Sales Forecasting using Machine Learning

Aparna T S
PG Scholar
Amal Jyothi College of Engineerin
APJ Abdul Kalam University
Kottayam, Kerala
aparnats2909@gmail.com

Gloriya Mathew
Assistant Professor,
Department of Computer Applicatio
Amal Jyothi College of Engineering
Kottayam, Kerala
gloriyamathew@amaljyothi.ac.in

In this article, [2] employ a range of forecasting techniques to project Amazon's future quarterly net sales based on historical quarterly statistics. It suggests three potential forecasting methods based on the historical data pattern: ARIMA, neural network auto regression, and Holt-Winters exponential smoothing.

Forecasting economic time series with significant trends and seasonal patterns is the focus of [3]. Major goal is to examine the forecasting accuracy of additive and multiplicative Holt-Winters models and to shed fresh light on the techniques employed in this way.

The autoregressive integrated moving average (ARIMA) time series forecast model and propose an unique framework for predicting COVID-19 infections is constructed in [4]. The ARIMA model is constructed with the best parameters, as shown by [4], and the accuracy of the model is 93.615.

The authors of [5] will use the ARIMA Box-Jenkins method to forecast sales of plastic manufacturing for 2015. The ARIMA algorithm in SAS will be used in this study to identify, estimate, and forecast time series models. The accuracy of outcome prediction is determined using the MAPE (Mean Absolute Percentage Error) score.

III. Methodology

The main aim of this system is to analyse and predict the future sales using time series forecasting models.

A. Data collection

During 10,000 product purchases over a four-year period are included in the dataset, which contains three main categories. Row ID, Ship Date, Ship Model, Country, City, State, Postal Code, Region, Product ID, Product Name, Quantity, Sales, Order Date, Category, Sub-Category, Segment, Order ID, Discount, and Profit are the features of this dataset.

B. Data Preprocessing

The dataset contained a number of factors, some of which were either useless or had negligible effects on product demand.

Steps involved:

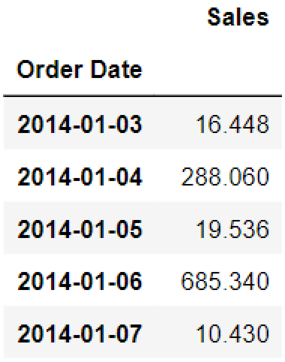
- Dropping unwanted columns.
- Checking for null values and dropping them.
- Converting the data type and sorting the data based on date.
- Grouping and summarizing the values based on the date.

As a result, only the Order Date and the Quantity sold were kept. The data was resampled using the average daily amount

DOI: 10.5281/zenodo.7961138

sales for that month and the month's beginning as the timestamp in order to obtain month-wise predictions.

Training and test data were separated from the processed data according to date. In this project, training will utilise 70% of the data, while testing will use 30%.



	Sales
Order Date	
2014-01-03	16.448
2014-01-04	288.060
2014-01-05	19.536
2014-01-06	685.340
2014-01-07	10.430

Fig.1.Sample Data

C. *Forecasting Models*

As a result, only the Order Date and the Quantity sold were kept. To get month-wise predictions, the data was resampled using the average daily amount sales for that month and the month's start as the timestamp.

1. *ARIMA, SARIMA*

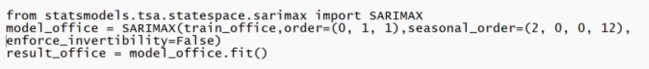
Seasonal ARIMA is used to train the model based on dates and category. It's a development of ARIMA, that includes a seasonality component. As a result, SARIMA is utilised rather than ARIMA because data is discovered to demonstrate seasonal.The "autoregressive" component of the equation accounts for the pattern of increase or fall in the data, the "integrated" part for the rate of change, and the "moving average" part for any noise between adjacent data points. Seven hyperparameters make up SARIMA: - The seasonal part is (P,D,Q)s while the non-seasonal part is (p,d,q).The general formula for SARIMA(p,q,r) x (P,Q,R,s) is given as:

$$\phi_P(B^s)\varphi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t$$

Based on the trend seen in the four years of training data, the model can predict demand for each product category in the next months. A time series is a collection of well defined data collected over a period of time at regular intervals. It should be highlighted that information gathered over an erratic time period does not constitute a time series. By examining the time series data, significant statistics and data features can be discovered. By applying the right model to it, it may be utilised to identify trends in the data that can be highly helpful for forecasting and monitoring the data points. The average of daily sales for each month has been utilised because the data has so many different categories and subcategories that dates may be challenging to handle. An estimator called AIC is used to select the best values for p, d, and q. The solution is more ideal when the AIC value is smaller. It is given as:

AIC= -2log(L)+2(p + q + k)

Where, L=Likelihood of data
K=intercept of ARIMA value



```
from statsmodels.tsa.statespace.sarimax import SARIMAX
model_office = SARIMAX(train_office,order=(0, 1, 1),seasonal_order=(2, 0, 0, 12),
enforce_invertibility=False)
result_office = model_office.fit()
```

Fig.2.SARIMA

2. *Exponential Smoothing*

Exponential smoothing is a prominent time-series forecasting technique. This model, which is similar to the ARIMA approach, can identify trends and seasonality. However, it gives each observation a weight that decreases exponentially rather than using the weighted linear sum of prior data points (Brownlee, 2018b). In other words, it gives the observations made through time less weight.

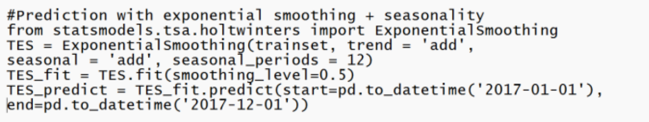
The three basic types of exponential smoothing can be categorised. For forecasting univariate data without a trend or seasonality, use Single Exponential Smoothing (SES). Double Exponential Smoothing (DES), which supports trends in time-series data and Triple Exponential Smoothing (TES), which can also manage seasonality.

The Triple Exponential Smoothing technique can be used to identify a seasonal trend in the time-series (also known as the Holt-Winters method). This model goes beyond DES by including extra hyperparameters like Gamma that control the effects of seasonal components. This approach, the most complex among exponential smoothing models, allows for the use of either linear or exponential seasonality. This technique can be used to set the seasonal cycle length, or L. (much like SARIMA). The seasonal period that repeats annually in this monthly dataset is 12. The levels of the series, growth, and seasonal component are represented mathematically by the variables y, bt, and ct in the following equations. The parameters α, β, and γ are usually limited between 0 and 1.

$$y = \alpha\ xt\ ct{-}L + (1 - \alpha)\ (\ yt{-}1 + bt{-}1)$$

$$bt = \beta\ (\ yt - yt{-}1) + (1 - \beta)bt{-}1$$

$$ct = \gamma\ xt\ yt + (1 - \gamma)ct\text{-L}$$



```
#Prediction with exponential smoothing + seasonality
from statsmodels.tsa.holtwinters import ExponentialSmoothing
TES = ExponentialSmoothing(trainset, trend = 'add',
seasonal = 'add', seasonal_periods = 12)
TES_fit = TES.fit(smoothing_level=0.5)
TES_predict = TES_fit.predict(start=pd.to_datetime('2017-01-01'),
end=pd.to_datetime('2017-12-01'))
```

Fig.3.Exponential smoothing

D. *Evaluation*

Based on the statistical metric known as Root Mean Square Error(RMSE), the models compared.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{pred,i} - y_i)^2}{n}}$$

The predicted amount is denoted by yi, while the predicted quantity with n variables is denoted by yj.

IV. Results and Discussions

This section discusses the findings. A trainset containing 70% of the dataset is used to train each model. The remaining information is utilised to test each model.

First, the proposed system built a simple moving average model with SARIMAX(0,1,1)x(2,0,0,12) ie, the hyperparameter value for SARIMA model is yielding the lowest AIC value 493.072. This is the most optimal option.

Then the proposed system the value for the next year as shown in Figure.2.

