IE 7275 – Data Mining in Engineering

Sales Forecast of Walmart

Shardul Jani – 001221826

Nagarjuna Sricharan Maram – 001212206

Table of Contents

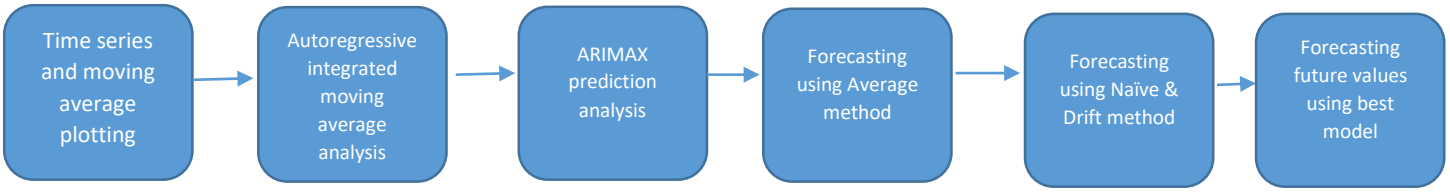
Abstract

Retail Industry is one of the biggest business sectors in the United States. Giant firms like Walmart and Costco are prominent companies in this field. The retail industry in the US contributes 2.2 trillion dollars an year to its overall GDP. But this trend is decreasing today due to several factors. Introduction of ecommerce has influenced the concept of buying and selling at a large scale. Companies like Amazon and other start-ups have entered into online retail business and changed the conventional Brick and mortar system.

This project aims to capture that essence by using the historical data from Walmart and then predicting future values for the sales. This techniques in a real-time world can help companies corrective measures with respect to inventory and personnel management. Here we have explored simplest techniques like moving average and drift method to complex procedures like ARIMA to build and test our models. We have then determined which model is the best fit for forecasting values.

Methodology

The basic out line of the methodology we have adopted is as follows:



- The dataset for 45 stores and 99 departments in each one is shortlisted into one store with 5 departments on the basis of maximum revenue contributed. The entire process is based on the assumption that, making marketing decisions based on performance of the bestselling departments.
- To understand the basic idea of the plot, we have plotted time series and moving average graphs at a factor of 3, different departments have shown different trends and we have selected different modelling techniques for them
- The first step involves performing adf, acf and pacf tests to the data sets for each of the departments to interpret the stationary characteristics of the data chosen. Based on the parameters obtained from the above tests ARIMA models are build and RMSE errors are computed.
- To understand the effect of holiday variable in the forecast we built ARIMAX models to find the significance of additional variable in the forecast process
- Average forecast method has been implemented to verify its performance as the time series plot implied data to be stationary at a small range and further RMSE has been computed by test data
- Naïve Method and Drift Method has been computed to check the effect of most recent values in the time series data and RMSE is calculated and prediction plot is built
- The individual data of each department has been divided into training and test data and all the above models are built on training set and the model is verified on test data to check its predicting capabilities.
- The RMSE computed for all the models with different departments has been considered as a decision parameter to select the best model for each of the department.
- The selected model have been run on the entire data of the departments to forecast the future values for the department.

Data Description and Preprocessing

With an aim to study the effect on sales on conventional brick and mortar business in comparison to online giants like amazon this project was undertaken. The dataset was obtained from kaggle.com and contains the weekly sales data of 45 stores located in the United States starting from February 2010 and ending in October 2012. Furthermore, each store has 99 departments and the weekly sales value for each department is given. The dataset also provides important weekly holidays(is holiday) and gives the value of sales for that particular week.

The table below shows the brief description of data

Store	Dept	Date	Weekly_Sales	IsHoliday
16	1	2/5/2010	12786.85	FALSE
16	1	2/12/2010	18305.6	TRUE
16	1	2/19/2010	13714.98	FALSE
16	1	2/26/2010	11182.72	FALSE
16	1	3/5/2010	10996.97	FALSE

In order to reduce the data complexity we have calculated the maximum sales for each of the 45 stores and we chose store number 20 which has a maximum sale over the other 44 store. This was done to get the best possible model for predicting the maximum sale as the store has more customers which could contribute to marketing decisions and additional profits. We then calculated the amount of sales for each department and chose the top 5 departments for our consideration which had the highest amount of sales. With an aim to get an effective visualization charts we converted weekly sales for this top 5 stores into monthly sales using Pivot tables. We also assumed that if any one week had a special holiday then the entire month should be true for Isholiday.

