|  |  |
| --- | --- |
| Internship Project Title | Forecasting System - Project Demand of Products at a Retail Outlet Based on Historical data |
| Name of the Company | TCS ION |
| Name of the Industry Mentor | Himdweep walia |
| Name of the Institute | B.K.Birla College,kalyan |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 11/09/2022 | 06/11/2022 | | 60 | |  | Google Colab, Numpy, pandas,statsmodel,seaborn,fbprophet,ARIMA |
| Milestone # |  | Milestone: | |  | | |

**Acknowledgements :**

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them. I would like to express my special thanks of gratitude to TCS ION team who gave me the wonderful opportunity to do a project on Forecasting System - Project Demand of Products at a Retail Outlet Based on Historical data and gave us all support and guidance which made me complete the project duly.

I am extremely thankful to them for providing such nice support and guidance which also helped me in doing a lot of research and I came to know about so many new things.

**Objective :**

A sales forecast is an important component of any business plan. For the sales rep, as well as the entire organization, a sales forecast aims to predict future sales and is used as the basis of planning time and resources. A good forecast should have several objectives, all directed at identifying what you will sell, when you will sell it and to whom.

The general process for ARIMA models is the following:

Visualize the Time Series Data

Make the time series data stationary

Plot the Correlation and AutoCorrelation Charts

Construct the ARIMA Model or Seasonal ARIMA based on the data

Use the model to make predictions

Forcasting using ARIMA model

**Introduction / Description of Internship :**

Sales prediction is an important part of modern business intelligence. It can be a complex problem, especially in the case of lack of data, missing data, and the presence of outliers. Sales prediction is rather a regression problem than a time series problem. Practice shows that the use of regression approaches can often give us better results compared to time series methods. Machine-learning algorithms make it possible to find patterns in the time series.

What is Forcasting?

Forecasting is the process of making predictions based on past and present data. Later these can be compared against what happens.

Introduction to Time Series Forecasting

A Time Series is defined as a series of data points recorded at different time intervals. The time order can be daily, monthly, or even yearly.

Time Series forecasting is the process of using a statistical model to predict future values of a time series based on past results.

Forecasting is the step where we want to predict the future values the series is going to take. Forecasting a time series is often of tremendous commercial value.

Forecasting a time series can be broadly divided into two types:

If we use only the previous values of the time series to predict its future values, it is called Univariate Time Series Forecasting.

If we use predictors other than the series (like exogenous variables) to forecast it is called Multi Variate Time Series Forecasting.

**Assumptions :**

Every year is different, so you need to consider any changing circumstances that could significantly affect your sales.

The biggest assumption in forecasting modeling is that the future will follow past trends

However, sometimes, people change their behaviors, and the models used to predict them may become invalid. And time is the most common reason that can make assumptions invalid. For example, if a model was built several years ago, it may not accurately predict today’s customers’ behavior as it changes over time.

Additionally, other factors that contribute to invalid assumptions in predictive modeling include:

Missing key variables

Significantly altered key variables

**How Can Predictive Modeling Help a Business?**

This type of modeling can help businesses:

Understand how well they’re performing

Get a clearer image of the competition

Build strategies to achieve a competitive advantage

Optimize their products and services

Eliminate potential risks

Understand their customers buying behaviors and preferences

Improve the customer experience

Retain customers

Reduce costs

Grow revenue

**Algorithms :**

Exploratory Data Analysis with Python (missing values, Correlation analysis ...)

Forecasting with ARIMA model

Libraries used: numpy, pandas, matplotlib, seaborn, statsmodel, ARIMA

**Forecasting Concept Image :**

Icon

Description automatically generated

**Internship activities:**

* Gone through all the contents in welcome kit and day wise plan.
* Attempted RIO pre-assessment and passed it successfully in the first attempt itself.
* Introduced myself in the Digital discussion room.
* Attended both webinar 1 and 2
* Gone through YouTube videos on ‘Forecasting Methods Overview’,’ Moving Averages’, ‘Time series forecasting’, and ‘ARIMA models’ given in the project reference material.
* Downloaded the dataset from Kaggle And starts working on it…
* Used Colab Notebook for coding and implement part.
* Take a reference from YouTube and Google.
* Imported the needed libraries.
* Worked on Data Cleaning part
* Worked on EDA (Exploratory data analysis).
* Then Worked on Model .
* Used ARIMA and SARIMA to Forecast sale by Year.
* Created time series models of this order with the help of SARIMAX function.
* Compared predicted results with the test set data points and evaluated the performance of my model with the help root mean square function.
* Done predictions of the data points in the unknown future.
* Built some deep learning models and compared their performance with SARIMA models and found that SARIMA models are better.
* Compared the sales happening in each product categories and plotted them.

