Proposal on Data Mining and Machine learning – 1

Veeresh Shivabasappa Kumbi – **X20165749**

MSc in Data Analytics – Jan 2021, Group A

MOTIVATION:

The basic idea to choose the domain of food industry is to dive-in and understand the basic knowledge on how the food industry runs also to have a fair knowledge on the customer demand, likes and dislikes on the cuisines they would love to be served. I was always astonished by the food culture. Starting up a food chain is always a difficult job. The establishment needs lot of skills and hands-on experience also with the business running everyday would be a challenge and an opportunity to learn something new. It also summons a result on few new things we learn and things need to be kept in mind at the end of the day. Starting from Licensing, real estate, hiring adequate man power, cost of the food, compete with the existing business holders and supply chain are the major key issues that are encountered.

Restaurant are classified into many different ways. The primary factors are usually food and cuisines that are being offered. Food Industry always majorly concentrates on food hygiene and health of the consumer. Employees face issues such as low pay. Long hours of work, minimal benefits, stress, discrimination and poor working conditions.

In the present situation effect of global pandemic high importance is drawn towards prevention of community transmission. Health commissions have implemented outmost rules and regulation to avoid spread of airborne diseases. They also recommend reduced dining capacity, Face masks, adequate ventilation, and flexible leave policies to workers. The dataset on which the test is been carried out deal with few of the key factors involved in the field of food industry.

Research Questions:

Dataset 1: Zomato Bangalore Restaurants.

- Can we predict the rating of the restaurant?

Dataset 2: Food Service Establishment Inspections: Beginning 2005 (INACTIVE)

- Can we predict the severity of violations during inspection of restaurant?

Dataset 3: Restaurant Revenue Prediction

- Can we predict the revenue generated by the restaurant?

- The above questions are often answered using the method of (Knowledge Discovery in Databases).

  Process of knowledge model preparation is mentioned below:
  • Data selection.
  • Pre-processing.
  • Transformation.
  • Data mining and
  • Interpretation/evaluation.

  Some Cross industry Standard process for data processing features, like data quality report, were also used for to possess clear picture on data.

Review on Datasets

1): Zomato Bangalore Restaurants.

Bengaluru is the city in Karnataka, India. Also, Bengaluru is recognized as IT hub and has population around 15Million and 27th largest city in the world. With growing population, it has a scope in almost every field. One of the most and important among them is food industry. Bangalore has a variety of food culture also most of the population is of people from other parts of India. Hence, this resulted in increase of restaurant number and restaurants of all over the world with different varieties of cuisines. This also substituted to what pays me a way to analyze the insights on facts and figures. Zomato is the online food delivery application through which the one can order food from their favorite restaurant and get it delivered at their door step. The dataset contains 17 columns and 56201 rows.

Kaggle – https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants have been collected.

HIMANSHU PODDAR – Courtesy. This dataset has been scarped and found to be correct by Zomato's website.

Here are the data set characteristics:

- url : This feature includes a restaurant link on Zomato.
- address : this Bangalore restaurant url
- name : This feature includes the restaurant's name
- online_order : whether or not online restaurant orders are available
- book_table : table book option available, or not
- rate : contain the overall restaurant rating out
- votes : includes aggregate restaurant upvote numbers
- phone : comprises the restaurant's phone number
- location : includes the restaurant's neighborhood
- rest_type : Type of restaurant.
- dish_liked : foods that are liked to be eaten in the dining area

- cuisines : food types, divided by comma
- approx_cost(for two people) : Approximate cost of meal for two people
- reviews_list : list of tuple restaurant ratings that consist of two values, ranking and customer opinion
- menu_item : includes menu list
- listed_in(type) : meal type
- listed_in(city) : includes the restaurant neighborhood

Target Variable – Rate: indicates the Rating.

2: Food Service Establishment Inspections: Beginning 2005.

Violations pose a high risk to food safety and this plays a vital role in food industry. Safety measures need to be carried out to avoid foodborne diseases. This dataset comprises of name and location of the inactive establishments and violations that were found at the time of inspections. Inactive establishment inspections include only those business that are no longer in business or which have been not operated for an extended time period. The dataset contains 30 columns and 678245 rows.

