A Predictive Approach to Supply Chain Analysis

Kavya PK[#1], Shraddha Bharadwaj[#2], Rohan Biju[#3]

Semester 5, CSE Department, PES University
Bengaluru, Karnataka, India

[1]kavyapk@pesu.pes.edu

[2]shraddha5bharadwaj@gmail.com

[3]rohan.biju@gmail.com

Abstract*— Over the last decade, with the onset of the internet, the outlook of retail and businesses has changed drastically, and multichannel (i.e, offline and online) retail has become extremely popular. It becomes essential to bring about coordination between technologies. However, the massive upsurge in both quantity as well as diversity of data have resulted in datasets that are no longer manageable manually or even by conventional management tools. The motivation for this study is to come up with unique implementations to maximise profits of an online retail store, by analysing the Supply Chain Management (SCM), forecasting demand using the massive datasets from previous sales in order to facilitate effective replenishment, ensuring sustainable development and subsequently predicting the time it takes for orders to reach customers. This research has been built on multiple other papers that have been done in the area, putting together theoretical concepts of SCM and predictive analysis.

Keywords*— Data Analysis, Supply Chain Management, Demand Forecasting, Trend Analysis, Regression, Random Forests, Univariate, Multivariate

I. Introduction

Today, E-commerce and Online Retail have become extremely popular, and more so during recent times and through the pandemic. Products are now widely accessible in the remotest of locations, and this has led to a massive surge in sales. Rising competition among various e-retail providers has also upped the importance of customer satisfaction, which depends prominently on appealing prices and timely delivery, which in turn depend on availability of stock. Furthemore, with the rising need of the world moving towards sustainable development, minimising loss and overstocking in warehouses becomes one of the primary points that can be acted upon.

The rise in online transactions and also the use of artificial intelligence and IOT(methods such as RFID) in retail locations has resulted in the availability of several new and potentially invaluable datasets, and new methods of data science, primarily predictive analytics, have been developed. The fundamental process of a supply chain however, remains unchanged. Most research today is focused on bringing together a combination of integrated supply chain management and data science concepts, and this is often referred to as *supply chain predictive analysis* and sometimes as *DPB (Data Science, Predictions and Big Data).*

"A definition of SCM predictive analysis : SCM predictive analysis is the application of quantitative and qualitative methods from a variety of disciplines in combination with SCM theory to solve relevant SCM problems and predict outcomes, taking into account data quality and availability issues."[3]

This paper aims to use concepts from existing research and optimize the final stages of a supply chain by predicting upcoming sales of products in a multi-purpose online store, and in turn use it to predict an optimized time of delivery (TOD).

II. Literature Survey

The advent of technology has made it clear that data analytics and successful businesses go hand in hand. Over the years, research has been done trying

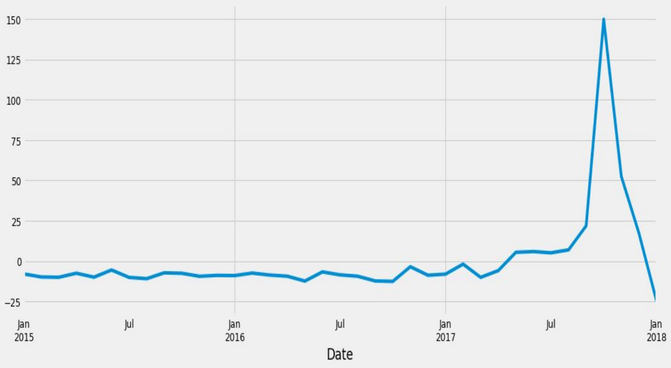


Fig 1 Overall Trend in Sales

TABLE I
FINDING THE OPTIMAL LAG VALUES

Dept. Name	Lag Order	AIC	BIC	HQIC
Fitness	9	23.95	24.01	23.97
Apparel	9	19.67	19.75	19.69
Footwear	9	20.34	20.65	20.44
Technology	9	17.37	17.47	17.40
Outdoors	9	20.72	20.87	20.77

to understand implications of research and analysis in business logistics, revealing that with the tremendous amount of data being generated every day, domain knowledge and analysis cannot be separated. More specifically, as the number of variables increases - i.e, as the use of a theoretical data mining proliferates, the chances of false positives increases exponentially, which results in wasted resources.

