Machine Learning Models on Sales Data

By

Kundan Surya Teja Nanda

Table of Contents

1. Introduction and Background

Sales here refers to the distributor sales, where we'll concentrate on buying and selling manufactured goods to consumers. Who are distributors? Distributors connect manufacturers with retailers, business, and other organizations in order to get products to the consumers. So, in order to make profits these distribution companies need to find what products are in demand in the present and in future so that they can buy these products from the manufacturers and sell them to the customers (customers here refers to retailers, business and other organizations). Wherever the sales are there is a need of predicting the future sales, so that the companies can plan what materials are to be purchased from then manufacturers. So, here is the need for machine learning algorithms and that's what we are going to see in rest of the report. The first half of the report concentrates on the first project I have done (i.e., 6 weeks of work, 40 hours per week) which is customer segmentation, the second project which I've done is Time series forecasting which involves in using the previous data to predict the future sales of the company.

2. Projects

2.1 Project – 1: Customer Segmentation

Success in marketing and sales depends on how good a company is understanding and analyzing their respective customers. If a company is good at understanding their customers, then it is well likely that the company can get consistent profits. Customer segmentation is method used in grouping the customers based on various aspects like their age, marital status, county, state, city, sales, profits, frequency, etc. By segmenting their customers, a company can get necessary information about their customers to describe them. Customers can be segmented into various types out of which there are four common ways of segmenting customers which are discussed below.

2.1.1. Types of Customer Segmentation

There are many types of segmentation available, but the most common ones are discussed below in brief.

1. Demographic Segmentation: If the grouping (segmentation) is done on the bases of age, gender, marital status, education etc., then the type of segmentation is known as the demographic segmentation.
2. Geographical Segmentation: If the grouping (segmentation) is done on the basis of country, region, county, state, city, town etc., then the type of segmentation is known as geographic segmentation.
3. Psychographic Segmentation: If the grouping is done on the basis of the class, personal life, lifestyle etc., then the type of segmentation is said to be psychographic segmentation.
4. Behavioral Segmentation: If the grouping(segmentation) is done based on the sales, profits, COGS, purchase history, etc., then such types of segmentation is commonly known as behavioral segmentation.

2.1.2. Choosing the right Segmentation method

In order to choose the right method to segment the customers we need to first look at the data we are provided with. The data which we are going to deal with for the rest of the report is basically a sales data of different customers for the years 2020 and 2021 as shown in the below figure (Figure 2.0.11).

2.1.3. Dataset

A dataset is an organized collection of data which might be in any format like .csv, .xml, .xlsx, etc. Dataset gives us the information about all the data we needed in two components i.e., rows and columns. The data in rows are known as entries and that in the columns are known as variables or sometimes as features. So, the below figure (Figure 2.0.11) shows us the basic sales data that we are going to use for the rest of the report.

