Forecasting sales of Walmart departments for marketing budget allocation

Course Name: FCAS_A
Group: A3

Team A3	
Gandharv Paliwal	61710328
Asmita Sonik	61710866
Aayam Ankan	61710631
Abhinav Mani	61710499
Swagnik Chatterjee	61710076
Samta Jhingan	61710109

1.1 EXECUTIVE SUMMARY

This project aims at providing the best possible sales forecasts for our client, 'Walmart'. Walmart is currently facing stiff competition from E-commerce companies like Amazon. It is commonly observed that customers visit the brick and mortar store to view the product and end up buying at Amazon at cheaper rates. Individuals prefer doing their repeat purchases online instead of visiting brick and mortar stores due to the associated convenience. This has led to year on year declining sales trend of brick and mortar channels.

This report proposes a solution to Walmart to increase the monthly sales of their brick and mortar channels. The solution suggests **forecasting next 2 months** of Walmart departments and using that information to appropriately allocate marketing budget each month (for sales promotions and advertisements) to different departments. Here, an underlying assumption that has been made is that the sales of Walmart departments will be positively impacted by allocating more budget. Hence, allocating more budget to the departments could be expected to generate more sales.

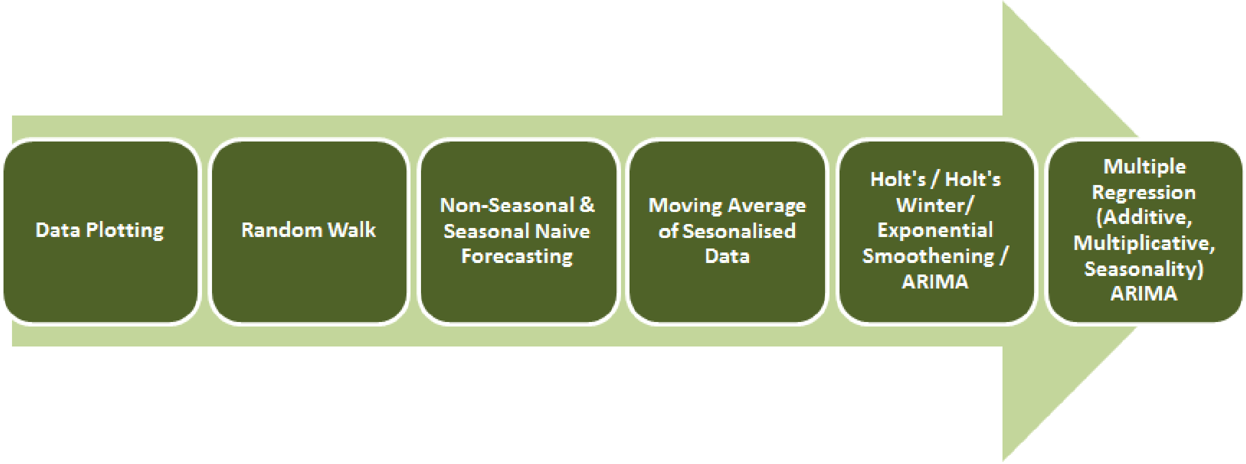
Additionally, this forecast can also help in planning budgets in a way that expenses could be minimised in non performing departments in the coming months which will help in affecting the bottom line in a positive way.

1.1.1 DATA DESCRIPTION

The weekly sales data of 45 Walmart stores located at different geographic regions was obtained online (Kaggle.com). The data contained weekly sales from Feb'2014 to Oct'2016 of 99 different departments. Here, the different departments included were Books, Clothing, Furniture, Electronics etc.

The weekly sales varied drastically from among different departments and can be seen in Figure 1 (appendix). As the dataset of different departments contains different/varying components, it became essential to obtain the best forecasting for each department individually.

