Sales Forecasting and Data Analysis of Product Based Company

Faiz Khwaja
Northeastern University
(khwaja.f@northeastern.edu)

Abstract: Sales forecasting and drawing business insights are essential components of any product-based company. Efficient ways of forecasting future sales and trends of a supply chain company is critical to maximize profit of the company. To achieve the best results in sales forecasting, new technologies are tested and brought into use in the industries. Machine learning techniques allows for predicting the number of products/services to be purchased during a defined future period. Forecasting models developed by ML can be very beneficial in enhancing the productivity of a traditional product-based companies by employing different ML models forecasting and using these models to improve the efficiency of the product-based company. In this report I would help expand our knowledge in the application of ML and statistical models in sales forecasting. Moreover, I would draw business insights and construct different visualization plots to analyze the data better.

Keywords: Industry 4.0; Product based Companies; Warehouse Management; Statistical Models; Machine Learning; Supply Chain Management.

1 Introduction

The world has been moving towards a digital future and the use of Machine Learning and Deep Learning technologies have played a major role in various Industries. Accurate Forecasting is vital for any product-based company's progress, it has a direct influence on the company's performance and growth. When it comes to supply chain(product-based) companies, the major problem they face is of managing their data and drawing insights from it. This data can be used for forecasting, improving productivity in the inventory and help in making informed decision about maximizing the profit of a company. But not using the right technologies for forecasting can derail their advancement. In this paper I have used Facebook's prophet model and SARIMA model to forecast sales of office supply and furniture of a particular product-based company. Moreover, I have evaluated these models on various metrics and drawn insights for the same. Although, the models and insights are for one particular company's data, these techniques can be transferred to any other product-based company's data.

2 Literature Survey

In 21st Century forecasting has been an essential component in the supply chain companies. Today, most of the product-based companies want to boost their sales as fast as possible to increase customer satisfaction and revenue. Some small-scale supply chain companies should employ ML and Deep Learning technologies effectively in order to quickly scale up and compete with some of the big supply chain companies.[1]. A case study on many Machine Learning (ML) models using real-world sales data from a mid-sized restaurant. Accuracy in sales forecasting provides a big impact.[2] and competitive advantage to companies. With the advancements in AI and ML Technologies, warehouse management system has changed the way companies send and receive orders.[3] Traditional systems can be problematic because as a company's volume order increases, the system of receiving and distributing orders becomes more complex. Companies often expand product offerings, meaning warehouse management systems must adapt to new inventory and the changing business.[4] A poor warehouse management system and forecasting technique can have deleterious effect on a company's profit, making company lose money by not accurately keeping track of orders. The package picking, sorting and dispatching operations are logistic processes.

3 Data

3.1 Dataset and labels

The dataset using for this project was obtained from Kaggle.[6]. There are 24 columns in the dataset. The timeseries date column used in this data set for forecasting is the 'Order Date'. The label i.e., the variable which is being forecasted into the future is the 'Sales' columns, which is the total revenue generated by the company. The resolution of the data is per day from January 2014 to December 2017.

3.2 Data Preprocessing and Cleaning

The resolution of the data was per day. The data was converted from per day sales to mean sales per month. This was done in order to ensure that there is less fluctuations in the data and models used for forecasting can fit giving a better accuracy. Moreover, there were 10 days for which the sales were not present. These days were dropped and not considered in forecasting sales. There was also data type conversion performed in order to ensure that data columns have the right data type. The 'Order Date' column was interpreted as object type, it was converted to 'datetime' format in order to avoid any errors while fitting the different models. Furthermore, the states were converted to their two-letter abbreviation so they can be plotted on that map. Also, month column was extracted from 'Order Date' to perform analysis on sales according to the month.

3.3 Columns for Data Analysis:

Out of the 24 columns of the dataset few of the columns were used to perform forecasting and the others were used to draw business insights and construct visualization plots. The 'Order Date' and 'Sales' column is used for forecasting. 'State' was used to plot mean sales per state on the map. 'Customer ID' was used to find the customer with most orders. 'Product Name' was used to plot count of products sold per state on the map.

