

COMPANY STOCK INVESTMENT

SUBMITTED BY: NANSHIN LAZAK

ACKNOWLEDGEMENT

Fliprobo Live Session expositions on Time series

Scikit-learn documentation

SME- Astha Mishra

Stackoverflow

Vidyaanalytics

Towards data science

INTTRODUCTION

The Company Stock Investment project attempts to display the statistics of the performance of the company’s investment in various sectors specifically Oil sector, gold sector, company stock itself and othershares Investments over a certain period of time (2009 – 2017). The analysis will be utilizing the past data to be able to predict the best sector which may be most lucrative and profitable for the company to invest in in the future years.

The domain related to this project is the business industry which has the aim of maximizing profits from its investments. Investors are always searching for sectors which will yield the optimal profit possible. This will avoid a waste of time and other resources in investing in sectors with less or without profit.

REVIEW OF LITERATURE

Time series analysis is a concept that tries to assess statistics with relation to time. It considers output over a period. Hence the time series analysis is usually implemented where only two features are available i.e time as the variable and the output as the target. The resources researched on this topic include;

Time series analysis in python by edureka YouTube video which explained what time series is, components of time series, when to apply time series, stationarity of data, ARIMA models and a demo forecast using the ARIMA model.

The complete guide to time series analysis and forecasting by towards datascience where time series was introduced, and a few time series models were also implemented.

MOTIVATION

This Project may serve as a guide for decision towards the most profitable sector to invest in. such business decisions are critical to every business because the risk involved is greatly reduced. Investment decision need substantial information else devastating consequences may evolve.

MODELLING

The ARIMA and SARIMAX models where used to test the most effective to be implemented for the forecast.

DATA SOURCES

The source of data is Fliprobo tech. it was obtained as the company stock investment project as a csv file.

DATA PROCESSING

Data had no missing values, however the original data was modified to have one output as the target therefore the other features were dropped to be tackled using a separate sheet. Three of the investment sectors are considered with the assumption that the forth is less relevant to the company.

HARDWARE AND SOFTWARE USED

A windows 10 laptop is utilized.

The software utilized for this project include;

The Python programming language

Jupyter Notebook which gives a platform for input and output of the programs.

Pandas library which is a tool for analyzing and manipulating data in the python programming language

Numpy library which is a tool for the analysis of numerical data in the python programming language

Matplot library which is a tool for analysis graphical data in the form of charts and plots

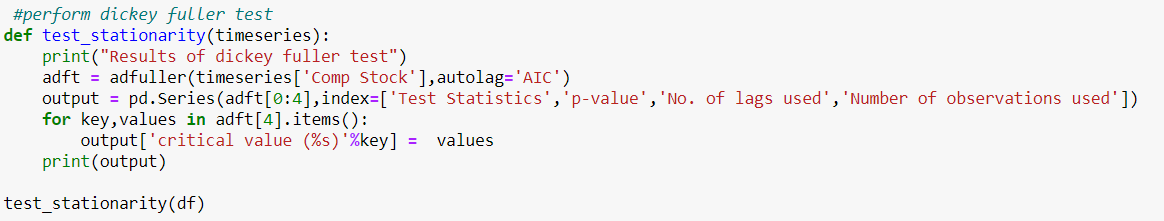
Statsmodels Library which is a tool for statistical analysis of the data.

Google: used for stuffing the internet for obtain information over grey areas that arouse in the cause of the project execution.

Microsoft Office Package specifically the Microsoft word used for the compilation of the project report

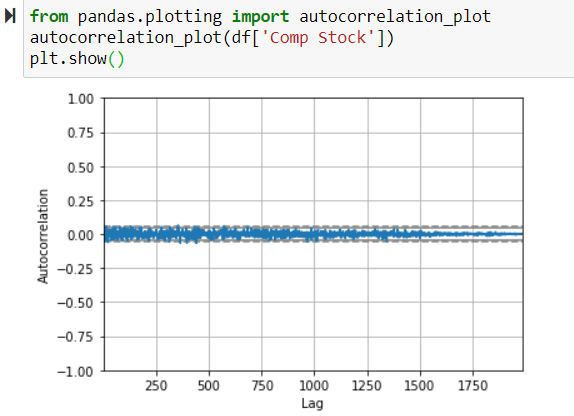
MODELS/DEVELOPMENT AND EVALUATION

These problem was solved by uploading the dataset set given unto the jupyter notebook, than visualizing the data and other details such as shape, description, datatypes, information, missing values etc. the data was plotted to check if it’s a trend, seasonal, cyclic or irregular. The plot revealed that the data was a trend and so the next step was to check whether the data is stationary. This is applied with the aid of a p-value result obtained from the augmented dickey fuller test (adfuller) imported from the statsmodel. A p-value less than 0.05 indicates that the data is stationary while greater than 0.05 indicates that the data is non-stationary. In this project, the p-values for Oil Investments, Gold Investments and Comp Stock were less than 0.05 to signify that the data is stationary.

