# Analysis and Prediction of Annual Sales

**A PROJECT DOCUMENT**

***Submitted by***

**SRI SURYA VAMSI SARIPUDI [UB50357055]**
**YASHWANTH REDDY KATAM [UB50361872]**

## ABSTRACT

With advent of the internet and the e-commerce, it has since been the choice of the large section of the consumers to order the commodities online through e-commerce platforms. However, there is a huge competition between the E-commerce providers in expanding their market and improving their business strategies. But many of the firms are unaware of the factors that are responsible in improving their business model. Hence, they're lacking behind benefiting from huge potential of market and in analyzing the possible pitfalls in the sales strategies. We would analyze the sales dataset of the company with strong foot-hold in the United states and have a aggressive strategy for expanding to other emerging economies. We would clean the dataset that is mostly collated haphazardly and by imposing few questions, we will extract the cues from the data set itself. Finally we would develop a training model that would suggest the near accurate solution for the hypothesis we base it upon, thus it is a major cost cutting step and will benefit the company and its employees to concentrate on the resources.

**Keywords:**

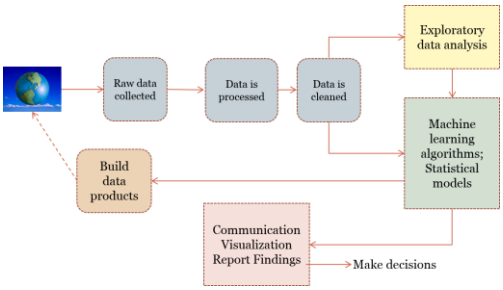
EDA – exploratory data analysis

LRM- logistic regression model

SVM- support vector machines

## INTRODUCTION

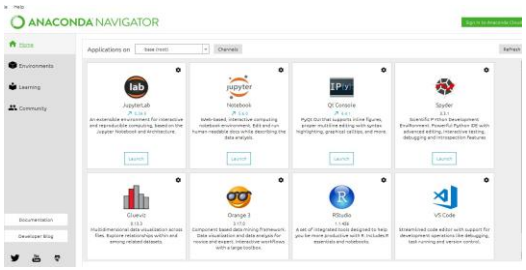
Forecasting sales is a common and essential use of machine learning (ML). Sales forecasts can be used to identify benchmarks and determine incremental impacts of new initiatives, plan resources in response to expected demand, and project future budgets. In this article, We will show how to implement 5 different ML models to predict sales. We have followed the below System Architecture Diagram in order to successfully get the end result.




## SOFTWARE REQUIREMENTS

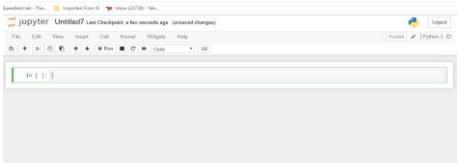
**Anaconda Navigator :-**

Anaconda Navigator is a desktop graphical user interface (GUI) included in AnacondaÂő distri- bution that allows you to launch applications and easily manage conda packages, environmentsand channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, macOS and Linux




**Jupyter notebook :-**

The Jupyter Notebook is an open-source web application used for creating and sharing docu- ments containing code,visualizations ,texts etc. It is used for data cleaning, transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.



**PHASE 0 (DATA COLLECTION) :-**

In Sales Analysis prediction problems huge amounts of data is available thus it is important to select the right data from the vast data science world. The data set is obtained from Kaggle Website, which is not a cleaned dataset and is used in this paper for the Sales analysis and prediction of future sales. This data set comprises 6 attributes and 7000 records or tuples.

| | Order ID | Name_Of_Purchase | Purchased_Quantity | Cost_Each | Order Date | Purchase_Address |
|---|---|---|---|---|---|---|
| 0 | 176558 | USB-C Charging Cable | 2 | 11.95 | 4/19/2019 8:46 | 917 1st St, Dallas, TX 75001 |
| 1 | NaN | NaN | NaN | NaN | NaN | NaN |
| 2 | 176559 | Bose SoundSport Headphones | 1 | 99.99 | 4/7/2019 22:30 | 682 Chestnut St, Boston, MA 02215 |
| 3 | 176560 | Google Phone | 1 | 600 | 4/12/2019 14:38 | 669 Spruce St, Los Angeles, CA 90001 |
| 4 | 176560 | Wired Headphones | 1 | 11.99 | 4/12/2019 14:38 | 669 Spruce St, Los Angeles, CA 90001 |

