IV Sem B. Tech - Computer and Communication Engineering

19CCE213 Machine Learning and Artificial Intelligence

Term Work

SALES PREDICTION

Prepared By:

Aksshaya R - CB.EN. U4CCE20005

Dhashyanth N - CB.EN. U4CCE20018

Hemchand R - CB.EN. U4CCE20024

Priyanga R - CB.EN. U4CCE20045

Department of Electronics and Communication Engineering Amrita School of Engineering, Coimbatore – 641112

2021 – 2022 (even)

Abstract:

Machine Learning is undergoing a huge transformation in recent times and has become a major contributor in real world scenarios. It can be seen in every field including education, healthcare, transport, entertainment and several more; In the fast-moving world traditional sales and marketing approaches find no place to match the pace of the competitive market as they don't have insights on customer's interest and sales prediction.

An Extensive study of sales prediction is done using Machine learning models such as: KNeighbors Classifier, SVM classifier, Naive Bayes, Decision Tree Classifier, Random Forest Classifier. The main aim of this project centres around finding a good model using supervised learning approach that can be used for future prediction with already available data.

Key Words- Machine learning models, Supervised learning approach, KNeighbors classifier, Random Forest classifier, SVM classifier, Naive Bayes, Decision Tree Classifier

Introduction:

One of the major objectives of this project work is to find out the reliable sales trend prediction which is implemented by using data mining techniques to get an insight on how a company should manage cash flow and resources. Sales prediction allows companies to predict achievable sales revenue, efficiently allocate resources and plan for future growth.

Today's business handles large amount of data. The volume of data is expected to grow further in an exponential manner.

Predicting sales manually could lead to drastic error and poor management in the organization and most importantly time consuming which is not desirable. Thus, using data mining and machine learning techniques are the need of the hour as these methods are less time consuming and more accurate and efficient

In our project, Supervised machine learning models are designed which can accurately predict the data collected from the previous sales of a Supermarket. These businesses are in need of new data mining techniques and intelligent prediction model of sales trends with highest possible level of accuracy and reliability. It allows companies to plan their business strategies effectively.

In this project, the dataset used consist of many attributes such as branch, customer type, city, gender, cost of goods sold, unit price and rating of each branch. Study has been made on how store sales are influenced by the above factors. This project report explains in detail the surveys, methodology that was undertaken and its results with a note on future improvement.

Survey:

The base paper used for reference, for this project are mentioned in this section.

1. Intelligent Sales Prediction Using Machine Learning Techniques-iCCECE_Paper_procedings. This paper used a dataset of an e-fashion store, which contained three consecutive years of sales data. To predict the sales of the e-fashion store, past sales record for three years from 2015 to 2017 were collected. It implemented data mining techniques like pre-processing, Outlier detection. Trends were analysed using different clustering techniques. Different classification algorithm was used for prediction and their performance analysed.

2. ’Walmart's Sales Data Analysis - A Big Data Analytics perspective’-In this study, inspection of the data collected from a retail store and prediction of the future strategies related to the store management is executed. Effect of various sequence of events such as the climatic conditions, holidays etc. can actually modify the state of different departments so it also studies these effects and examines its influence on sales.

3. ’Sales Prediction System Using Machine Learning’ In this paper, the objective is to get proper results for predicting the future sales or demands of a firm by applying techniques like Clustering Models and measures for sales predictions. The potential of the algorithmic methods is estimated and accordingly used in further research

4. ’Intelligent Sales Prediction Using Machine Learning Techniques’ This research presents the exploration of the decisions to be made from the experimental data and from the insights obtained from the visualization of data. It has used data mining techniques. Gradient Boost algorithm has been shown to exhibit maximum accuracy in picturizing the future transactions

These papers helped in moving on with our project.

Methodology:

In order to predict the sales different data mining techniques were undertaken. The library functions that were used are seaborn, os, datetime, SciPy.

