VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

Submitted towards lab component of the course

Data Mining Technique

ITE2006

Under the guidance of

Dr. Ranichandra C

Associate Professor, SITE

Submitted by

Bhaumik Tandan (19BIT0292)

Dhaval Mavani (19BIT0316)

Parinita Parihar (19BIT0287)

Review 1

Abstract

The project aim is to predict the restaurant rating based on the various factors. Online customer feedbacks are taken as an integral part for making a decision before purchasing any product. Present and future generation is the journey of e-commerce platform is booming in world.

Sharing on the internet is something we usually do. Giving a review is a useful activity through which other people on the internet can find out something else and see opinions about things. The usual things reviewed by someone in the form of experiences, places, objects, and others. Give a review we usually use text to explain something that we experience with an item, place, or event that we normally experience. Customer satisfaction is an opinion between expectation and reality obtained by consumers. Giving a review is also a useful activity so that other customer on the internet can find out something else and see opinions about things and its satisfaction.

Commonly, most people express their opinion through social media on the review platform like Zomato, Yelp, etc. Customer reviews on online media like Zomato become important as it might increase the popularity of something. Zomato is a site where someone can give a review of a restaurant, how the restaurant is and someone's opinion about the restaurant. Restaurant customer satisfaction can be analysed by their review on Zomato. Sometimes, restaurants see the reviews in Zomato, but they didn’t get if the reviews are positive or negative to their restaurants. Review on Zomato is still in the form of text and can be classified with positive, negative, or neutral with their ratings.

The food chain industry is a very competitive one and lack of research and knowledge about the competition usually leads to the failure of many such enterprises. The principal issues that continue to produce difficulties to them include high real estate expenses, escalating food costs, fragmented supply chain, over-licensing, and even after that restaurateur does not know whether the business will develop or not. This project aims to solve this problem by analysing ratings, reviews, cuisines, restaurant type, demand, online ordering service, table booking, availability of the restaurant and make the machine learning model learn these and predict ratings of new restaurant and how positive and negative reviews should be expected. This project considers the data of the restaurant from all over India from Zomato as an example for showing how our model works and can help a restaurateur choose the location and cuisine which will give it better ratings, reviews, and make the business more profitable.

LITERATURE SURVEY

Research Paper Title: Predict User Ratings based on Review Texts

Date of Publishing: December, 2016

Survey: A new trend has emerged with emergence of e-commerce website: feedback and reviews, providing platform to market analysis. This project makes use of such review texts in order to perform an inference-based analysis and thereby, train the model to be able to predict the ratings based on these texts. The objective of the project is to implement a model that predicts user ratings based on the review texts. The obtained data-set is in the JSON format containing feature vectors and, KNN algorithms are used. The model has predicted in some cases the exact star rating associated with the review rating online. As the goal of the project is to predict user ratings based on review texts, the prediction includes the review class identification to fall into three broad categories as mentioned earlier. They are, bad (1 star), average (3 stars) and good (5 stars).

Research Paper Title: Uncovering Business Opportunities from Yelp

Date of Publishing-2016

Review: The authors created a model that would predict the success of a business based on business type and location. From Yelp, aspects such as ratings and check-ins were utilised. Data mining processes and Random Forest machine learning algorithm was utilised in creating a model. A user interface was successfully developed to identify business opportunities on a map, but it was difficult to draw accuracy from the Yelp data. Still, further iterations, using various modelling methods, drew better results according to the visualisations shown in the paper.

Research Paper Title: Predicting Restaurants' Rating And Popularity Based On Yelp Dataset

Date of Publishing-January,2016

Review: This paper analyses the Yelp dataset and reveals the relation between restaurants' quality and services and how it helps to attract more customers. This project works on predicting ratings and popularity change of restaurants based on features such as location, opening hours, price level, food type, service provided along with review text, time and rating. The model makes predictions on both the Yelp rating and popularity change. The paper deals with the problem of overfitting with "sequentialfs" function in MATLAB. Overall, the paper chooses 42 features for rating predictions and 28 features for popularity change predictions. The paper uses several machine learning methods including logistic regression, Naive Bayes, Neural Network, and Support Vector Machine (SVM) to make relevant predictions. While logistic regression performs better than the others in terms of performance metrics such as F1 score and accuracy the paper does state it requires improvement of data and methodologies.

