**PREDICTIVE & DESCRIPTIVE ANALYSIS ON SALES DATA**

Name: Gurpreet Kaur

Student ID: 0776519

Supervisor’s Name: Savita Seharawat

Capstone Project

Data analytics For Business | St. Clair College Ace Acumen

[Year]

Table o Contents

[**ABSTRACT:** 1](#_Toc101429457)

[**KEYWORDS:** 1](#_Toc101429458)

[**RESEARCH QUESTIONS:** 2](#_Toc101429459)

[**TOOLS:** 3](#_Toc101429460)

[**GITHUB ACCOUNT INFORMATION:** 3](#_Toc101429461)

[**Introduction:** 3](#_Toc101429462)

[**Literature Review:** 5](#_Toc101429463)

[**Methodology:** 8](#_Toc101429464)

[Figure 1. Methodology 9](https://d.docs.live.net/67e089ca5d286eea/Documents/Capstone%20Project(Final%20Report).docx#_Toc101429473)

[**DATA DICTIONARY:** 10](#_Toc101429476)

[**Detailed Data Dictionary:** 14](#_Toc101429477)

[**Data Cleaning:** 14](#_Toc101429478)

[Table 3. Missing Values 15](#_Toc101429479)

[Figure 3. Region vs Customers Count 16](#_Toc101429480)

[Figure 4. Gender Vs Car\_ probability 17](#_Toc101429481)

[Figure 5. Occupation Vs Customers Count 18](#_Toc101429482)

[Figure 6. Online Vs Customer Count 19](#_Toc101429483)

[Figure 7. Whether the customer has bought the target product or not 20](#_Toc101429484)

[Figure 8. Whether the customer has online shopping experience or not 21](#_Toc101429485)

[Figure 9. Total count of Male & Female 22](#_Toc101429486)

[Figure 10. Total count of married & unmarried customer 23](#_Toc101429487)

[Figure 11. Highest & lowest family income of customer 24](#_Toc101429488)

[**Experimental Design** 25](#_Toc101429489)

[**Train-Test-Split Method** 25](#_Toc101429490)

[**Modelling** 25](#_Toc101429491)

[**Cross Validation:** 26](#_Toc101429492)

[**K-Folds Cross Validation** 26](#_Toc101429493)

[Table 5. Accuracy of Models corresponding to kfolds\_5 27](#_Toc101429494)

[**Stratified K Fold:** 27](#_Toc101429495)

[Table 6. Accuracy of Models cross ponding to Stratifiedkfold\_5 27](#_Toc101429496)

[**Repeated Random Train-Test-Split** 27](#_Toc101429497)

[Table 7. Accuracy of Models cross ponding to Repeated Random Train-Test-Split 28](#_Toc101429498)

[**Confusion Matrix Corresponding to Random Forest Classifier Algorithm** 29](#_Toc101429499)

[Figure 12. Confusion Matrix Corresponding to Random Forest 29](#_Toc101429500)

[Figure 13. ROC Curve 30](#_Toc101429501)

[Classification Report 30](#_Toc101429502)

[Table 8. Classification Report 30](#_Toc101429503)

[**Conclusion:** 31](#_Toc101429504)

[**Reference** 31](#_Toc101429505)

## **ABSTRACT:**

We have selected Sales Datafrom 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 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 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. Rather than spending money to marketplace a brand-new product to each purchaser within the organization's database, the organisation can analyze which patron segment is probable to shop for the product and then marketplace the product to handiest that precise segment 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.

## **KEYWORDS:**

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

## **RESEARCH QUESTIONS:**

1. By predictive analysis find out whether the customer has bought the target product or not?

In this pie chart Y shows that the percentage of customer who bought their target product and N shows that the percentage of customer who do not purchase their target products. It is clear from the pie chart that the percentage of both type customer who purchase their order product or who do not buy their products are almost same with small difference. As 50.1% customer received their order products and 49.9% do not bought their target products. So, from this percentage corresponding to both type customer its is predicted that company does not have more profit. As half of the count of total customer do not satisfy from the service of company because they do not buy their target product which they want to buy.