	BillingDate	sales_order_id	sales_order_sequence	shipment_number	Sales	Cogs	GrossProfit	LoadFactor	sale_type	invoice_term	Cdc_cust_codeNK	Cdc_seq_numNK	Cdc_cust_key	Cdc_cust_name	Cdc_shipto_pay_
0	1/6/2020 0:00	6589657	1	1	193.800	139.635550	54.164450	1.2090	01 Warehouse	0.5%10PROX11	3AXIKANKS1	1	11063	3 AXIS INC	0.5
1	1/24/2020 0:00	6687701	1	1	43.750	31.392889	12.357111	0.3020	01 Warehouse	0.5%10PROX11	3AXIKANKS1	1	11063	3 AXIS INC	0.5
2	4/30/2020 0:00	7216718	1	1	43.104	29.056060	14.047940	0.2562	01 Warehouse	0.5%10PROX11	3AXIKANKS1	1	11063	3 AXIS INC	0.5
3	4/30/2020 0:00	7216790	1	1	44.350	31.897293	12.452707	0.3020	01 Warehouse	0.5%10PROX11	3AXIKANKS1	1	11063	3 AXIS INC	0.5
4	3/11/2020 0:00	6906515	1	1	32.300	23.513378	8.786622	0.2015	01 Warehouse	0.5%10PROX11	3AXIKANKS1	1	11063	3 AXIS INC	0.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
740756	2/19/2021 0:00	8799842	4	1	188.160	126.757765	61.402235	0.4704	01 Warehouse	NET30	YUTCONGARKS	1	19440	YUTZY CUSTOM STRUCTURES	
740757	2/19/2021 0:00	8799842	5	1	497.700	327.607764	170.092236	1.1970	01 Warehouse	NET30	YUTCONGARKS	1	19440	YUTZY CUSTOM STRUCTURES	
740758	2/19/2021 0:00	8800038	1	1	820.800	568.262234	252.537766	2.1750	01 Warehouse	NET30	YUTCONGARKS	1	19440	YUTZY CUSTOM STRUCTURES	
740759	2/19/2021 0:00	8800038	1	1	300.960	208.362819	92.597181	0.7975	01 Warehouse	NET30	YUTCONGARKS	1	19440	YUTZY CUSTOM STRUCTURES	
740760	2/19/2021 0:00	8800038	1	1	410.400	284.131117	126.268883	1.0875	01 Warehouse	NET30	YUTCONGARKS	1	19440	YUTZY CUSTOM STRUCTURES	

Figure 2.0.11

So, here all the rows from 0 – 740760 are the entries and all the columns starting from the billing date to Unnamed: 32, as shown in the below figure (Figure 0.2.12), are our features. As, we can see the dataset we have is a sales dataset and we don't have any information in any of the columns regarding the age, marital status or class, personal life so we cannot do any of the Demographic or the Psychographic Segmentations. Hence, we are proceeding with Behavioral Segmentation. So, now we have the data, and we also know that we are going to do Behavioral Segmentation, we can now jump into the next step which is data preprocessing.

2.1.4. Data Preprocessing

The live data or real-world data or the raw data which we have can be inconsistent, incomplete or may have many missing values due to human errors. So, using such data may lead to low efficiencies. Turing the raw data or the live data into a machine understandable and useful format is known as data preprocessing. Hence, data preprocessing is very important to obtain the required efficiency and accuracy. There are four steps in data preprocessing:

1. Data Cleaning: The data can be noisy, or it can have a lot of missing values. We can deal with missing values either by replacing them with the most occurring values or the mean, median or mode values. The data is said to be noisy if the data is having some irrelevant data in between. This irrelevant data can either be deleted or it can be replaced.

2. Data Integration: This step involves consolidating the data, visualizing the data and propagating the data.

3. Data Reduction: This step involves in reducing the unwanted data in the dataset and thus increases the accuracy and reduces the cost. Dimensionality reduction, feature subset selection, and numerosity reduction are the parts of data reduction process.

4. Data Transformation: This step involves in converting all the values into single dimension. Strategies like data smoothing, data aggregation, data normalization etc., comes under this stage.

Here, for the data we have we can see (Figure 0.2.12) that there are few null values in few columns and as it is a sales data, we have chosen to delete the entries with null values.

```
BillingDate                     0
sales_order_id                  0
sales_order_sequence            0
shipment_number                 0
Sales                          77
Cogs                            0
GrossProfit                    77
LoadFactor                      0
sale_type                       0
invoice_term                    0
Cdc_cust_codeNK                 0
Cdc_seq_numNK                   0
Cdc_cust_key                    0
Cdc_cust_name                   0
Cdc_shipto_pay_terms_code   47152
CustomerSegmentation       740488
Cdc_Item                      109
Cdc_Size                    10615
Cdc_ItemDesc                    0
Pg_Type                         0
Pg_Major                        0
Pg_MajorDescription             0
Pg_Mid                          0
Pg_MidDescription               0
Pg_Minor                        0
Pg_MinorDescription             0
Pg_MajorMidMinor                0
BranchId                        0
BranchName                      0
Region                          0
RVP                             0
GM                              0
GeoCodedAddress                 0
GeoCodedCity                    0
GeoCodedState                   0
GeoCodedZip                     0
GeoCodedCounty                  0
GeoCodedGPSLat                  0
GeoCodedGPSLon                  0
DayShortName                    0
FiscalYear                      0
FiscalQuarterKey                0
FiscalWeekOfYearKey             0
FiscalDayOfYear                 0
IsHoliday                       0
IsWeekday                       0
Unnamed: 46                740642
dtype: int64
```