The table below gives the idea of conversion of data into months

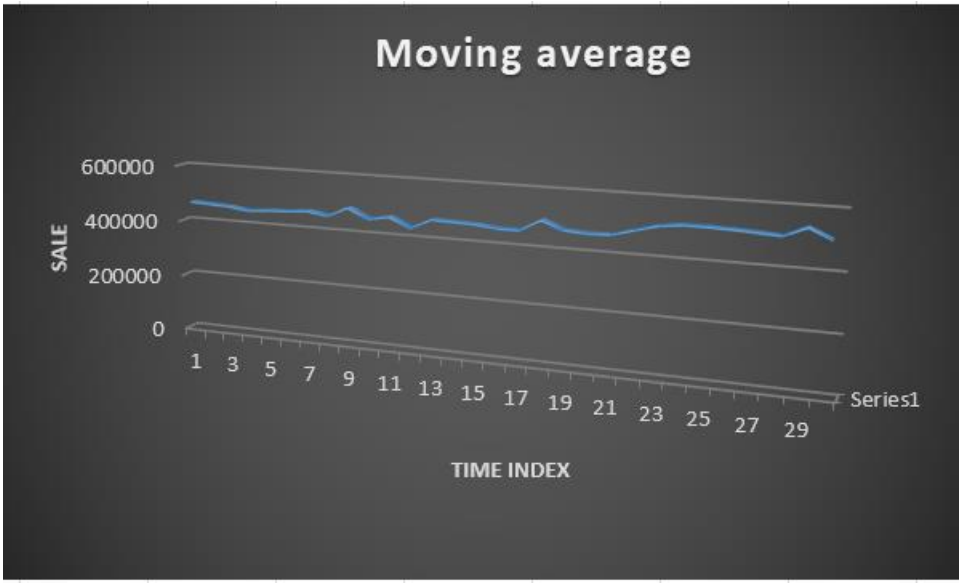
Month	Sale	Holiday
Feb-10	433203.4	TRUE
Mar-10	435312.1	FALSE
Apr-10	536099	FALSE
May-10	421895	FALSE
Jun-10	425340.1	FALSE
Jul-10	506941.3	FALSE
Aug-10	441467.9	FALSE
Sep-10	433409	TRUE
Oct-10	527226	FALSE

By reducing the date set to one which with the highest sales will enable to build time series models on the best performing departments which contribute to maximum revenue and can give tentative analysis about the forecast for other departments in the same store and other stores as well.

Data Visualization

Time series plot for department 38

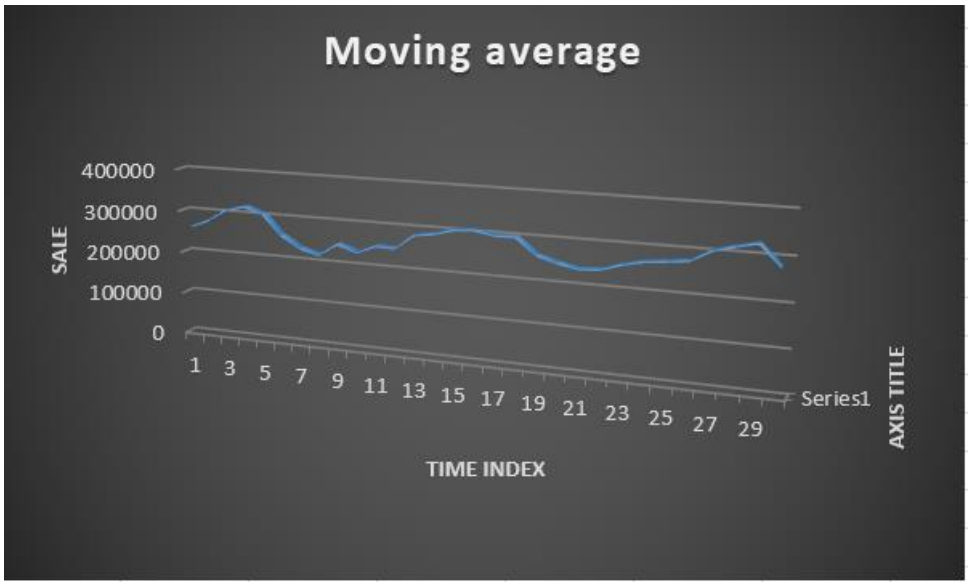
Moving average plot for department 38



The time series plot for department 38 implies that the revenue is mostly in the range of 400K to 600K and moving average plot infers that the data is mostly stationary.