Research Paper: Reviews, Reputation, and Revenue: The Case of Yelp.com

Date Of Publishing-2016

Review: The author attempted to identify a significant connection between Yelp reviews and restaurant revenue. Ultimately, the author identified a positive correlation between Yelp reviews and revenue of independent restaurants. The author also noted a specific lack of correlation between Yelp reviews and the revenue of chain restaurants.The model used KNN model to try and evaluate and predict the reviews.

Research Paper Title: Restaurant Revenue Prediction using Machine Learning

Date of Publishing-June,2016

Review: This paper explores how opening a new restaurant outlets is subjective in nature based on personal judgement and how to predict across geographies and cultures. The model employs a supervised learning algorithm will construct complex features using simple features such as opening date for a restaurant, city that the restaurant is in, type of the restaurant (Food Court, Inline, Drive Thru, Mobile), Demographic data (population in any given area, age and gender distribution, development scales), Real estate data (front facade of the location, car park availability), and points of interest including schools, banks. The paper applies concepts of machine learning such as support vector machines and random forest on these parameters, it predicts the annual revenue of a new restaurant which would help food chains to determine the feasibility of a new outlet. The dataset used in this paper is obtained from Kaggle and consists of a training and test set with 137 and 100,000 samples respectively. The paper experimented with SVM and Random forest and compared their accuracy in terms of estimation.

Research Paper Title: Rating Prediction with Sentiments and Topic Models

Date of Publishing: May, 2017

Review: The aim of the research is to determine rating inference rather than just labelling it as positive or negative on the basis of sentiment-words in the textual reviews. First, the topic-model is trained using LDA. Then the trained topic model is used alone and then, also used by combining with baseline sentiment model for rating prediction. The results show that there is inconsistency at assigning ratings among authors, the cross-author divergence. And also, ratings not entirely supported by the text. Adding personalization data might be helpful in performance improvement.

Research Paper Title: Service Rating Prediction and Recommender System-A Survey

Date of Publishing: July, 2017

Review: Due to increasing access to mobile phones in India, amounting to 300 million(approx.) users has resulted into increase online opinions, reviews and suggestions. We need this huge data to be recommended to users with the help of geographical location and social network. We use location-based services (LBS), recommended system (RS) and finally, location-based service recommended system (LBSRS). Additionally, the personalisation and personal interest factors are identified with the user latent features. This combines the individual preference and social interaction to rank the services. As a result, this project provides a review on problems like data-overloading, appropriate service recommendation with help of LBS.

Research Paper Title: Data Mining for Predicting Customer Satisfaction in Fast-Food Restaurants

Date of Publishing: May, 2017

Review: A fast-food restaurant in Indonesia has been chosen as case study. Multiple data mining techniques are used. Some of the algorithms used are: regression analysis, co-joint analysis, descriptive statistic, correlation analysis, factor analysis and structural equation modelling. This study provides some findings that are useful for the marketers, policy makers who have interest in customer satisfaction research. For practical implication, an understanding of underlying determinant factors of customer satisfaction could strengthen the competitive advantage of the fast-food restaurant.

Research Paper Title: Restaurant Customer Rating Prediction

Date of Publishing-2017

Review: The paper explores how the restaurant rating is based on many attributes like the food quality, prize, ambience of the restaurant, if the restaurant has online delivery system, if the restaurant has table booking etc. It analyses how these factors will affect the ratings as customers will consider these factors to dine in their favourite restaurant. Ultimately, the Customer Relationship Management aspect is explored to figure out its relation with the ratings. The paper uses a Zomato Dataset which has the features such as Restaurant Name, Restaurant ID, City, Address, Cuisines, Cost for two people, Has table booking, Has online delivery, Is delivering now, Switch to order menu, Prize range, Aggregate rating, Rating colour, Rating text and votes. The paper explores sentiment analysis on the reviews and applies this on a Support Vector machine model and proceeds to interpret the results which show decent accuracy in estimation of ratings on the testing dataset.

Research Paper Title: Predicting Yelp Food Establishment Ratings Based on Business Attributes

Date of Publishing-2018

Review: The authors aimed to utilise script analysis to predict helpfulness rating of online reviews. Human annotators were asked to highlight important phrases of reviews along with that phrases were added to a script lexicon. The paper used a Yelp dataset to identify a correlation between rating and the social factors of a user's physical location. They worked to identify a relationship between rating and the distance a user is from their home; they also investigate the connection between the ratings of similar items by users that are friends. They initially inferred that users 5 tend to rate items higher the further away they are from their home, or "activity centre". Text regression was performed to predict review helpfulness based on words in the lexicon. Compared against a Baseline model and a bag-of-words model, the scripts enriched model was found to predict helpful reviews quicker and more accurately.