**PHASE 1 (DATA CLEANING/PRE-PROCESSING) :-**

This phase is an very important step which prepares the dataset for the modelling part. It
Significantly improves performance metrics and boosts prediction accuracy. It consists of detecting outliers. There can be different kinds of outliers which mislead the data and gives us unnecessary results. So it is best if these tuples are removed from the dataset as they will introduce noise in the modelling phase leading to incorrect predictions. Missing values are also treated in this phase. It is better to replace missing fields with any sort of measures of central tendency. After detecting outliers and treatment of missing values feature selection is done, some

columns are added to the existing raw dataset for exploring the dataset in coming phase i.e the EDA/ Exploratory Data Analysis. After completion of the PHASE 1 the dataset is as follows :-

| | Order ID | Name_Of_Purchase | Purchased_Quantity | Cost_Each | Sales | Order Date | Purchase_Address | Month | City | Hour | Minute |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 176560 | Google Phone | 1 | 600.00 | 600.00 | 2019-04-12 14:38:00 | 669 Spruce St, Los Angeles, CA 90001 | 4 | Los Angeles | 14 | 38 |
| 4 | 176560 | Wired Headphones | 1 | 11.99 | 11.99 | 2019-04-12 14:38:00 | 669 Spruce St, Los Angeles, CA 90001 | 4 | Los Angeles | 14 | 38 |
| 18 | 176574 | Google Phone | 1 | 600.00 | 600.00 | 2019-04-03 19:42:00 | 20 Hill St, Los Angeles, CA 90001 | 4 | Los Angeles | 19 | 42 |
| 19 | 176574 | USB-C Charging Cable | 1 | 11.95 | 11.95 | 2019-04-03 19:42:00 | 20 Hill St, Los Angeles, CA 90001 | 4 | Los Angeles | 19 | 42 |
| 30 | 176585 | Bose SoundSport Headphones | 1 | 99.99 | 99.99 | 2019-04-07 11:31:00 | 823 Highland St, Boston, MA 02215 | 4 | Boston | 11 | 31 |
| 31 | 176585 | Bose SoundSport Headphones | 1 | 99.99 | 99.99 | 2019-04-07 11:31:00 | 823 Highland St, Boston, MA 02215 | 4 | Boston | 11 | 31 |
| 32 | 176586 | AAA Batteries (4-pack) | 2 | 2.99 | 5.98 | 2019-04-10 17:00:00 | 365 Center St, San Francisco, CA 94016 | 4 | San Francisco | 17 | 0 |

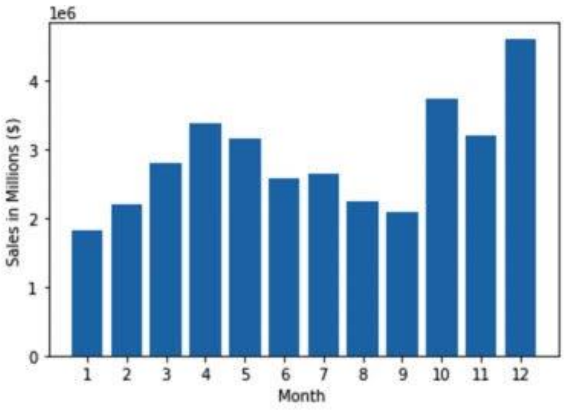
**PHASE 2 (Exploratory Data Analysis / EDA) : -**

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. Based on the hypothesis we built we have performed 13 EDA tasks which are displayed as follows :-

**1.Operation Name: Magnitude of sales on monthly basis.**

Reason for doing: To get the overall picture of the sales happening over the months, such that the inventories will be updated.

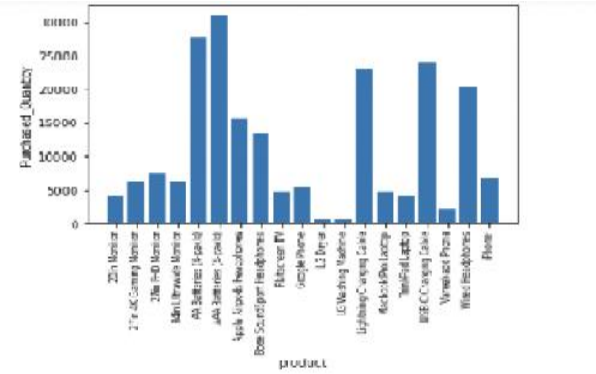
Outcome: A general idea about the sales figure was developed and we were able to figure out maximum and minimum sales figures.

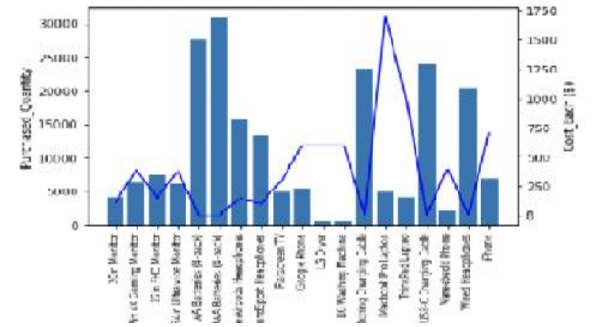



**2.Operation Name: Product and sales correlation**

Reason for doing: To uncover the structural relation between each product and it's sales occured, so as to update the firm regarding its future purchases.

Outcome: A relationship was established between the considered elements with the help of Histogram plot and was able to find out the correlation between sales and product.