I. Data Collection

Dataset of a supermarket is taken for this project. This dataset consists of information about Invoice ID, Branch City, Customer type, Gender, Product line (beauty, electronics, sports etc), Unit price, Quantity, Tax 5%, Total, Date, Time, Payment, cogs (cost of goods sold), gross margin percentage, gross income and Rating.

II. Data Pre-processing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

Normally this data is not always clean and formatted. There can be noise, missing data, inaccurate data that must be removed and rectified. This process has been done making the data ready to be trained in a model.

III. Outlier detection

The main focus is on the quality of the data, especially because Outliers can skew results. Outliers are values falling at least 1.5 × IQR above the third quartile or below the first quartile. This outlier detection can also be done using IRQ.

IV. Correlation

Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it’s a measure of how things are related. The study of how variables are correlated is called correlation analysis. It important in real life because the value of one variable can be predicted with the help of other variable.

It is given by

$$Correlation = \frac{Cov\ (x, y)}{\sigma x * \sigma y}$$

Heat Map: The correlation matrix is generated showing the correlations between the attributes with a positive correlation

V. Classification

Classification in machine learning and statistics is a supervised learning approach in which the model learns from the data given to it called the training data and makes new observations on the test data or groups data into classes.

This is used for sales prediction. In this project, part of the data is used to train the model, and the rest is used to verify the result, by comparing the predicted sales with the existing data. The following Classifiers are used: KNeighbors Classifier, SVM classifier, Naive Bayes, Decision Tree Classifier, Random Forest Classifier.

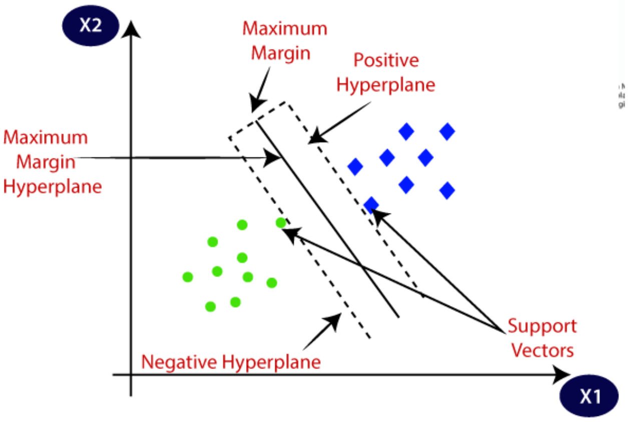
Further the performance of classifier is checked based on the accuracy, precision, F1 score. Confusion matrix as well as AUC- ROC curve is generated to check which classifier model’s predicted values is almost closer to the existing value.

1. **KNeighbors Classifier:**

One of the simplest Machine Learning algorithms based on Supervised Learning technique. This classifier Stores all the available data and classifies a new data point based on the similarity. Also called a **lazy learner algorithm** because it does not learn from the training set, but stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

2. SVM classifier:

It is a supervised machine learning approach where it finds a hyperplane that best separates the two classes by finding the maximum margin between the hyperplanes meaning to find the maximum distances between the two classes.



3. Naive Bayes:

This classifier will have to predict X [where X is a vector consisting of ‘n’ attributes] belongs to a certain class. The class delivering the highest posterior probability will be chosen as the best class. The attribute’s value is assumed to be independent of one another conditionally.

Following is the Bayes theorem is used to implement Naive Bayes classifier.

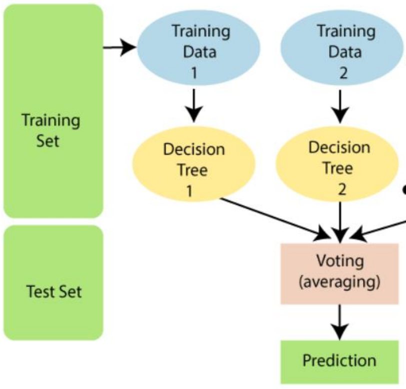
$$P(C_i|\, x_1, x_2 \dots, x_n) = \frac{P(x_1, x_2 \dots, x_n|C_i).\, P(C_i)}{P(x_1, x_2 \dots, x_n)} \; for\; 1 < i < k$$

4. Decision Tree Classifier:

The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute. This project used ecision tree with Information gain.