Figure 0.2.12

2.1.5 K-Means Clustering

After preprocessing the data in order to segment the customers we have used K-means clustering algorithm. This algorithm takes in the data as an input and chooses a random number of points known as

centers and calculates the distance of each point in the dataset from these centers and randomly moves these centers until each of the centers best fits the data points. So, in order to find the number of centers we have to plot the elbow curve as shown below (Figure 2.10.3).

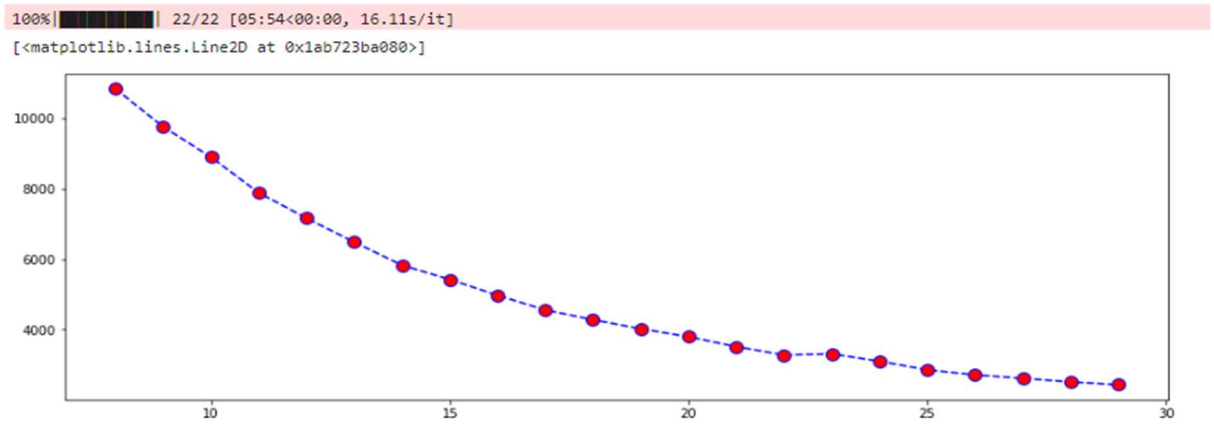


Figure 2.10.3

As we can see from the above elbow curve the number of clusters or the number of centers can be any number between 15 to 20. Hence, we chose 16 as the number of clusters. Therefore, after plotting the data points we get the below graph (Figure 2.10.4) with the datapoints clustered accordingly.