Research Paper Title: Data mining of restaurant review

Data of Publishing-2018

Review: The research paper explores how features are extracted from dataset and classification model is created. This paper is an attempt to create a trigger model to improve restaurants based on Zomato dataset. The paper uses the Zomato dataset and works on Naïve Bayes classification algorithm to create a model after following the various pre-processing steps. The paper analyses the relation between different columns in the dataset to gain relevant features and then apply on the algorithm.

Research Paper Title:Predicting the Growth of Restaurants Using Web Data

Date of Publishing: May, 2018

Review:With the rise of data mining techniques in risk and growth field, researchers have also turned their focus towards identifying the reasons of failure of restaurants. Many factors such as competition, business network, characteristics of entrepreneurs etc... are considered in this research. Use of random forest classifier (RFC), multi-layer perception (MLP), neural networks, and also logistic regression (LR) is stated. Therefore, RFCs, MLPs and LRs are tested and challenged with the goal to estimate a binary outcome, i.e. non growing v/s growing restaurants. The results suggest, that RFCs with an average accuracy of 68% outperform MLPs and LRs.

Research Paper Title:Review Rating Prediction Using Yelp Dataset
Date of Publishing: February, 2019

Review:Yelp is a review website that allows users to rate various business categories. Since there are hundreds and millions of reviews, we need to prevent customers from reading a random review and creating bias decision by using ML algorithms. To train our prediction model, linear classification algorithms such as logistic regression, support vector classification and naïve Bayes multi-class classification algorithm are used. The result show that Naïve Bayes' and SVMs provide output with better accuracy. The output are received as form of star rating with 94% when Naïve Bayes' is used.

Research Paper: Predictive Analytics Using Text Classification for Restaurant Ratings

Date Of Publishing-2018

Review: The paper attempts to analyse the Yelp dataset to create a model which can provide ratings on the restaurants. The paper proposed a system for predictive analysis consists of language features and embedded classification models. The paper explored Natural language processing and how words can be used as features. It explored the subsequent techniques of converting words into features using N-grams and performing sentiment analysis on these features. The paper tried using Naïve Bayes (NB), recurrent neural network (RNN) and Support Vector Machine (SVM). The results showed that the multinomial NB model has better results, indicating that the dataset tends to have multinomial distribution. Hence with fine tuning of parameters an accurate model for detecting ratings can be done.

Research Paper Title:Restaurants Rating Prediction using Machine Learning Algorithms

Date of Publishing: September, 2019

Review:The rating is the most important feature of any restaurant as it is the first parameter that people look into while searching for a place to eat. In this model, various restaurants records with features like the name, average cost, locality, whether it accepts online order, can we book a table, type of restaurant are considered. It portrays the quality, hygiene and the environment of the place. Some of the algorithm used here is SVMs, KNN, Decision Tree, Random Forest. The result showed that the algorithm ADA Boost (DT) predicted the result with maximum accuracy amounting to 83%. [1]. This paper discusses the feasibility of combined scheme involving Discrete Wavelet Transform (DWT) and Singular value Decomposition (SVD) for content authentication in video media content. This paper involves use of two watermarks one being the original watermark selected for embedding