5. Random Forest Classifier:

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset instead of relying on one decision tree.

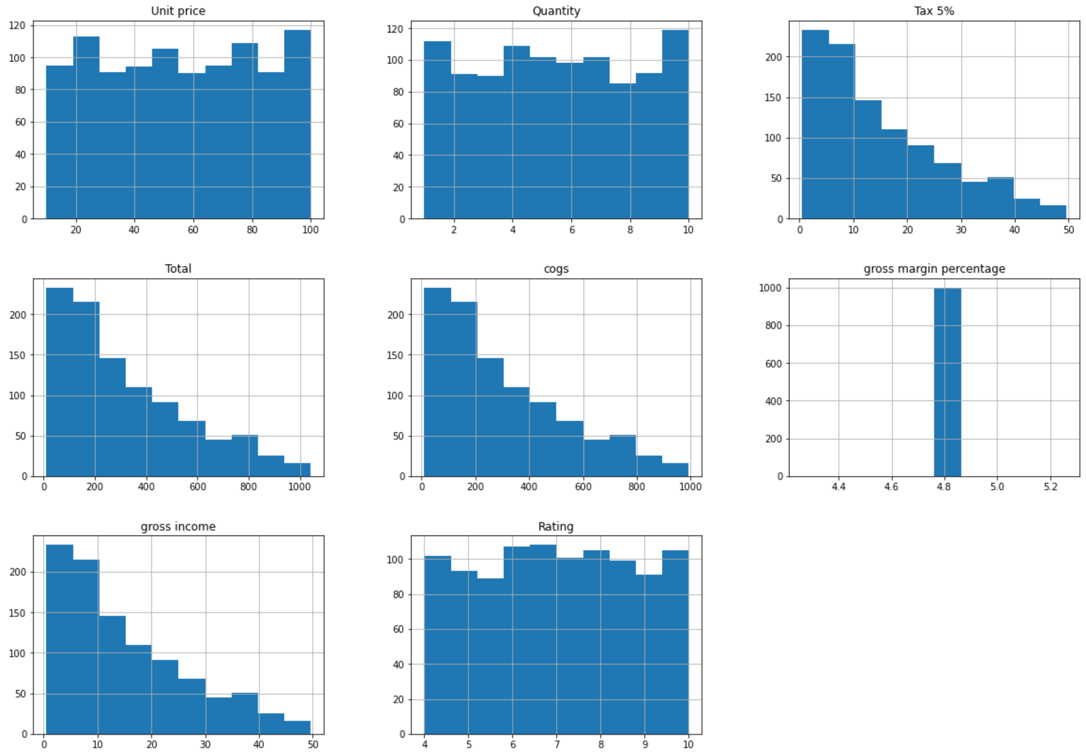


6. Extra Trees Classifier:

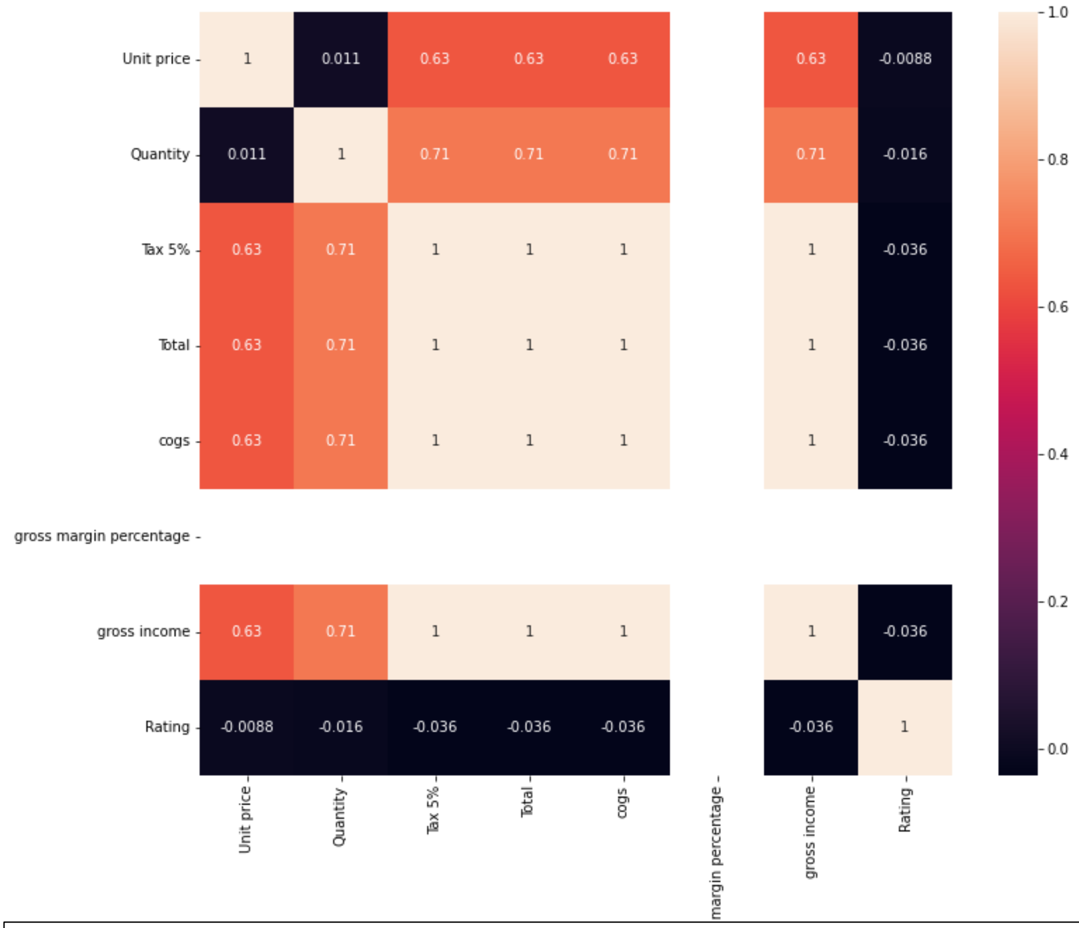
Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output its classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple de-correlated decision trees.

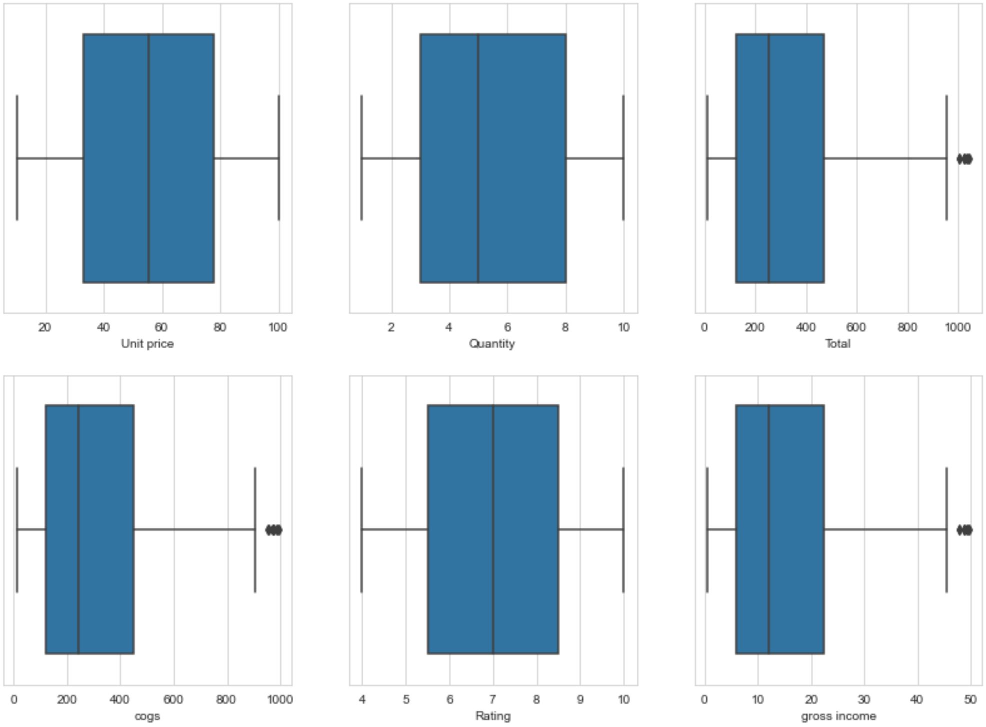
Result:



Fig–1: Bar plots of various attributes, x-axis represent bins, y-axis represents frequency



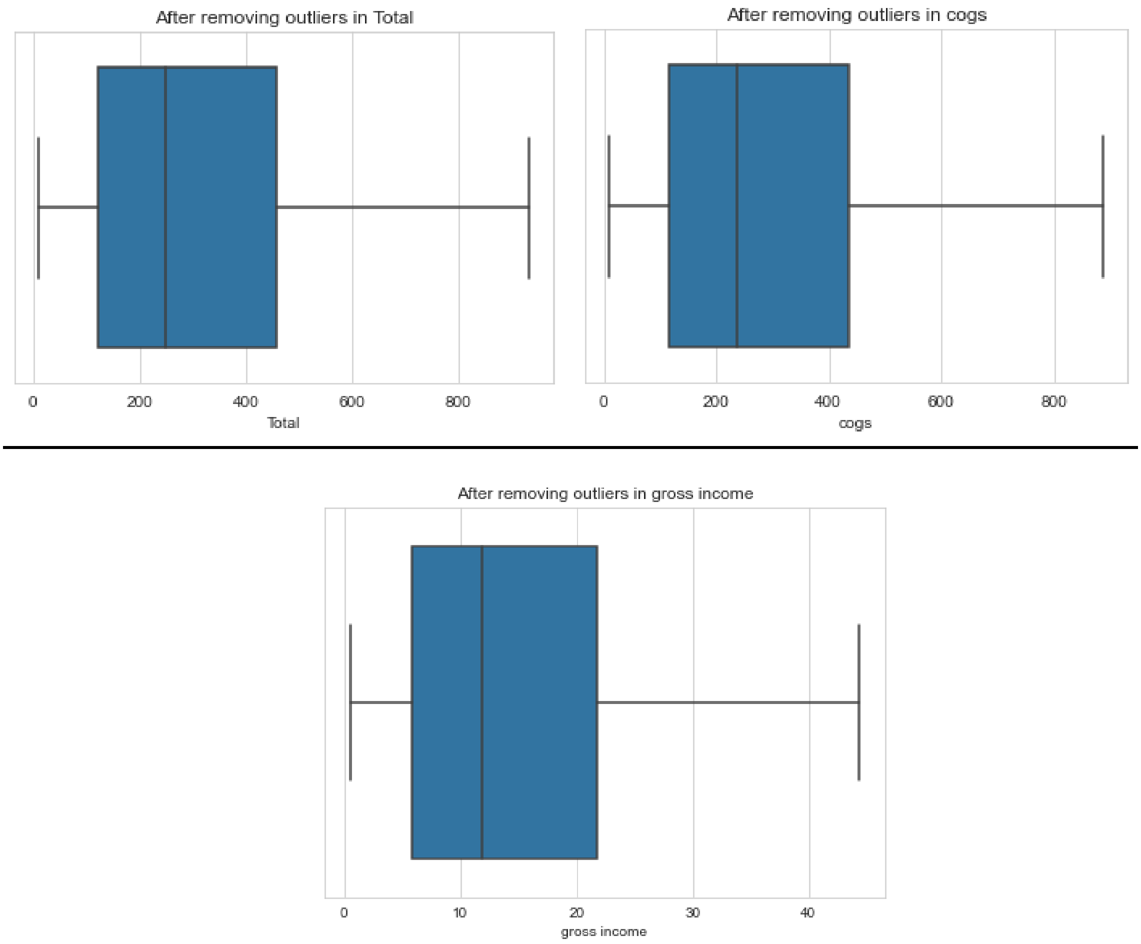
Fig–2: Heat map representing the correlation between each attribute