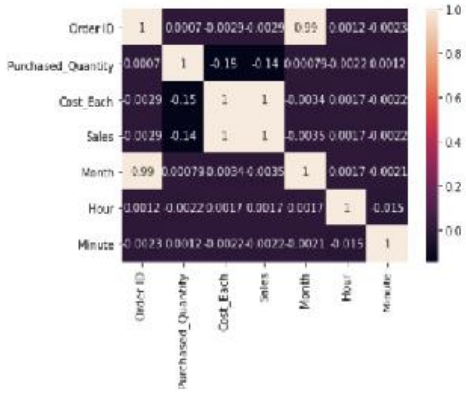


**3.Operation Name: General Correlation using heat-map.**

Reason for doing: To develop the correlation matrix among all the elements of the dataframe, that give further insights into behaviour of the data elements, on mutual dependency.

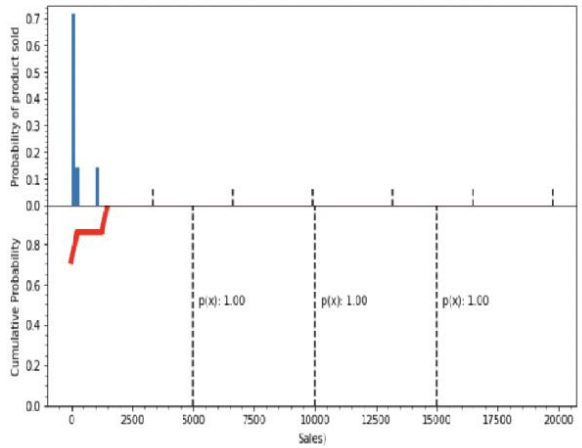
Outcome: We could built a general correlation, that enabled us to witness the degree of mutual dependencies.




**4.Operation Name: Cumulative Probability**

Reason for doing: To find out the cumulative probability that exists between the Product and it's sales figures.

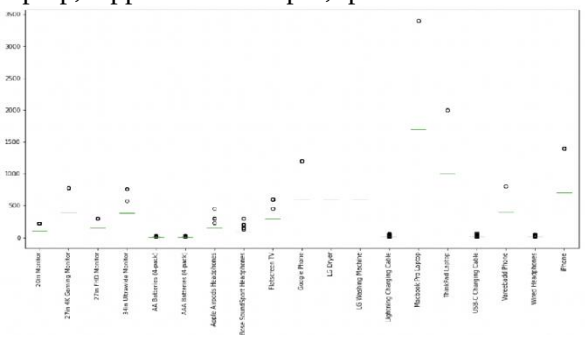
Outcome: The plot helped us identify the narrowness in the window of the price range, as the large chunk of sales occured in the very small window.




**5.Operation Name: Multi-varite Distributional data Summarisation based on products**

Reason for doing: By plotting the multivarite nature of products on x-axis, we want to notice any abnormal behaviour in the sales of any product.

Outcome: We were able to find out the abnormal price versus purchase relation in the products as google phone,thinkpad laptop, Apple Macbook pro, iphone.



**6.Operation Name: Statistical Calculation of Mean, Standard Deviation and corresponding Z Score**

Reason for doing: To define the behaviour of each individual purchase based on the z score and isolate the purchases with higher z- score .

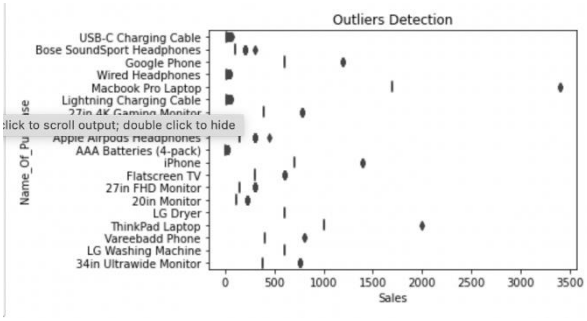
Outcome: We were able to find the z-score values of all purchases and specifically isolated those with a zscore>3, this helps in the identificaion of the higher average purchases.