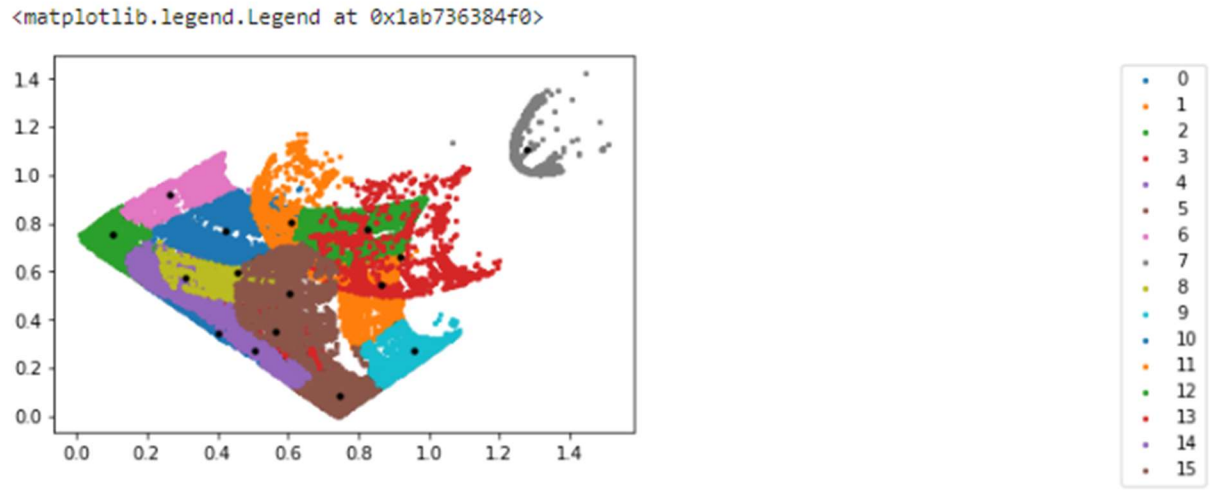


Figure 2.10.4

Using the above the data the frequency, recency and monetary values are calculated. Frequency, monetary and recency are the marketing analysis tool used by the industries to measure their customers spending habits. Recency says how recent a customer made a purchase. Frequency says how frequently a customer is making a purchase. Monetary says how much customer is spending. The monetary vs recency clusters as shown in the figure below (Figure 2.10.5).

2.1.6 Observation

From the above graph we can say that the customers in the cluster-0 had made recent purchases and had an average monetary value and the customers in cluster-1 are least recent and had high monetary. The customers in the clusters 0 and 1 are shown below (Figure 2.10.6)

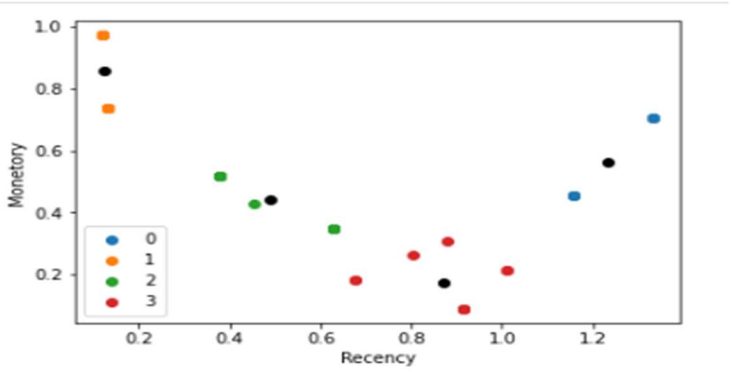


Figure 2.10.5

Pg_MajorMidMinor	R	F	M	label
11CYP	0.50	0.25	0.25	0
11OTH	0.50	0.00	0.00	0
23MIS	0.50	0.00	0.00	0
23REM	0.50	0.25	0.25	0
31FIR	0.75	0.00	0.00	0
41LVL	0.50	0.00	0.00	0
42LMB	1.00	0.00	0.00	0
92MIS	0.75	0.00	0.00	0
93BLD	0.50	0.00	0.00	0

Figure 2.10.6

Pg_MajorMidMinor	R	F	M	label
11CED	1.0	1.00	1.00	1
11SPF	1.0	0.75	1.00	1
11TRT	1.0	1.00	0.75	1
12OSB	1.0	1.00	1.00	1
13INS	1.0	0.75	1.00	1
21HWD	1.0	1.00	1.00	1
22FIB	1.0	1.00	1.00	1
22PAN	1.0	1.00	0.75	1
22PLD	1.0	1.00	1.00	1
22PLI	1.0	1.00	1.00	1
33FCM	1.0	1.00	1.00	1
33VNY	1.0	1.00	1.00	1
42IJO	1.0	0.75	1.00	1
42LVL	1.0	1.00	1.00	1
51SWD	1.0	1.00	1.00	1
53MIS	1.0	1.00	1.00	1

2.2 Project – 2: Time series forecasting

Time series forecasting helps industries or organizations to make predictions scientifically based on historically time stamped data. Time series forecasting basically involves building models by using the historical data to predict the future data which can be used by the companies or organizations to plan their strategies and to make decisions accordingly. The most important part of time series forecasting lies where there is no future data available at the time where the companies deal with the present data, so companies need to carefully analyze and choose the models which performed better on previous data.

