Sales Prediction And Analysis Using Time Series Analysis

**ABSTRACT**

Time series techniques are one amongst the commonly used method that we encounter knowingly or unknowingly almost every day. Research frequently uses data (i.e. marketing mix variables) that is equally spaced over time. Here, statistical techniques are best suited to find variables dependency on time. It is applied for causality analysis as well, but they are apt for discovering the effects of a particular event with time, which makes it a fantastic algorithm for marketers. This paper starts with the important concepts, statistic theory and demonstrates its current use in marketing research. We use mathematical analysis to predict order for a brand or product within the market using scanner data from major retail stores. We can approach these prediction tasks using different methods counting on the specified quality of the prediction, length of forecast period, and, of course, the time within which we've to settle on features and tune parameters to get appropriate results. Most data scientists will come across this algorithm within their learning journey and the knowing to create a working model is a crucial in data science field. The basic agenda is to directly show a time-series like a mixture of different patterns like daily basis, weekly basis and yearly, alongside an overall trend. For example, the demand might rise within the summer and reduce in the winter, but have an overall decreasing trend as you increase the energy efficiency of your home. This model can display both patterns/trends and help in prediction making by supported observations. Some methods in this algorithm are Autoregression (AR), Moving Average (MA), ARMA, ARIMA (Integrated) and SARIMA (Seasonal).

**KEYWORDS**

Time Series Analysis, ARIMA, Sales Prediction, SARMA

**INTRODUCTION**

Time series is a series of data (points) indexed in time order. These points are indexed in equal intervels of time.

Despite being a strong tool, this technique is not used that often in research by marketers. the main reasons for this reluctance, they mention the supply of quality time series, the unavailability of an efficient software, absence of data and a reluctance to use secondary data for modelling customer’s behaviour. At an equivalent time, they announce that wide use of your time series remains to return with advances in information technology, software development and an increasing number of educational studies dedicated to the topic. This project presents uni-and multi-variate analysis techniques applied to plug research forecasting. The remainder of the project is organized as follows: first we present the essential statistical analysis, starting with univariate ARIMA models.

Time series (data) have a natural characteristic ordering. This makes the examination unmistakable from cross-sectional investigations, during which there's no common requesting of the perceptions (for example clarifying individuals' wages by respect to their separate training levels, where the people's information may be entered in any request). This investigation is moreover particular from spatial information examination where the perceptions regularly identify with topographical areas (for example representing house costs by the circumstance additionally on the grounds that the natural attributes of the houses). A stochastic model for statistics will by and large mirror the very actuality that perceptions estimated in time will be more firmly related than perceptions further separated. also, time arrangement models will regularly utilize the common single direction requesting of your time all together that qualities for a given period will be communicated as getting in how from past qualities, rather than from future qualities.

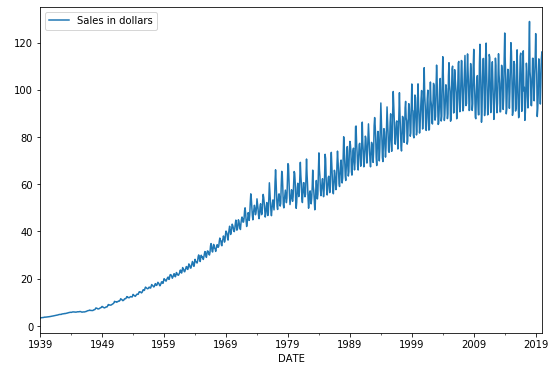


Fig 1. This is an example of time series data plotted in a line graph

**EXISTING SYSTEM APPROACH**

Before understanding ARIMA concept and its application it is important to know a few basic aspects of ARIMA:

The ARIMA model includes a class of statistical models for analyzing and predicting values in time series data. It directly relies and trains itself to a set of points or values in time series data, and intrinsically provides an easy yet reliable method for creating skillful statistical forecasts and models. ARIMA stands for AutoRegressive Integrated Moving Average.