| | Order ID | Name_Of_Purchase | Purchased_Quantity | Cost_Each | Sales | Order Date | Purchase_Address | Month | City | Hour | Minute | State | zscore |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 176565 | Macbook Pro Laptop | 1 | 1700.0 | 1700.0 | 2019-04-24 10:38:00 | 915 Willow St, San Francisco, CA 94016 | 4 | San Francisco | 10 | 38 | CA | 4.549171 |
| 85 | 176639 | Macbook Pro Laptop | 1 | 1700.0 | 1700.0 | 2019-04-28 15:14:00 | 853 Cedar St, San Francisco, CA 94016 | 4 | San Francisco | 16 | 14 | CA | 4.549171 |
| 89 | 176643 | Macbook Pro Laptop | 1 | 1700.0 | 1700.0 | 2019-04-27 21:32:00 | 373 Adams St, Boston, MA 02215 | 4 | Boston | 21 | 32 | MA | 4.549171 |
| 92 | 176646 | Macbook Pro Laptop | 1 | 1700.0 | 1700.0 | 2019-04-22 07:14:00 | 657 Adams St, Portland, OR 97035 | 4 | Portland | 7 | 14 | OR | 4.549171 |
| 135 | 176687 | Macbook Pro Laptop | 1 | 1700.0 | 1700.0 | 2019-04-30 10:25:00 | 121 Wilcox St, Los Angeles, CA 90001 | 4 | Los Angeles | 10 | 25 | CA | 4.549171 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 185619 | 259041 | Macbook Pro Laptop | 1 | 1700.0 | 1700.0 | 2019-09-13 23:14:00 | 220 Sunset St, New York City, NY 10001 | 9 | New York City | 23 | 14 | NY | 4.549171 |
| 185739 | 259157 | Macbook Pro Laptop | 1 | 1700.0 | 1700.0 | 2019-09-12 05:01:00 | 326 Forest St, Seattle, WA 98101 | 9 | Seattle | 5 | 1 | WA | 4.549171 |
| 185888 | 259299 | Macbook Pro Laptop | 1 | 1700.0 | 1700.0 | 2019-09-30 23:59:00 | 249 Chestnut St, Los Angeles, CA 90001 | 9 | Los Angeles | 23 | 59 | CA | 4.549171 |
| 185898 | 259308 | Macbook Pro Laptop | 1 | 1700.0 | 1700.0 | 2019-09-26 19:18:00 | 912 8th St, Boston, MA 02215 | 9 | Boston | 19 | 18 | MA | 4.549171 |
| 185915 | 259324 | Macbook Pro Laptop | 1 | 1700.0 | 1700.0 | 2019-09-29 13:37:00 | 326 North St, San Francisco, CA 94016 | 9 | San Francisco | 13 | 37 | CA | 4.549171 |

**7.Operation Name: Detection of Outliers**

Reason for doing: To identify outliers if any, in our data set and discard them in order to prevent abnormal behaviour in our data.

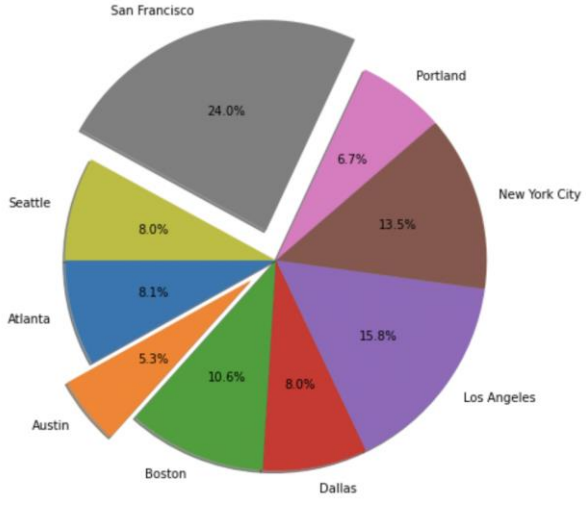
Outcome: We were able to identify few outliers especially in products as wired headphones, LG washing machine etc.




**8.Operation Name: Covariation in the sales frequency among different cities.**

Reason for doing: To find out the respective behaviour of sales frequency among the different cities.

Outcome: We have found out the data and plotted as histogram and piechart, this helps the firm to strategically locate their inventories by forecasting the future sales figures.




**9.Operation Name: Identifying the skewness in the data elements**

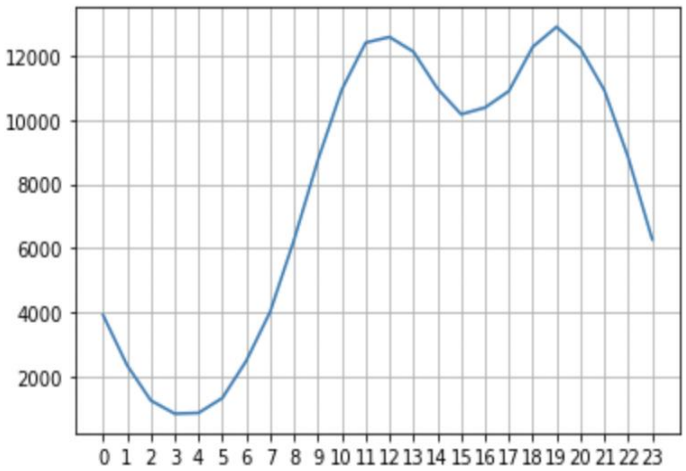
Reason for doing: To identify the skewness of the data elements to identify the probability distribution.

Outcome: We were able to identify that our data is positively skewed, i.e heavily right skewed.


**10.Operation Name: Develop Ordering Hours and Sales Frequency relation**

Reason for doing: TO find out the customer purchasing behaviour at different hours of the day.

Outcome: We were able to observe that 11.am and 7.pm are two most busiest hours, the firm was informed so they can pitch up the endorsements during these hours




**11.Operation Name: Combination of products ordered together¶**

Reason for doing: To identify the structural relation in the customer's behaviour, while purchasing a group of products.