The use of both qualitative and quantitative predictive analysis calls for data captured at multiple points in the supply chain as well as consumer sentiment, which can be used in forecasting inventory, transport and manpower, designing solutions that use applied probability along with effective data mining, and in optimization of implementable solutions [3]. One study gives interesting insights about direct and indirect relationships between the supply chain flow, customer services and finance, and talks about the importance of an integrated supply chain strategy where customer service is the yield of the entire system, achieved by synchronizing the requirements of the final customer and reaching a balance between high customer service and cost of the final customer and reaching a balance between high customer service and costs[5].mDecision Variables and Performance measures deeply affect the accuracy of the problem at hand [6].

Efficient Supply Chain Management (SCM) has been an area of interest for several years, with a significant number of works in academic literature, each focusing on approaches ranging from regression to deep learning methods and neural networks. Data Preprocessing, Dimensionality reduction, Effective Sampling and Feature engineering are highlighted to be important steps in developing an effective model. Data gathered in the field is often found to be skewed, and techniques such as upsampling has been used effectively in some studies to improve accuracy of models, along with methods like target transformation [7]. In [1], the products were classified into four categories in order to be able to analyze the correlation of the sale of the products by channels, and then pre-processed to remove the outliers identified through statistical methods.

The problem of Demand Forecasting in order to predict Stock replacement has been approached in several ways by different people. In [1], the author describes combination of K-means clustering and ANNs (NAR and NARX) , and the model has been validated using the by the R2 coefficient of determination to observe the adjustment of the forecast equation and the Mean Squared Error (MSE) to analyze the accuracy measure, which

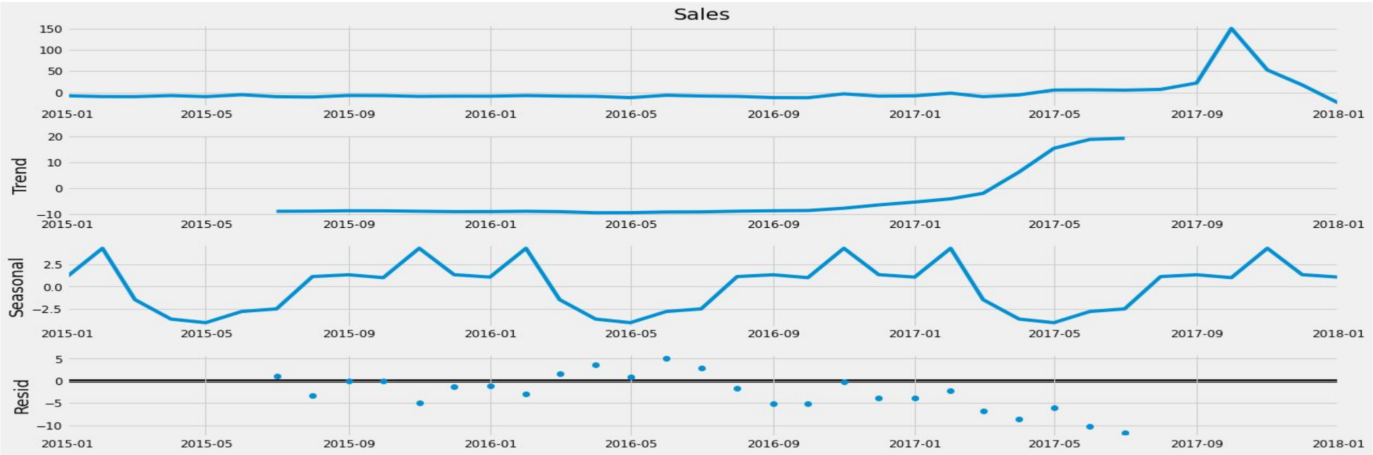


Fig 2 Decomposition of Sales Time Series

revealed that NARX presented a good performance for all the products.

Another paper [7] compares the use of Logistic Regression, Random Forests and AdaBoost to first classify products into categories, and then compare their F scores, to find that random forests work best in product classification.

III. Problem Statement

The data set that has been used in this research consists of a DataSet of Supply Chains used by the company DataCo Global describes Provisioning , Production , Sales , Commercial Distribution.

Supply chain data has a vast scope for optimisations in various fields such as transportation, inventory management and human resources.