This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

* AR: Autoregression. The autoregression (AR) method models or forecasts the next value in the sequence as a linear function of the observations at prior time steps. It finds the dependent relationship between an observation and a few number of lagged observations.
* I: Integrated. The making use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) so as to form the statistic stationary.
* MA: Moving Average. A model uses the dependency between an observation and a residual error from a moving average model applied to lagged observations and predicts the next value in the sequence as a linear function of the residual errors from a mean process.

Seasonality, trend, and noise are the three main components in ARIMA . These parameters are labelled p, d and q.

1. ‘p’ is that the parameter related to the auto-regressive aspect of the model, which includes past values. for instance, forecasting that if it rained tons over the past few days, you state its likely that it'll rain tomorrow as well.
2. ‘d’ is that the parameter related to the integrated a part of the model, which effects the quantity of differencing to use to a statistic. Imagine an example of this as forecasting that the quantity of rain tomorrow to be almost like the quantity of rain today, if the daily amounts of rain are similar over the past few days.
3. ‘q’ is that the parameter related to the moving average a part of the model.

**PROPOSED SYSTEM APPROACH**

1. Packages- These are the packages I imported before executing code.



Fig.2

1. The Dataset - I have used the dataset of a retail store from 1970-2010. The main parameters or features of this dataset is “Order Date” and “Sales”. The sales have been recorded on every 1st day of the month. Using past data of the set, I have trained and tested the model. In the end, I predicted the future sales of the store as well.



Fig 3. A small sample of the dataset and how to load the data on Python.