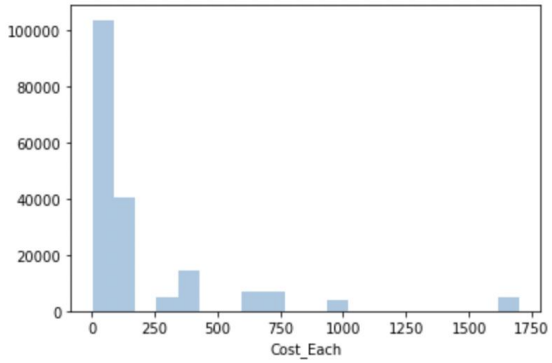
Outcome: The results that follows, showcases the combination of products that were frequently brought together, this will help the firm in giving relevent suggestions to customers based on the products purchased.

```
('iPhone', 'Lightning Charging Cable') 1005
('Google Phone', 'USB-C Charging Cable') 987
('iPhone', 'Wired Headphones') 447
('Google Phone', 'Wired Headphones') 414
('Vareebadd Phone', 'USB-C Charging Cable') 361
('iPhone', 'Apple Airpods Headphones') 360
('Google Phone', 'Bose SoundSport Headphones') 220
('USB-C Charging Cable', 'Wired Headphones') 160
('Vareebadd Phone', 'Wired Headphones') 143
('Lightning Charging Cable', 'Wired Headphones') 92
@@The sorting of products using a pivot table@@:

<bound method DataFrame.sort_values of Name_Of_Purchase  20in Monitor  27in 4K Gaming Monitor  27in FHD Monitor  \
Sales               110.740968              390.866382        150.849141

Name_Of_Purchase  34in Ultrawide Monitor  AA Batteries (4-pack)  \
Sales                        381.096588               5.157137

Name_Of_Purchase  AAA Batteries (4-pack)  Apple Airpods Headphones  \
Sales                          4.49304                151.080455

Name_Of_Purchase  Bose SoundSport Headphones  Flatscreen TV  Google Phone  \
Sales                             100.98052       301.1875    600.760181

Name_Of_Purchase  LG Dryer  LG Washing Machine  Lightning Charging Cable  \
Sales                600.0               600.0                  16.02614

Name_Of_Purchase  Macbook Pro Laptop  ThinkPad Laptop  USB-C Charging Cable  \
Sales                   1701.439458      1000.474491             13.080457

Name_Of_Purchase  Vareebadd Phone  Wired Headphones      iPhone
Sales                 400.581114         13.053619  700.716165  >
```

**12.Operation Name: Determining the density distribution for cost_ Each variable.**

Reason for doing: This is to plot the cost_each variable data against the density distribution.

Outcome: We observed that the prices between 0(Dollars)- 150(Dollars) has a lot of density while compared to the other prices. This helps the firm to introduce more products ranging between the above mentioned price range which eventually increases their sales.




**Model Training and Performance Report**

Factors considered in measuring performance are accuracy, sensitivity, specificity and precision . Accuracy in the sense how good the classifier is performing sensitivity means how perfect the classifier is with respect to positive entries. There are 2 kinds of tuples ,positive and negative tuples. positive tuples obey some specific rules whereas negative tuples do not.Factors taken into consideration true positive(TP) tuples under consideration is perfectly positively classified i.e expected and observed tuple were positive. True negative(TN):if tuple was observed negative and expectation was same False positive(FP): data tuples were mistakenly classified as positive however expected outcome was the opposite . False negative(FN) were actually supposed to be classified as positive but were shown to be negative. We can calculate the performance measures of the created classifiers by using the following equations representing the relationship b/w the tuples to the performance measuring factors.

**Performance Metrics**

**1.Accuracy**

It is the ratio of the correctly classified tuples to the entire collection of them. Accuracy answers the following question: How many customers did we correctly label out of all the customers.

Accuracy=(TP+TN)/(TP+FP+FN+TN).

**2.Precision**

Precision is defined the ratio of the correctly +ve tuples classified by the model to all +ve entries. Precision answers the following: How many of those classified as churners will actually churn.

Precision = TP/(TP+FP)

**3.Recall**

Recall is the ratio of the correctly +ve classified by our model to all customers that will churn. Recall answers the following question: Of all the customers who will churn, how many of those were correctly predicted. Recall = TP/(TP+FN)

**4. F-1 Score**

F1 Score takes into account both precision and recall metrics. It is calculated as the harmonic mean(average) of the precision and recall. F1 Score is best in case of imbalances between precision and recall.
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

**5.Specificity**

Specificity is the correctly identified -ve tuples classified by the model to all customers will not churn. Specifity answers the following question: Out of all the customers which are not going to churn, how many of those were correctly predicted.