This project concentrates on Inventory management from the perspective of the distributor, and aims to produce visualisations and statistical models in such that the statistics behind it are abstracted and the real world meaning is effectively conveyed.

Aim: To perform statistical analysis and visualisation on product popularity, profit, sales and supply, and draw insights to relate various attributes in the dataset.

To perform Category wise Demand Forecasting of products- using both Univariate and Multivariate Time Series Analysis, and compare these results with that of a Random Forest Regressor.

IV. Proposed Solution

Overview: The primary objective is to meaningfully use present data to forecast the demand for each of the departments so that the provider is always one step ahead of the customer. Data Preprocessing and EDA involved data cleaning, selection of appropriate attributes, and normalisation. The solution that has been developed consists of 3 different models - a random forest regressor, a univariate forecasting model that predicts Sales from its trend alone, and finally a multivariate time series model that takes into account the effects of attributes that primarily affect sales. These models have been compared to reach a final conclusion on which approach can be applied for this dataset.

A. Data Cleaning and Exploratory Data Analysis

Through EDA (primarily Pearson's Correlation), it was found that the dataset has as many as 23 categorical variables whose effect is not visible. So, it was concluded that Feature Engineering, more specifically Feature Selection, would be extremely valuable in order to extract maximum valid information for the time series model. Forecasting the future demand for products of each department based on purchase history, delivery history and other attributes that have a pla-usible connection to Sales would be optimal to perform any trend Projection and visualisation. Since the dataset has 53 attributes, experiments on the regression model were done only after Attribute selection.

B. Random Forest Approach

A random forest is an ensemble based approach, that averages the results of several decision trees, each built on a different subset of the dataset, making it a much more powerful approach than a standalone decision tree. It gives a meta estimator that fits a number of decision tree classifiers on various subsamples of the dataset. RandomForestRegressor from sklearn’s ensemble library has been used in this solution, considering attributes that most likely have an effect on Item Order Quantity and Sales. Keeping in mind the fact that decision trees are prone to overfitting, the optimal max-depth was found after several rounds of testing. The model has been evaluated using standard parameters - RMSE and MAE.

C. Univariate Time Series Analysis using SARIMA

Time series analysis is used to extract meaningful insights from statistical analysis of a time series data.

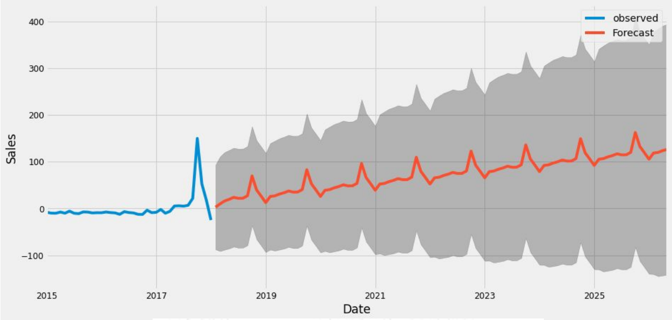
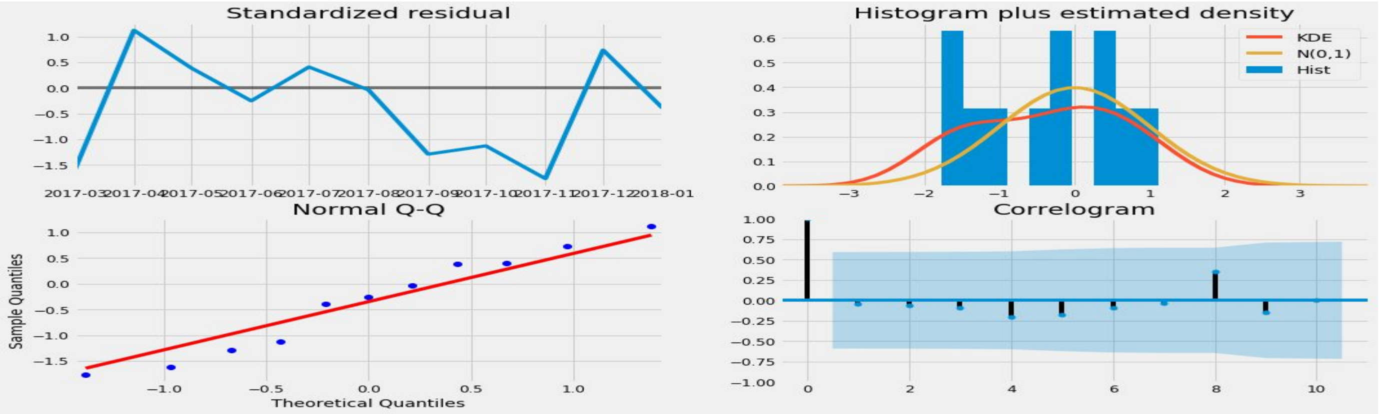