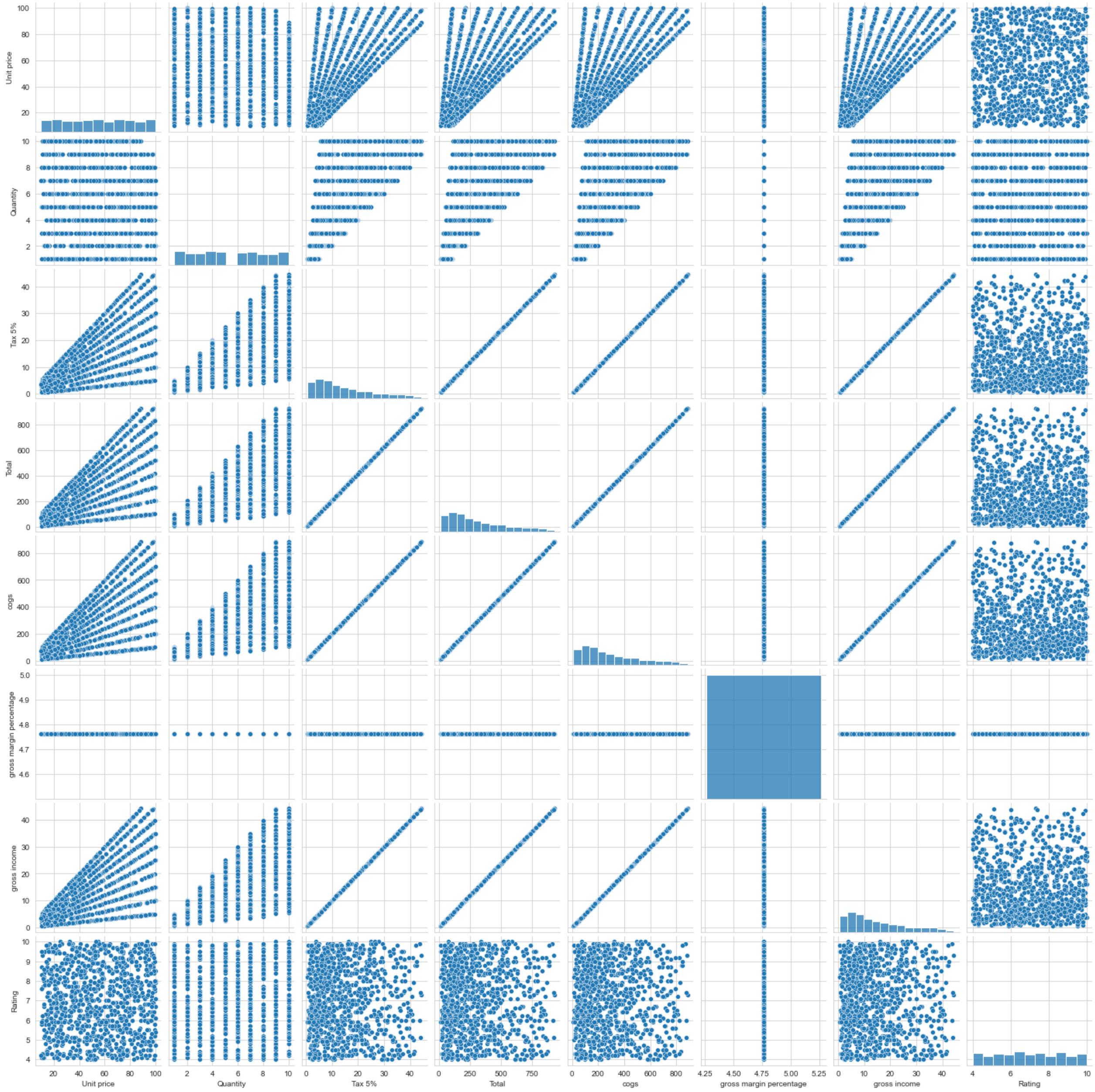
Fig–3: Box plot with outliers

From the above plot we can infer that total, cogs (cost of goods sold), gross income have outliers. They can be removed using IQR.