Existing Systems

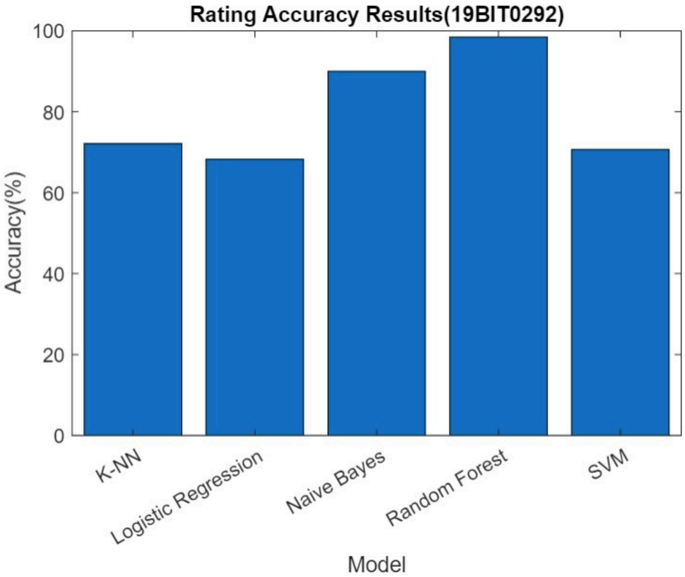
Due to the rich information contained in the Yelp dataset, many past research and projects tried to use it to predict ratings of restaurants and to evaluate the future development. For example, Kong, Nguyen and Xu in the paper Predicting International Restaurant Success with Yelp, Stanford University classified restaurants based on cultural categories and analyzed international restaurants success mostly with Gaussian Discriminant Analysis (GDA). Several other previous papers focused on the sentiment analysis with text content from Yelp. A few papers combined the customer reviews and ratings together to conduct sentiment analysis, while others mainly used matrix factorization to analyze text information and predict Yelp ratings. Other than Yelp reviews, Tang, Qin, Liu and Yang in their paper User Modeling with Neural Network for Review Rating Prediction. IJCAI. 2015 introduced neural network to predict restaurant reviews.

Gap Identified

The paper claimed that matrix-vector multiplication would be more effective than vector concatenation when considering text analysis. So far, most research works on text analysis of customer reviews, but leaves out other features in Yelp Dataset Challenge.

To have an overview of how different methods perform, we have summarized the accuracy for all the previous methods in Table 1. For the rating prediction, we can see that linear regression perform comparatively worse than the Native Bayes and Random Forest. Clearly from the bar plot we can see how the accuracy for Random Forest Algorithm is the highest achieved and is significantly higher than Naïve Bayes as well. One possible explanation is that the Random Forest method apply to a wider range of problems, and are more robust to problematic model specifications. For example, in the Naive Bayes model, we need to assume conditional independence among the independent variables, which clearly cannot be the case. Another possible explanation is that we don't have enough input features to make a decent prediction. A more complicated method might involve more noise and thus yield low accuracy when the information is limited.

Rating Prediction	Accuracy
Logistic Regression	68.283%
SVM	70.673%
K-NN	72.123%
Naïve Bayes	90.016%
Random Forest	98.451%



Made using MATLAB

Proposed Method

Zomato has changed the way people browse through restaurants. It has helped customers find good places with respect to their dining budget. Various researches and students have published related work in national and international research papers, thesis to understand the objective, types of algorithm they have used and various techniques for pre-processing and feature selection.

For the pre-processing stage we need to apply basic statistical functions as they might be needed for data cleaning purposes. Also, we need to know about Exploratory data analysis, Feature engineering through introduction of various functions using python modules, outlier analysis and removal, text cleaning and encoding.