Time series plot for department 94

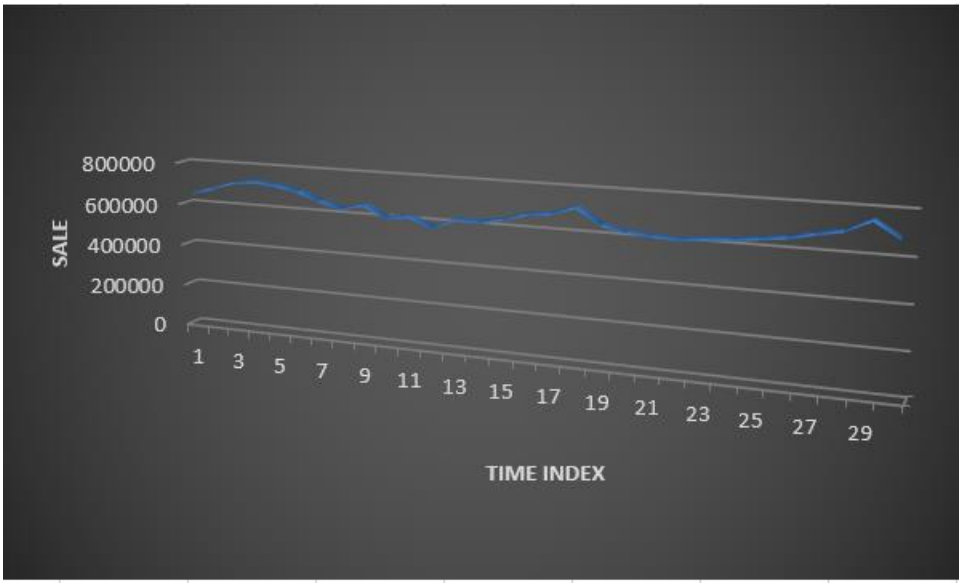
Moving average plot for department 94



The time series plot of department 94 implies that the revenue is not stationary and is constantly fluctuating between 100K and 400K. Most of the sale is happening during summer season. The moving average also shows a cyclic trend.

Time series plot for department 95

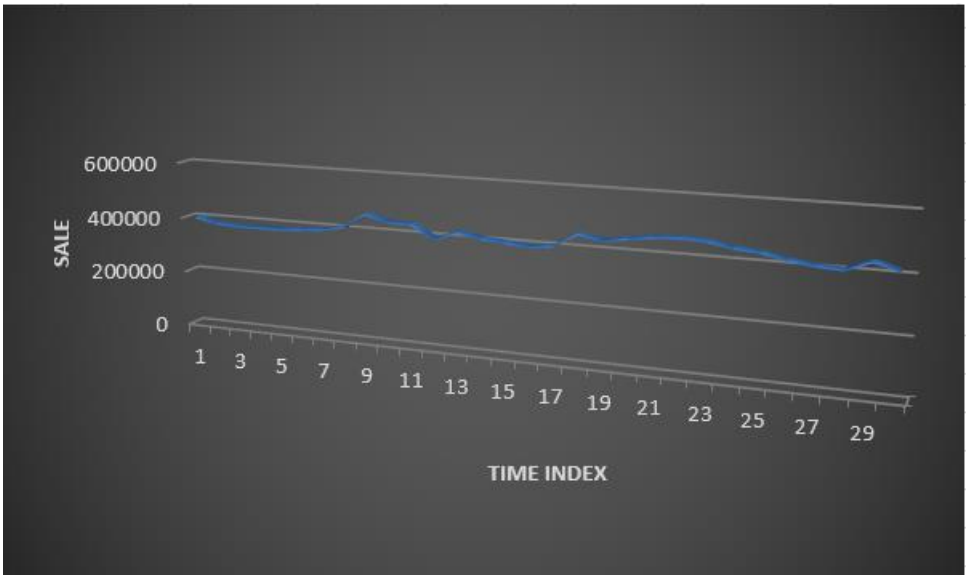
Moving average plot for department 95



The moving average and time series plots suggest that the entire dataset is cyclic and the moving average analysis shows that the entire data is between the range of 600K – 800k

