TIME SERIES ANALYSIS: FORECASTING WITH SARIMAX MODEL AND STATIONARITY CONCEPT

Malde Ritik Vimal
Department of computer science
Somaiya Vidyavihar University
Mumbai, India

Shaikh Mohammad Bilal Naseem
Assistant Professor
Dept of Computer Science/IT
Somaiya Vidyavihar University
Mumbai, India

ABSTRACT: **Time series models have been the basis for any study of a behavior of process or metrics over a period of time. Machine learning is a subset of artificial intelligence (AI) that provides system the ability to automatically learn and improve from experience. The applications of time series models are manifold including sales forecasting, weather forecasting etc. Goal of this research is to explain the uni-variate forecasting, concept of stationarity, Auto Regressive Integrated Moving Averages model and Seasonal Auto Regressive Integrated Moving Averages model with eXogenous factor. To keep a track of sales of a company is the major key to gain profits in future and to prevent the further losses. As well as for stock holders it is supreme to keep an eye on stocks of the company to prevent their losses and gain significant profit.**

Keywords- Machine Learning, Time series, Uni-variate forecasting, stationarity, Sarimax.

I. INTRODUCTION

Machine learning is the study of computer algorithms which improves automatically through the experience. It is a subset of Artificial Intelligence. A time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Time Series analysis can be useful to see how a given asset, security or economic variable changes over time. This research is the parametric approach assumed that the underlying stationary stochastic process has a certain structure which can be described using a small number of parameters, for example using an autoregressive or moving average model.To keep track of sales and prediction of sales it is important for preventing the losses which were made in past and uplift the sales of a company by making smart decisions using the statistical data. Stock price prediction is a method to determine the value of company, so one can buy or sell their stocks accordingly to gain the significant profits out of it[5]. Datasets used are furniture sales dataset and stock price of a company which differ in the stationarity. Furniture sales dataset is the stationary dataset and stock price dataset is non-stationary [6]. For all this we have developed few machine learning models using time series analysis. The main aim of the paper is to make more understanding about the time series analysis, uni-variate forecasting, stationarity concept with the examples of sales and stock price prediction.

II. PROBLEM DEFINITION

STATEMENT OF PROBLEM

For Sales dataset 1. To build and forecast the furniture sales with the dataset consists of daily sales data of a furniture company. Apply time series and build the model to predict the future sales of the company. The dataset contains past 4 years data.

For Stocks dataset 2. To build and forecast the Stock value of a company with the help of Stock price dataset which consists of 5 attributes of stock values of the company. The 5 attributes are open, high, low, close and volume. You will need to find the target variable, prepare the data for analysis and build the model for the prediction of the future stock value of the company.

Time series analysis comprises of the methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Evaluation will move forward in 5 steps which consist of data exploration & preparation, feature engineering, model comparison, model selection and visualization.

III. LITERATURE SUREVEY

Table No. 1

Sr. No	Paper Title	Author	Year	Advantages	Limitations
1.	**Build foundation for time series forecasting**	Ajay Tiwari	2020	Tutorial Data preparation for time series.	Not mentioned what and how to do after preparing the data for time series.
2.	**Time series forecasting**	Richa Kumar	2019	In dept explanation of EDA.	Lacking in explanation for modeling part.
3.	**Forecasting of demand using ARIMA model**	Jamal Fattah, Latifa Ezzine, Zineb Aman, Haj El Moussami, Abdeslam Lachhab	2018	Very useful for the manager of store as demand is a major function for managing a supply chain.	More information required for the implementation part.
4.	**Investigating the pattern of Inflation in Nigeria using SARIMAX model**	Oladapo Ifeoluwa , Ajewole K P, Ayanlowo E A	2020	Model explanation is done very well	Language used is very difficult to understand and can be more simplified.

As shown in table no.1, Ajay Tiwari has Explained the data preparation in "**Build foundation for time series forecasting**" which gives idea for to developer about how to prepare the data for machine learning models. Richa Kumar in "**Time Series Forecasting**" gave the information about the Exploratory data Analysis which developer needs to convert the raw data into useful data. Jamal Fattah, Latifa Ezzine, Zineb Aman, Haj El Moussami, Abdeslam Lachhab in "**Forecasting of demand using ARIMA model**" explained the ARIMA model and build a forecasting model of demand which is useful for the manager of store to mange the supply chain. Ayanlowo E A, Oladapo Ifeoluwa, Ajewole K P in "**Investigating the pattern of Inflation in Nigeria using SARIMAX model** " used Seasonal ARIMA for building a Inflation model which gave the idea of seasonality in the time series data of Inflation in Nigeria.