2.2.1 Usage of time series forecasting

Any data which is consistent for the past few years can be used to run the time series forecasting models on them in order to predict the desired quantity in the future. Time series model can be used in various industries and organizations few of them are:

- Weather forecasting – Here we have the data of weather like climatic conditions, rainy days, sunny days, any climatic calamities like earthquakes, tsunamis, volcanic eruptions etc. in the past few years and this data is used in forecasting the weather for the future months or years.
- Finance forecasting – Here we have the financial data like the sales, profits, and their revenue and the causes for each of the uplift and downfall of the respective financial data for the last few years and this data is used to predict how the company's financial status could be in the future.
- Retail forecasting – In this forecasting data for the past years retailers is collected and used to predict the future retails of the company.

  As we are dealing with the customers, sales and the profits of the company we are ultimately doing the financial forecasting in this project.

2.2.2 Dataset used

The data set used here is the same dataset used in the above project. As we can see from the above figure (Figure 2.0.11) the data we have is more of a sales data and which is more consistent for the past few years. So, it is likely that we can run timeseries models on this data to forecast the sales for the next few months.

2.2.3 Data Preprocessing

The data preprocessing is the same as project – 1. The only thing changes here is the selection of features. In time series forecasting month and year column is the most important one without that we cannot predict future. As the data is mostly about sales, I'm only considering the sales column as my other feature. Here I have converted the daily data into a month and year format as shown below (Figure--2.21).

2.2.4 Data Visualization

Data visualization helps in transforming the data into visual context which helps in understanding the data in more accurate way. The visualization can either be in a map form or a graph form whichever is feasible to identify the patterns and outliers from the data. Data visualization can be a bar graph, line graph, pie chart, scatter plot etc. So, from the data set we have the line graph something looks like below in the Figure 2.22.

data1

	BillingDate	Sales	Cdc_cust_codeNK
0	1/6/2020 0:00	193.800	3AXIKANKS1
1	1/24/2020 0:00	43.750	3AXIKANKS1
2	4/30/2020 0:00	43.104	3AXIKANKS1
3	4/30/2020 0:00	44.350	3AXIKANKS1
4	3/11/2020 0:00	32.300	3AXIKANKS1
...	...	...	...
740756	2/19/2021 0:00	188.160	YUTCONGARKS
740757	2/19/2021 0:00	497.700	YUTCONGARKS
740758	2/19/2021 0:00	820.800	YUTCONGARKS
740759	2/19/2021 0:00	300.960	YUTCONGARKS
740760	2/19/2021 0:00	410.400	YUTCONGARKS

740761 rows × 3 columns

Before converting

data1_month_year

	month_year	Sales
0	2019-12	1.435609e+06
1	2020-01	2.211844e+07
2	2020-02	2.090936e+07
3	2020-03	2.448294e+07
4	2020-04	2.309090e+07
5	2020-05	2.401166e+07
6	2020-06	2.486582e+07
7	2020-07	2.521118e+07
8	2020-08	2.682968e+07
9	2020-09	2.652159e+07
10	2020-10	2.559344e+07
11	2020-11	2.415977e+07

After converting billing column to Month-Year format

Figure 2.21

As we can see from the figure 2.22 the trend is and increasing one from the year 2019 December to 2021 December. So, from the visualization graph it is evident that the data is consistent and is good to use the time series forecasting models on this data.