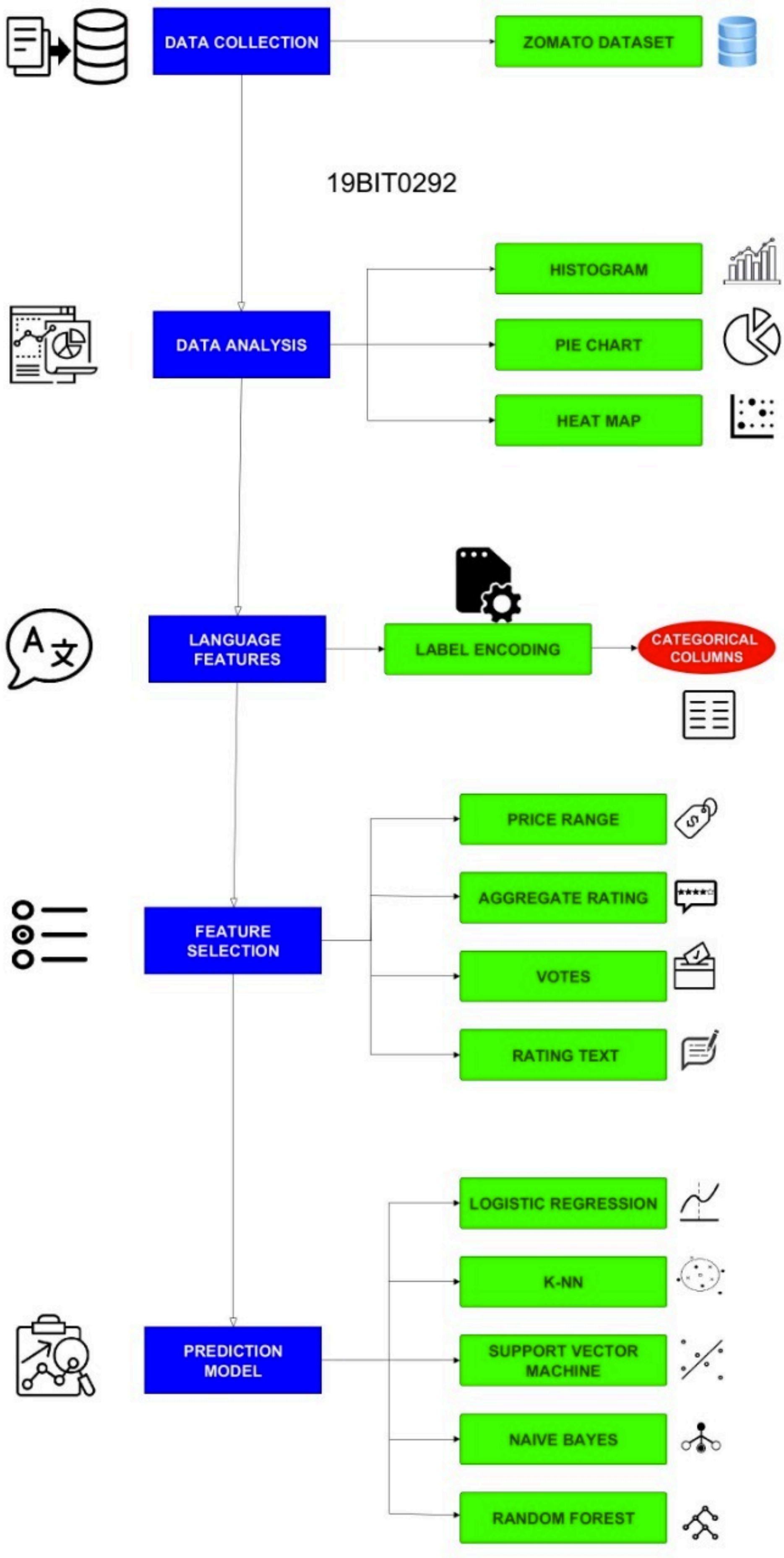
User reviews are an integral part of web services like TripAdvisor, Amazon, Zomato and Yelp, where users can post their opinions about restaurants and services through reviews consisting of free-form text and a numeric star rating, usually out of 5. These online reviews function as the 'online word-of-mouth' and a criterion for consumers to choose between similar restaurants. Studies show that they have a significant impact. The process of predicting a relationship for a generic user (but for a specific product/business) is called Review Rating Prediction. We aim to create a model to accurately predict the ratings of the restaurants based of this dataset.

The proposed system for predictive analysis consists of language features and embedded classification models. The dataset (Zomato Dataset) is split into test and training data. The first step is to generate the ground truth, and is performed based on the training reviews and ratings. We perform Exploratory data Analysis and gain an insight into the contents of dataset and their relation to each other through various graphs and charts. The next step in the model was to perform Data Cleaning, Feature Selection and Feature Engineering. In this step we also deal with missing data. In the process of building models, we extract:

(a) Statistical features of users and reviews

(b) Language features extracted from the review corpus. The next step was to test and evaluate the various classifiers which were used. We also decided to test and check for overfitting and find out the trend of the training vs true error to pick the best model.

Flowchart



Made using yEd Graph Editor

The dataset used in the project gives a fair idea about the factors affecting the establishment of different types of restaurant at different places in country aggregating the rating of eac h restaurant. The Dataset contains all details of the restraint listed on the Zomato website as of March 2019. The dataset contains 9554 rows. In order to build the final model, four phases were followed (1) exploratory data analysis, (2) data cleaning, (3) feature engineering and (4) model creation.

(1) The data analysis and exploration of the dataset consisted of Displaying the ratings of restaurants by country code and sorting them according to that criteria. We explored the data further by calculating the percentage of restaurants that were not rated and compared these restaurants in terms of percentage to whole of the dataset. From this analysis we explored and analysed that lower priced food was often not rated. Therefore, by similar case imputation, we replaced the 'not rated' with the average rating of lower priced food. We also noted that lower priced food is considered average for the most part, while more expensive food is rated more highly. Expensive food seems rarer than low cost food. We further explored the data with the help of various charts and compared the relationships of the various attributes of the dataset.