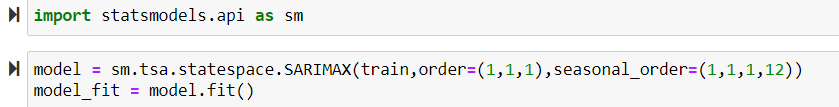


Timeseries models require that data must be stationary before analysis can be implemented. The Time series model applied for this project is the Auto Regressive Integrated Moving Average (ARIMA) which is a combination of two model, the Auto Regressive and the Moving Average.

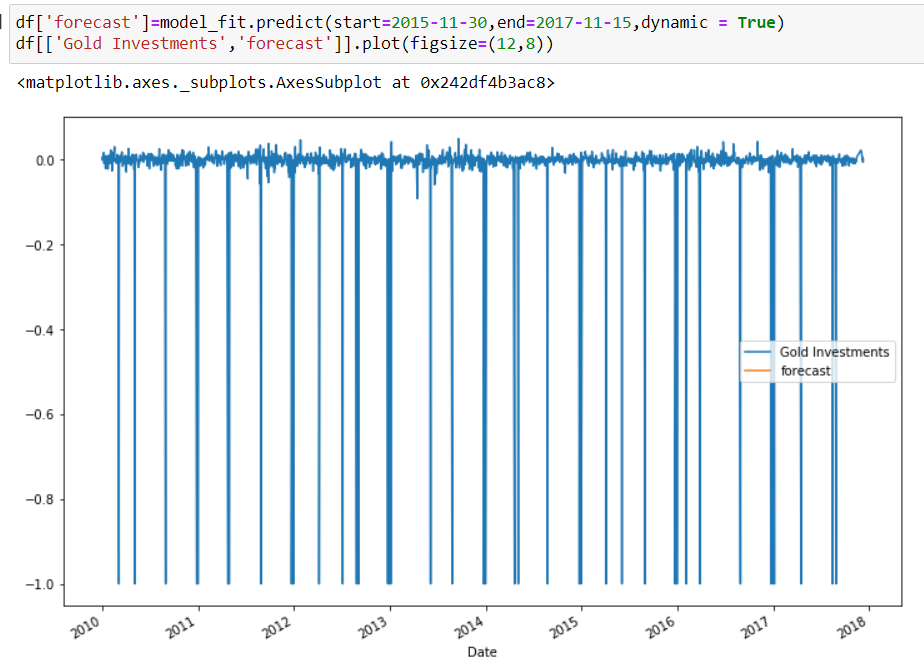
ARIMA comprises of the Partial autocorrelation(PACF) tool which can be imagined as the correlation between the series and its lag and the Autocorrelation(ACF) tool which tells how many Moving Averages terms are required to remove any autocorrelation in the stationarized series. The sample autocorrelation plot and the sample partial autocorrelation plot are compared to the theoretical behavior of these plots when the order is known.