The Dataset is been taken from –

https://data.world/healthdatany/aaxz-j6pj

Target variable – Critical violation indicates the type of type of Violation.

----

3 : Restaurant Revenue Prediction.

Compared to industries, like everything in this industry it is hard to predict the revenue a firm is been generating. Revenue massively varies across types of restaurants, sizes, regions, and service models. A food court and drive thru as well as drive thru and mobile restaurants these different models are hard to be even compared.

This Dataset was taken from Kaggle:

https://www.kaggle.com/c/restaurant-revenue-prediction/data

This file has 2 .csv files.
a) train.csv – Contains raw data of 137 restaurants: 43 columns and 137 rows.
b) test.csv – contains sample data of 10000 restaurants: 42 columns and 10001 rows.

Here are the data set characteristics:

- Id: Same Restaurant.

- Opening day: restaurant opening date

- Town: town in which there is a restaurant. Please notice that Unicode is in the names.

• Town group: city type. Large cities, or other towns.

- Type: Restaurant type.
- Type: DT: Drive Thru, MB: mobile, FC: Food Hall, IL: Inline

• P1, P2 – P37: • P1. Three types of such obscured data are available. GIS applications from third-party vendors are processed demographic data. This includes people from all regions, distribution of age and gender, scales of growth. The property details primarily apply to the location's m2, the front façade and availability of car parks.

Target variable – Revenue: indicates the Revenue generated.

Machine Learning Methods:

Dataset 1: Zomato Bangalore Restaurants.

Linear Regression - is a basic paradigm for learning. The association between the data points and the least square lines can be found. Simplicity and interpretability are good.

Dataset 2: Food Service Establishment Inspections: Beginning 2005.

Support Vector Machine (SVM): SVM is a learning model. It tries to find the separation between different groups in the data and then classifies new data accordingly.

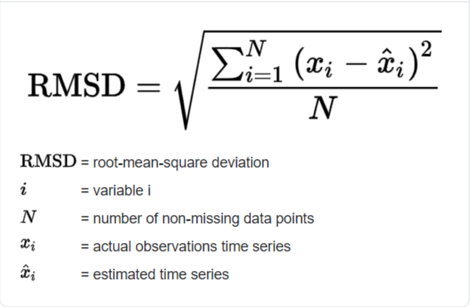
Decision Tree: The Decision Tree is a controlled model of schooling. It operates by studying the rules and patterns in the data characteristics and forecasts the variable.

Dataset 3: Restaurant Revenue Prediction.

Linear Regression is a basic form of learning. The correlation between population data points and the lowest square lines is used. Simplicity and interpretability are good.

Ridge regression is a multi-coordinate study of multiple regression data. As multi-collinearity arises, the smallest square figures are unbiased, but the variances are large enough that they are not valid.

Evaluation Methods

- **RSME (Medium Root Error):** is the square root of the square mean of the total error. It is a fair indicator of precision but just to compare the prediction mistakes of the various models.
  - 

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2}{N}}$$

RMSD = root-mean-square deviation

i = variable i

N = number of non-missing data points

x_i = actual observations time series

$\hat{x}_i$ = estimated time series

- **Accuracy:** It is a measure of nearest value achieved compared to specified value.
- **Precision:** Is a measure of closeness to each other. Precision is reciprocal of the variance.
- **Recall:** Is the ratio of the true positive count to the sum of true positive and false negative counts.
- **F- Score:** factors are taken into account in calculating the weighted average of accuracy and recall by the False Positive counts and False Negative counts.

References:

1) A case study on Zomato – The online Foodking of India
Panigrahi A, Saha A, Shrinet A, Nauityal M, Gaur V. A Case study on Zomato – The online Foodking of India. J Manag Res Anal. 2020;7(1):25-33.

2) Effectiveness of Public Health Interventions in Food Safety: A Systematic Review
Sandra Isaacs, Paul Krueger
(https://www.researchgate.net/publication/13629609_Effectiveness_of_Public_Health_Interventions_in_Food_Safety_A_Systematic_Review)

3) Predicting Future Visitors Of Restaurants Using Big Data.
YUEHUI ZHANG, Chu Luo
(https://www.researchgate.net/publication/328903381_Predicting_Future_Visitors_Of_Restaurants_Using_Big_Data)