# CAPSTONE PROJECT

Individual sales dataset

Name = Anukriti
Student Id = 0775876

Table of Contents

## ABSTRACT

We have selected Sales Data from sales datasets**.** This dataset includes about 40,000 records and 15 attributes. Each record corresponds to a customer information like (gender, education, house Value, age, region, fam income, region, marriage, children, occupation, car probability, house own, flag (whether the customer purchased the target productor not) and online (whether the consumer had online shopping experience or not). This dataset provide helps to organizations to better understand their customer's needs and makes it easier for organizations to modify products according to the specific needs, behaviors, and interests of customers. When organizations satisfy customers specific needs according to their demands it helps companies to increase there productive of different products in the entire market and helps to gain more and more profit. Because companies' whole profit and loss is depending upon customers demand if company fulfill the demand of customers, it definitely gains profit if it do not fulfill need of their customers than it will gain loss. This dataset is uncleaned there are some missing values in the dataset. It contains character and numerical data type. We will use some method to clean our dataset to make it stronger and more valuable to perform different types of models to collect different results. We will also use predictive and descriptive analysis to discuss about how many customers received their products according to their demands and we will also find in which direction company sales trend moved upward or downward.

KEYWORDS:

descriptive analysis ,Classification, Regression, Clustering, predictive analysis, Data Cleaning.

## RESEARCH QUESTIONS:

1. Find out the customer buy the target product or not by using predictive analysis?
2. Find out the number of buyers who have online shopping experience or not?
3. By using descriptive analysis find out Find out number of male and female in the dataset?
4. Find out sales increase by married or unmarried?
5. How many customers have highest or lowest family income?

## TOOLS:

Python is use to do all data visualization.

## GITHUB ACCOUNT INFORMATION:

Anukriti GitHub account link below:

https://github.com/Anukriti0775876

### Introduction

Individual company sales dataset is plays important role in marketing sector. Sales are sports associated with promoting the number of products bought in a given targeted time period. Sellers without difficulty achieve consumer product evaluations from online reviews so that the see the competitive products in the market and enhance the productivity of their products to get more profit. The delivery of a company for a price is also considered a sale. The seller, or the corporation of the goods or services, completes a sale in response to an acquisition, appropriation, requisition, or an immediate interaction with the customer on the aspect of sale. There is a passing of call (property or ownership) of the item, and the settlement of a charge, in which settlement is reached on a fee for which switch of ownership of the object will arise. The issuer, not the consumer,

typically executes the sale and it is able to be completed previous to the responsibility of charge. In the case of indirect interaction, a person who sells items or company on behalf of the proprietor is called a shop clerk or saleswoman or salesperson, however this frequently refers to someone promoting items in a shop/maintain, wherein case different terms are also commonplace, which includes shop clerk, keep assistant, and retail clerk.

This dataset includes about 40,000 records and 15 attributes. Each record corresponds to a customer information like (gender, education, house Val, age, region, fam income, region, marriage, children, occupation, car probability, house own, flag (whether the customer purchased the target productor not) and online (whether the consumer had online shopping experience or not). This dataset provide helps to organizations to better understand their customers' needs and makes it easier for organizations to modify products according to the specific needs, behaviors, and interests of customers. When organizations satisfy customers specific needs according to their demands it helps companies to increase their productive of different products in the entire market and helps to gain more and more profit. Because companies' whole profit and loss is depending upon customers demand if company fulfill the demand of customers, it definitely gains profit if it do not fulfill need of their customers than it will gain loss. This dataset is uncleaned there are some missing values in the dataset. It contains character and numerical data type. We will use some method to clean our dataset to make it stronger and more valuable to perform different types of models to collect different results. We will also discuss about how many customers received their products according to their demands and we will also find in which direction company sales moved. Because now it is difficult to say that about sales company move upward or downward.

As we selected this dataset from Kaggle so there are some operations are performed on this dataset but those are not same as we expected to operate on our project. No one else selected this dataset

to perform different operation. Our dataset will best fit on predictive analysis as we think to predict different outputs from this dataset to know about company sale in the market. Our research is worth because now a days sales data is very popular in the market to predict better future results of products by using present and past result. Due to which every organization will receive positive results.

## Literature Review

Individual company sales data is done by Mickey (2019), Rondinelly Oliveira (2021), Guozhen Chang, Fabio Traverso (2020), MichelGadomsky, nadia hajrasi on Kaggle. D. J. Dalrymple, "Sales forecasting methods and accuracy," Business Horizons, vol. 18, pp. 69–73, 2006. C. Dellarocas, X. Zhang, and N. F. Awad, "Exploring the value of online product reviews in forecasting sales: the case of motion pictures," Journal of Interactive Marketing, vol. 21, no. 4, pp. 23–45, 2007.A. Tony, P. Kumar, and S. Rohith Jefferson, "A study of demand and sales forecasting model using machine learning algorithm," Psychology and Education Journal, vol. 58, pp. 10182–10194, 2021. F. Zhu and X. Zhang, "Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics," Journal of Marketing, vol. 74, no. 2, pp. 133–148, 2010. It is based on customer information like the are male or female, married or unmarried, have any children or not, income status. The pattern agency does an excessive extent of industrial agency, so it runs industrial company statistics reports to beneficial aid in selection aid. Many of these reviews are time-based totally and non-volatile. That is, they analyze past facts tendencies. The enterprise agency masses information into its facts warehouse regularly to collect records for those reviews. These critiques encompass annual, quarterly, monthly, and weekly earnings figures thru product.

Identifying competition of a business enterprise in particular requires their analysis and assessment the use of the received enterprise facts to discover their organization positions. Relevant identity

methods can consequently be identified from the exclusive perspectives, consisting of resource- and net-based totally strategies.

Within the same market, one-of-a-kind organizations can offer comparable merchandise. If they compete, they may possibly have incredibly comparable products, similar technology, and similar marketplace strategies. Therefore, the sellers have to put more effort to sell their products to compete their competitors. In past organization used different technique of investigating competitors calls for the usage of a questionnaire, due to the fact the product is in the long run aimed toward customers who're most strongly privy to the opposition. Using questionnaires can achieve direct information from the front-line market consumers. This method is the most direct, however it takes a long time and consumes many assets. Furthermore, facts comments can be incomplete, causing records errors that result in faulty conclusions.

Sales Company used many different techniques to achieve their goals like unsupervised approach based totally on a multi-strategy getting to know algorithm for figuring out competitors in a in the market with similar products. In their company competition evaluation, the positioning of the competition changed into crucial. It became determined that the organization compete with their competitors in better results when their competitors clearly positioned and sales company make their products with more advance and better results. In individual company sales dataset prediction models based on online shopping have been lots better than the ones the use of offline shopping customers, along with general employer market place price.