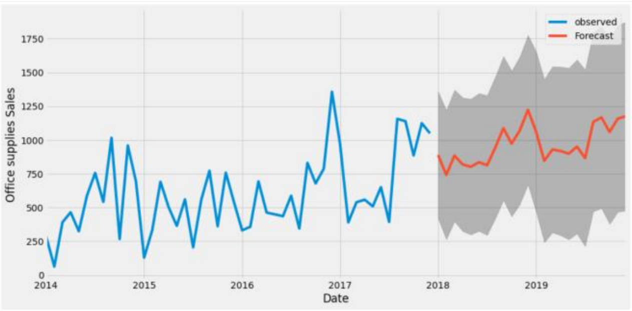


Fig.4 Forecasting using SARIMA

Another traditional time-series forecasting method used in this study is called Triple Exponential Smoothing (TES). Although TES has done better, SARIMA still outperformed it. Figure.3 illustrates the TES sales forecasting.

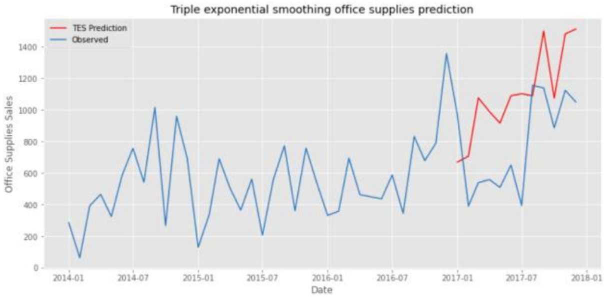


Fig.5 Forecasting using TES

V. Best model selection

In order for retailers to properly manage their inventory, it will be preferable if the forecast for the quantity of things sold is more precise. In order to compare the models, we are utilising RMSE values. We discovered that the RMSE value for the ARIMA model was 303.63 whereas the RMSE value for the Exponential Smoothing model was 411.87 for the same test data of the office supplies category.

Therefore, SARIMA model with RMSE of 303.63 can be classified as an acceptable model.

Model	**MSE**	**RMSE**
SARIMA	92188.29	303.63
TES	169634.81	411.87

Fig.6 Prediction results

VI. Conclusion and Future Work

Supply chains must include sales forecasting in order to improve and update stock, boost sales, cut costs, profit, and customer loyalty. Thus, models like the ones discussed here can be used to find patterns in historical data, whether they are basic or complicated, that may not be immediately obvious to us. The patterns for the upcoming years can then be predicted using this information. The benefit of forecasting is knowing how many units customers may purchase to maintain the production level. It aids the e-commerce platform in overcoming the decline in sales and boosts their profit margin.

Based on the RMSE values, ARIMA models perform well for our dataset. Despite the dataset only having a few product categories, the model can be trained to predict any category with the correct training data. More features, such as social media feedback, economic studies, shopping trends, location-based demographic data, goods, etc., can improve the model's effectiveness. This model works effectively for ordinary, everyday commodities because it primarily considers seasonal changes in demand.

References

[1] Ranjitha P, Spandana M, "Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms", Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)

[2] Balpreet Singh, Pawan Kumar, Dr. Nonita Sharma, Dr. K P Sharma, "Sales Forecast for Amazon Sales with Time Series Modeling", 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)

[3] Susana Lima, A. Manuela Gonc¸alves, Marco Costa, "Time Series Forecasting using Holt-Winters Exponential Smoothing: an Application to Economic Data", Proceedings of the International Conference of Computational Methods in Sciences and Engineering 2019 (ICCMSE-2019)

[4] Leila Ismail, Member, IEEE, Shaikhah Alhmoudi, and Sumyah Alkatheri, " Time Series Forecasting of COVID-19 Infections in United Arab Emirates using ARIMA **",** 2020 International Conference on Computational Science and Computational Intelligence (CSCI)

[5] Baihaqi Siregar, Erna Budhiarti Nababan, Alexander Yap, Ulfi Andayani, "Forecasting of Raw Material Needed for Plastic Products Based in Income Data Using ARIMA Method", 2017 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)