**Methodology:**

ARIMA(Autoregressive Integrated Moving Averages)

In statistics and in time series analysis, an  ARIMA( autoregressive integrated moving average) model is an update of ARMA (autoregressive moving average). The ARMA consists of mainly two components, the autoregressive and moving average; the ARIMA consists of an integrated moving average of autoregressive time series.  ARIMA model is useful in the cases where the time series is non-stationary. And the differencing is required to make the time series stationary.

Mathematically we can represent the formula like this.



Where

* C is an intercept of the ARMA model.
* is the first difference operator.
* ʏ is the time lags

In the ARIMA model, we have to consider three values which we also need to give in our parameters while implementing it. Therefore, we can represent it by (p, d, q).

* P = lags in the autoregressive model.
* D = differencing / integration order.
* Q =  moving average lags.

So more formerly if we are saying that ARIMA(1,1,1) which means ARIMA model of order (1, 1, 1) where AR specification is 1, Integration order or shift order is one and Moving average specification is .1

## SARIMAX

SARIMAX(Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) is an updated version of the ARIMA model. ARIMA includes an autoregressive integrated moving average, while SARIMAX includes seasonal effects and eXogenous factors with the autoregressive and moving average component in the model. Therefore, we can say SARIMAX is a seasonal equivalent model like SARIMA and Auto ARIMA.

Another seasonal equivalent model holds the seasonal pattern; it can also deal with external effects. This feature of the model differs from other models. For example, in a time series, the temperature has seasonal effects like it is low in winter, high in summers. Still, with the effect of external factors like humidity, the temperature in winter is increased and also due to rain, there is a chance of lower temperature. We can’t predict the exact value for these factors if they do not appear in a cyclic or any seasonal behaviour. Other models are not capable of dealing with this kind of data.

In the SARIMAX models parameter, we need to provide two kinds of orders. The first one is similar to the ARIMAX model (p, d, q), and the other is to specify the effect of the seasonality; we call this order a seasonal order in which we are required to provide four numbers.

(Seasonal AR specification, Seasonal Integration order, Seasonal MA, Seasonal periodicity)

Mathematically we can represent the model like this.

Text, letter

Description automatically generated

**Where:**

Text

Description automatically generated

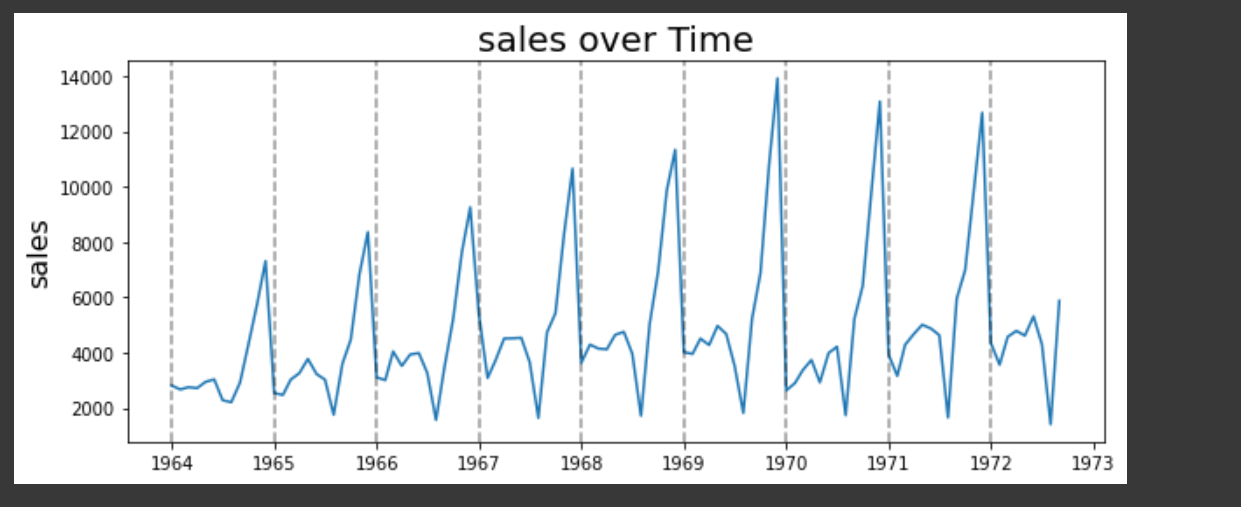
**How ARIMAX Works? How it is different from ARIMA?**

* ARIMAX is nothing different from ARIMA. It is just an extension. You have to call ARIMA function and add one more parameter "exog".
* ARIMAX = ARIMA + X, where X=exogeneous variable
* E.g. ARIMAX = ARIMA(endog = Passenger count, order = (p,d,q), exog = Weather)

How SARIMAX Works? How it is different from ARIMA ?