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	...	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales
7981	CA-2014-103800	2014-01-03	2014-01-07	Standard Class	DP-13000	Darren Powers	Consumer	United States	Houston	...	77095	Central	OFF-PA-10000174	Office Supplies	Paper	Message Book, Wirebound, Four 5 1/2" X 4" Form...	16.448

Fig.: 1 Dataset Sample

4 Methodology for forecasting

4.1 SARIMA Model:

SARIMA model stands for seasonal autoregressive integrated moving average. It is similar to a basic ARIMA model with slight difference that SARIMA model also incorporates the seasonality component. In order to implement SARIMA for the data set used in the project. I have used 'SARIMAX' function of Python's 'statsmodels' package. The parameters of SARIMAX chosen for office data were (1,0,1) for ARIMA part and (1,0,1,12) for the seasonal part. Similarly, parameters of SARIMAX chosen for furniture data were (2,0,2) for ARIMA part and (1,0,[1],12) for the seasonal part. These parameter values were chosen as they gave the least AIC, BIC and error values. SARIMAX model was trained on two years of data and remaining one year of data was used for testing the model. 'auto_arima' function in Python was used to find these parameters. Moreover, SARIMA model was used since there was apparent seasonality in the data. Moreover, it was confirmed that seasonality is present in the dataset by using 'seasonal_decompose' function of statsmodel to plot the seasonality of the plot. Other models like ARIMA and ARMA were tested against this model while selecting a statistical model for this dataset. However, SARIMA model performed the best out of these models.

4.2 Facebook's Prophet Model:

"Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well." [5]. In order to fit prophet model to the data, I have directly used Facebook's open-source prophet package in Python. Since, prophet model captures seasonality well and the data used in this project is strongly seasonal it was chosen for the dataset used in this project. All the default parameters were used to fit the model on the dataset as prophet model is intelligent enough to detect the trends in the data and choose the best parameters. Similar, to the SARIMA model used in this project prophet model was trained on 2 years of data and remaining 1 year of data was used for testing the dataset. Other models like LSTM and Deep Neural Network were tested against Facebook's Prophet model while selecting models for this project. However, prophet model performed the best among all these models in term of giving least error out of all these models.

5 Experimental Results and Analysis

5.1 Root Mean Squared Error

The following table shows the error percentage for Facebook's prophet model and SARIMA model for the two categories.

	Prophet Model	SARIMA Model
Furniture	24.5%	32.0%
Office Supply	51.02%	34.8%

Fig.: 2 Percentage Error

From Table it can be concluded that Prophet model should be used for forecasting of furniture whereas SARIMA model should be used for office supply.

5.2 AIC and BIC for SARIMA model

The AIC and BIC value obtained for SARIMA model are in the following table. The model fits when these values are least of different combinations of the model parameter.

	AIC	BIC
Furniture	338.64	346.88
Office Supply	510.44	518.44

Fig.: 3 AIC and BIC

5.3 Mean Sales Per State:

Mean Sales of the State is maximum in the state of Wyoming (WY). The maximum mean revenue generated was in the state of Wyoming which was around 1603.136(in dollars).

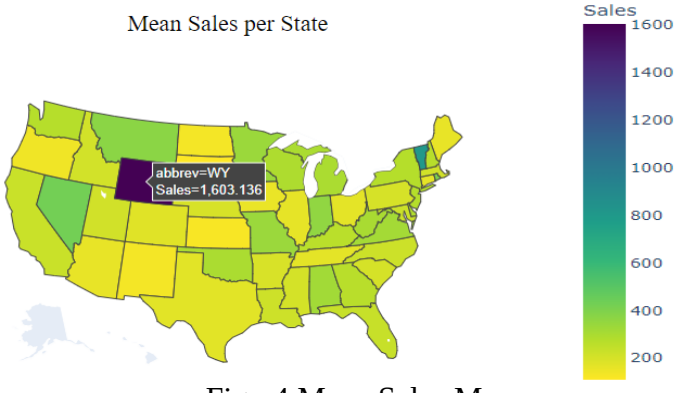

Fig.: 4 Mean Sales Map

5.4 Number of products sold per state:

The maximum amount of product sold is in the state of California. Total number of products sold in California were 2001.