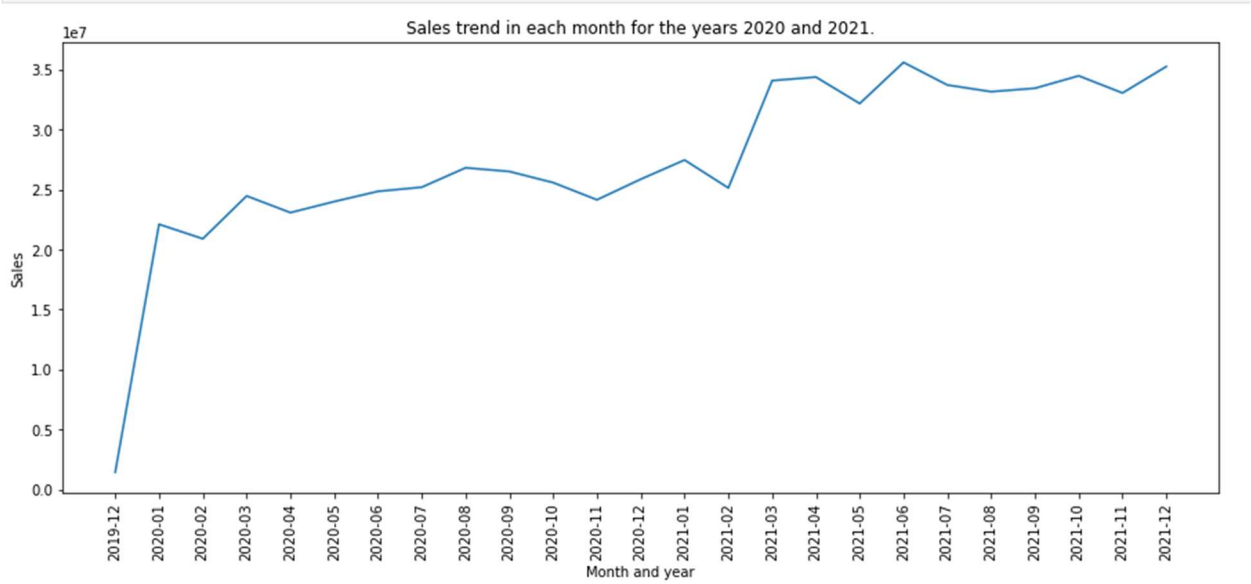
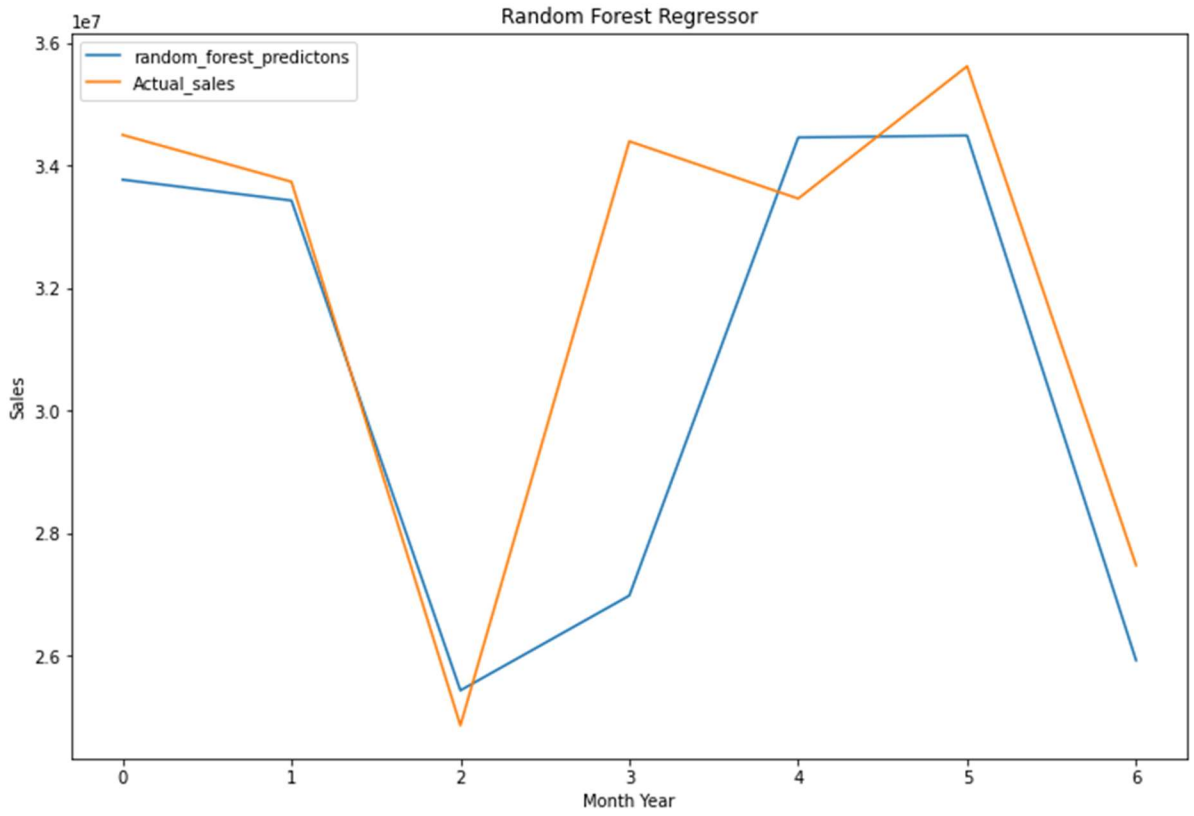