1. Some basic info about the dataset:

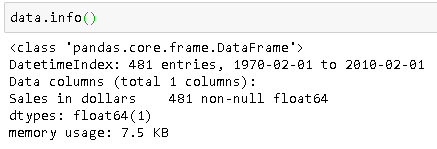


Fig 4

Next I plotted the data to visualise it and for better understanding,

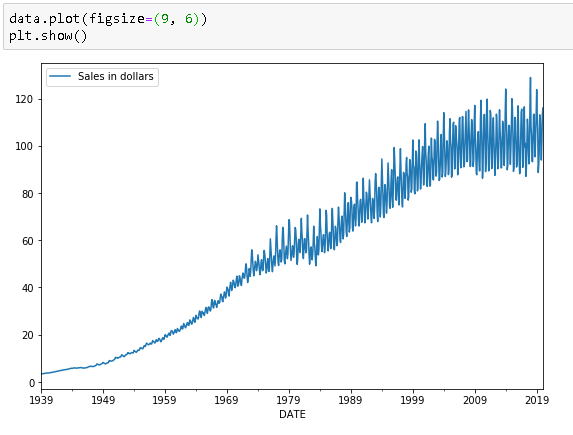


Fig 5. Line Plot

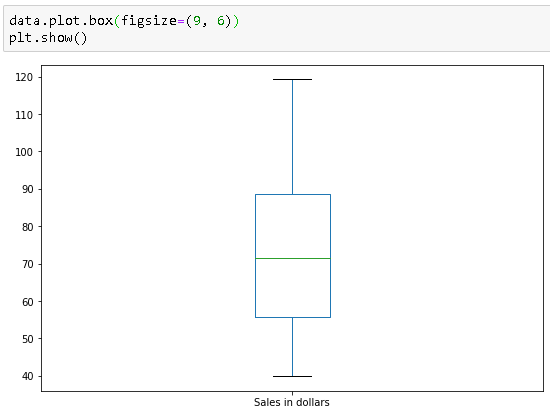


Fig 6. Box Plot

1. Fitting time series data with a seasonal ARIMA model,

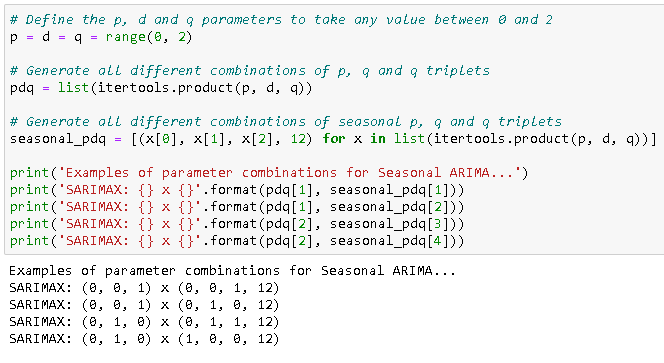
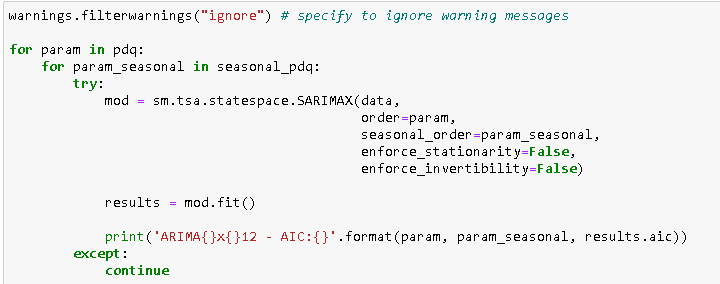
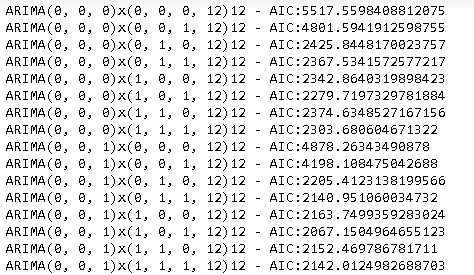


Fig 7. Initializing ARIMA model with output below.

Now, I found out the AIC (Akaike Information Criterion) value, which is conveniently returned with ARIMA models fitted using statsmodels. The AIC measures how well a model fits the info while taking under consideration the general complexity of the model. A model that matches the info alright while using many features are going to be assigned a bigger AIC score than a model that uses fewer features to realize an equivalent goodness-of-fit. To get an accurate model, it is important to select the least AIC value as reference.

Fig 8.

It’s Output:



**…**

**…**

**…**

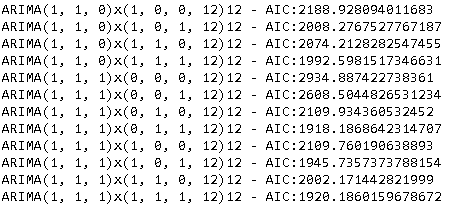


Fig 9.

After analysing the output, I considered ARIMA(1,1,1)x(0, 1, 1, 12)12 for making the model as a it has the least AIC value.

Fitting Time Series Model

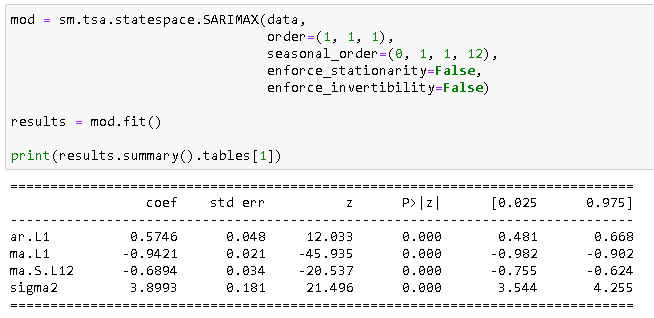
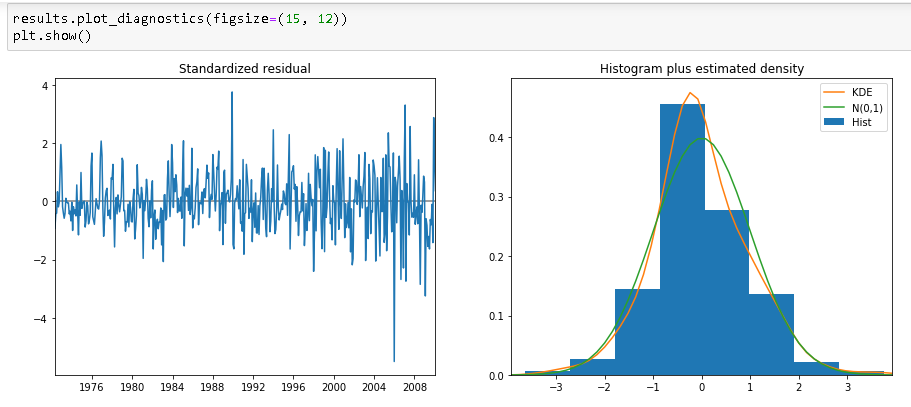