Fig 4 Forecast of Sales from 2019



The key idea is to predict future results from present data. This approach focuses on a univariate model that is based on the assumption that the dependent variable (Sales) is affected only by time, and therefore predicts results by interpreting patterns of change of single (scalar) observations recorded sequentially over equal time increments (Days).

The seasonality, trend and residual components of the data has been visualised using time-series decomposition. It was observed that there is a seasonality component cyclic behaviour between May-May of each year accounting for summer and holiday sales. So, the solution implemented is one of the most commonly used methods for time-series forecasting, known as SARIMA(Seasonal Autoregressive Integrated Moving Average. The ideal p,d,q values (Trend autoregression order, Trend difference order, Trend moving average order) for the SARIMA model was reached through a grid search , so as to reach the performance for the model. This was done by finding the model that yields the lowest Akaike Information Criteria (AIC) value. The fitted model summary and the diagnostics were also plotted.

D. Multivariate Time Series Analysis

Univariate analysis provides a holistic view of the

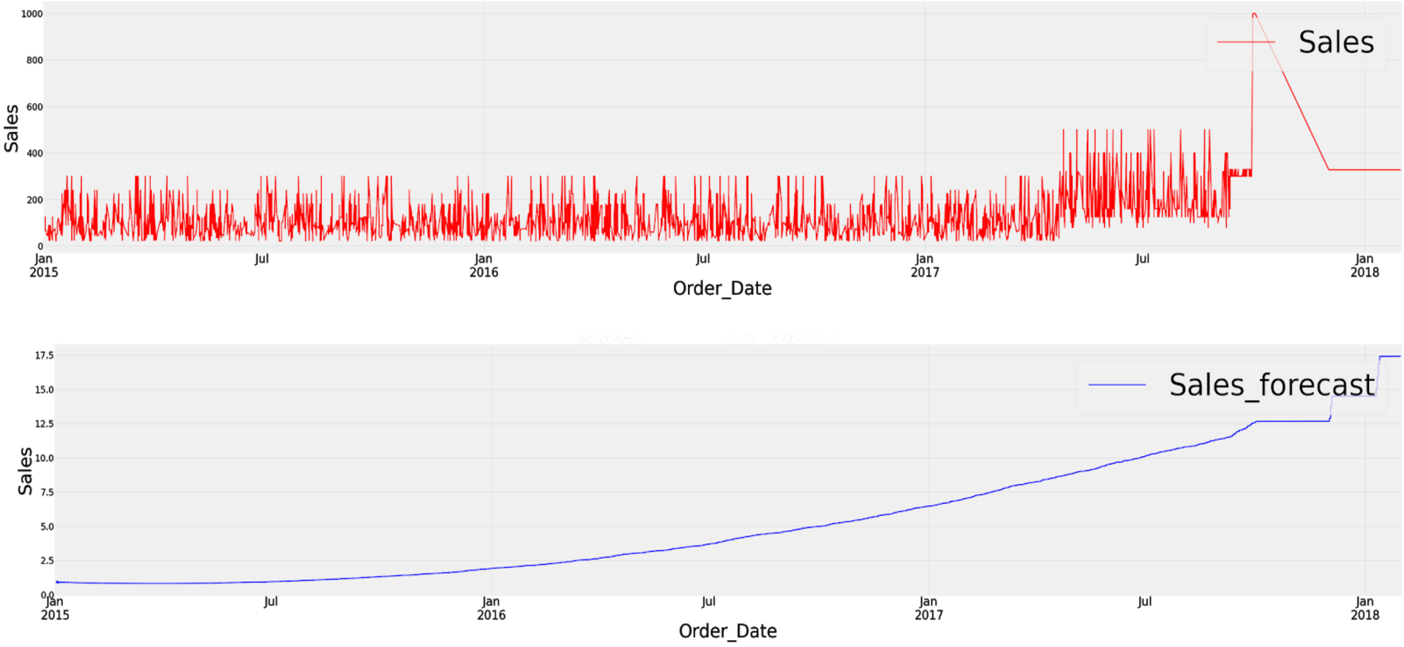


Fig 5 Forecast vs Actual Sales for Department Fitness

TABLE II
DURBIN WATSON TEST RESULTS

Dept. Name	Sales per customer	Benefit per order	Days for shipping
Fitness	2.04	2.06	2.05
Apparel	2.04	2.06	2.06
Golf	2.04	2.05	2.05
Technology	2.06	2.06	2.06
Outdoors	2.06	2.06	2.04

sales for the company and an idea of the expected revenue. But it does not account for the various departments, delivery and the customer traction they bring in. Multivariate Time series analysis focuses on forecasting the sales for each department by taking into account all the correlated variables that affect Sales in the given time series - Product Category Id, Customer Id, Department,Days for shipment (scheduled), Order Item Profit Ratio, Product Id, Item Discount Rate, Item Product Price.

The Forecasting algorithm used is Vector Auto-Regression (VAR) since the dataset has two or more time series that influence each other. That is, the relationship between the time series involved is bi-directional. Attributes that have a Pearson's correlation above 0.5 for with Sales have been chosen in order to forecast sales and demand (Product Quantity) for each department using the multivariate model. The time series for this dataset resembles a typical AR(p) model, whose equation looks something like this:

A Cointegration test was performed, giving a statistically significant relationship for the selected attributes and its order, i.e, if there exists a linear combination of them that has an order of integration (d) less than that of the individual series. The training and testing data was then initialised to all the records of the department so as to not miss any trends or seasonality that would be missed by splitting the data, as the number of records per department is very less compared to the entire dataset.

The Augmented Dickey-Fuller Test (ADF Test) was implemented to check if the dataset is stationary. If not, the dataset was different twice to make it stationary and to make sure that series in the VAR models have the same number of observations.

To select the right order of the VAR model, increasing orders of the VAR model and Lag order were tested, and the order with the least AIC was picked to find the best fit comparison. Once the VAR model was trained with selected order(p), The Serial Correlation of Residuals (Errors) was listed using Durbin Watson Statistic, calculated as:

$$DW = \frac{\Sigma_{t=2}^{T}((e_t - e_{t-1})^2)}{\Sigma_{t=1}^{T} e_t^2}$$

Forecast plots were generated after bringing back the data to its original scale, by de-differencing it as many times as differenced earlier - this is done using the inverse transformation function.

Finally, the model is evaluated by plotting Forecast and actual values against each other and by computing a comprehensive set of metrics, namely, the MAPE, MAE, MPE, RMSE, corr and minmax.

V. EXPERIMENTAL RESULTS

A. Random Forest Regressor

The optimal max-depth was found after several rounds of testing, and finally set to 10. The random state was set to 40. The model has been evaluated using standard parameters - RMSE and MAE, for two of the demand forecasting attributes- Sales and Item Order Count. The results obtained were very satisfactory - Sales recorded an MAE of 0.20207 and an RMSE value of 1.845, Item Order Count recorded an MAE value of 0.00758, and an RMSE of 0.00375.

TABLE III
EVALUATION PARAMETERS FOR THE MULTIVARIATE MODEL ACROSS DEPARTMENTS

Dept. Name	RMSE	MAE	Corr	MinMax
Fitness	66.99	0.82	0.372	1.04
Apparel	398.64	374.98	0.006	0.98
Footwear	817.64	665.00	0.028	1.00
Technology	239.77	192.29	0.035	0.94
Outdoors	408.00	287.4	0.005	-7.32

B. Univariate Time Series Analysis

Overall, the forecasts align with the true values very well, capturing the seasonality toward the end of the year 2017. SARIMAX(1, 1, 1)x(1, 1, 1, 12) yields the lowest AIC value of 151.78 and is therefore taken.
The fitted model summary and the diagnostics reveal that the residuals, even if not perfect, are normally distributed.