First the interquartile range for the data is calculated. This is multiplied by 1.5. The Value we get by adding this with Third quartile and subtracting with first quartile are Upper Limit and Lower Limit respectively. Values outside upper and lower limits are removed.



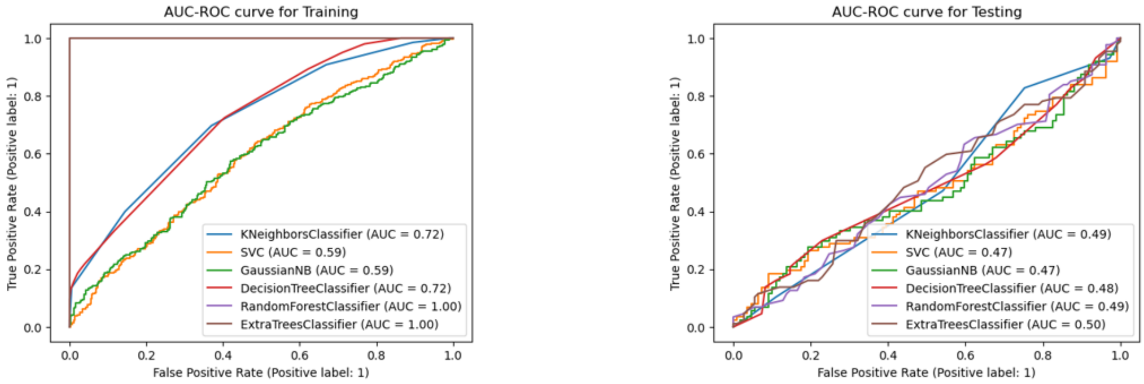
Fig–4: This box plot is plotted after outlier removal