* SARIMAX is nothing different from ARIMA. It is just an extension with seasonal order. You have to call SARIMAX function and add one more parameter "seasonal\_order".
* SARIMAX = Seasonal + ARIMAX, where seasonal order also considered
* E.g. SARIMAX = sm.tsa.statespace.SARIMAX(endog = Passenger count, order = (p,d,q), seasonal\_order = (P,D,Q,s), exog = Weather).fit()

**Project Code Conclusion:**

****

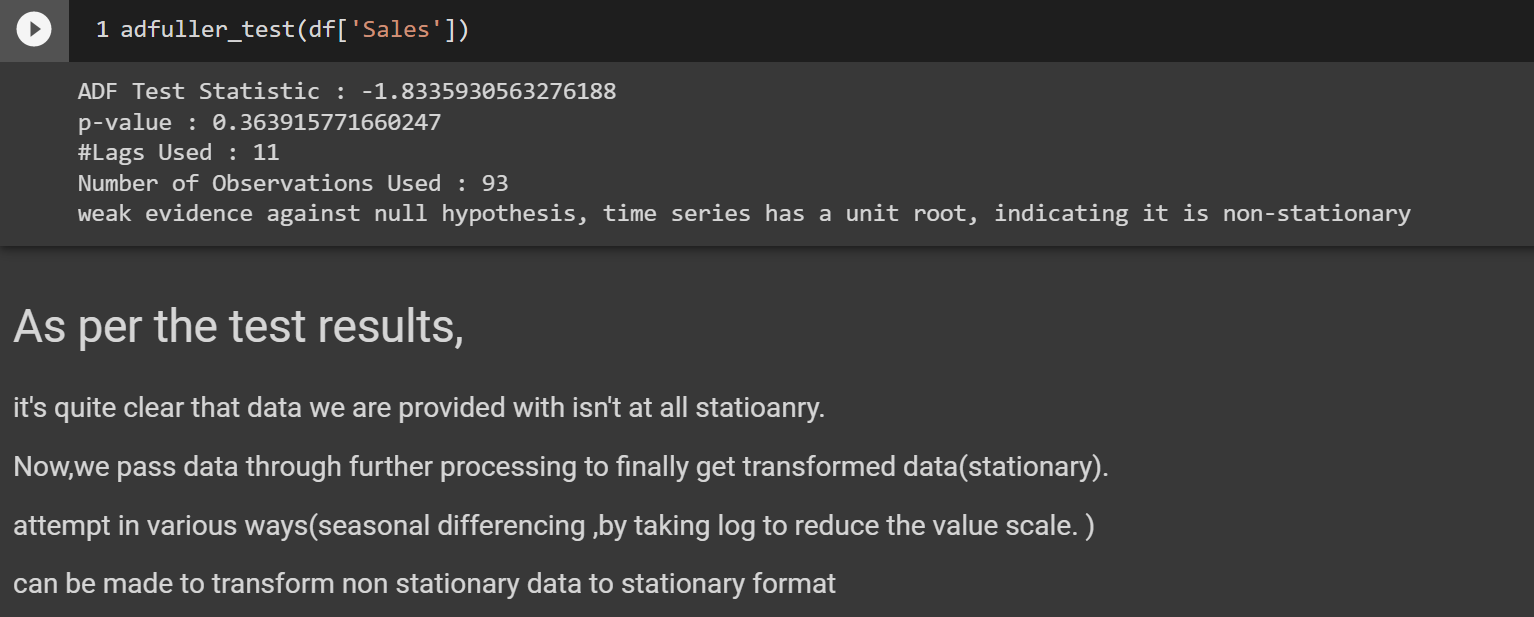
we can easily see the seasonality effect in our time series. In the above image, we have extracted the seasonality from the time series.

To perform forecasting using the ARIMA model, we required a stationary time series. Stationary time series is a time series that is unaffected by these four components. Most often, it happens when the data is non-stationary the predictions we get from the ARIMA model are worse or not that accurate.