Time series plot for department 90

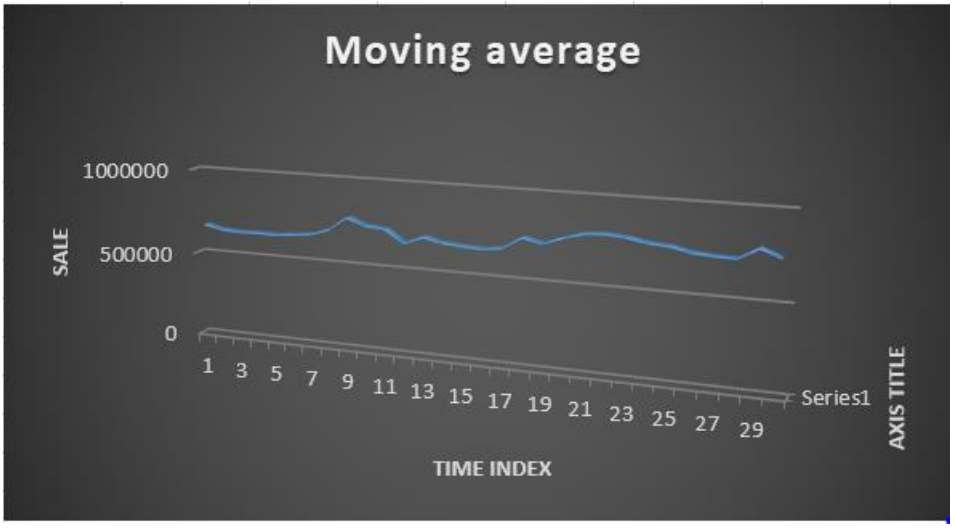
Moving average plot for department 90



The time series and moving average plot implies that the average revenue generated is 400K and mostly cyclic.

Time series plot for department 92

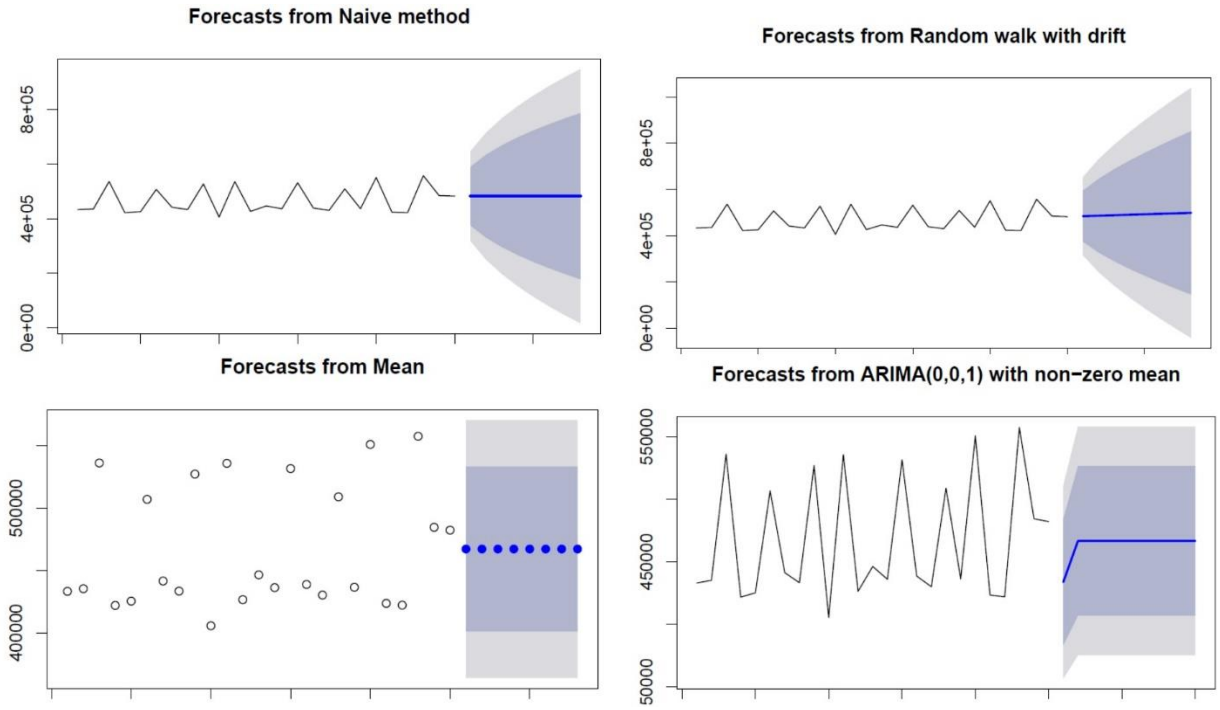
Moving average plot for department 92



The plots above describe the data to be stationary and mostly ranging between 600K to 800K

