictor model should be trained, we will need to know our dataset in and out. We need to find what factors actually influence the success of a restaurant.



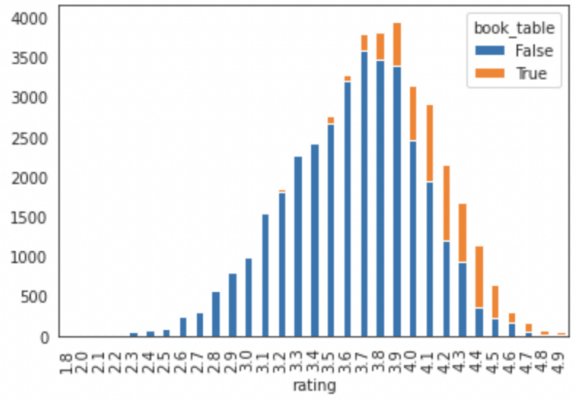


Proportion of restaurants that allow booking tables

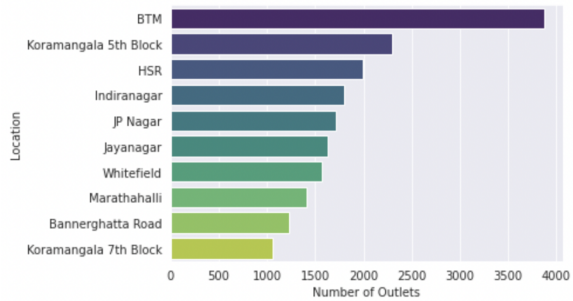
These plots tell us which restaurant types are in the most demand, we can infer that people prefer quick bites, over any other kind of restaurant styles. Also, most of the restaurants now allow online orders, especially with the rise of food delivery services.



These are the most heavily expanded restaurant/cafe chains in Bangalore



This plot gives a great example of how ratings are distributed across restaurants which allow booking tables in advance. It can be observed that restaurants that allow table bookings tend to have higher ratings.



This plot here shows how the restaurants and/or cafes are based/located across the entire city. BTM layout turns out to be the most popular location, with close to 4000 restaurant outlets. Koramangala 7th Block however, turns out to be the least popular despite having over a 1000 outlets.

B. MODELS

Now comes the important part of building the models, the accuracy measure that is chosen is also very important. We settled on the R-squared score for our accuracy measure as R-squared is often more informative than other measures like MSE, MAPE in regression analysis[5].

1. Linear Regression

Linear Regression is the most basic and common type of model used for predictive analysis. The overall idea of linear regression is to examine if a set of predictor variables does a good job in predicting an outcome variable and in what way do they impact the outcome variable. Given that the variables in our dataset were not highly correlated, Linear Regression was a good starting point for our analysis.

2. Decision Tree Regressor

In this model, Decision trees are used to build regression models in the structure of a tree. Basically, it breaks down a dataset into smaller subsets while at the same time, an associated decision tree is incrementally developed. It's quite popularly used as it can handle numerical as well as categorical data. We wanted to build a model which could perform better than a classical regression model, since our data was found to be quite non-linear.

3. Random Forest Regressor

The relative success of Tree-based regressors motivated us into examining the Random Forest Regressor in the hope that an ensemble of trees can give better results[6]. A random forest is a meta estimator that fits a number of decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

4. Extra Tree Regressor

Extra Trees is an ensemble algorithm that combines the predictions from many decision trees. It is related to the widely used random forest algorithm. We used this model since it can often achieve equal or better performance than the random forest algorithm, although it uses a simpler algorithm to construct the decision trees used as members of the ensemble.

5. Gradient Boosting Regressor

On observing the improvement in accuracy of the bagging ensemble models above, we decided to look into boosting algorithms. Gradient Boosting Regressor builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

6. XGBoost Regressor

Extreme Gradient Boosting (XGBoost)[7] is a tree based ensemble machine learning algorithm which has higher predicting power and performance and it is achieved by improvisation on the Gradient Boosting framework by introducing some accurate approximation algorithms. XGBoost expects to have the base learners which are uniformly bad at the remainder so that when all the predictions are combined, bad predictions cancel out and the better one sums up to form the final good predictions. Our search for a well-performing model that was not prone to overfitting ended here due to XGBoost's high performance and reliability.