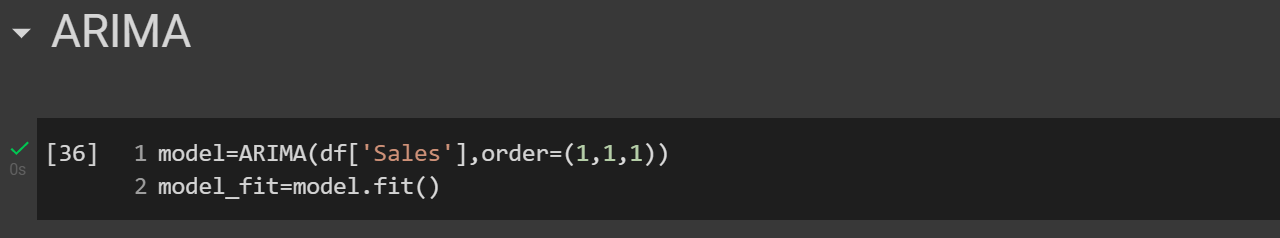
If the data is not stationary, we can do one thing: either make the data stationary or use the SARIMAX model.

we can perform the ADfuller test, a test based on hypothesis, where if the p-value is less than 0.05, then we can consider the time series is stationary, and if the P-value is greater than 0.05, then the time series is non-stationary.

Performing the adfuller test on data.



Importing the ARIMA model.



Text

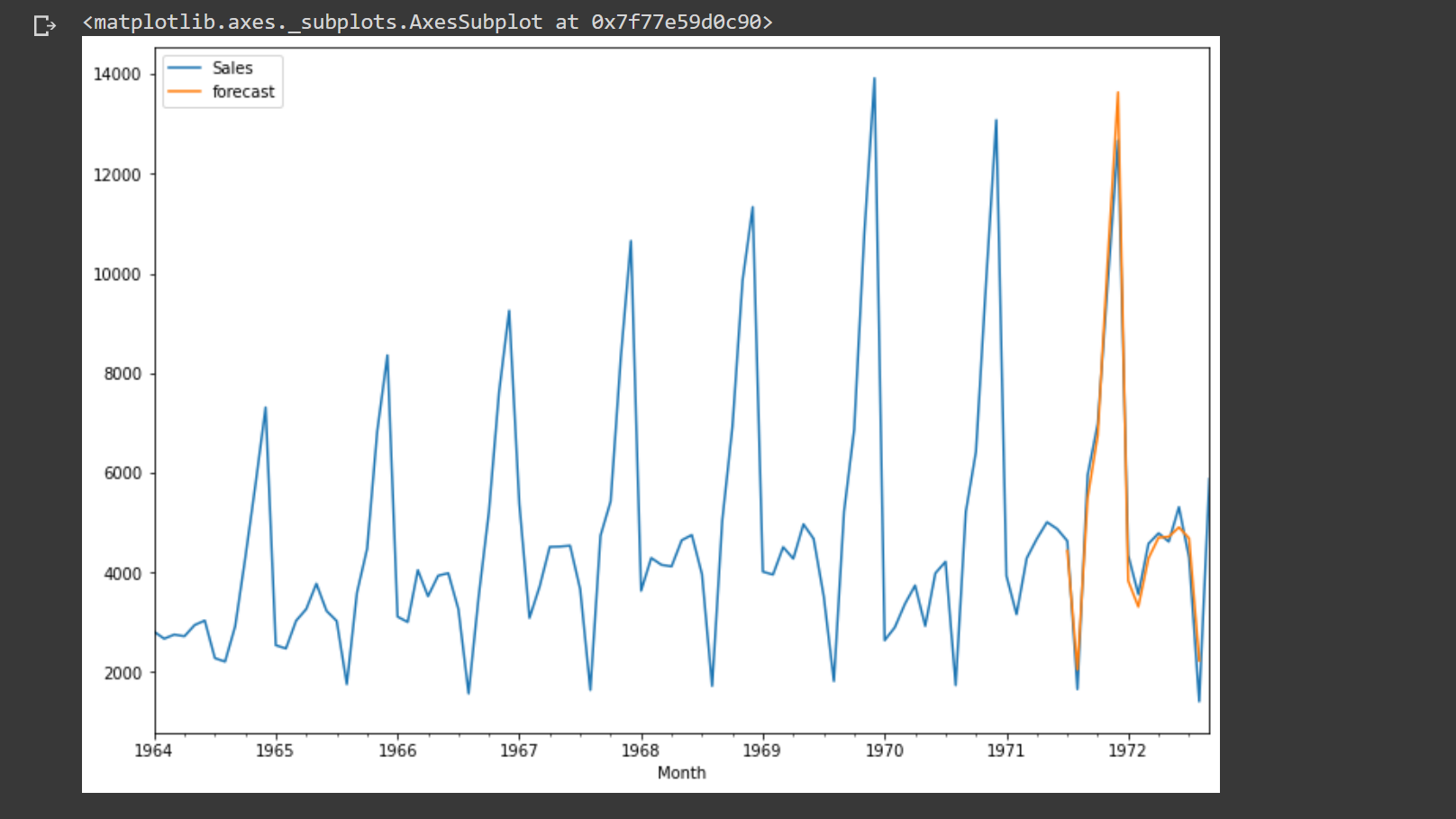
Description automatically generated

Graphical user interface, chart, histogram

Description automatically generated with medium confidence

Here we can easily see the results we have got by the model is very unsatisfactory. This is because we have fit the model with a non-stationary time series. Without the stationary data, the model is not going to perform well.