With the facilitation of the internet, social and consumption activities have regularly shifted to online platforms, and social systems and on-line-shopping systems have emerged as essential entertainment and consumption channels. Relying on community consumer facts, such as social activity facts and person on line remark data on online income systems, it may serve all social

contributors via statistics extraction and mining. In sales dataset some user receives there ordered products and some consumers do not receive the target product. After analysis it will be predict in which company sales trend move forward and downward. From martial column it will be analysis which age group customer are more interested in this individual company sales data. Consumer reviews are specially supplied in text shape on each purchasing and social media sites. The incorporation of consumer statement records into predictive fashions requires the translation of textual records into specific variables, which, in flip, requires evaluation of the characteristics of client evaluations. Many scholars have carried out large studies within the area of purchaser commentary research and have accomplished excellent consequences. Specifically, client evaluate studies has targeted on empirical evaluation and sensible strategies. In phrases of empirical evaluation, this has in particular included the traits and values of purchaser evaluations, customer-institution characteristics, and product advantages and disadvantages. In phrases of generation, the research has centered on consumer evaluation characteristic extraction, assessment sorting and show, review sentiment analysis, and more.

Consumer evaluations normally are reflective of product characteristics. After a review location is examined, next customer-buy choices could be prompted with the aid of those comments. Thus, the characteristics of client evaluations will usually have an effect on product sales. However, fine reviews had a larger impact than did negative ones when the reviews were from the consumers' friends. The consumer text content of social media become analyzed, showing that their feedback covered now not best fine and negative emotions, however additionally notes about balance, warnings, happiness, peaceful components, etc. Furthermore, particular emotional dimensions have been acknowledged to have an impact on purchaser choices. With the deepening of this research, researchers have found that purchaser evaluations now not only replicate the attitude of

clients, but they also predict the characteristics of dealers from assessment functions. Through online opinions, consumers' attributes approximately services and products may be obtained. The connection among consumer online reviews conducts, pleasure, and call for within the lodging-sharing resort industry. When customers are at unique sharing degrees, the elements that affect customer pride and call for are distinct. In sales dataset there are some duplicate values which will create errors. Unfortunately, the differences are only pondered by means of textual content characters. This issue reasons mistakes whilst determining the similarity of comments and competitors. From family income variable it will determine that in this sales dataset from which class customers included from their family income status.

## Methodology

In the initial stage, The dataset is unclean and contain some count of missing values and duplicate values. On the first phase of analysing, cleaning will take place to perform further operations. In this we will use regression model to find out relationship between dependent and independent variables. Descriptive and predictive analysis will also take place to find out how many customers are satisfied related to the sales of individual company's sales. We will also use classification and clustering in this. We will collect different results related to sales of this data set by using these model.

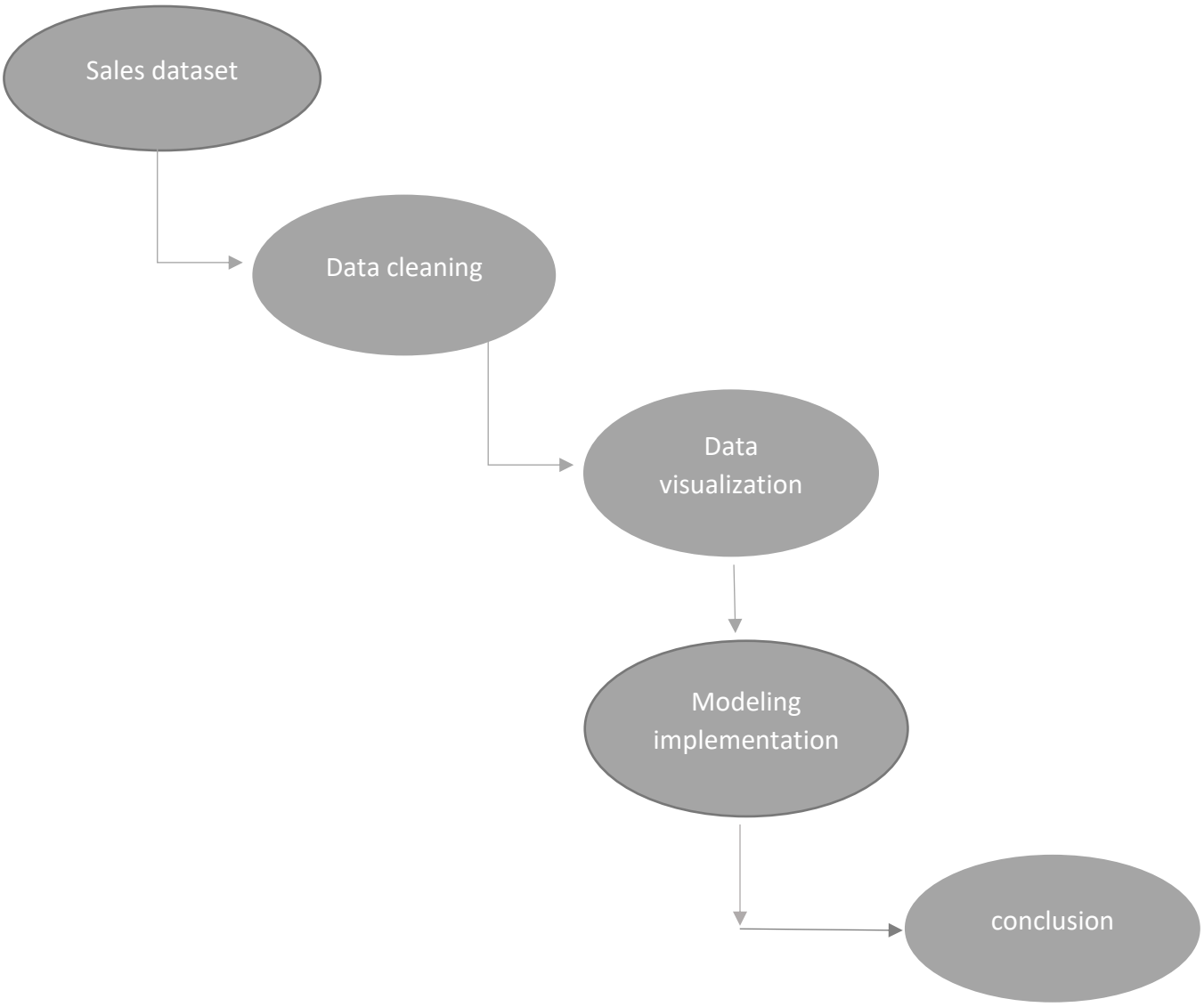
Sales dataset
Data cleaning
Data visualization
Modeling implementation
conclusion