Figure 2.22

2.2.5 Times Series Forecasting Models

There are a number of time series forecasting models but only a few are being used here which were seemed to be performing better

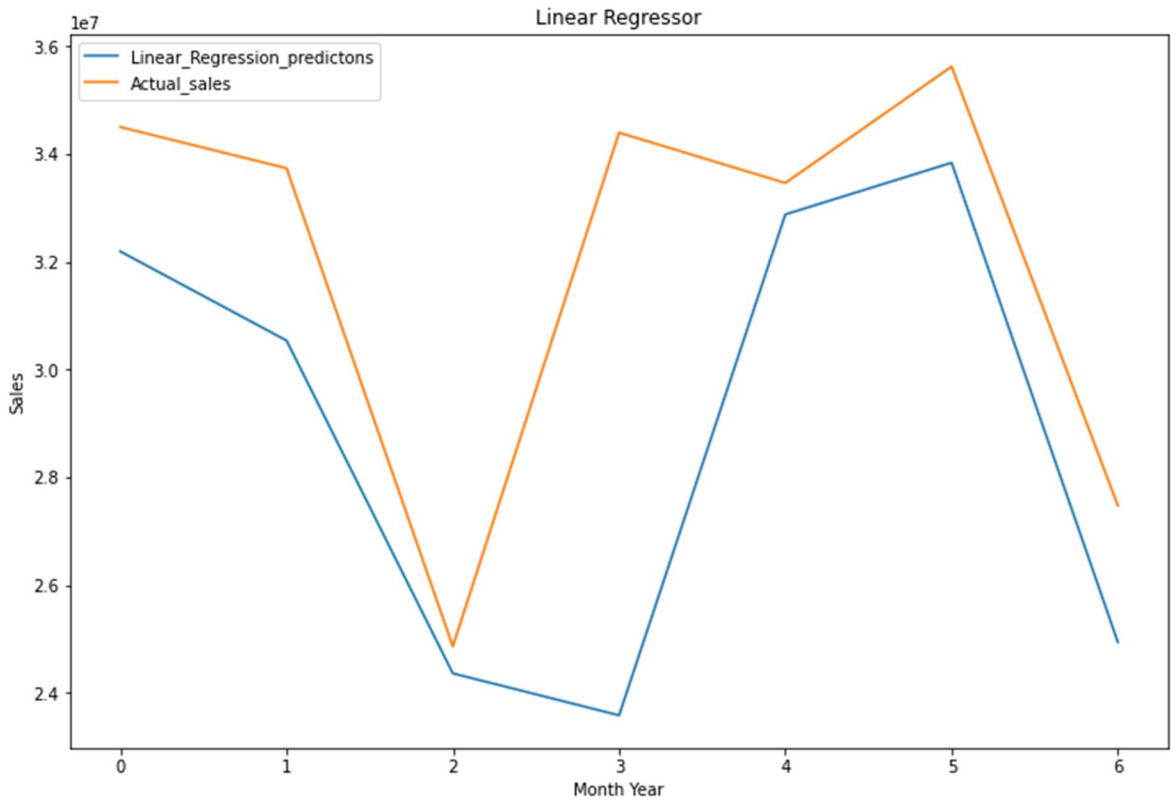
1. Random Forest Regressor: Random Forest regressor is an ensemble learning technique which takes multiple algorithms and put together on a model. This type of prediction uses trees to predict the data which is why it is more accurate in prediction. The comparison graph for the predicted and actual sales are as shown below.