IV. METHODOLOGY

PROPOSED METHODOLGY

- **For Sales Dataset – Stationary dataset**

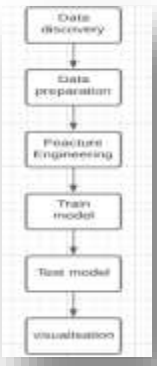


Figure No. 1

To build a machine learning model, there are few steps to be followed. Data discovery, data preparation, feature engineering, training model, testing model and visualization.

Step 1. Data discovery

Data is the most important part of the machine learning. Data discovery means the process of collecting data from various sources and finding out the features of the data.

Figure No.2

In figure no.2 we loaded our dataset to our python frame work and understanding the data, i.e. how many columns are string object, integer or any other.

Step 2. Data Preparation & Cleaning

Data preparation and cleaning involves the removal of null data, replacing the null data with mean and corrections in the data. In short, data preparation is to make data efficient from the raw and poor quality of the data.

Figure No.3

In fig.no 3 we dropped the columns except the variable and target columns. In the above code we also checked for null values and cleaning our dataset.

Step 3. Feature Engineering

Feature engineering means to extract the features from the data with the help of knowledge of the domain or field. Feature Engineering uses different methods or test to find the dataset is fit for building models. For uni-variate time series we need the data to be stationary means the "Standard deviation" and "mean" should not change with the time.

In figure no.4, we are decomposed our data to find out the characteristics of the data. Time series data has 3 characteristics i.e. "trend", "seasonal", "residual". Trend deflects according to time in upward or downward direction. Seasonal is the line which give the idea about the data is repeated in period of time.

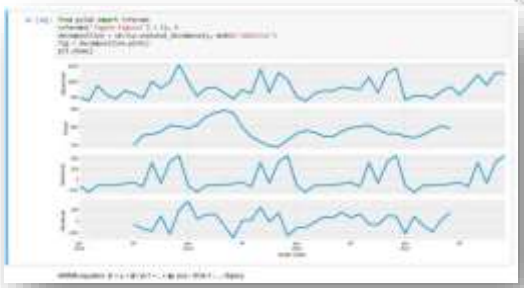


Figure No.4

Then in the feature engineering we check for the stationarity of the data. We used "dickey-fuller test" to check the stationarity of the dataset. In "dickey-fuller test" it uses the hypothesis testing to check the stationarity. Sales dataset is stationary, so we can build the forecasting models with help of this data as shown in the figure no.5.

Figure No.5

Step 4. Model Selection and Training

In time series there are many different models, but from figure no.4 we know our data is in seasonal format. Hence we used the SARIMAX model (Seasonal Auto Regressive Integrated Moving Average model with eXogenous factors). SARIMAX model not only works with three parameter p, d, q, but also requires another set of parameters for the seasonality aspects i.e. p, d, q, s. p means the order of autoregressive, q is the order of moving average and d is the number of differencing as shown in fig. no 6.

Figure No.6

Figure No.7

In figure no.6, using grid search we found out the input parameters for SARIMAX model. And in figure no.7 we input the parameters into the model.

Step 5. Testing of Model

In the figure no.8, the model is tested on the last part of data to find out the accuracy of the forecasting model. The orange line is the static prediction of the sales to check the model is valid or not. To find out the accuracy or line of bestfit of our model we use MSE (mean squared error) which should be close to zero and non- negative value.

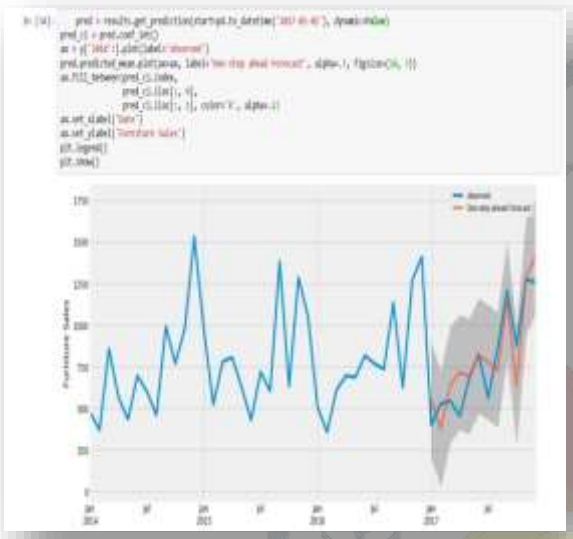


Figure No.8

Step 6. Visualization and Final Forecasting

In this step we will use our tested model to forecast the future sales of the company. In the figure no.9, we forecasting the next 4 years of sales.