Figure 1

## DATA DICTIONARY

### Categorical Attributes:

| Attribute Name | Description | Data Type | NO. of Levels | Value Counts |
|---|---|---|---|---|
| | | | | |
| **Flag** | Customers buy their target productor or not | category | 2 | Y 13334<br>N 10196 |
| **Gender** | Gender of the customer (Male or Female) | category | 3 | M 14186<br>F 8938<br>U 406 |
| **Education** | Educational degree of customer | category | 5 | 1. HS 50382. Some College 6546<br>3. Bach 5987<br>4. Grad 3910<br>0. <HS 2049 |

| Attribute Name | Description | Data Type | NO. of Levels | Value Counts |
|---|---|---|---|---|
| **Age** | Age of the customer according to their group | category | 7 | 5_<=55 6089<br>4_<=45 4986<br>6_<=65 3986<br>1_Unk 3061<br>3_<=35 2515<br>7_>65 2306<br>2_<=25 587 |
| **Online** | Customer have online shoping experience or not | category | 2 | Y 16516<br>N 7014 |
| **Customer psychology** | Customer psychology based on the area of residence where they live | category | 10 | H 382 |

| Attribute Name | Description | Data Type | NO. of Levels | Value Counts |
|---|---|---|---|---|
| **Marriage** | Customer's marital status, customer is married or not | category | 2 | Married 19266<br>Single 4264 |
| **Child** | Consumer have children or not | category | 3 | Y 11174<br>N 7307<br>U 5049 |
| **Occupation** | Information about customer's career | category | 6 | Professional 9818<br>Sales/Service 6626<br>Blue Collar 3983<br>Retired 2018<br>Farm 169 |
| **Mortgage** | Information about customers | category | 3 | 1Low 16458<br>3High 3838<br>2Med 3234 |

| Attribute Name | Description | Data Type | NO. of Levels | Value Counts |
|---|---|---|---|---|
| | have any housing loan or not | | | |
| **House owner** | Customers have their owns house or not | category | 2 | Owner 18778<br>Renter 4752 |
| **Region** | Information about area where customer lived | category | 5 | South 9174<br>West 5128<br>Midwest 4883<br>Northeast 4207<br>Rest 138 |
| **Family income** | Information about income of customer's family | category | 13 | E 4881<br>F 4099<br>G 2606<br>D 2469<br>H 1602<br>C 1357<br>A 1180<br>B 1118 |

| Attribute Name | Description | Data Type | NO. of Levels | Value Counts |
|---|---|---|---|---|
| | | | | L 1082<br>I 1077<br>J 1076<br>K 970<br>U 13 |

Table 1. Categorical Attributes

## Numeric Attributes:

| Column1 | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| House value | 39945 | 307597 | 422342 | 0 | 81455 | 215133 | 394067 | 9999999 |
| Car probability | 39945 | 3 | 3 | 0 | 1 | 3 | 5 | 9 |

## Detailed Data Dictionary:

To assigned the correct datatype and to check the categorical attribute I take one variable . After that I check the number of levels in every attribute and also count Its values on each level.

There I used five number summary and it is mean, minimum, maximum, count, standard deviation. It use For numerical attributes.

## Data Cleaning:

1. Duplication of values in the dataset
2. Missing values in the dataset

(Duplication of values in the dataset):55 duplicate values can be see in this dataset. the drop duplicated command used for drop these values from data set.

(Missing values in the dataset): The total missing values in this dataset are counted as 18106. Further in each category such as in House owner have 3369 missing values, There are 735 null values in education attribute, and marriage attribute 14002 missing values. For data cleaning these values are filled by 0.

| **flag** | **0** |
|---|---|
| gender | 0 |
| house Val | 0 |
| age | 0 |
| online | 0 |
| customers | 0 |
| child | 0 |
| occupation | 0 |

| | |
|---|---|
| mortgage | 0 |
| region | 0 |
| Car probability | 0 |
| Fam income | 0 |
| education | 735 |
| House owner | 3369 |
| Marriage | 14002 |

## Data Visualization:

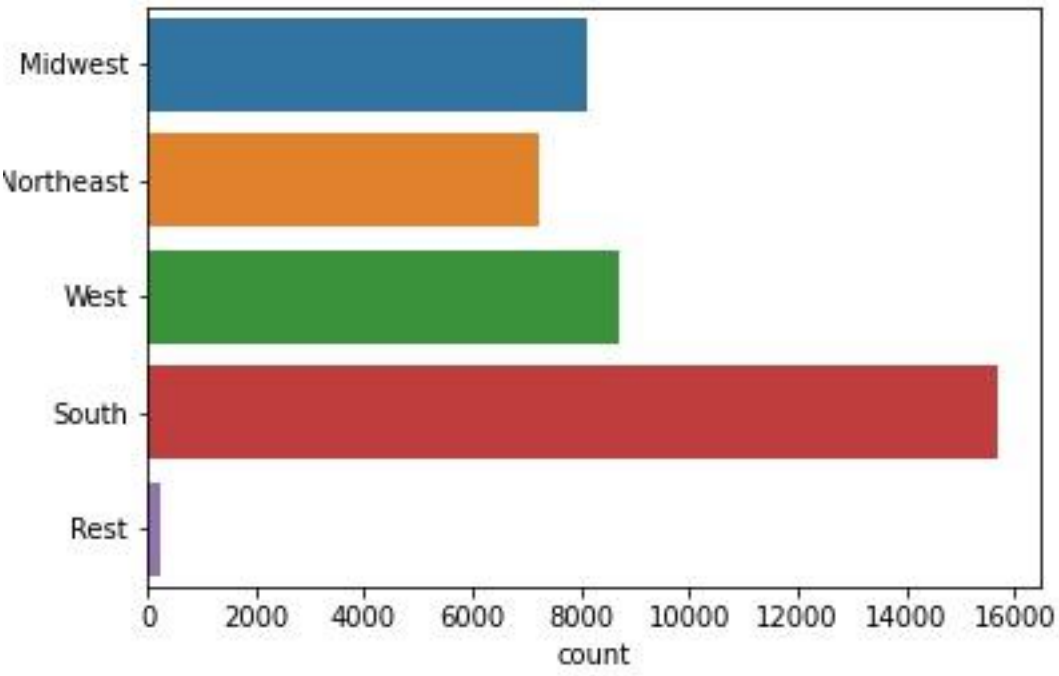


Figure 2. Region vs Customers Count

This bar graph depicts  the information about customers count from different region. The highest count 15676 which accounts in south region. The count of customers from Midwest and west is approximately same which count as 8107 from Midwest and 8725 from west. The count of rest of region is 245.