1. Find out how many consumers have online shopping experience and how many do not have online experience?

In this dataset we find the consumers who have online shopping experience which are denoted by Y and who do not have online experience those are denoted by N by plotting the pie chart. With the help of pie chart, it is found that 50.4% customers have online shopping experience and customers with the percentage of 49.6% do not have online shopping experience in this dataset. So, the number of counts of both customers is almost equal.

1. Find out total number of male and female in the dataset by descriptive analysis?

In this bar graph, M represent male, F female and U unknowns. It is found that more male count included in sale of products as the count of male is highest (22500) as compared to other two categories female and unknown. Females have approximately (17000) count in this dataset who bought the products which are on second position. Whereas, there are some unknown values in gender attributes the count of unknown is approximately 150. Least count of gender is denoted in the category of unknown.

1. By predictive analysis find out which type of consumers increased the sales of products married and unmarried in the dataset?

In married and unmarried bar graph it is found that more count of customers is married in this dataset that is approximately 21000 who purchased more products from the company and increased the sale of company. The count of unmarried customers in this dataset is 500 which is less as compared to married category.

1. Find out how many customers have highest and lowest family income?

From this bar graph it clear that in this dataset there are 13 levels of family income in which customers are divided. The customer who has highest income the found in level E and the count of those customers are 8400. Whereas, the customer who have least family income level they are found in the level U. The count of customer who have least family income is 200.

## **TOOLS:**

Whole data visualization will be done by using python.

## **GITHUB ACCOUNT INFORMATION:**

Gurpreet Kaur GitHub account link below:

<https://github.com/GurpreetKaur519>

**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, FabioTraverso (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:**

Firstly, our dataset is uncleaned like there are some missing values and duplicate values. First, we will clean our dataset to perform father operations. In methodology part we will use regression model to find out relationship between dependent and independent variables. We will also use descriptive and predictive analysis to find out how many customers are satisfied from the sales of individual company sales data. On Individual Company Sales Data we will also use classification and clustering. By using these different models, we will collect different results for sale

## 

**Data Visualisation**

## 

**Data Pre-processing**

## 

**Experimental Design**

## 

**Modelling Implementation**

## 

## 

## Figure 1. Methodology

## 

**Conclusion**

## **DATA DICTIONARY:**

Table 1. Categorical Attributes

| **Attribute Name** | **Description** | **Data Type** | **NO. of Levels** | **Value Counts** |
| --- | --- | --- | --- | --- |
| **Flag** | Consumer has purchased the target productor or not | category | 2 | Y 13334  N 10196 |
| **Gender** | Gender of the customer whether he is male or female | category | 3 | M 14186  F 8938  U 406 |
| **Education** | Educational degree of customer | category | 5 | 2.Some College 6546  3.Bach 5987  1. HS 5038  4.Grad 3910  0.<HS 2049 |
| **Age** | Age of the customer by group | category | 7 | 5\_<=55 6089  4\_<=45 4986  6\_<=65 3986  1\_Unk 3061  3\_<=35 2515  7\_>65 2306  2\_<=25 587 |
| **Online** | Customer have online experience or not | category | 2 | Y 16516  N 7014 |
| **Customer psychology** | Customer psychology based on the area of residence  where customer live | category | 10 | C 5251  B 5151  E 3749  G 2324  F 2189  D 1316  J 1201  I 1163  A 804  H 382 |
| **Marriage** | Customer marital status whether customer married or unmarried | category | 2 | Married 19266  Single 4264 |
| **Child** | Consumer has children or not | category | 3 | Y 11174  N 7307  U 5049 |
| **Occupation** | Information about customer career | category | 6 | Professional 9818  Sales/Service 6626  Blue Collar 3983  Retired 2018  Others 916  Farm 169 |
| **Mortgage** | Information about customer has any housing loan | category | 3 | 1Low 16458  3High 3838  2Med 3234 |
| **House owner** | customer owns a house or not | category | 2 | Owner 18778  Renter 4752 |
| **Region** | Information about area where customer lived | category | 5 | South 9174  West 5128  Midwest 4883  Northeast 4207  Rest 138 |
| **Family income** | Information about customer family income | category | 13 | E 4881  F 4099  G 2606  D 2469  H 1602  C 1357  A 1180  B 1118  L 1082  I 1077  J 1076  K 970  U 13 |

## 

## Table 2. Numerical 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:**

For Categorical Attributes first I took the one variable at a time to check the datatype and assigned the correct datatype. Then I check the number of levels corresponding to each attribute and count of values corresponding to each level.