Fig 10. Creation the model with its output. Output summaries the model.

Generating model diagnostics and investigating for any unusual behaviour



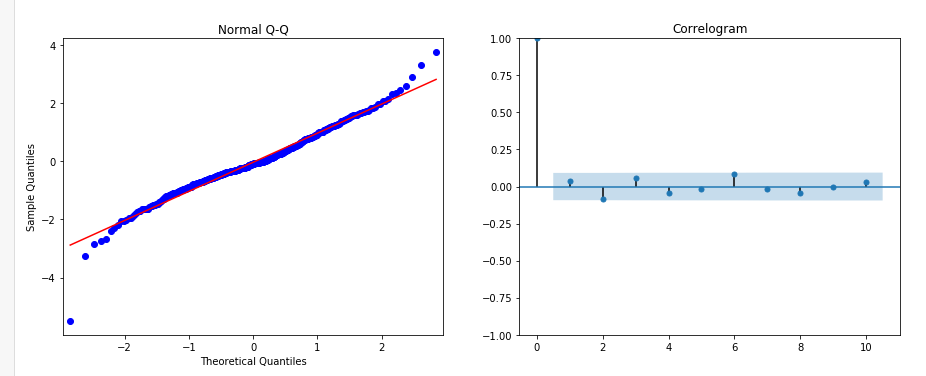
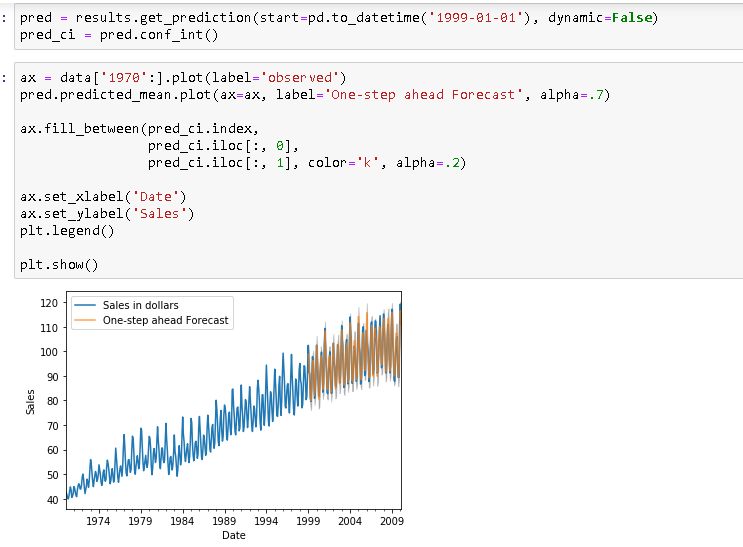


Fig 11. Checking for usual behaviour by visualization.