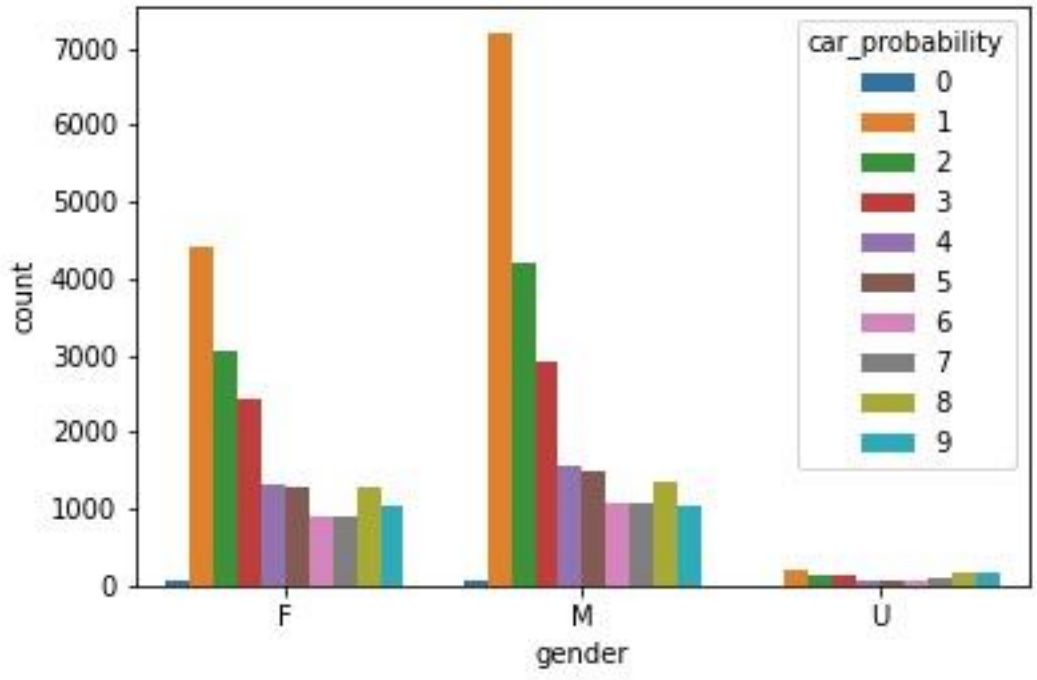


Figure 3. Gender Vs Car_ probability

this  bar graph represents the information of car probability. This graph shows the probability of female and male have car. As highest probability in both genders having a car is one and shows as orange in graph . Having 4 to 9 car probability  is approximately looking similar. unknown have the lowest car probability between these three sections.

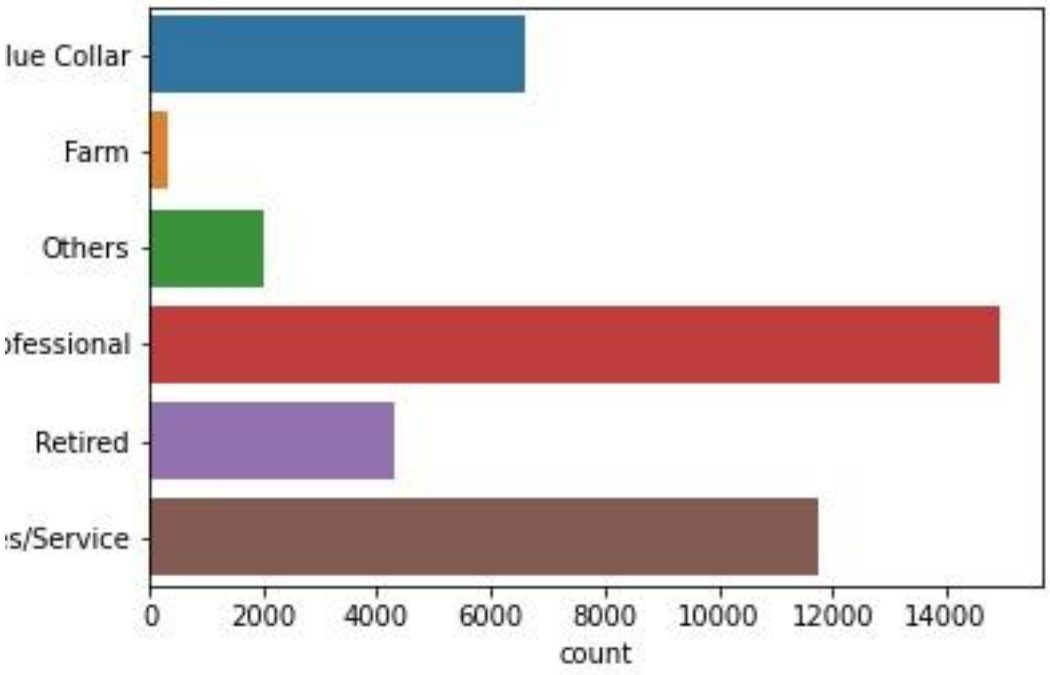


Figure 4. Occupation Vs Customers Count

This graph represents the information of customer's profession. As the highest count of customers lies in 'professional' category and the count is 15000. After that the sales/service category is stand on second highest category which have 11500 numbers of customers. The number of blue collar customers are accounted as 6300 and farmers buy less products than all attributes and 'other' are the second last in this field.