Specificity = TN/(TN+FP)

**Confusion Matrix :**

Confusion matrix is simple tool that is used to check whether tuples belonging to a class are perfectly classified.It is used to describe the perfromance of the model using a set of test data for which the true labels are known.

| | positive(actual) | negative(actual) |
|---|---|---|
| positive(predicted) | true positive(TP) | false positive(FP) |
| negative(predicted) | false negative(FN) | true negative(TN) |

Confusion Matrix

**PHASE 3 (Modeling /Application Of Machine Learning Algorithms) :**

**Train/Test split**

The dataset is used for analysis is generally divided into two parts ,training part and test part. The training set consists of known output value which is used by the model to learn form this data in order to classify unknown data later on. The test dataset is used to test the efficiency and performance of our model. In Python this is done using the scikit-learn library which consists of traintestsplit method. This method is imported from the library to split the dataset into two parts. Before importing this method some important libraries must be imported first. Which are as follows: Panda library is use to load the dataset file into a pandas data frame for working on data, Matplotlib pyplot,seaborn can be imported for data vizualization. It is used for plotting graphs of the data, The test-size=0.3 tells that 30 percent of data is used for testing the model and the remaining 70 percent is for training the model.

**Modelling**

After splitting the dataset into traintest split ,it is now the time to apply the necessary/required machine learning algorithms to predict the accuracy of the applied machine learning model and visualize the results and predict our assumed hypothesis using the required plots .Here we have used 5 suitable machine learning models on our dataset and plotted the results and determined the accuracy as well as the area under ROC curve. We have also answered the built hypothesis using these models.Here are the results and performances of machine learning algorithms after application :-

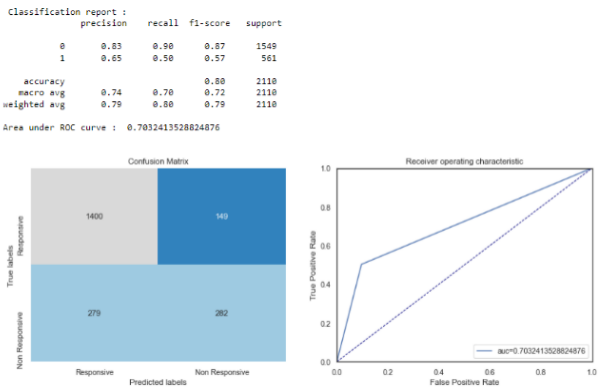
**1. LOGISTIC REGRESSION ALGORITHM**

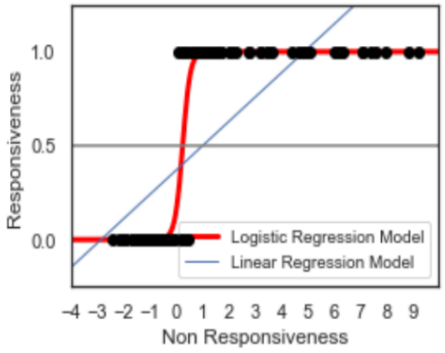
Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

**Hypothesis :-**

The promotional offers by he firm are very common, but the customer's resposiveness towards the promotional offers is what a question to be answered. So in the effort to determine the right offers to be endorsed, there arises a necessity to find the resposiveness of customer towards the offer using the exisitng data, by feeding to machine learning algorithm and building a model.

**Results :-**



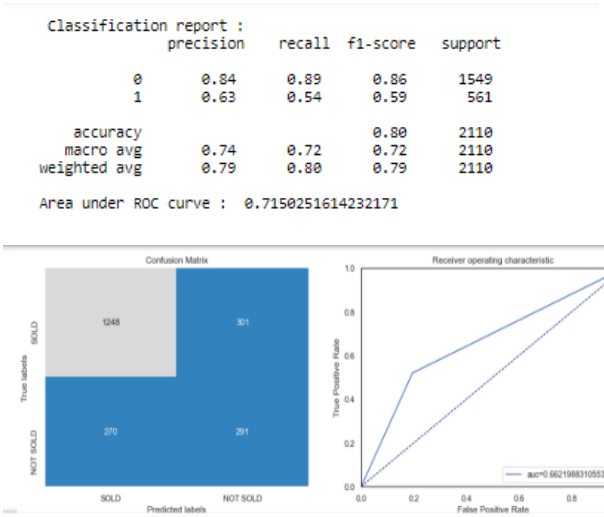

## 2. DECISION TREE ALGORITHM

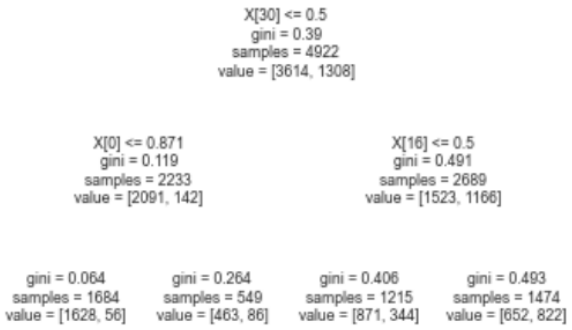
The decision tree classifiers organized a series of test questions and conditions in a tree structure.In the decision tree, the root and internal nodes contain attribute test conditions to separate recordes that have different characteristics. All the terminal node is assigned a class lable Yes or No. Once the decision tree has been constructed, classifying a test record is straightforward. Starting from the root node, we apply the test condition to the record and follow the appropriate branch based on the outcome of the test. It then lead us either to another internal node, for which a new test condition is applied, or to a leaf node. When we reach the leaf node, the class lable associated with the leaf node is then assigned to the record.