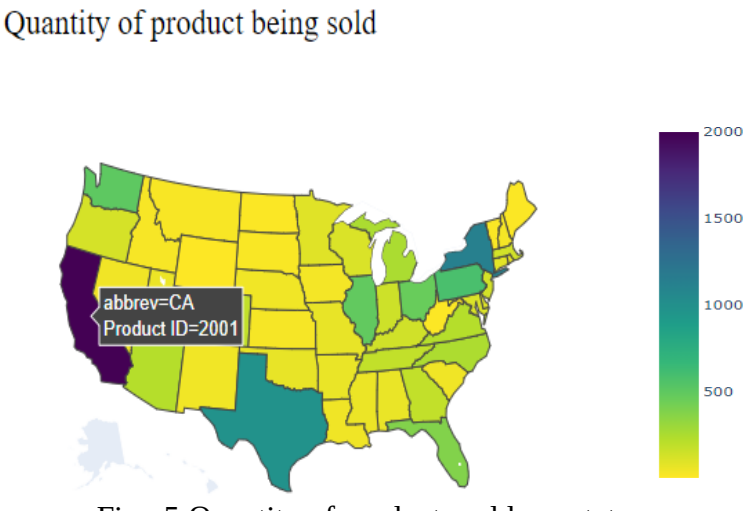

Fig.: 5 Quantity of products sold per state

5.5 Sales per month

November and February are the months with most and least sales respectively.

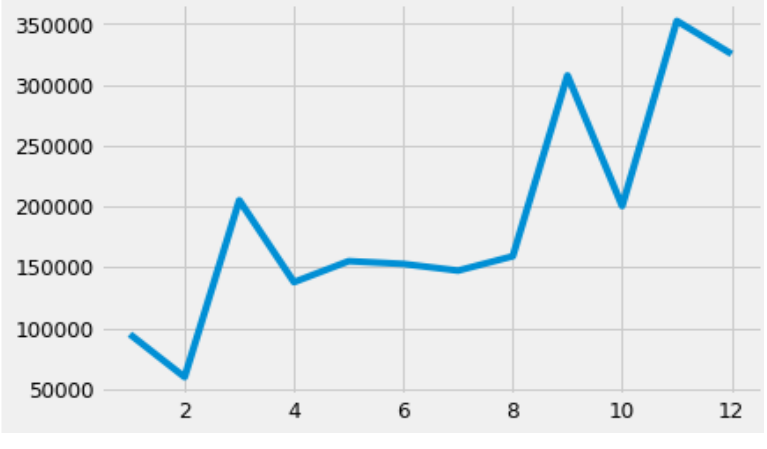

Fig.: 6 Sales Per Month

5.6 Forecasting trend:

Forecasting trend for both the models is increasing until 2024 indicating that demand for both furniture and office supplies will go up. Moreover, the slope for the office supply is much more than furniture indicating office supply will be more in demand in future.