For numerical attributes I used five number summary that is mean, minimum, maximum, count, standard deviation.

## **Data Cleaning:**

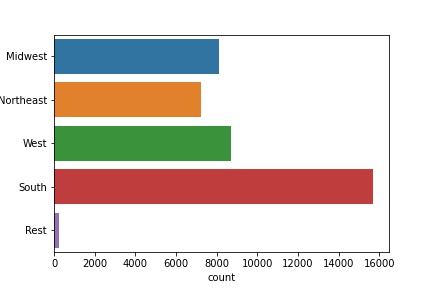
Duplication of values in the dataset: In this dataset there are 55 duplicated values. These values are dropped from the dataset by using the drop duplicated command.

Missing values in the dataset: There are total 18106 missing values in this dataset. In education attribute there are 735 null values, in House owner 3369 missing vales and in marriage attribute 14002 missing values. These values are filled by 0 for data cleaning.

## Table 3. Missing Values

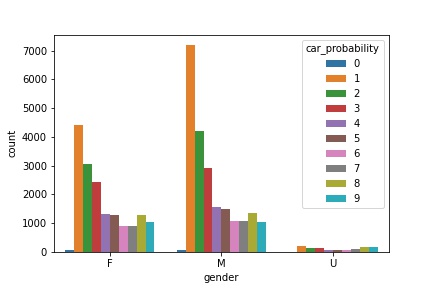
| **Missing Values** | **Count** |
| --- | --- |
| education | 735 |
| House owner | 3369 |
| marriage | 14002 |

**Data Visualization:**

****

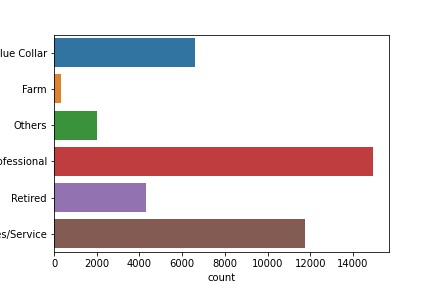
## Figure 3. Region vs Customers Count

This bar graph represents the information cross ponding to customers count from different region. The highest count of customer belongs to south region that is 15676. The count of customers from Midwest and west almost same with small difference that is 8107 from Midwest and 8725 from west. The count of rest of region is 245.

****

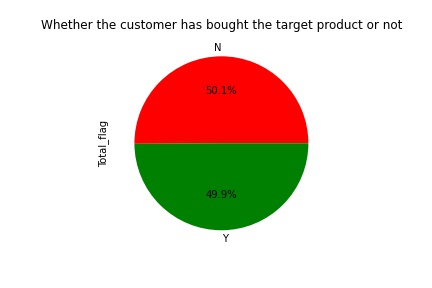
## Figure 4. Gender Vs Car\_ probability

This bar graph shows the information gender cross ponding car probability. This graph shows the count of female, male cross ponding to the count of car probability. As the car probability of 1 in both female, male is highest. And the probability of 4 to 9 cars in female, male is almost same. There are some unknown values in gender attribute the count of car probability is low as compared to male and female.



## Figure 5. Occupation Vs Customers Count

This graph represents the information of customers occupation in different fields. As the highest count of customers count lies in the professional category that is 15000. the sales/service category is second highest category with 11500 counts of customers. In blue collar the count of customer is 6300 and the count of customers in in farm and others is lowest as compared to other four categories.



## Figure 7. Whether the customer has bought the target product

In this pie chart Y shows that the percentage of customer who bought their target product and N shows that the percentage of customer who do not purchase their target products. It is clear from the pie chart that the percentage of both type customer who purchase their order product or who do not buy their products are almost same with small difference. As 50.1% customer received their order products and 49.9% do not bought their target products. So, from this percentage corresponding to both type customer its is predicted that company does not have more profit. As half of the count of total customer do not satisfy from the service of company because they do not buy their target product which they want to buy.



## Figure 8. Whether the customer has online shopping experience or not

In this dataset we find the consumers who have online shopping experience which are denoted by Y and who do not have online experience those are denoted by N by plotting the pie chart. With the help of pie chart, it is found that 50.4% customers have online shopping experience and customers with the percentage of 49.6% do not have online shopping experience in this dataset. So, the number of counts of both customers is almost equal.