**Hypothesis :-**

The products available in the firm's inventory need to be updated according to sales dynamics of the particular product. There always arises a need to predict the sales dynamics in order for the company to mobilize the inventory according to the predicted future sales of the product. Here we would like to model a machine learning algorithm based on the assumption that a particular product, say a google phone and predict its sales by training the model.

**Results :-**




## 3. RANDOM FOREST ALGORITHM

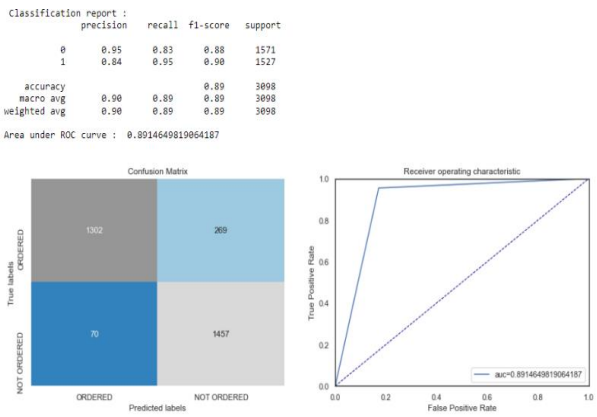
Random Forest build multiple decision trees and merge them together to get a more accurate and stable prediction. Random forest is a kind of supervised machine learning algorithm based on collective learning .Collective learning

is a type of learning where you join different types of algorithms or same algorithm numerous times to form more powerful prediction model. The random forest algorithm combines many similar algorithms multiple decision trees, resulting in a trees forest, hence the name Random Forest. For both regression & classification tasks, the random forest algorithm can be used. This algorithm is not prejudiced or biased, as, it consists of multiple decision trees and each tree is basically undergoes training under some sort of data. In other words, this supervised learning algorithm relies on the vote of all the trees and thus features with maximum votes is selected and thereby reducing the overall biasness of the system. This algorithm is quite sturdy and robust , I . e Even in case some new or unknown data point is introduced in the dataset the overall efficiency and performance of the algorithm is less affected since newly introduced sample might impact not more than 1 or 2 trees.

**Hypothesis :-**

When it comes to sales on the e-commerce platform, most of the times, it is not just about a individual product, but a right combination of products that can be offered to the customers, which will act as reckoning factor to buy such a combination.

**Results :-**

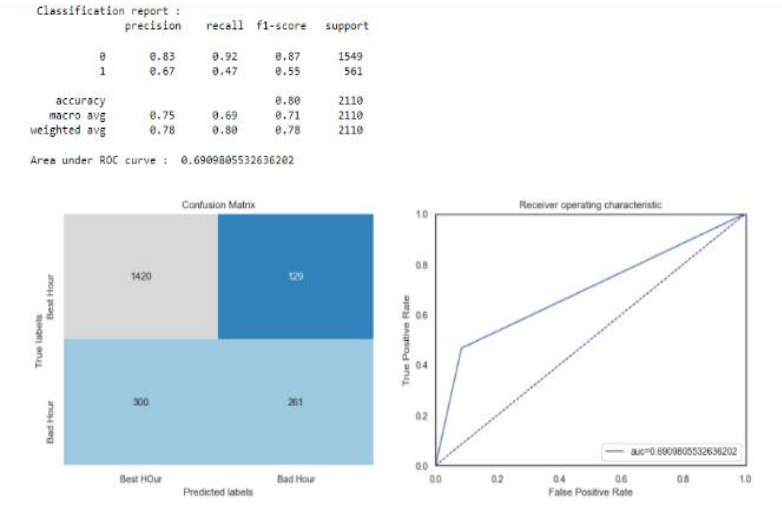


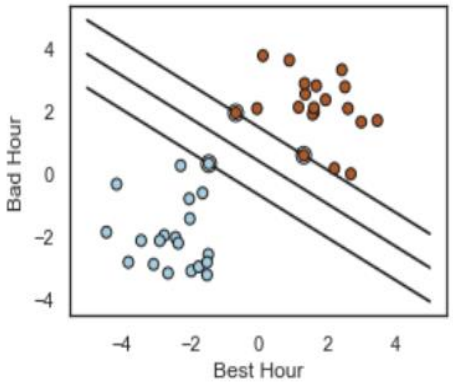

## 4. SVM ALGORITHM

This is a kind of classification algorithm, where the independent data and dependent data points are basically are separated by a line or a hyperplane depending upon whether the SVM is linear or non linear. The separation line or plane is selected such that; the two sides formed from the division by the line or plane makes 2 classes. When an unknown tuple/data comes its task is to correctly predict and identify which side/class of the line it belongs. The margin between the hyperplane and the support is kept as large as it can be possible In order to reduce the errors and improve the overall efficiency.