Model Formulation and Implementation

Department 38



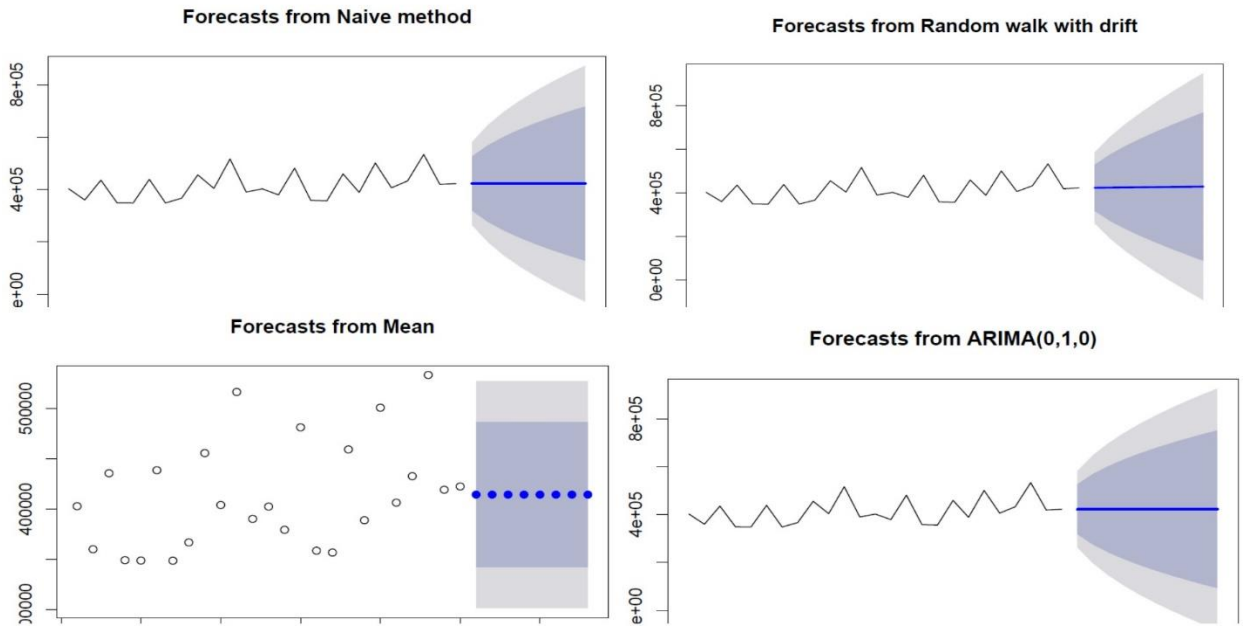
Analysis

Based on the adf test, we have received a P-value (< 0.05) suggesting the time series to be a stationary, On further analysis with acf and pacf tests, we have decided the order to be (0,0,1). The model built with ARIMA has fit well and gave good predictions with a confidence interval in the plot. The ARIMAX model with an additional variable in the analysis has not shown much significanct effect in the model because the model was capturing the effect already due to variation in the data at different intervals.

The models built using Naïve , average and drift methods have proved more efficient by providing lesser RMSE error. The plots and prediction ability was analysed to select one model with the one providing least RMSE error.

Based on the plots and RMSE error, we have decided to forecast sales for next 4 months using drift method.

