****

**School of Business**

**OPIM 5671 – Data Mining and Business Intelligence**

**Forecasting the Future Sales of Superstore**

**Group - 3**

**Aniket Jana**

**Pratiksha Nimesh**

**Kaman Dhami**

**Meghna Ashok**

**Haneesha Balireddy**

**Kushagra Goyal**

**Introduction**

Super Store is a tiny retail establishment in the United States of America. Their customers are the public, corporations, and home offices, and they sell furniture, office supplies, and technology products.

Super Store's sales, profit, and regional statistics are all included in this data collection.

**Background and Problem Statement**

Time series analysis plays an important role in helping different business entities get an idea of how good their sales are by implementing sales forecasting based on historic data in today's contemporary times, where every business is highly dependent on its data to make better decisions for developing business.

The Superstore manager wants to forecast the future sales so that required action such as stocking specific items from the warehouse, increasing the manual labor, etc. can be taken to meet the requirements. Moreover, the manager would like to evaluate the sales pattern and trends in the superstore data.

**Data Columns**

* **Row ID:** It depicts the serial number of the transaction in our record.
* **Order ID:** It is the unique id of the order and changes with every order.
* **Order Date:** It is the date at which the order is placed.
* **Ship Date:** It is the date at which the order is delivered.
* **Ship Mode:** It is the mode in which the shipment is done.
* **Customer ID:** It is the unique id assigned to every customer to distinguish among the customers.
* **Customer Name:** The name of the customer who placed the order.
* **Segment:** It is the segment for which the order is placed.
* **Country:** The country of the customer who placed the order.
* **City:** The city of the customer who placed the order.
* **State:** The state of the customer who placed the order.
* **Postal Code:** The postal code of the customer who placed the order.
* **Region:** The region of the customer who placed the order.
* **Product ID:** The product id of the product which is ordered.
* **Category:** The category of product for which the customer has placed the order.
* **Sub-Category:** The subcategory of product for which the customer has placed the order.
* **Product Name:** The product name for which the customer has placed the order.
* **Sales:** The sales the store has accumulated through this order.

**Super Store Data Exploration**

We tried to explore the data on our own category wise through excel just to find out the general insights and got to know some key points:

* Maximum sales records are for ‘Office Supply’ category.​
* Average sale is maximum for ‘Technology’.​

​

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

When we tried to split these day wise, we got to know an insight that maximum average sales happen on Thursday.​

Chart, bar chart

Description automatically generated

**Data Preparation and Forecasting**

* We did the time series exploration on the complete data set firstly.
* When done exploration we found out that the complete data is a white noise and modelling cannot be performed on the entire data.
* So we Segregated data based on Category column:
* Furniture
* Office Supplies
* Technology
* Now we again performed time series exploration on the above categories.
* We took the sales day wise and aggregated the sales amount based on the order date.
* Then at last modelling and forecasting was performed on the data.

**Time Series Exploration**

As we have segregated our superstore data into three categories based on ‘**Category’** variable. We decided to do the time series analysis on an individual category, this way we would be able to identify if we can go ahead with time series forecasting model or not, and if we can model then which type of model should be used.

1. **Furniture data set**

Below is the plot between Sale amounts and month (ranging from year 2015 to 2018)

Chart, line chart

Description automatically generated

We can see that there is a nice seasonality for this time series:

Chart, histogram

Description automatically generated

However, there is no Trend for it, the trend line is flat.

Chart, line chart, histogram

Description automatically generated

From white series noise probability test, we can see that the p-value is very low (<0.05). Hence, we reject the Null Hypothesis (H0: Series is White noise), so we can proceed with the time series modeling for ‘Furniture’ category.

Graphical user interface, chart, histogram

Description automatically generated

Then, we performed ADF unit test to find out if the series is Stationary or not:

Table

Description automatically generated

From above results we can confirm that the series is Stationary, hence we can proceed with the ARMA models.

From autocorrelation plot (PACF) we can see that there is one spike at lag 12. So, we can proceed with auto regressive order 1 model. Later we will see which model performs best for the ‘Furniture’ category.

1. **Office Supplies data set**

Below is the plot between Sale amounts and month (ranging from year 2015 to 2018)

Chart, line chart

Description automatically generated

We can see that there is a nice seasonality for this time series:

Chart, histogram

Description automatically generated

However, the trend line doesn’t give much information.

Chart, line chart

Description automatically generated

From white series noise probability test, we can see that the p-value is very low (<0.05). Hence, we reject the Null Hypothesis (H0: Series is White noise), so we can proceed with the time series modeling for ‘Office Supplies’ category.