**Hypothesis :-** In our previous EDA step, we found out the best hours in a day to advertise the products of the firm, which will enhance the sales magnitude of the products. We would try to employ the machine learning model on the best hours of advertising to predict the sentiment of the customer towards buying the product.

**Results :-**






## 5. NAIVE BAYES ALGORITHM

The Naive Bayesian classifier is based on Bayes' theorem based on conditional probability. This model is quite easy to deploy and build, as it does not require any complicated parameter estimation iteratively and thus making it fairly convenient for larger data samples. Despite its simplicity, the Naive Bayesian classifier often does predicts unknown class labels surprisingly well compared to even some of the most sophisticated classification algorithms .A classifier built on this model performs comparatively better than
other models like logistic and linear regression and requires less time in training data and testing the model. It performance is better in the case of categorical data as
compared to numerical data. This model makes it easy to predict class labels in a very short time with good
accuracy. When multiple class prediction is required it performs well in that case as well. Another restriction of Naive Bayes is the assumption of independent predictors. In
real life, it is almost impossible that we get a set of predictors which are totally independent.

**Hypothesis :-**

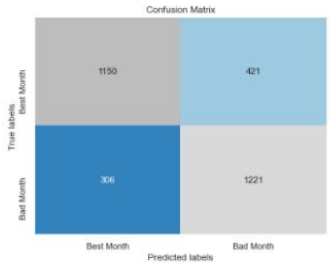
The firm needs to track its sales on a monthly basis for the better mobilisation of its resources to increase the sales. The data though gives maximum sales on a month basis, it is again a task of bigger scale to determine the month where there can be higher sales in upcoming years.Thus, to verify and predict the month with maximum sales is what we will try to determine using machine learning model.

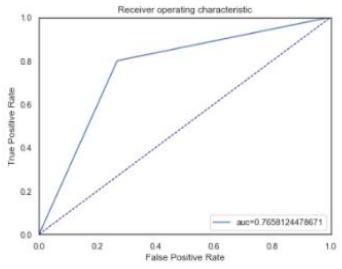
**Results :-**

Classification report :

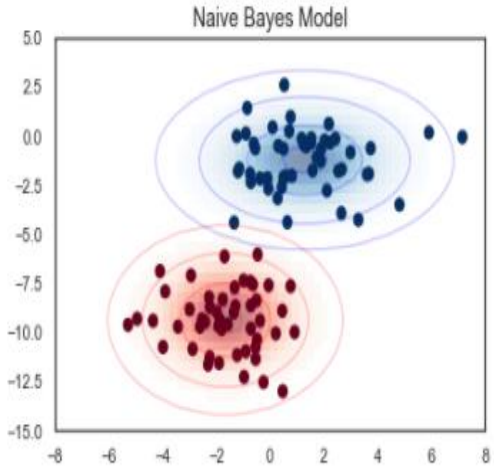
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.79 | 0.73 | 0.76 | 1571 |
| 1 | 0.74 | 0.80 | 0.77 | 1527 |
| accuracy | | | 0.77 | 3098 |
| macro avg | 0.77 | 0.77 | 0.77 | 3098 |
| weighted avg | 0.77 | 0.77 | 0.77 | 3098 |

Area under ROC curve : 0.7658124478671




Out[30]: [(-8.0, 8.0), (-15.0, 5.0)]




## Conclusion

This document gives a detailed picture of the analysis and prediction of sales data from given raw sales data set. The dataset in order to do analysis has undergone multiple cleaning methods. Further, Exploratory Data Analysis has been performed on the data and lot of quality metrics have been determined and an attempt to find answers from the data based on our hypothesis, in final phase we tried to build machine learning models based on the hypothesis made in the EDA step, prediction was done based on the model output. A lot of effort has been involved in developing the models. As it is always, not everything is perfect, we would advice the future exploration be more focused towards the incoming huge chunk of data, and relevant algorithms be applied.

## REFERENCES


[1] Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari. "Evaluation of machine learning models for Sales prediction" , 2017 International Conference on Inventive Computing and Informatics (ICICI), 2017

[2] zhenyu chen ; zhiping fan(2011) building comprehensible customer sales prediction model.

[3] Saad Ahmed Qureshi ; Ammar Saleem Rehman ; Ali Mustafa Qamar; Aatif Kamal(2013) . Annual Sales prediction model using machine learning

[4] Adnan Amin ;Changez Khan ;Imtiaz Ali ;Sajid Anwar(2014) Sales Prediction in E-Commerce Industry: With and without Counter

[5] Alpaydin, E. (2010). *Introduction to Machine Learning.* London, England: The MIT Press.

[VIII] Archaux, C., Laanaya, H., Martin, A., & Khenchaf, A. (2004). An SVM based Sales Detector in E-Commerce. *IEEE*