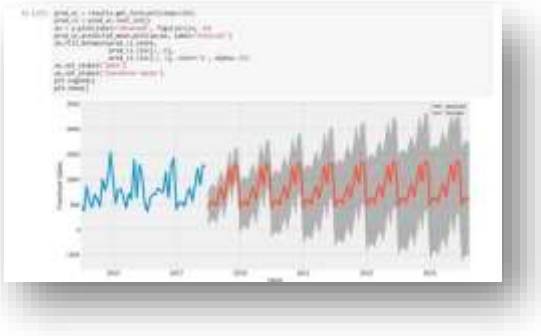


Figure No.9

- **For Stocks Dataset – non-stationary data**

Step 1. Data Discovery

In figure no.10, we loaded the data and understood the features of the data for forecasting future stock price. And in figure no.11 is the plotted raw data for better understanding

Figure no.10

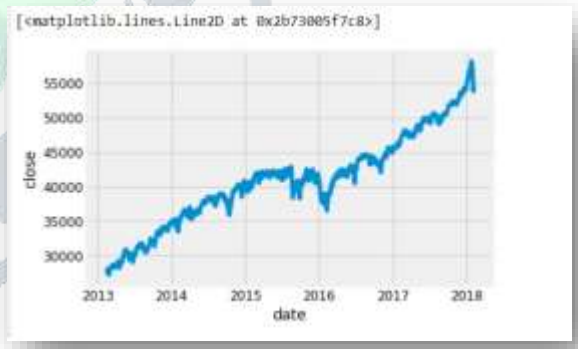


Figure No.11

Step 2. Data Preparation and Cleaning

In the figure no.12, we checked for the null values and dropped the unused columns.

In figure 13, we change the date to index and prepared the time series data.

```
cols=['open','high','low']
s_data.drop(cols, axis=1, inplace=True)
s_data.isnull().sum()
```

```
date      0
close     0
volume    0
Name      0
dtype: int64
```

Figure No. 12

```
s_data=s_data.groupby('date')['close'].sum().reset_index()

s_data=s_data.set_index('date')
s_data.index
```

```
DatetimeIndex(['2013-02-08', '2013-02-11', '2013-02-12', '2013-02-13',
               '2013-02-14', '2013-02-15', '2013-02-19', '2013-02-20',
               '2013-02-21', '2013-02-22',
               ...
               '2018-01-25', '2018-01-26', '2018-01-29', '2018-01-30',
               '2018-01-31', '2018-02-01', '2018-02-02', '2018-02-05',
               '2018-02-06', '2018-02-07'],
              dtype='datetime64[ns]', name='date', length=1259, freq=None)
```

Figure No.13

Step 3. Feature Engineering

For stocks dataset we used rolling statistics for checking the stationarity of data. Rolling statistics is very useful for time series. It's kind of window which will perform the operations on the data on specified size. Here we found the data is non-stationary.

```
# determing rolling statistics
rolmean=s_data.rolling(window=12).mean()

rolstd=s_data.rolling(window=12).std()
print(rolmean, rolstd)

#plot rolling statistics

orig= plt.plot(s_data, color='blue',label='original')
mean= plt.plot(rolmean, color='red', label='rolmean')
std= plt.plot(rolstd, color='black', label='rolstd')
plt.legend(loc='best')
plt.title('Rolling mean and std')
plt.show(block=False)
```

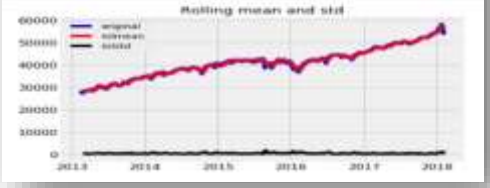


Figure No. 14

There are two ways of removing stationarity from time series data i.e. differencing and transformation. In our data we used the differencing method, it is the process of subtracting the value of observation with another observation over a period of time, as shown in the figure no.15. And plotted the graph.

```
# making data stationary with differencing method(d)
s_data['close']=s_data['close'] - s_data['close'].shift(1)

plt.plot(s_data)
```

```
[<matplotlib.lines.Line2D at 0x2b72eba1ec8>]
```

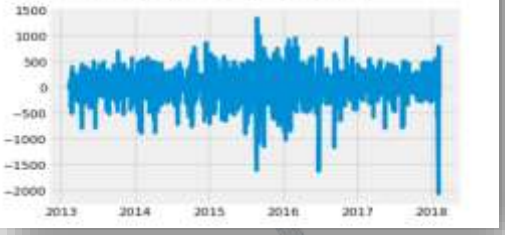


Figure No.15

Then we decomposed to data to find out the seasonality and trend of the dataset. The data is seasonal in nature as show in the figure no.16.