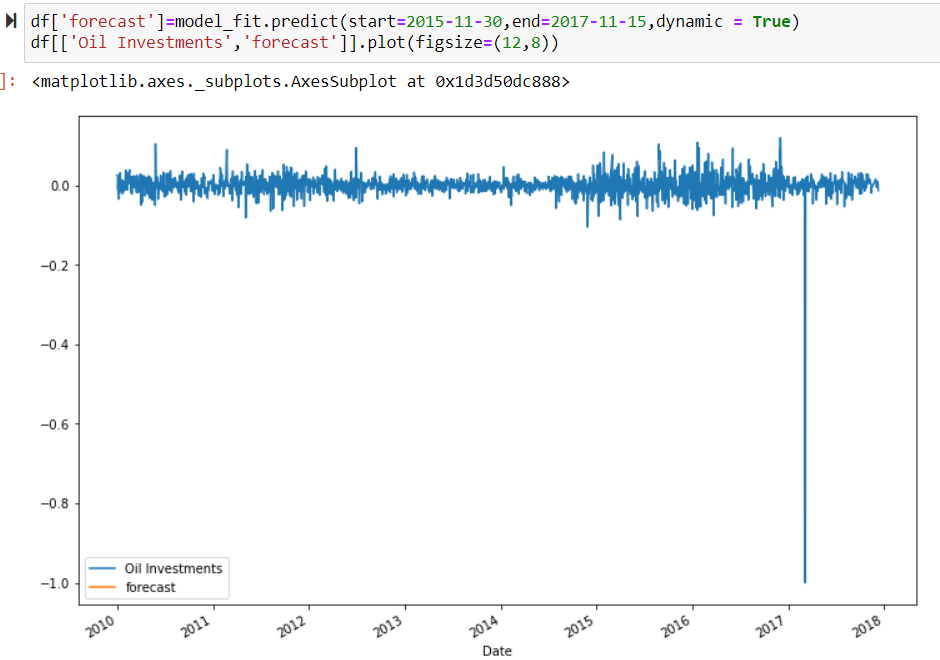


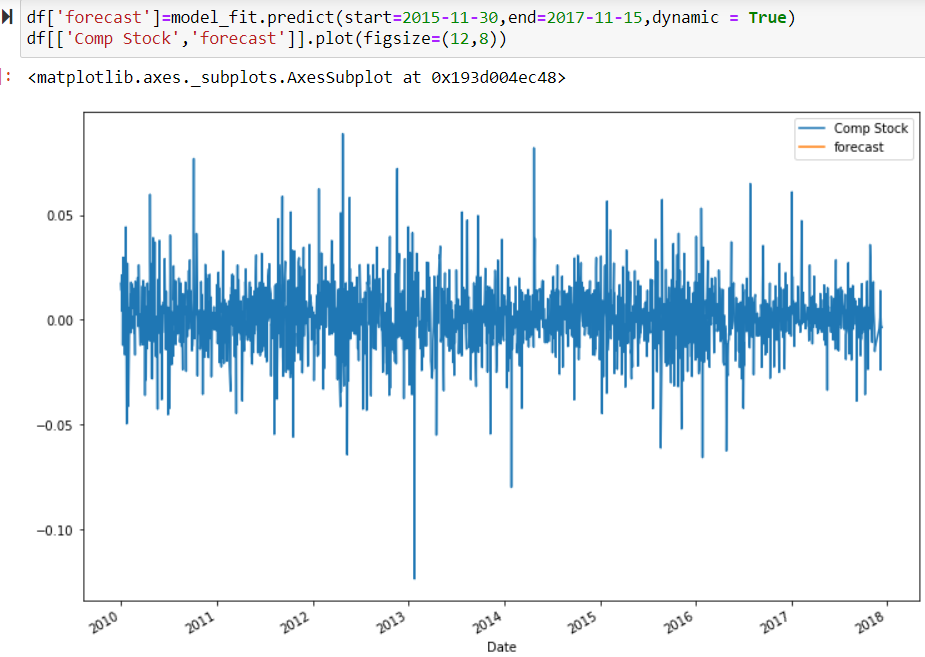
The SARIMAX which is a version of the ARIMA model used when the data is seasonal is imported from the Statsmodels and fitted to yield a summary after which the forecast is implemented and shown on a plot.



After fitting the model the trained model the test data is applied for a forecast for Gold, Oil and Comp Stock investments







A new dataset is then created to make predictions for a range of say 24 months (2018-2020). A plot of the model results indicates the forecasted output along side the initial data.

is less than the ‘Critical Value’, the null hypothesis can be rejected and it can be

said that the series is stationary.

The two major reasons for the time series to be non-stationary is the trend and the

seasonality. The series is made stationary by estimating the trend and seasonality and

eliminating them from the series. For this purpose the logarithmic transformation and

differencing methods are applied.

Step 3. Deduction of Optimal Parameters

ACF and PACF are used to determine the suitable model parameters.

Step 4. Model Validation

This step involved validating the model using statistics and conﬁdence intervals and

tracking of model performances.

Step 5. Forecast

The best model obtained is implemented on the series and used to forecast the future

values. The values are reverted back to the original scale.