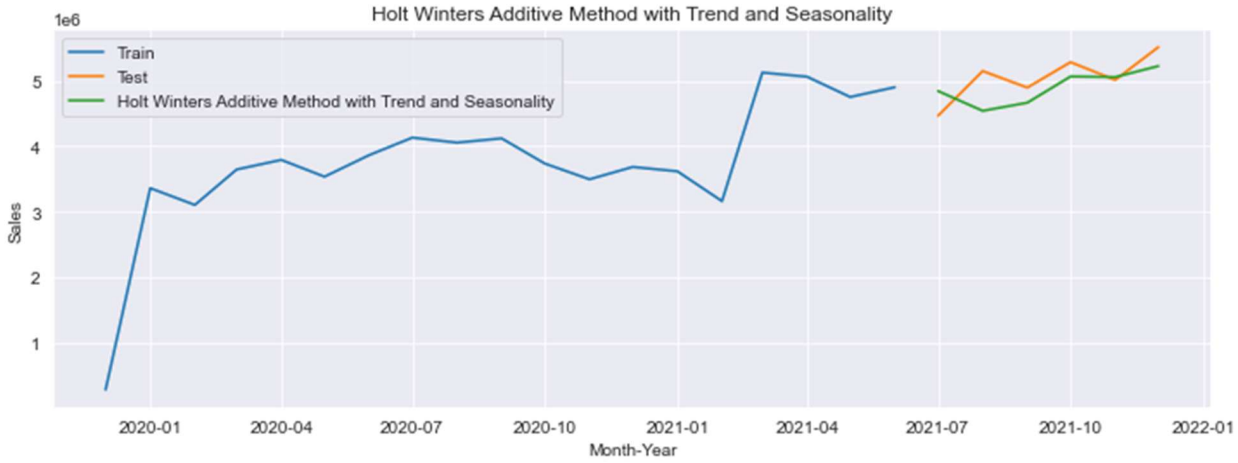


As we can see from the above graph the predictions are almost accurate. The root mean squared error for r andom forest regressor is 2942834.19

2. Linear Regressor: Linear regressor shows the linear relationship between the two variables. The compa rison graph for the actual and predicted sales using linear regressor is as shown below. As we can see the predictions are not that accurate when compared to the above random forest regressor. The root mean squ ared error for the linear regressor is 4514618.12

3. Hot Winters Additive Method with Trend and Seasonality: Holt winters additive method with trend and seasonality is an extension of holt winters exponential smoothing method that captures the seasonality an d does the predictions. The below graph shows the comparison between the actual and predicted sales and also the accuracy for this method is 94%.





	Method	RMSE	MAPE	Accuracy
0	Holt Winters Additive Method with Trend and Se...	341062.97	5.88	94.12

So, as we tried various other models the above-mentioned models performed the best and out of them Hot winters method has the best accuracy in predicting the sales for the given dataset.

3. References

1. https://numpy.org/
2. https://pandas.pydata.org/
3. https://matplotlib.org/
4. https://seaborn.pydata.org/
5. https://scikit-learn.org/stable/index.html
6. https://learn.g2.com/data-preprocessing#:~:text=Data%20preprocessing%20is%20the%20process%20of%20transforming%20raw,more%20complete%20and%20efficient%20to%20perform%20data%20analysis.
7. https://towardsdatascience.com/what-is-a-data-set-9c6e38d33198
8. https://bluelinxco.com/
9. https://www.tableau.com/learn/articles/time-series-forecasting