Importing the SARIMAX model.

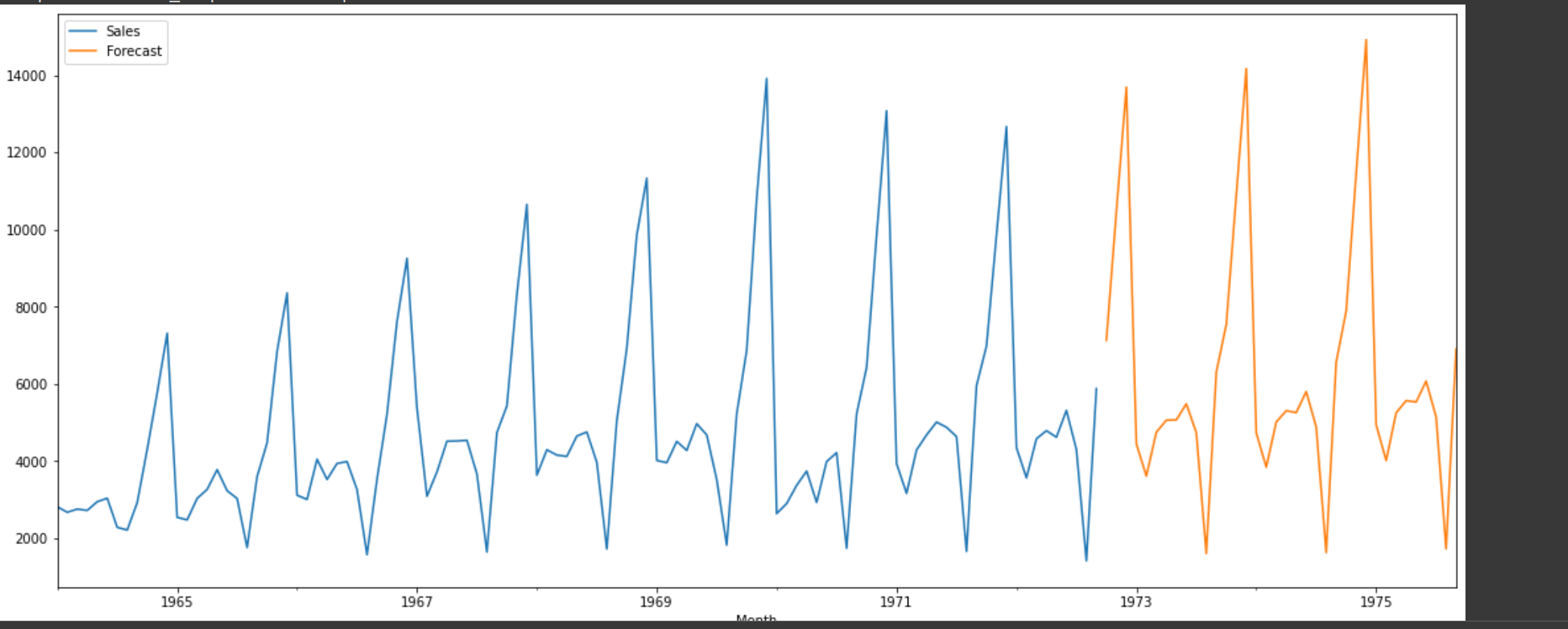


Here in the graph, we can see the results: the forecasting line is almost lying on the given values for this model. We didn’t even require the differencing method. Using this model now, we can predict the future values too.

A screenshot of a computer

Description automatically generated with medium confidence

We can see that the model has predicted the values without compromising with the seasonality effects and exogenous factors. And the trend line is almost going as usual as it was going in past years.



Now you can see, We have predicted the SALES for the next Three Years

Challenges & Opportunities:

* Since the technology sales data is “white noise” I wasn’t able to create SARIMA based models on that.
* I was able to create SARIMA models for both office supplies sales data and furniture sales data.
* I also tried to build some deep learning models for this data, but later found SARIMA models have less root mean square error value than those deep learning models. So concluded that SARIMA models are better to apply in this dataset.