(2) Data Cleaning step of the model consisted of dealing with the attributes not important for the prediction of the final output and hence these attributes were dropped. We analysed the dataset for ASCII variable and URL’s and other noise but found out that the dataset was clean enough to be used for prediction without much data cleaning.

(3) Feature Engineering and Feature selection: In this step we analysed all the restaurants that were not rated and were given a rating based on their price range. As an aggregate rating of 2.5-3.5 translates to a rating of "Average", the "Not Rated”, the text can be replaced with "Average" in the dependent variable. The next step in feature engineering was dealing with the 'Rating Text' as it Is given as a string. This needed to be encoded into numeric data so that the algorithm can perform calculations. This was done using the LabelEncoder tool from sklearn. We classified the rating text into the form where excellent = 1, very good = 4, good = 2, average = 0, and poor = 3.

(4) Having pre-processed the data, We now provided the data to different classifiers and see which one performs better in creating a model of classification for this data. We compared performance using the cross-validation. Python has the cross_val_score class from the

sklearn.model_selection library to perform cross validation. We utilized supervised learning techniques for classifying high-dimensional data in order to predict.

(a) Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature.

(b) K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition as a non-parametric technique.

(c) Support vector machines (SVM) represents examples as points in space for non- linear classification. Different kernels lead to differences in the performance based on the dataset.

(d) Naïve Bayes (NB) is a probabilistic model with maximum likelihood to classify categories. We use NB to compare the results of different distributions.

(e) Random forest (RF) is an ensemble technique that fits a number of decision tree classifiers on various sub-samples of the dataset. We have experimented with different number of decision trees as estimators.

DATASET DESCRIPTION & SAMPLE DATA

The dataset used in the project is the Zomato Dataset. The dataset gives a fair idea about the factors affecting the establishment of different types of restaurant at different places in country aggregating the rating of each restaurant. The Dataset contains all details of the restraint listed on the Zomato website as of March 2019. The dataset contains 9554 rows with the following features:

- Restaurant Id: Unique id of every restaurant across various cities of the world
- Restaurant Name: Name of the restaurant
- Country Code: Country in which restaurant is located
- City: City in which restaurant is located
- Address: Address of the restaurant
- Locality: Location in the city
- Locality Verbose: Detailed description of the locality
- Longitude: Longitude coordinate of the restaurant's location
- Latitude: Latitude coordinate of the restaurant's location
- Cuisines: Cuisines offered by the restaurant
- Average Cost for two: Cost for two people in different currencies
- Currency: Currency of the country
- Has Table booking: yes/no
- Has Online delivery: yes/ no
- Is delivering: yes/ no
- Switch to order menu: yes/no
- Price range: range of price of food

- Aggregate Rating: Average rating out of 5
- Rating color: depending upon the average rating color
- Rating text: text on the basis of rating of rating
- Votes: Number of ratings casted by people Sample of the dataset:

REFERENCES

(1) Uncovering Business Opportunities from Yelp, Gianni Passerini

(2) Predicting Restaurants' Rating And Popularity Based On Yelp Dataset, Yiwen Guo, ICME, Anran Lu, ICME, and Zeyu Wang

(3) Reviews, Reputation, and Revenue: The Case of Yelp.com, Michael Luca

(4) Restaurant Revenue Prediction using Machine Learning, Prof. Nataasha Raul, Yash Shah, Mehul Devganiya

(5) YELP RATING PREDICTION WITH SENTIMENT AND TOPIC MODELS, Ying Liang

(6) Service Rating Prediction and Recommender System-A Survey, T. Sivakumar , Sneha Prakash

(7) Data mining for predicting customer satisfaction in fast-food restaurant, Bayu Adhi Tama

(8) Predicting Yelp Food Establishment Ratings Based on Business Attributes,Peter Mark Shellenberger Jr

(9) PREDICTING THE GROWTH OF RESTAURANTS USING WEB DATA, Yiea-Funk Te, Daniel Müller, Sebastian Wyder, Dwian Pramono

(10) Predictive Analytics Using Text Classification for Restaurant Inspections, Zhu Wang, Booma Sowkarthiga Balasubramani, Isabel F. Cruz

(11) Restaurants Rating Prediction using Machine Learning Algorithm, Atharva Kulkarni, Divya Bhandari, Sachin Bhoite

(12)