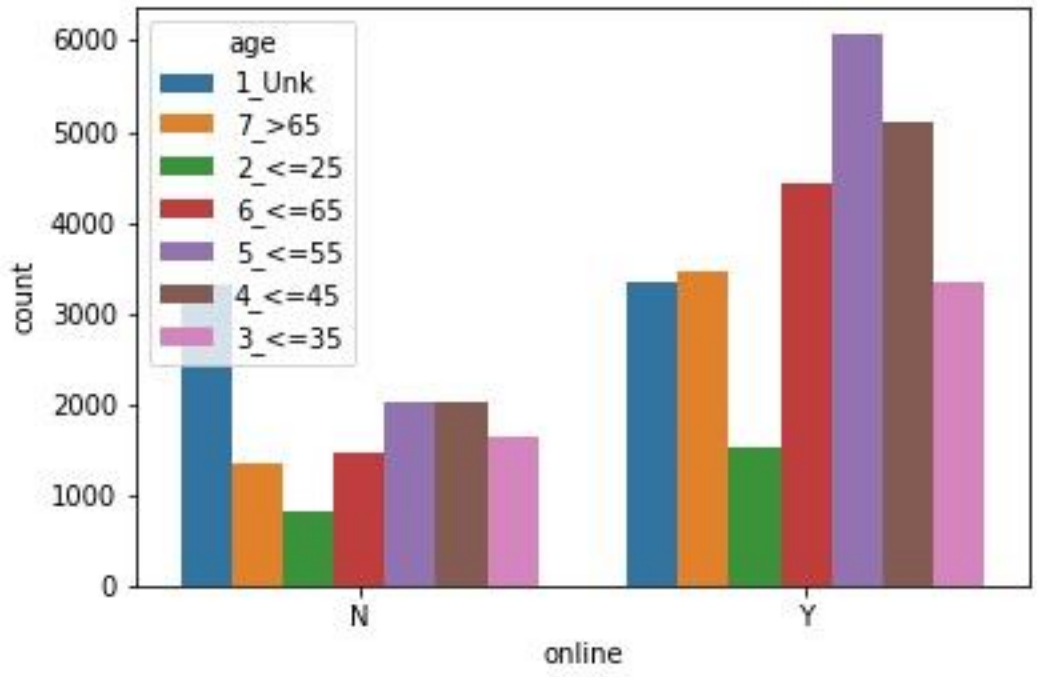


Figure 5. Online Vs Customers Count

This bar graph illustrate the information about how many customers have online shopping experience and vice versa according to their age group. It is crystal clear that number of customers have online shopping experience is more.

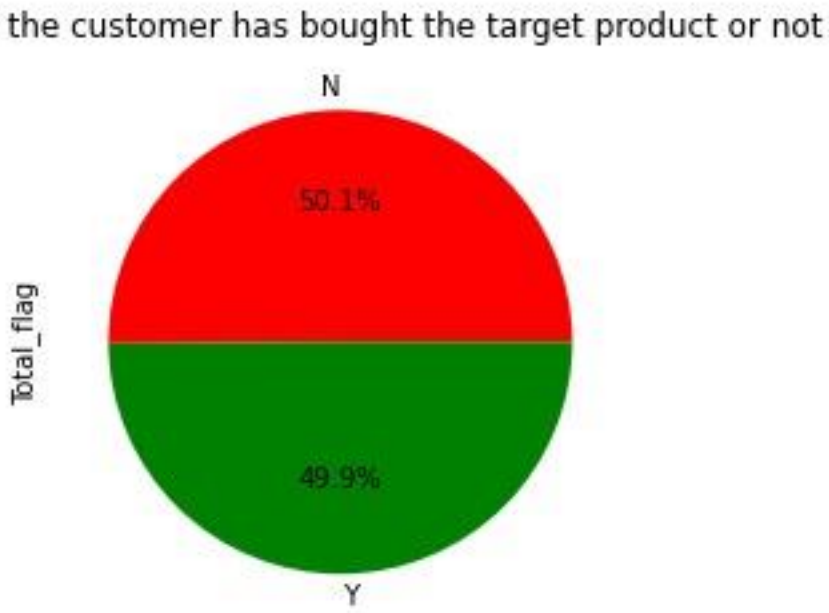


Figure 6.

In this picture the percentage of buyers who bought their target product is denoted by ‘Y’ and the percentage of customer who do not purchase their target products is denoted by ‘N’ and the percentage is approximately same. Due to this result it is crystal clear the company do not have more profit because of half of the customer do not satisfy from the services by company because they do not buy their target product which they want to buy.


Figure 7.

In this illustrated picture the ‘Y’ is consumers who have online shopping experience and ‘N’ who do not have online. It can be clearly found that 50.4% customers have online shopping experience on the other hands 49.6% do not have any online shopping experience.

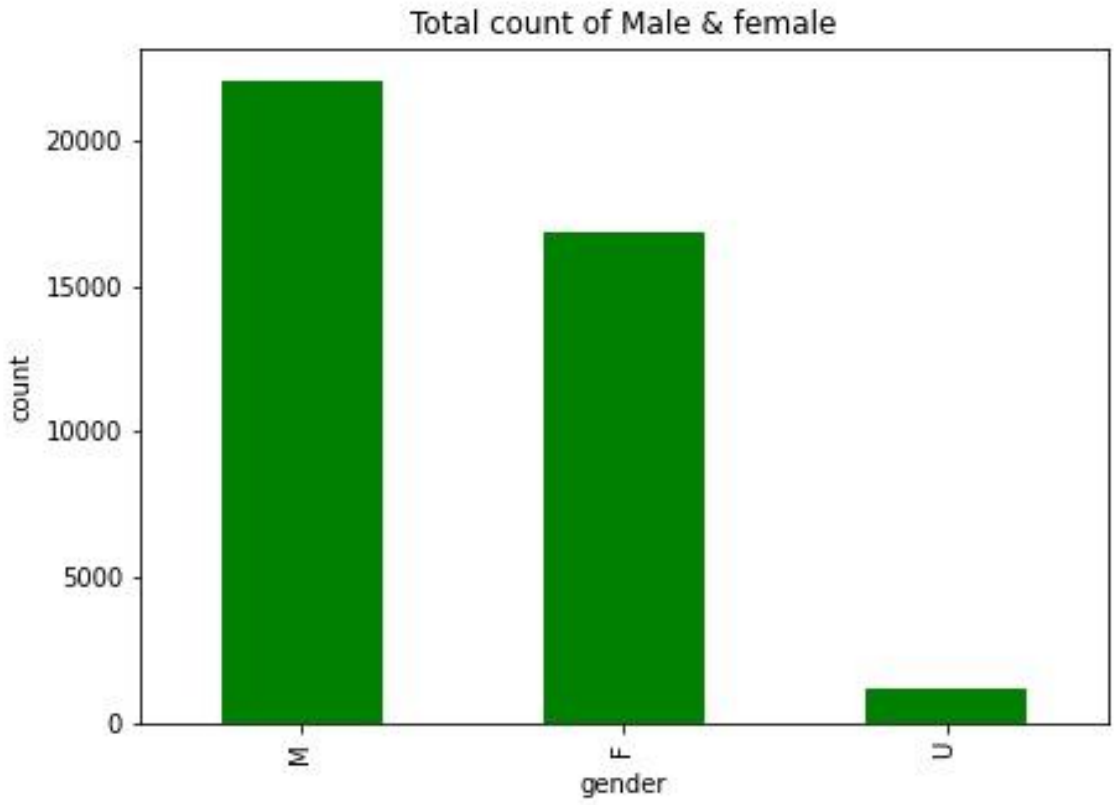


Figure 8. Gender Vs Count

It can be clearly seen that ‘M’ is the number of male, F female and U unknowns. It is found that male buy more products count is highest (22500) as compared to other two categories female and unknown. Females lies on second position with (17000) count. After that, I can be accounted some unknown values in these three gender and the count is approximately 150. Unknown comes on last rank.