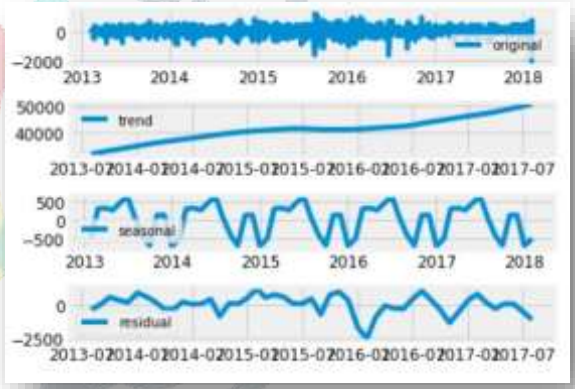
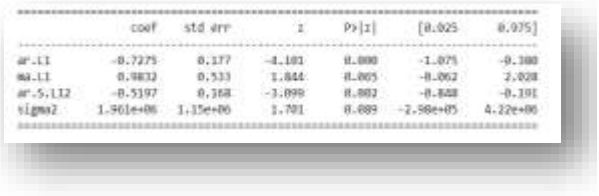


Figure No.16

Step 4. Model Selection and Training

According to the figure no.16, we know our data is seasonal in nature so we selected SARIMAX model. Trained the model with the data as shown in the figure no.17.



	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7275	0.177	-4.101	0.000	-1.075	-0.380
ma.L1	0.9832	0.533	1.844	0.065	-0.062	2.028
ar.S.L12	-0.5197	0.168	-3.099	0.002	-0.848	-0.191
sigma2	1.961e+06	1.15e+06	1.701	0.089	-2.98e+05	4.22e+06

Figure No.17

Step 5. Testing of Model.

In figure no.18, the MSE (Mean Squared Error) of the model came to 1.51 which is fit for the forecasting the future values of stocks.

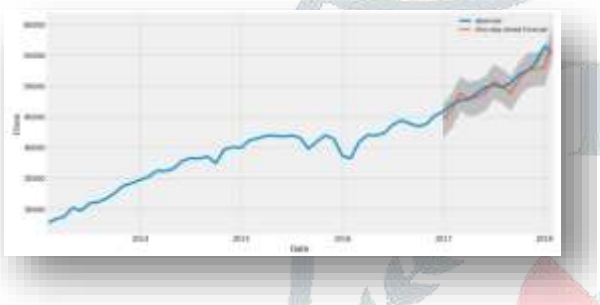


Figure No.18

Step 6. Visualization and Final Forecasting

In the figure 19, forecasting done for the next 4 years for the data.

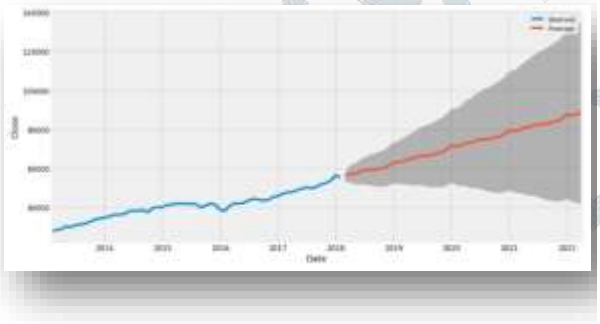


Figure No.19

V. CONCLUSION

The main goal of this paper is to give the better explanation on Time Series with the help of examples. Our research explains the stationarity concept, uni-variate time series analysis, and steps for building a machine learning model. With the help of our research one can have a clear idea about the time series analysis and forecasting methods. Our research paper will be helpful for the ones who are interested in studying data over time.

VI. ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my **prof. Shaik Mohammed Bilal N** to encourage and guide me to do this research work on "**Time Series Analysis: Forecasting with SARIMAX model and Stationarity Concept.**". The project received whole hearted assistance, inspiration, encouragement and valuable guidance in all phases.

VII. FUTURE SCOPE

We can build models for weather forecasting, gold price etc. It would be more useful with the front and general model which can forecast any kind of time series.

VIII. REFERENCES

[1] Jamal Fattah, Latifa Ezzine, Zineb Aman, Haj El Moussami, Abdeslam Lachhab, "Forecasting of demand using ARIMA model", Internal Journal of Engineering Business Management, Oct 2018 - https://www.researchgate.net/publication/328633706_Forecasting_of_demand_using_ARIMA_model

[2] Richa Kumar, "Time Series Forecasting", Academia, 2019 - https://www.academia.edu/44101487**.**

[3] Ayanlowo E A, Oladapo Ifeoluwa, Ajewole K P, "Investigating the pattern of Inflation in Nigeria using SARIMAX model", IJIRMPS International Journal, 2020- https://www.academia.edu/44210567

[4] Ajay Tiwari "Build foundation for time series forecasting", Towards data science, Jul 12,2020 –

https://towardsdatascience.com/time-series-forecasting-968192b3781a

[5] Sales tracking, Commence –https://rb.gy/o3dh3u

[6] Kaggle Dataset - https://www.kaggle.com/datasets

[7]Super_Store.csv, Imarticus Learning -https://rb.gy/f8yozu