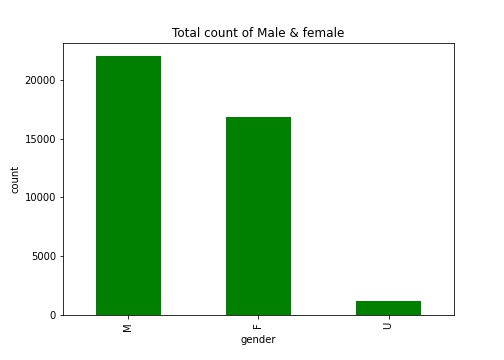
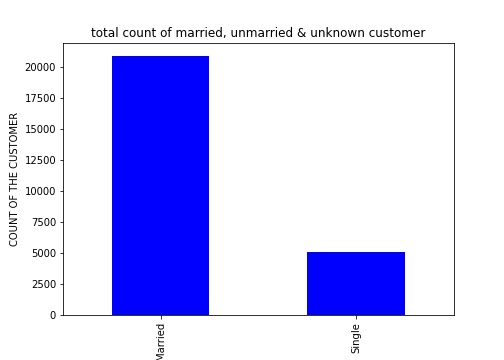


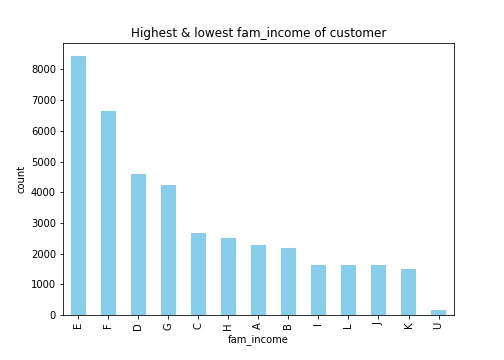
Figure 9. Total count of Male & Female

In this bar graph, M represent male, F female and U unknowns. It is found that more male count included in sale of products as the count of male is highest (22500) as compared to other two categories female and unknown. Females have approximately (17000) count in this dataset who bought the products which are on second position. Whereas, there are some unknown values in gender attributes the count of unknown is approximately 150. Least count of gender is denoted in the category of unknown.



## Figure 10. Total count of married & unmarried customer

In married and unmarried bar graph it is found that more count of customers is married in this dataset that is approximately 21000 who purchased more products from the company and increased the sale of company. The count of unmarried customers in this dataset is 500 which is less as compared to married category.



## Figure 11. Highest & lowest family income of customer

From this bar graph it clear that in this dataset there are 13 levels of family income in which customers are divided. The customer who has highest income the found in level E and the count of those customers are 8400. Whereas, the customer who have least family income level they are found in the level U. The count of customer who have least family income is 200.

## **Experimental Design**

## **Train-Test-Split Method**

In train-test-split method the entire dataset is partitioned into training and testing sets. The training set contain 70% records and, in the testing, set it contain 30% records. Then I apply the under-sampling technique to make my target variable balance.

## **Modelling**

In the modelling part there are four methods Logistic Regression, KNN, Naive Bayes, random forest that I have applied on my dataset.

Logistic Regression: This model is supervised learning classification algorithm. It is used to predict the value of target variable.

KNN: The KNN stands for “K-Nearest Neighbour”. It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.

Naive Bayes:  Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors.

Random Forest: It is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Table 4. Accuracy of Models cross ponding to Train Test Split

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

To check the accuracy of my dataset corresponding to each model. To find the best fit model for the dataset. I have found that the random forest is the best fit model for my dataset with highest accuracy 66% as compared to other models.

## **Cross Validation:**

## **K-Folds Cross Validation**

The K-Folds technique Divide the dataset into k-folds in a series. At random, I divided the records into 5 folds. The four folds are then used for training part, and the fifth fold is used for testing. Repeat till each fold has been used as a take a look at set. Then upload together all the outcomes and calculate the average. That may be the version's metric of success.

## Table 5. Accuracy of Models corresponding to kfolds\_5

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

## 

## **Stratified K Fold:**

This pass-validation object returns stratified folds and is a variation of K-Fold. The folds are made by using preserving the share of samples in every class. I divided the data into 5 stratified folds. The 4 folds are then used to healthy the version, and the 5th fold is used to check it. Repeat until each fold has been used as a take a look at set. Then add collectively all of the effects and calculate the common. That might be the version's metric of success.