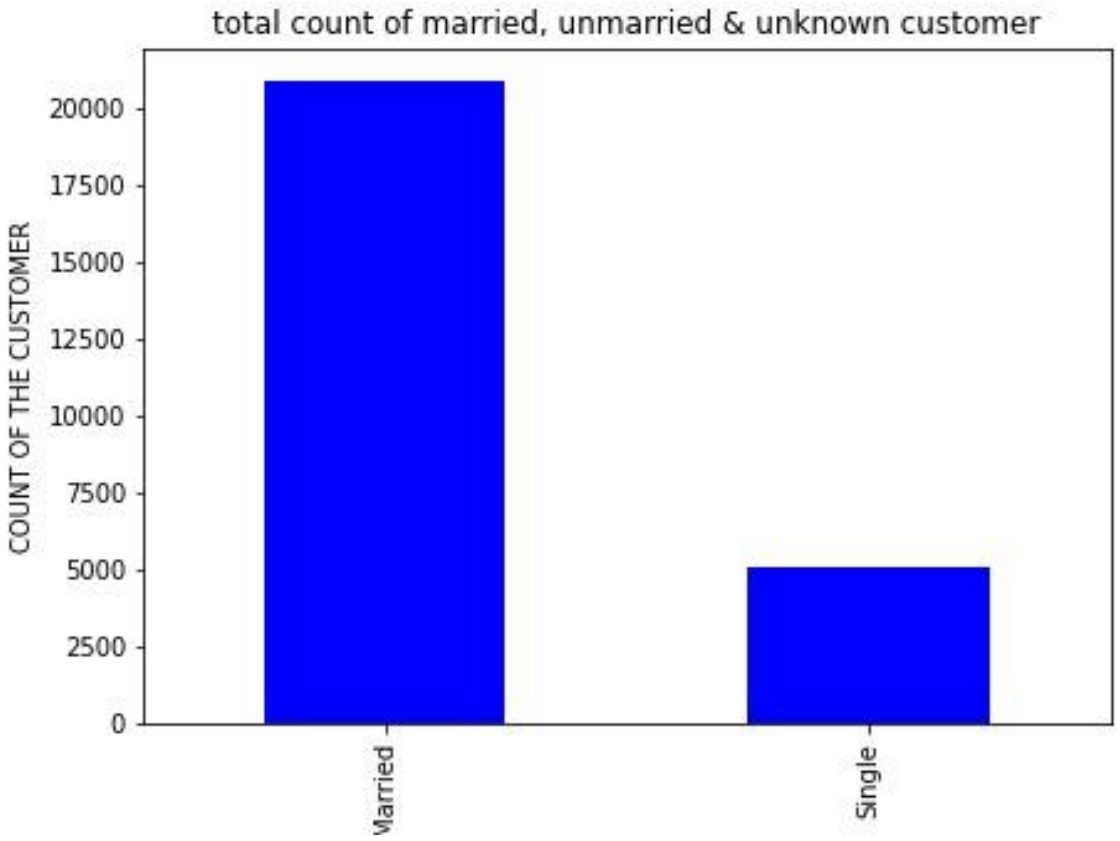


Figure 9. Martial Status Vs Customers Count

In this graph it can be see that more people are married and they buy more products from company and because of them company's sales increases . The count of unmarried customers in this dataset is 500 which is less as compared to married category.

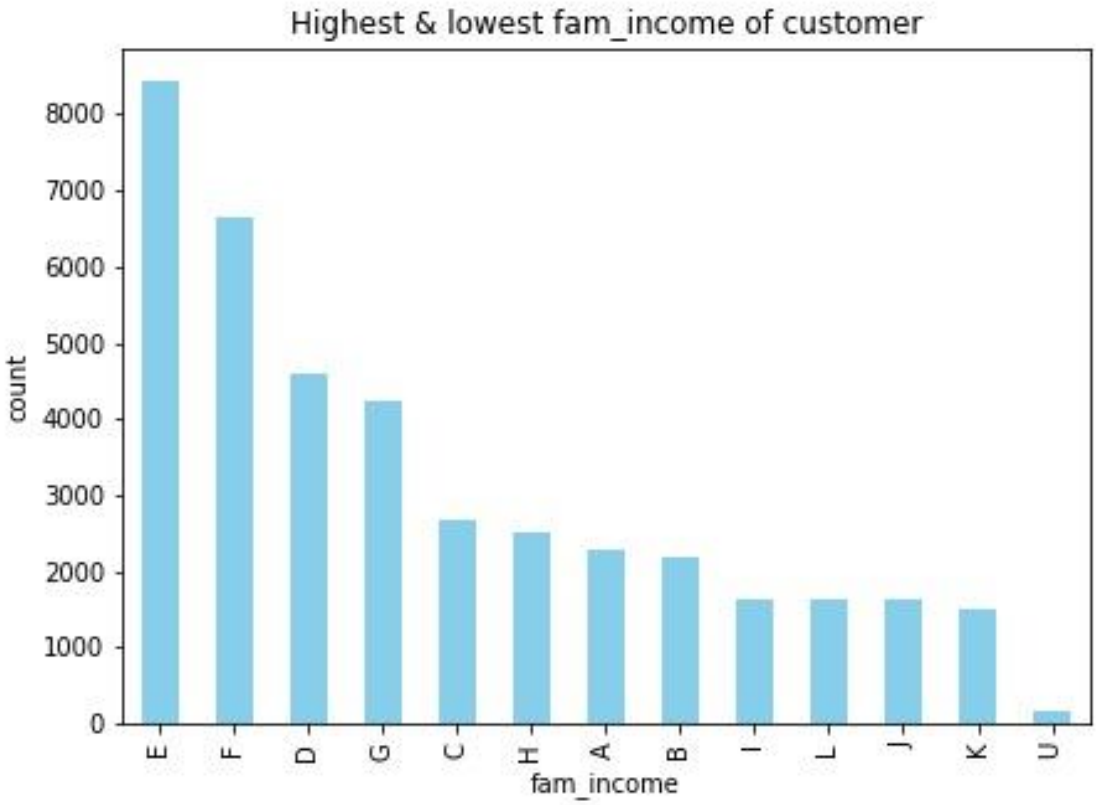


Figure 10. Family Income Vs Count of Customers

In this bar graph it is crystal clear, there are 13 level of family's income of customer. The customer of highest income comes in level 'E' and the count is 8400.on the others hand, the customer of least family income level comes in the level 'U'and the count of income is 200.