Department 90



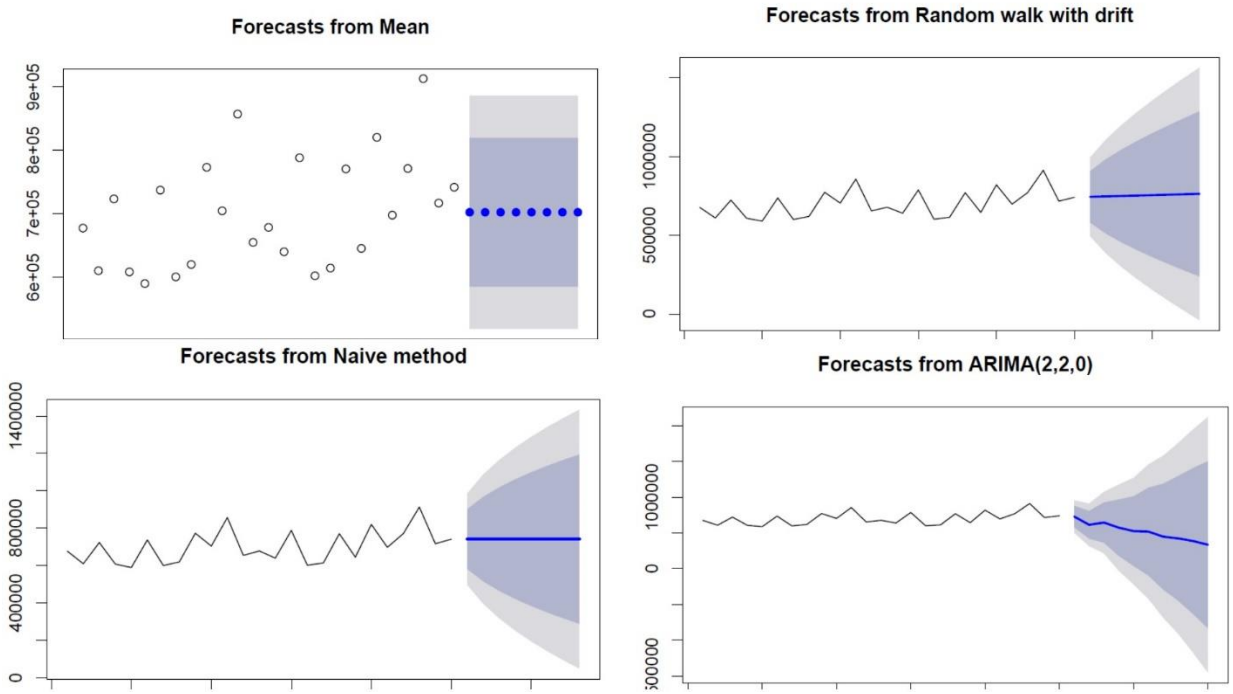
Analysis

The adf test has produced a P-value (> 0.05) suggesting the time series to be not stationary.To make it stationary, we had differneciated 2 times to stabilize the data.On performing the adf test again we have received a P-value (< 0.06) On further analysis with acf and pacf tests, we have decided the order to be (0,1,0). The model built with ARIMA has fit well and gave good predictions with a confidence interval in the plot. The ARIMAX model with an additional variable in the analysis shown the highest RMSE error implying that the variable is not adding any significance , infact skewing the series in the wrong direction

The models built using Naïve , average and drift methods have proved more efficient by providing lesser RMSE error. The Naïve and Average method performed equally well and the drift method has provided more RMSE error compared to Naïve and Avergae methods. The plots and prediction ability was analysed to select one model with the one providing least RMSE error.

Based on the plots and RMSE error, we have decided to forecast sales for next 4 months using Naive method.