Graphical user interface, application

Description automatically generated

Then, we performed ADF unit test to find out if the series is Stationary or not:

Table

Description automatically generated

From above results we can confirm that the series is not Stationary, hence we can’t proceed with the ARMA models.

1. **Technology data set**

Below is the plot between Sale amounts and month (ranging from year 2015 to 2018)

Chart, line chart

Description automatically generated

We can see that there is a nice seasonality for this time series:

Chart, histogram

Description automatically generated

However, there is no Trend for it, it’s like what we had for ‘Office Supplies’ category.

Chart, line chart, histogram

Description automatically generated

From white series noise probability test, we can see that the p-value is very high (>0.05). Hence, we fail to reject the Null Hypothesis (H0: Series is White noise). As the series is complete white noise, we can’t proceed with modeling for ‘Technology’ category. In such scenarios, mean value of the series is the forecasted value.

Graphical user interface, application

Description automatically generated

**Modeling and Forecasting**

1. **Forecasting of Furniture Category**

From our initial superstore sales data, we extract our furniture data separately by executing the following query.

Graphical user interface, text, application

Description automatically generated

We then aggregate the resulting query by Monthly intervals, and Accumulate on Average Sales, using ‘Time Series Data preparation’.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface

Description automatically generated

We then perform forecasting of Furniture Average Sales on the below resulting data:

Table

Description automatically generated

**Note:**

We choose to accumulate ‘Average’ Sales, as opposed to ‘Total/Sum’ of Sales, since our data has multiple Sales entries for inconsistent dates, i.e. Some days have multiple entries, and some days have no entries. So, to ensure consistency and avoid unique behavior, we are aggregating the monthly average of Sales.

**Checking for Stationarity**

Table

Description automatically generated

From the above Augmented Dickey-Fuller test or ADF test, we can see that since the F test P-values are constant, and Tau test P-values are <0.001, we can reject the null hypothesis (H0) that the data is not stationary.

Since it is stationary data, we proceed with SARIMA model implementation.

Note that from the time series exploration analysis, the model has strong seasonality, because of which Seasonal ARIMA or SARIMA are implemented.

**SARIMA Model Implementation:**

**Model comparison:**

We ran multiple SARIMA models with different values of pd, q, P, D and Q to identify the best model.

Our results are as follows:

|  |  |  |
| --- | --- | --- |
| Model | AIC | SBC |
| (1,0,0) (1,0,0) | 662.7234 | 668.337 |
| (1,0,2) (1,0,2) | 675.959 | 689.0574 |
| (1,0,1) (1,0,1) | 660.0262 | 669.3822 |
| (1,0,1) (1,1,1) | 492.0802 | 499.9978 |
| (1,1,0) (1,1,0) | 490.9189 | 495.5849 |
| (1,0,0) (1,1,0) | 489.5722 | 494.3228 |
| (1,1,1) (1,1,1) | 485.8904 | 493.6671 |
| (1,0,1) (1,1,1) | 492.0802 | 499.9978 |
| (2,1,0) (2,1,0) | 489.8776 | 497.6543 |
| (0,1,0) (0,1,0) | 501.3034 | 502.8587 |
| (2,0,1) (2,0,1) | 663.4496 | 676.548 |
| (2,1,1) (2,1,1) | 491.4271 | 502.3145 |

From the above models,

We analyzed the two models with the lowest AIC and SBC values and compared their white noise residual plots.

1. SARIMA (1,1,1) (1,1,1)

Chart, box and whisker chart

Description automatically generated

1. SARIMA (1,0,0) (1,1,0)

Chart, box and whisker chart

Description automatically generated

We can see that SARIMA (1,0,0) (1,1,0) has a better white noise residual graph, which means the model is well capturing the signal in the series and all that is remaining is white noise. Hence, we proceed with SARIMA (1,0,0) (1,1,0) forecasting.

**SARIMA (1,0,0) (1,1,0) Model results:**

Due to some technical issues with evaluating the data in SAS Studio, we calculated the final model Accuracy measures through Python.

**Accuracy Measures:**

Graphical user interface, text, application

Description automatically generated

Our model has a forecast of MAPE = 1.05%, indicating that it is a good fitting model with high predictive power.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, chart

Description automatically generated

Table

Description automatically generated

The model has AIC of 489.57 and SBC of 494.32.

Table

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Note that the white noise probability graph represents clear white noise.

Chart, line chart, histogram

Description automatically generated

The QQ plot has most of the points falling on the line, and the distribution is close to Normal.

Chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

We predicted the next 2 year Average sales, as seen below:

The first 11 observations are forecasts from the holdout sample, due to which we can compare the actual value and the forecasted value for each observation.

The next 12 observations are the forecasts for data that was not used to build the model.

Table

Description automatically generated

The below plot shows the forecasts for the next 2 years, including the holdout sample data.

Chart, line chart

Description automatically generated

The below plot shows the predictions of the full time series including the hold out sample forecast.

Chart

Description automatically generated

1. **Forecasting of Office Supplies Category**

From the ADF test, in the F test, we can see that the values are not constant, indicating that the data is non-stationary data.

Table

Description automatically generated

Hence, we use the **Exponential Smoothening Method** to forecast the sales data.

We will try both the **Additive seasonal exponential smoothening** model and **Winters Additive method** and see which data gives the best model based on the AIC value of each model.

Go to Tasks and Utilities and select Modeling and Forecasting to forecast the data.



Table

Description automatically generated

A window will pop up wherein the **Data** tab, give the below details:

Graphical user interface, text, application, email

Description automatically generated

Select **Total sales** as the Dependent values since we want to predict the sales and the **Order data** as the Time ID. SAS automatically converts the properties of the Time ID accordingly.

For example, SAS automatically detects that the interval is monthly in the given data.

Graphical user interface, application

Description automatically generated

To Select the Model, go to the **Model** tab and select **Exponential Smoothening** as the forecasting model type, and in the model settings, select **Additive seasonal exponential smoothening.**

Graphical user interface, text, application, email

Description automatically generated

We will predict the sales over the next two years, so we set the **number of periods to be forecasted** as 24

The **forecast confidence interval** is 95% and the **Number of Periods to hold back** is 12.

Graphical user interface, application

Description automatically generated

Let’s save the Output data in the work library.

Graphical user interface, table

Description automatically generated

**Run the model.**

We can see the snapshot of the data in the below two tables.

Table

Description automatically generated

Since the prediction errors of the fitted model are white noise, it means that the model has done a great job of explaining the variance in the dependent variable and is a good model.

Chart, bar chart, histogram

Description automatically generated

From the below forecast of Total sales, we can see that the trend and seasonality of the forecasted data are the same as the historical data.

Chart

Description automatically generated

 From WORK.OUTSTAT we can obtain the AIC and SBC values of the model.

Table

Description automatically generatedTable

Description automatically generated

Now let’s use the Winters additive model:

Graphical user interface, application

Description automatically generated

We will predict the sales over the next two years, so we set the **number of periods to be forecasted** as 24

The **forecast confidence interval** is 95% and the **Number of Periods to hold back** is 12.

Graphical user interface, application

Description automatically generated

We save the output data set in the work library.

Graphical user interface, table

Description automatically generated with medium confidence

Since the prediction errors of the fitted model are white noise, it means that the model has done a great job of explaining the variance in the dependent variable and is a good model.

Chart, bar chart, histogram

Description automatically generated

For the Winters Additive model, we can observe a slight trend in the forecasted data, and we can also observe that the seasonality is continuing from the historical data.

Chart

Description automatically generated

From WORK. OUTSTAT1, we can obtain the AIC and SBC values.

Table

Description automatically generated with medium confidence

To know which model is the best out of the two, we use AIC values to compare them. The model with the lowest AIC value is the best.

We can observe that the additive seasonal exponential method is the best of the two as the AIC value of this model with the fit data is lower.

Hence **Additive Seasonal Exponential Method** is the best to forecast the Office Supplies.

1. **Forecasting of Office Supplies Category**

From the below Correlations Graph, we can see that the data is white noise.

Graphical user interface

Description automatically generated

 If a time series is white noise, it is a sequence of random numbers and cannot be predicted. Hence, we cannot build a model with the data.

**Conclusion**

**Furniture**

* As the data was stationary, we ran ARIMA model.
* The forecasted values have the seasonality component retained.
* The period between September and January has high peaks.
* We should run Christmas and New Year Festive Sales to garner more traction.

**Office Supply**

* Here, the data was non-stationary, we ran exponential smoothening.
* The forecasted values retained the same seasonality as well.
* There are multiple peaks in the forecasted period mainly on August, October, and November.
* Run Promotional Campaigns during and before the above months to improve sales

**Technology**

* Model was not run due to White noise.
* This could be due to the following two reasons:
* The data has a constant mean.
* The values are skewed due to products with huge price differences. (Ex. Mobile vs. Accessories)

**References**

* <https://www.kaggle.com/rohitsahoo/sales-forecasting>