## Train-Test-Split Method

The whole dataset is separated into two parts which are training and testing sets by using train-test-split method. In which training set have 70% records and testing set have 30% records. Further I apply the under-sampling technique which is use to make the target variable.

## Modelling

In the modelling part there are four methods which are as follow:-

1.Logistic Regression

2. KNN

3. Naive Bayes

4. random forest

These four models are supervised learning classification algorithm.

Logistic Regression:  This model is used to predict the value of target variable. It is use to check the probability of target variable.

KNN: The KNN means “K-Nearest Neighbour”:- In this the number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.

It is use to check the similarity. It is a machine learning algorithm which can be used to solve classification as well as regression problem statements.

Naive Bayes:  It is based on Bayes' Theorem, It can create quick predictors because of it is fast machine learning model. It is use to check the strong independent assumptions between the features.

Random Forest: This is used in Classification and Regression problems. Due to it we can builds multiple decision trees on different samples by the result it can be find that which model is fit by seen accuracy. Which have highest result of accuracy then the model is best fit. In my dataset random forest is the best fit model.

| Column1 | Model | Train test split |
|---|---|---|
| 0 | Logistic Regression | 66.59166667 |
| 1 | KNN | 56.98333333 |
| 2 | Naive Bayes | 43.06666667 |
| 3 | random forest | 66.20833333 |

Table 2

### K-Folds Cross Validation:

The K-Folds operation which  is used to minimize the unfairness of the version. It partitioned the dataset into k-folds. On the random state, record is split into 5 folds. The first four folds are use for training and the fifth fold is used for testing. For the looking at a set Repeat it till each fold. At all five outputs are upload to get accuracy and it can be the version's metric of success.

| Column1 | Model | Train test split | kfolds_5 |
|---|---|---|---|
| 0 | Logistic Regression | 66.59166667 | 39.49689 |
| 1 | KNN | 56.98333333 | 11.13009 |
| 2 | Naive Bayes | 43.06666667 | 14.01036 |
| 3 | random forest | 66.20833333 | 50.48972 |

Table 3

### Stratified K Fold:

The stratified K fold is do the extension of cross validation technique for classify the problems . The preserving of share of samples in every class use to make the folds in it. In this data is divided in 5 stratified folds. In which, for healthy the version first 4 folds are use, and For further checking 5 fold is consider. Repeat it again and again until each fold have used as a take a look at set. Thereafter add all effects and calculate the common and it can be the version metric.

| Column1 | Model | Train test split | kfolds_5 | Stratifiedkfold_5 |
|---|---|---|---|---|
| 0 | Logistic Regression | 66.59166667 | 39.496886 | 55.20891182 |
| 1 | KNN | 56.98333333 | 11.130088 | 56.22795792 |
| 2 | Naive Bayes | 43.06666667 | 14.01036 | 54.53710675 |
| 3 | random forest | 66.20833333 | 50.489719 | 65.12576833 |

Table 4

### Random Train-Test-Split

This operation is use to merge the k-fold-cross-validation approach with typical train-test-splits. There the random divides of the information in the training-check set is done by me which is same as the move-validation technique, Thereafter repeating of process take place in splitting and for get the output. Here I divided the statistics into five Repeated Random Test-Train Splits.

| Column1 | Model | Train test split | kfolds_5 | Stratifiedkfold_5 | RRTestTrainSplits_5 |
|---|---|---|---|---|---|
| 0 | Logistic Regression | 66.59166667 | 39.49689 | 55.20891182 | 56.0113852 |
| 1 | KNN | 56.98333333 | 11.13009 | 56.22795792 | 56.28842505 |
| 2 | Naive Bayes | 43.06666667 | 14.01036 | 54.53710675 | 54.08349146 |
| 3 | random forest | 66.20833333 | 50.48972 | 65.12576833 | 65.17267552 |