The Univariate model clearly captures sales seasonality. Forecasting further out into the future gives less confidence in the values, which is reflected by the confidence intervals generated by the model, which grow larger.

The line residual plot is random around the value of 0 and does not show any trend or cyclic structure. It is observed that the Residual Histogram and Density Plots distribution do have a Gaussian look, showing an exponential distribution with some asymmetry. The Q-Q plot scatterplot shows that the actual and the forecast are of the similar distribution. In addition to this, the correlogram shows that serial correlation in data does not change over time.

C. Multivariate Time Series Analysis

VAR models have been built for each department with the same set of time series data attributes with training and testing data to be the entire dataset. The AIC, BIC and HIQC of the predictors chosen are shown in the Table I. The Durbin Watson test for the forecaster of each department is in Table II and their statistical error estimates are given in the table Table III. All of these results helped in choosing the optimal parameters for the final model.

When exploring the data it was observed that Apparel and outdoor industries have close to 1/3rd records each .Examining the Sales vs Order Date graph describes a spike in sales during the periods of October to end of January which may account for the holiday season. RMSE for indicates that the model was able to forecast the average daily sales in the test set within that value of the real sales . It was found that the departments Golf, 'Fan Shop ,Book Shop Discs Shop, Pet Shop, Health and Beauty were unable to be modelled due to lack of enough time-series data to extract trend or seasonal components, or did not pass the cointegration test i.e, all the data in the time series that was to be forecasted were not autocorrelated.

VI. Conclusions

This study has drawn several interesting insights from the dataset at hand. After comparing the results of the two completely different approaches - Random Forest Regression, and the Univariate and Multivariate Time Series models, it was observed that the Random Forest Regression gives optimal results. However, it was inferred that this approach has a strong chance of overfitting, and the results obtained for varying trends of data in future years may not be successfully predicted.

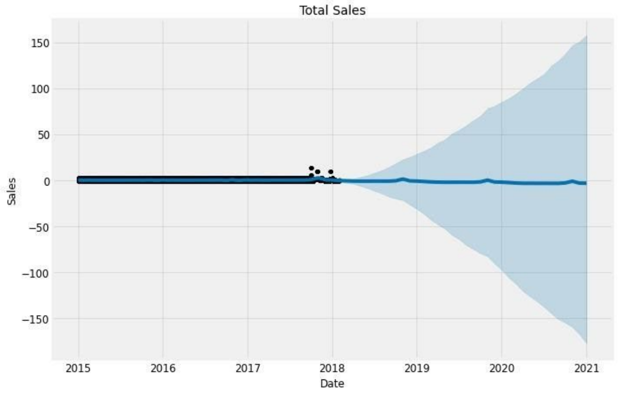


Fig 6 Univariate Sales forecast using Prophet

Although the dataset has a great number of features, it was concluded that for this specific use case, i.e, Demand Forecasting, results from the Univariate Time Series approach supersedes that of the Multivariate model, as these additional factors (other than time) have uneven effects across different departments of the store, resulting in dramatically different results. The univariate model gives results that take into account the seasonal variation in sales, and predicts results upto the next few years with decent accuracy.

Future Scope of this project involves improving the multivariate model by further tuning hyper parameters, and performing Cointegration test department wise to infer which attributes significantly affect each department, and perform customised time series analysis on each department. Further experiments using Deep Neural Networks, and other regression models can be performed in place of Random Forests, and the effects of each can be compared.

The scope of improvement for this project is immense, with a wide variety of possible approaches, each worthy of a deep dive research.

VI. Contributions of members

Each team member was involved in all phases of ideation and decision making, as well as report making. Specifically, the models were coded as follows:

Cleaning and PreProcessing- Kavya, Shraddha
Univariate models - Shraddha, Kavya
Multivariate models - Shraddha
Random Forest - Rohan, Kavya
Testing and Debugging - All

VI. References

[1] M. M. Pereira and E. M. Frazzon, “Towards a Predictive Approach for Omni-channel Retailing Supply Chains,” *IFAC-PapersOnLine*, vol. 52, no. 13, pp. 844–850, Jan. 2019.

[2] E. Hofmann and E. Rutschmann, “Big data analytics and demand forecasting in supply chains: a conceptual analysis,” *The International Journal of Logistics Management*, vol. 29, no. 2, pp. 739–766, Jan. 2018.