Department 92



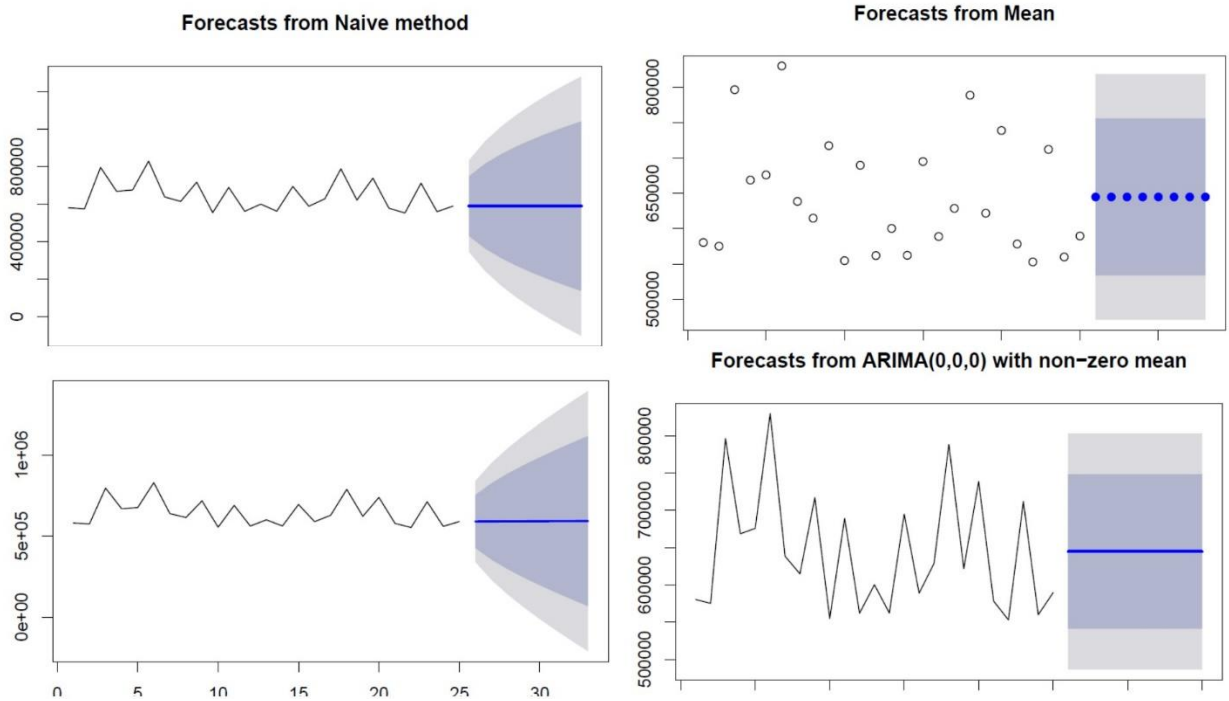
Analysis

Based on the adf test, we have received a P-value (> 0.05) suggesting the time series to be not stationary.So, we computed to difference two times to convert the data to stationary.On performing the adf test again, the P-value was less than 0.05 indicating that the data is stationary. On further analysis with acf and pacf tests, we have decided the order to be (2,2,0). The model built with ARIMA has not performed well and gave maximum RMSE. The ARIMAX model with an additional variable in the analysis has shown much significance effect in the model as it provided lower RMSE error.

The models built using Naïve , average and drift methods have proved more efficient by providing lesser RMSE error. The plots and prediction ability was analysed to select one model with the one providing least RMSE error.

Based on the plots and RMSE error, we have decided to forecast sales for next 4 months using Naive method.