Table 5

## Confusion Matrix Corresponding to Random Forest Classifier Algorithm

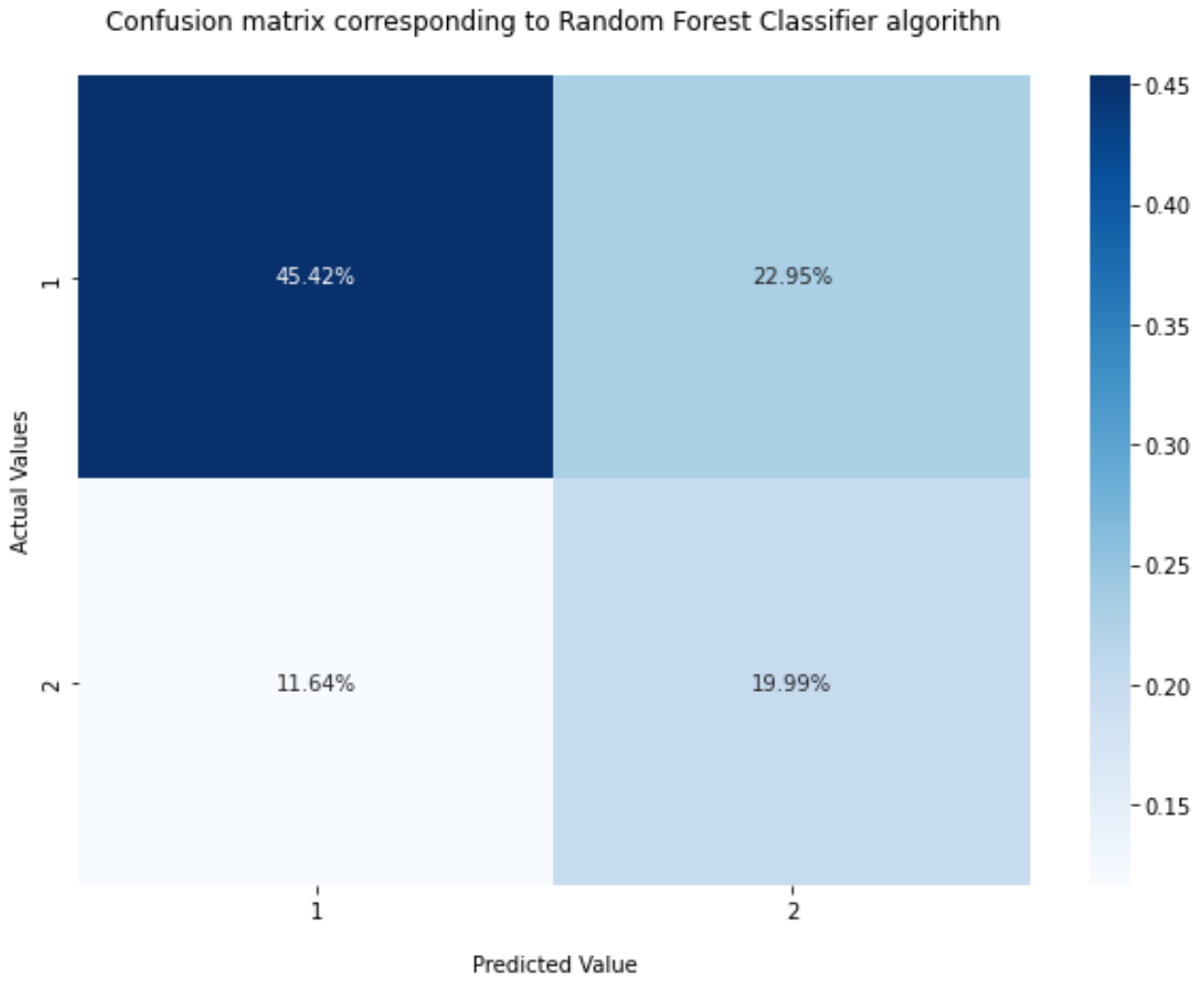


Figure 11. Confusion Matrix Corresponding to Random Forest

It can be clearly seen that there are two sections actual values and predicted values. If our actual value is 1 and predicted is also 1 then our accuracy is 45.42% and its true. On the other hand if actual value is 1 but predicted is 2 then its false. Afterthat if actual is 2 (11.64%) and predicted is also 2 (19.99%) then its true on the other hand if actual is 2 (11.64%) but predicted is 1 then its false.

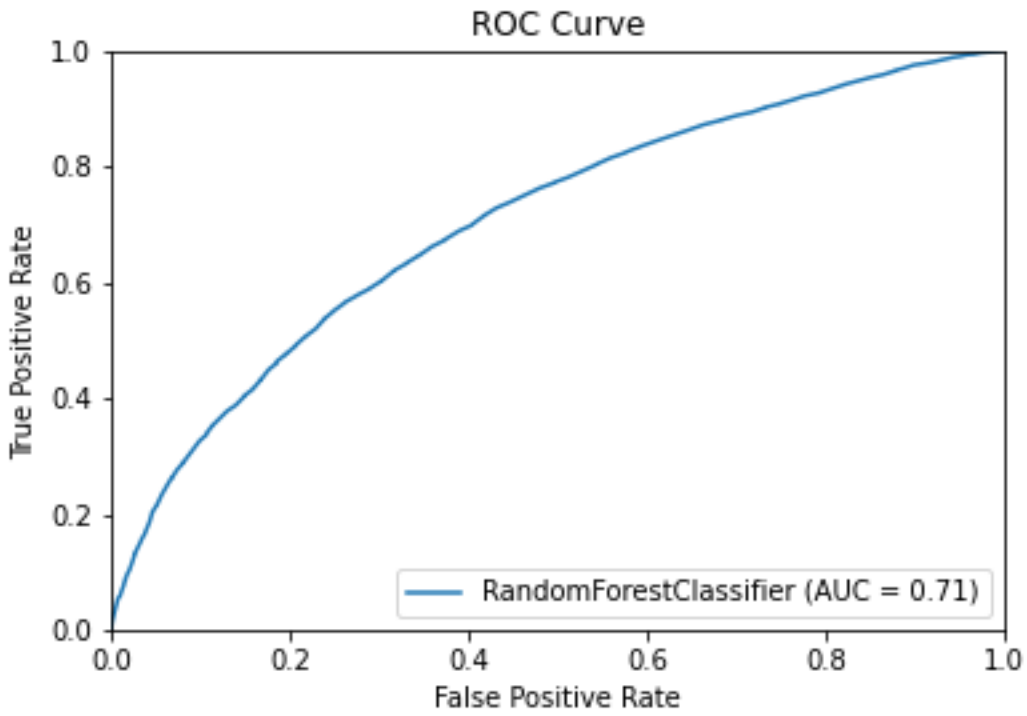


Figure 12

The Receiver operating characteristics is a metric to evaluate classifier output quality. The accuracy is 0.71.In figure 13, area under the curve is accurate.

Classification Report

| Column1 | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.79 | 0.67 | 0.72 | 8193 |
| 2 | 0.46 | 0.63 | 0.53 | 3791 |

Table 6

Precision – It represents predicted value are true which correspond to target variable

Recall –It reveals actual value that are true which correspond to target variable

F1 score:- it is weight It is positive class in binary classification of f1 which scores of each class for the multiclass task.

## Conclusion:

In conclusion, sales data show zig-zag trend which show 50-50 result. So, I found that half of the buyers get their target product. The random forest is the best fit model because it have the highest accuracy than other.

## Reference:


https://www.kaggle.com/datasets/mickey1968/individual-company-sales-data

D. J. Dalrymple, "Sales forecasting methods and accuracy," Business Horizons, vol. 18, pp. 69–73, 2006.

C. Dellarocas, X. Zhang, and N. F. Awad, "Exploring the value of online product reviews in forecasting sales: the case of motion pictures," Journal of Interactive Marketing, vol. 21, no. 4, pp. 23–45, 2007.

A. Tony, P. Kumar, and S. Rohith Jefferson, "A study of demand and sales forecasting model using machine learning algorithm," Psychology and Education Journal, vol. 58, pp. 10182–10194, 2021.

F. Zhu and X. Zhang, "Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics," Journal of Marketing, vol. 74, no. 2, pp. 133–148, 2010.

C.-P. Wei, Y.-M. Chen, C.-S. Yang, and C. C. Yang, “Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews,” Information Systems and E-Business Management, vol. 8, no. 2, pp. 149–167, 2010.

A. Tony, P. Kumar, and S. Rohith Jefferson, “A study of demand and sales forecasting model using machine learning algorithm,” Psychology and Education Journal, vol. 58, pp. 10182–10194, 2021.

F. Zhu and X. Zhang, “Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics,” Journal of Marketing, vol. 74, no. 2, pp. 133–148, 2010.

C.-P. Wei, Y.-M. Chen, C.-S. Yang, and C. C. Yang, “Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews,” Information Systems and E-Business Management, vol. 8, no. 2, pp. 149–167, 2010.