[3] M. A. Waller and S. E. Fawcett, “Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management,” *Journal of Business Logistics*, vol. 34, no. 2, pp. 77–84, Jun. 2013.

[4] A. Ilic, T. Andersen, and F. Michahelles, “Increasing Supply-Chain Visibility with Rule-Based RFID Data Analysis,” *IEEE Internet Comput.*, vol. 13, no. 1, pp. 31–38, Jan. 2009.

[5] S. K. Vickery, J. Jayaram, C. Droge, and R. Calantone, “The effects of an integrative supply chain strategy on customer service and financial performance: an analysis of direct versus indirect relationships,” *J. Oper. Manage.*, vol. 21, no. 5, pp. 523–539, Dec. 2003.

[6] B. M. Beamon, “Supply chain design and analysis:: Models and methods,” *Int. J. Prod. Econ.*, vol. 55, no. 3, pp. 281–294, Aug. 1998.

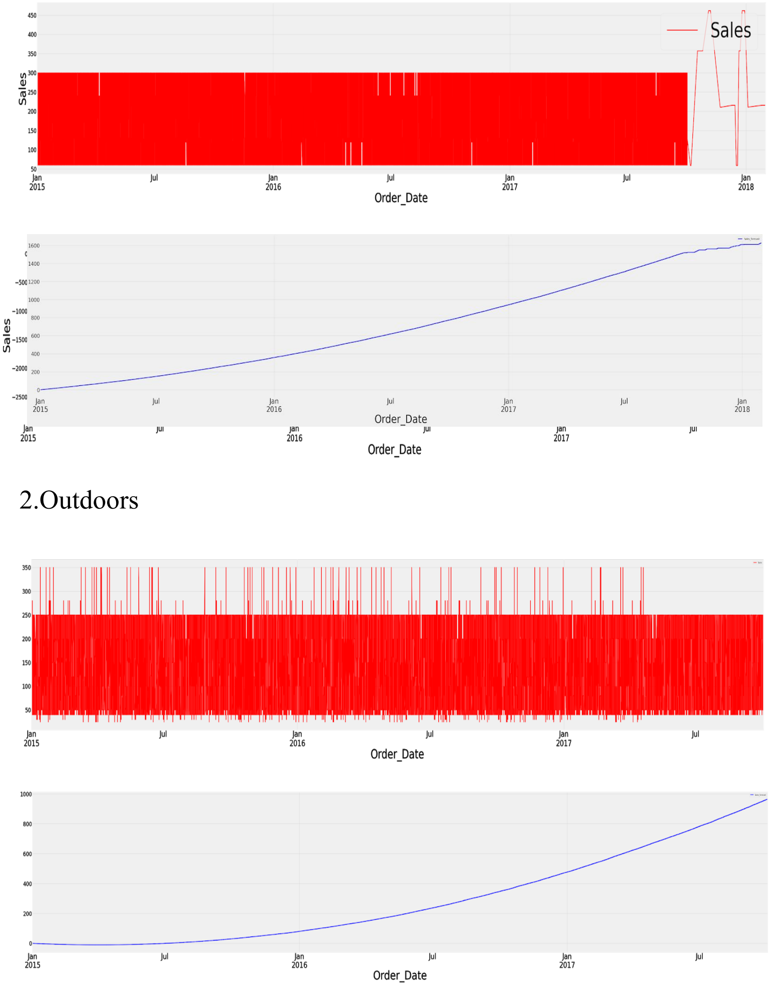
[7] P. Saboo, S. U. Arakeri, S. D. Mogali, S. S. Patra, Z. Ahmed, and M. A. Lanham, “Leveraging Insights from ‘Buy-Online Pickup-in-Store’ data to Improve On-Shelf Availability,” [Online]. Available: http://matthewalanham.com/Students/2020/ICDATA20_Leveraging%20Insights%20from%20Buy-Online%20Pickup-in-Store%20data%20to%20Improve%20On-Shelf%20Availability.pdf.

[8] https://www.kaggle.com/sukanthen/e-commerce-multi-output-models-project-cse07

VII. Appendix

Below are some department-wise sales vs Forecast Graphs that have not been included in the main graph.

1. Footwear



2.Outdoors

Fig 8 Forecast vs Actual Sales