IV. RESULTS

Compiled Table with R-squared scores of each model

	Model	Score
5	XgBoost	0.941137
3	Extra Tree Regressor	0.931554
2	Random Forest	0.929950
1	Decision Tree	0.888937
4	Gradient Boost	0.514598
0	Linear Regression	0.283244

Our results after training the Model showed that the XGBoost Regressor worked the best with our data on a 4:1 train:test split. This model fits around 94% of the dataset perfectly.

This takes us to the final part of our study, i.e., predicting the ratings of new unregistered restaurants given their location, cuisine style, restaurant type etc. We created our own dataset for this purpose, adding newer restaurants which were not in the previous dataset and tried to predict the ratings for the same.

name	online_order	book_table	votes	location	rest_type	cuisines	cost	type
Madras Meal Company	True	False	400	Banashankari	Casual Dining, Takeaway, Delivery	South Indian, Chettinad	400.0	Buffet
Dwarkamai	True	False	291	Banashankari	Casual Dining	South Indian, Chinese, Street Food	690.0	Buffet
Great Indian Khichdi	True	True	1279	J P Nagar	Casual Dining, Delivery	North Indian, Healthy Food	420.0	Buffet
Shiv Sagar Signature	False	False	705	J P Nagar	Casual Dining	South Indian, Fast Food, Chinese	239.0	Buffet
Vesuvio	False	True	200	Bannerghatta Road	Fine Dining	Italian	1000.0	Dine-out

Five more similar entries were added and the ratings were predicted using the model. The model gave very good and realistic results, which are captured below:

```
[4.0, 3.7, 4.2, 4.0, 4.0, 3.2, 3.5, 3.9, 3.8, 4.2]
```

These ratings are pretty realistic for the data provided, and also pretty accurate for almost all of the restaurants provided in the dataset.

V. CONCLUSION

Various models were tried and tested in our project in our quest to find the best performing model, out of which, XGBoost Regressor was the one which performed the best for our predictor.

Thus, this predictor can be helpful for up and coming restaurants to analyze their chances of success based on various factors. It can also predict the rating they may/will receive depending on their ideas for restaurant style, cuisine, location etc.

VI. REFERENCES

[1] PP Sahoo, MS Saunil, R Chandra, L Xiang, "Analyzing success in the Restaurant Industry", *wiki.smu.edu.sg*, 2015

[2] Jeong-Hee Hwang, Kyu-Suk Chung, Myoung-Soo Kim, "A Study on the Success Factors for the Restaurant Service Industry: Moderating Effects of Restaurant Types and Locations", *Asia-Pacific Journal of Business*, Vol. 9, No. 3,pp. 11-24 , Sept. 2018, doi: 10.32599/apjb.9.3.201809.11

[3] Dr. D. Singh, S. Srivastav, S. Shree, S. Shrivastava, S. Sharma, "Analysis of Online Food Delivery Process by Zomato using Data Science", *International Journal for Modern Trends in Science and Technology*, 6(8S),pp. 125-127, 2020, doi: 10.46501/IJMTSTCIET24

[4] Amin Mahmoudi, "Identifying Biased Users in Online Social Networks to Enhance the Accuracy of Sentiment Analysis: A Behaviour-Based Approach",arXiv, 2021, arXiv: 2105.05950

[5] Chicco D, Warrens MJ, Jurman G. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation", *PeerJ Computer Science*, 7:e623, 2021, doi:10.7717/peerj-cs.623

[6] Ulrike Grömping, "Variable Importance Assessment in Regression: Linear Regression versus Random Forest", *The American Statistician*, Vol. 63, No. 4, pp. 308-319, 2009, doi:10.1198/tast.2009.08199

[7] Krishnapuram B., Shah M., " XGBoost: A Scalable Tree Boosting System", *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, August 2016, doi: 10.1145/2939672.2939785

VII. CONTRIBUTIONS

Data Preprocessing and Cleaning: Abhijith S., Aniruddha Krishna Jha

EDA and Visualisation: Animesh Khare, Abhijith S.

Building Models : Aniruddha Krishna Jha, Animesh Khare

For more details on the study, you can visit the following links:

Kaggle Notebook with the complete dataset: https://www.kaggle.com/aniruddhakj/zomato-success-factors-rating-predictions

Github repository with the analysis and review: https://github.com/aniruddhakj/PredictorModelforRestaurantsBengaluru