6 Result Analysis

The electricity consumption in IIT(ISM) during the period July 2004 to June 2009 is

depicted in Figs. 1and 2represents the constituents of the time series viz. are trend,

seasonality and residual values. It can be observed that the electricity consumption data

contains both an overall upward trend and has a seasonality to it and thus, seasonal

ARIMA was used for forecasting.

The presence of trend and seasonality makes the data non-stationary and the same

can be conﬁrmed by the rolling statistics and Dickey-Fuller test on the electricity

consumption data as illustrated in Fig. 3and Table 2respectively. Although slight

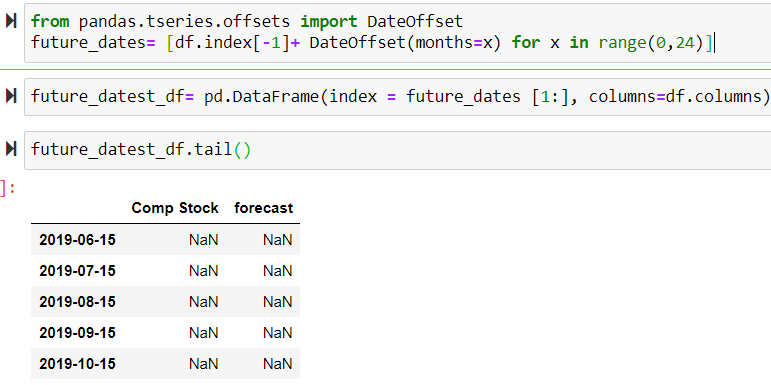
change in standard deviation is seen, but it can be clearly observed that the mean is

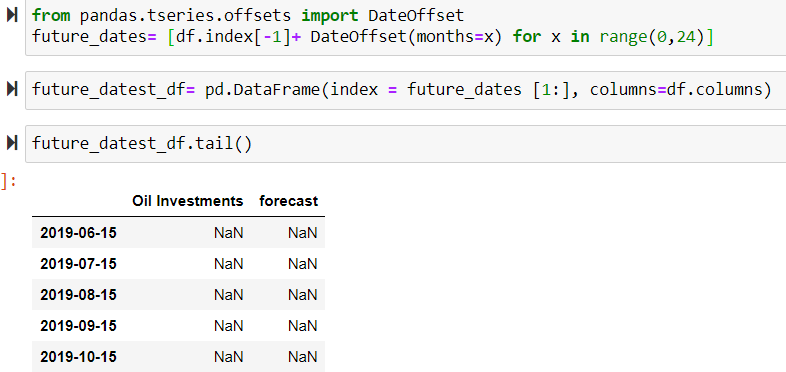
varying with time. Also, the test statistics conﬁrm the same since it is greater than the

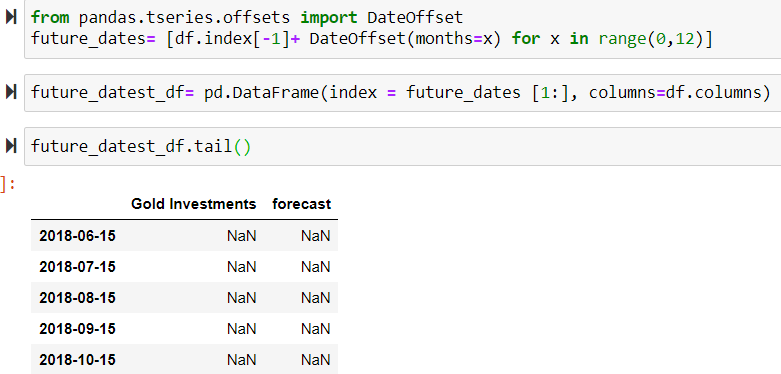
the critical values.

Fig. 1. Electricity consumption data for period 2004–09.

330 P. K. Jain et al







Unfortunately, the output of my predictions are yeilding NaN values this could be probably as a result of noise inherent in the dataset, hence the model is unable to perform on future dates.

CONCLUSION

In this article, we used a company’s (2009-2017) investment record on the SARIMAX model to predict what sector will be most lucrative in the future years.

* I learnt how to implement the ARIMA model
* The ARIMA can also be referred to as the Box-Jenkins Method
* The time series train-test split procedure is appropriate when you have a very large dataset, to train part of the data, test the trained data and make predictions.
* I did not clearly comprehend how to use the scikit-learn machine learning library to perform the train-test split procedure.
* I did not also comprehend why my output yielded the NaN values.