Department 95



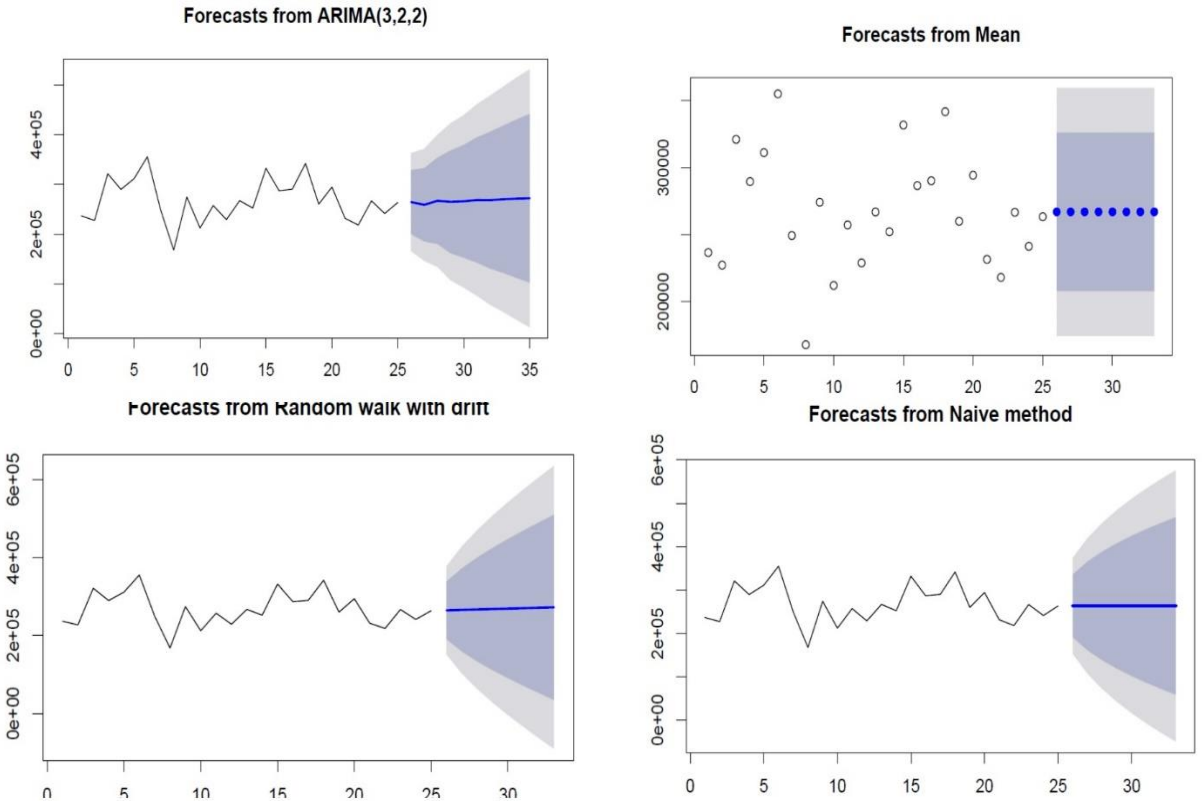
Analysis

Based on the adf test, we have received a P-value (> 0.05) suggesting the time series to be not stationary, On performing double differenciation, the time series became statiionary and the p value also became < 0.05. with acf and pacf tests, we have decided the order to be (0,2,0). The model built with ARIMA has fit well and gave good predictions with a confidence interval in the plot. The ARIMAX model with an additional variable in the analysis has shown much significanct effect in the model because the model was capturing the effect already due to variation in the data at different intervals. The RMSE error also has shown almost equall to the ARIMA model

The models built using Naïve , average and drift methods have proved not performed efficiently by providing highest RMSE error. The plots and prediction ability was analysed to select one model with the one providing least RMSE error.

Based on the plots and RMSE error, we have decided to forecast sales for next 4 months using ARIMA method.

Department 94



Analysis

Based on the adf test, we have received a P-value (> 0.05) suggesting the time series to be not stationary, On differentiating twice the time series became stationary.Further analysis with acf and pacf tests, we have decided the order to be (3,2,1). The model built with ARIMA has fit well and gave good predictions with a confidence interval in the plot. The ARIMAX model with an additional variable in the analysis has not shown much significanct effect in the model because the model was capturing the effect already due to variation in the data at different intervals.

The models built using Naïve , average and drift methods have proved more efficient by providing lesser RMSE error. The plots and prediction ability was analysed to select one model with the one providing least RMSE error.

Based on the plots and RMSE error, we have decided to forecast sales for next 4 months using ARIMA method.

The models built using the training data of each of the 5 departments are tested on the test data to compute the RMSE error. Below is the table of RMSE results for each department with different methods

Department	Arima	Arimax	Average	Naïve	Drift
92	213797.6	213797.6	89582	76996	78276
94	43213	47020.89	46058	47714	46221.86
95	84502	98095.62	87929	119826	118685.1
38	77480	75775	67693	59231	57165
90	435138	586181	42160	42112.4	43189.28

Results and Conclusion

On the basis of above results from several models, the RMSE errors are calculated and best models are selected to forecast future predictions. Below is the forecast made for the future of each department using the best model respectively.

Month	Dept 94	Dept 92	Dept 38	Dept 95	Dept 90
Nov-12	256072.8	710487.3	460421.1	652660.3	397748.5
Dec-12	249396.9	710487.3	461245.9	652660.3	397595.3
Jan-13	251359.6	710487.3	462070.7	652660.3	397442.3
Feb-13	252040	710487.3	462895.4	652660.3	397288.9

- The same process can be used to all the departments of each store to analyze and implement suitable model to forecast the future sales.
- In order to reduce complexity, the top selling departments can be given most priority in each store and provide flexible marketing budgets to them and have a fixed budget for underperforming departments.
- As customer priorities change regularly, it's ideal to consider short term forecast for sales. Hence we have considered only 4 months to predict the future sales.

References

- https://www.otexts.org/fpp/
- Data mining for Business analytics – Galit Shmueli , Peter C. Bruce , Nitin R.Patel