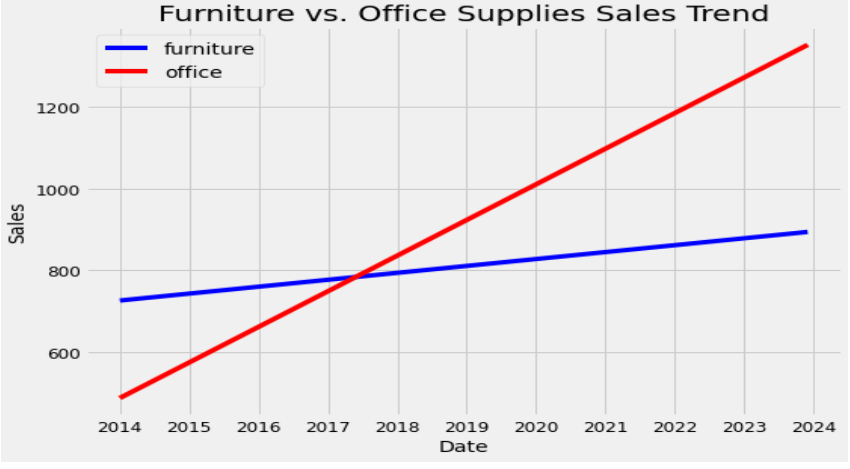

Fig.: 7 Forecasting Trend

5.7 Future forecasts

According to future forecasts obtained by the two models for the office supply and furniture, the estimate sales for office supply is around $1760 and for furniture supply is around $1320

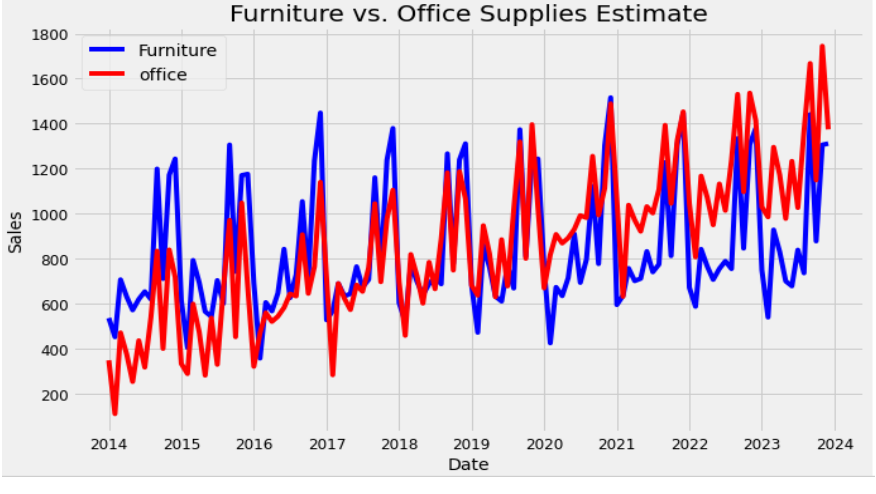

Fig.: 8 Future Forecasts

5 Conclusions

This report is able to show the potential Machine Learning and statistical models that can be used by product-based companies while performing forecasting. By investigating the problem and performing a literature review on the systems as well as Industry 4.0 technologies, it can be concluded that the use of right kind of Machine Learning models for forecasting can provide a competitive advantage to product-based companies. It not only helps in improving future sales but also helps in making informed decision about the company's inventory. There's still a lot of potential solutions that can come up with newer and newer models coming up every year. So, this study can be further upgraded to reap better results.

Supply Chain companies should start looking at Machine Learning Integration in their operations thus, improving efficiency and grow their company in the market at an exponential rate.

6 References

[1]. Schmidt, A.; Kabir, M.W.U.; Hoque, M.T. Machine Learning Based Restaurant Sales Forecasting. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 105-130. https://doi.org/10.3390/make4010006

[2]. S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, 2018, pp. 53-58, doi: 10.1109/iCCECOME.2018.8659115.

[3]. Anđelković, Aleksandra, and Marija Radosavljević. "Improving order-picking process through implementation of warehouse management system." Strategic Management-International Journal of Strategic Management and Decision Support Systems in Strategic Management 23.1 (2018).

[4]. Reza Toorajipour, Vahid Sohrabpour, Ali Nazarpour, Pejvak Oghazi, Maria Fischl, Artificial intelligence in supply chain management: A systematic literature review, Journal of Business Research,
Volume 122, 2021, Pages 502-517, ISSN 0148-2963, https://doi.org/10.1016/j.jbusres.2020.09.009.

[5]. Facebook's prophet link-https://facebook.github.io/prophet/

[6]. www.kaggle.com/datasets/jr2ngb/superstore-data