Forecasting and validation



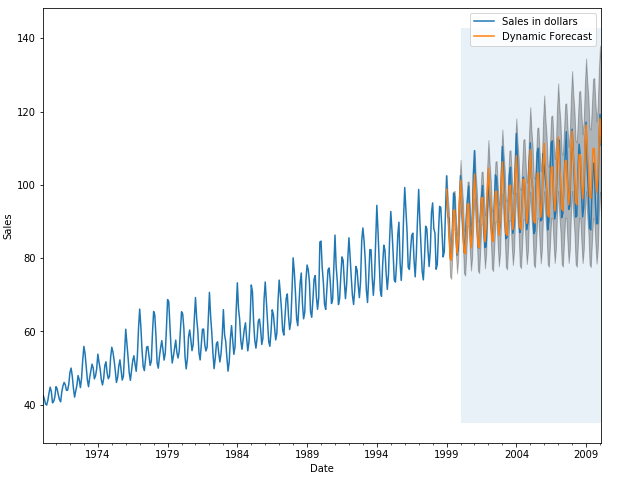


Fig 12. The data taken is from 1970 and the prediction starts from 1999 to 2010.

Calculating MSE (Mean Square Error)

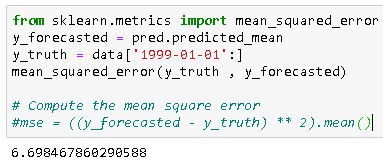


Fig 13. MSE Value

Visualizing Forecasts and Predicting future Sales.

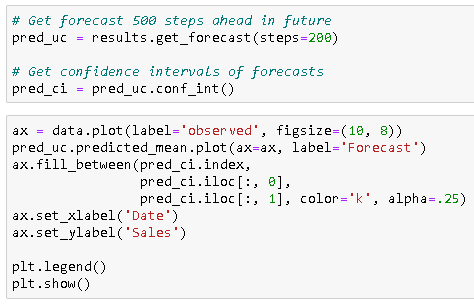


Fig 14. Code for predicting the next 200 steps.

Output

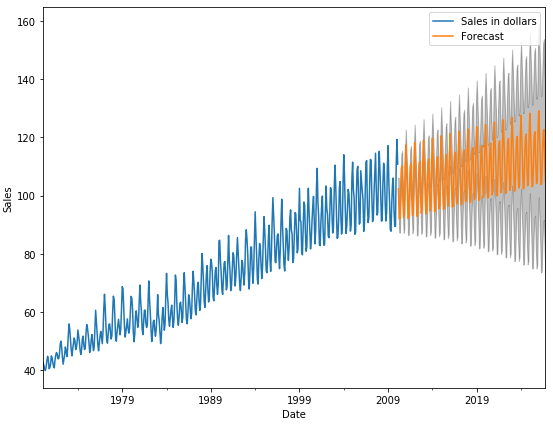


Fig 15. Line graph for the future prediction.

**FUTURE SCOPE**

The traditional prediction models considered for proposing hybrid models are ARIMA, GARCH and ANN models. Prediction accuracy are often improved by verifying with other traditional prediction models like SVM, fuzzy, HMM, and other techniques. during this thesis, MA filter based preprocessing technique is taken into account . Many other preprocessing techniques exist within the literature which may experimentally verified for better prediction accuracy. Pre-processing techniques can also be devised aptly to extend the prediction accuracy to suite the prediction model considered. The PI technique considered in devising the prediction model are often modified so on include a strategy to decide on the amount of partitions rather than trial and error. Also, the prediction performance are often studied by the inclusion of covariates, meaning development of multivariate models, but using minimum number of model parameters.

**Conclusion**

In this paper, I described the way to implement a seasonal ARIMA model in Python. I made extensive use of the pandas and statsmodels libraries and showed the way to run model diagnostics, also as the way to produce forecasts of the sales.

Here are a couple of other things that one could try:

1. Change the beginning date of the dynamic forecasts to ascertain how this affects the general quality of your forecasts.
2. Try more combinations of parameters to ascertain if it improves the goodness-of-fit of your model.
3. Select a special metric to pick the simplest model. for instance, I used the AIC measure to seek out the simplest model, but seek to optimize the out-of-sample mean square error instead.

For more practice, also attempt to load once more series dataset to supply your own forecasts.