## Table 6. Accuracy of Models cross ponding to Stratifiedkfold\_5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |

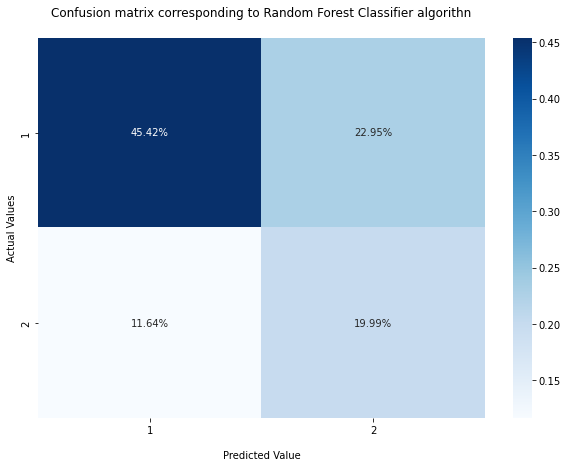
## **Repeated Random Train-Test-Split**

This method combines the k-fold-cross-validation approach with typical train-test-splits. I create random divides of the information in the training-check set, similar to the move-validation technique, and then repeat the procedure of splitting and trying out the algorithm many times. I divided the statistics into five Repeated Random Test-Train Splits.

## Table 7. Accuracy of Models cross ponding to Repeated Random Train-Test-Split

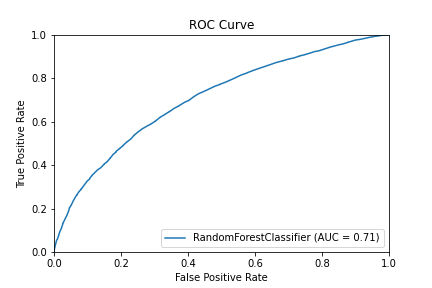
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column1** | **Model** | **Train test split** | **kfolds\_5** | **Stratifiedkfold\_5** | **RRTestTrainSplits\_5** |
| 0 | Logistic Regression | 66.59166667 | 39.49689 | 55.20891182 | 56.0113852 |
| 1 | KNN | 56.98333333 | 11.13009 | 56.22795792 | 56.28842505 |
| 2 | Naive Bayes | 43.06666667 | 14.01036 | 54.53710675 | 54.08349146 |
| 3 | random forest | 66.20833333 | 50.48972 | 65.12576833 | 65.17267552 |

## **Confusion Matrix Corresponding to Random Forest Classifier Algorithm**

****

## Figure 12. Confusion Matrix Corresponding to Random Forest

When the target variable's actual value is 1 and the predicted value is also 1 and 45.42% of observation fall into first quadrant. 22.95% of observations fall into the second quadrant; whilst the target variable's actual value is 1 and predicted value is 2. 11.64 % of observations fall into the third quadrant; and while the target variable's actual value is 2 and predicted value 1, 19.99% of observations fall into the fourth quadrant. The actual and predicted values are each 2.



## Figure 13. ROC Curve

Receiver operating characteristics it is metric to evaluate classifier output quality. The accuracy is 0.71. The area under curve is right.

## Table 8. Classification Report

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

Precision - it shows the predicted value that are true cross ponding to target variable

Recall - it shows the actual value that are true cross ponding to target variable

F1 score – It is positive class in binary classification or weighted average of the f1 scores of each class for the multiclass task.

## **Conclusion:**

In sales dataset I have found that in which direction company sales moving upward and downward. And which type customers increase the sale of company. So, I found that half of the customers receive their target product. So, needs to put more efforts to increase sale and satisfied their customers. In the modelling part I found that the random forest is the best fit model for this dataset with highest accuracy as compared to other models.

## **Reference**

Guozhen Chang, FabioTraverso, MichelGadomsky, nadia hajrasi.

Mickey. (2019) Knowledge creation in sales marketing using machine learning techniques

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

Rondinelly oliveira Knowledge creation in sales marketing using machine learning techniques

https://www.kaggle.com/code/rondinellyoliveira/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.