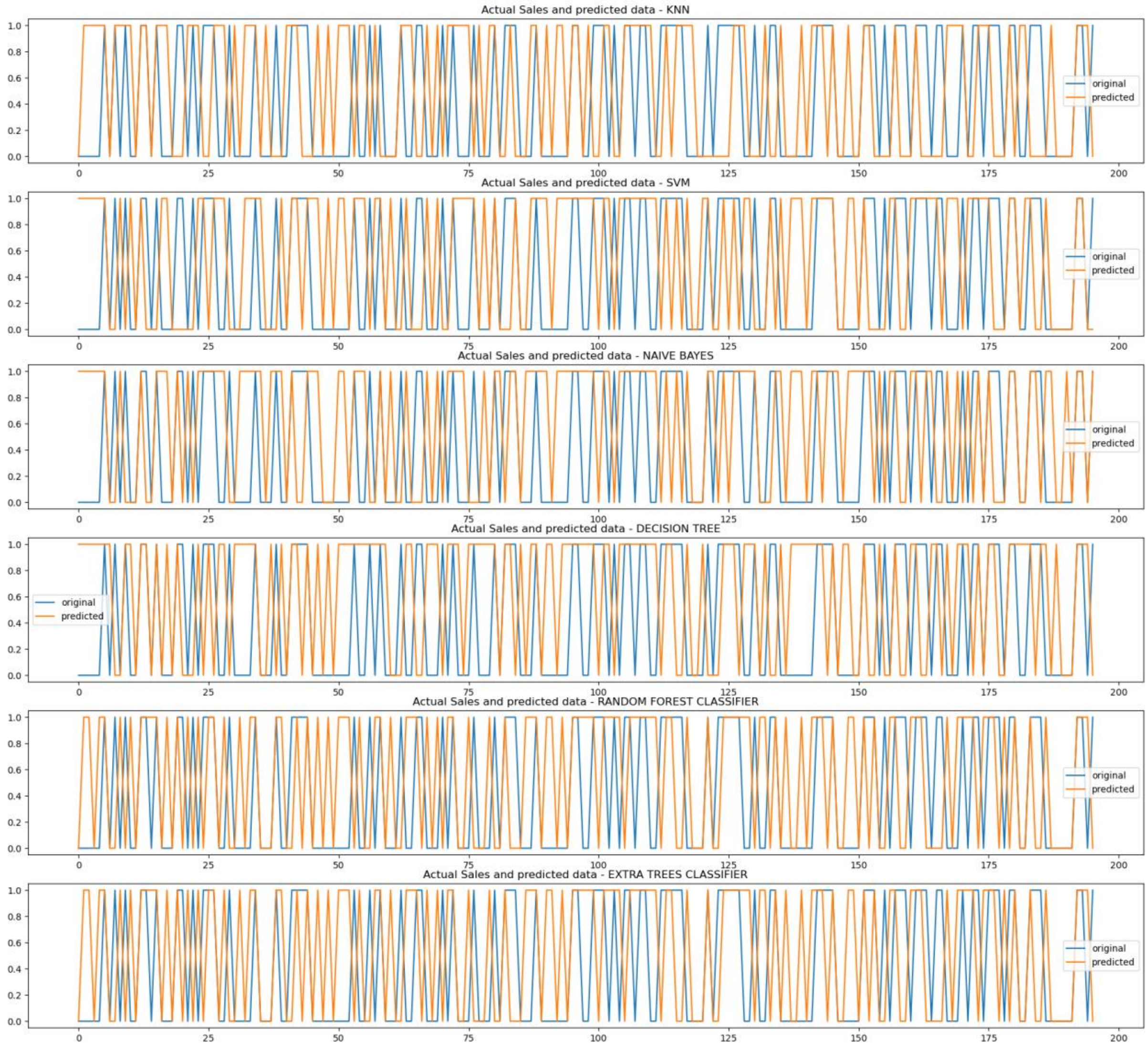
Fig–5: pair plot for the dataset

A pair plot plots the pairwise relationship in a dataset. The pair plot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column.

Pair plot is used to understand the best set of features to explain a relationship between two variables.



Fig–6: AUC – ROC curve for training and testing score



Fig–7: Actual and prediction sales prediction of different classifiers

Future Improvement:

In this project, prediction based on gender was done. But other attributes also can be considered for prediction, so that much clear picture can be obtained regarding the sales. This will help the company to prepare for budgeting, for stocking up products depending on the demand etc.

Conclusion:

It can be concluded that an intelligent sales prediction system is required for business organizations to handle enormous volume of data. Decisions can be done only based on speed and accuracy of data processing techniques. Machine learning approaches highlighted in this project will be able to provide an effective way for decision making and predicting. For this particular dataset Extra Trees classifier has shown great accuracy in its predicted values.

Reference:

[1]. Intelligent Sales Prediction Using Machine Learning Technique, Sunitha Cheriyan, 2018 IEEE. [https://ieeexplore.ieee.org]

[2]. B.Sri Sai Ramya, K. Vedavathi, An Advanced Sales Forecasting Using Machine Learning Algorithm, International Journal of Innovative Science and Research Technology, May 2020. [https://www.ijisrt.com/assets/upload/files/IJISRT20MAY134.pdf]

[3]. SINGH MANPREET, BHAWICK GHUTLA, REUBEN LILO JNR, AESAAN FS MOHAMMED, AND MAHMOOD A. RASHID. "WALMART'S SALES

[4]. DATA ANALYSIS-A BIG DATA ANALYTICS PERSPECTIVE." IN 2017 4TH ASIA-PACIFIC WORLD CONGRESS ON COMPUTER SCIENCE AND ENGINEERING (APWC ON CSE), PP. 114-119. IEEE, 2017.

[5]. Sekban, Judi. "Applying machine learning algorithms in sales prediction." (2019)

[6]. Ragg, Thomas, Wolfram Menzel, Walter Baum, and Michael Wigbers. "Bayesian learning for sales rate predictionfor thousands of retailers." Neurocomputing 43, no. 1-4 (2002): 127-144