1.1.2 FORECASTING METHODOLOGY


Data Plotting
Random Walk
Non-Seasonal & Seasonal Naive Forecasting
Moving Average of Sesonalised Data
Holt's / Holt's Winter/ Exponential Smoothening / ARIMA
Multiple Regression (Additive, Multiplicative, Seasonality) ARIMA

Stepwise modelling procedures

The above mentioned procedure was run to shortlist different departments and fit and analyze the different models to obtain the best fitting model. To obtain the best fitting model following things were analyzed:

1. Model with minimum MAPE in Training and Validation datasets
2. Model with no overfitting
3. For Multiple Linear Regression p-values,variable selection(Cp=#predictors+1) and adjusted R sq. were also checked. However, model variables were selected based on judgement taken from these parameters in order to obtain best forecasting model.
4. Plotting of actual and forecasts of different models was also done to select the best model.

The best fitting models obtained for different departments were:

Department #	Models	Training MAPE	Validation MAPE	Remarks
56	Multiple Linear Regression with multiplicative seasonality & AR(1)	4.96%	7.68%	Input variables: 11 Monthly dummy variables, no. of holiday weeks in the month Output variables: ln(Monthly Sales)
3	Moving Average of De-seasonalized data	5.446%	12.489%	Seasonality index to de seasonalize data, which was then fed to the model. Forecasts were re seasonalized for result analysis

16	Multiple Linear Regression With Additive Seasonality & Arima	6.51%	8.30%	Input variables: 11 Monthly dummy variables, no. of holiday weeks in the month, Output variables: Monthly Sales
67	Holt's Winter with multiplicative seasonality ARIMA AR(3)	3.016	10.024	Default Alpha(level)=0.2, Beta (Trend) =0.15, Gamma(Seasonality)=0.05 (Best output values) Period = 12, ARIMA = 3

For department no. 56, multiple linear regression with multiplicative seasonality and AR(1) Arima model, gave the best fit with minimum overfitting. Also, Ad. R.sq. of the model was 0.99, variables were selected based on variable selection and p-values (p-value<0.05). Similarly department 16 gave best results with Multiple Linear Regression with additive seasonality.

For department no. 3 best forecasts were obtained using moving average of deseasonlized data. Even though multiple linear regression with multiplicative seasonality and Arima modelling was giving better results in terms of MAPE the model was not selected as there was greater overfitting with the difference between training and validation MAPE approximately 10%.

Department 67, gave best results with Holt's Winter Multiplicative seasonality model with AR(3). Below figure 2 (Appendix), depicts Lag(3) correlation in the errors of Holt's Winter multiplicative seasonality, which was removed using ARIMA(3) modelling.

2.1 RECOMMENDATIONS:

- The modelling can be extended to all the departments which are not having random walk and budget can be allocated based on predicted sales.
- In order to reduce complexity, modelling can be done only for the top selling departments and budget can be allocated based on predictions. For the other departments a fixed budget can be allocated.
- Possibility of developing an interface to generate forecasts for all the departments, along with optimal marketing budget allocations.
- Possibility of forecasting using econometric models especially the impact of competitive pricing from major competitors like Amazon.

- As consumer preferences change drastically over the years, optimal to do role forward forecasting with 2months prediction horizon. Prediction horizon 2 months as Arima is used.

3.1 TECHNICAL SUMMARY

3.1.1)Data Preparation: The data was available at Weekly level with the information pivoted down in rows instead of columns as shown in the snapshot of the executive summary. This was a challenge as there was no way to estimate if data was missing for any one of the departments in any of the months in any store. Also, it was extremely cumbersome to estimate the number of departments per store and the sales information available in this layout. As a result, we had to perform certain steps in order to make the data useful and informative and in the format that could be used for analysis. Steps involved in the data preparation:

A. Sum the data in different weeks to months by summing the sales across the weeks to bring the data to monthly level. This can be done in the excel or using TIBCO spotfire.

B. Pivot up the data at Store_ID - Department_ID level with the months in the columns instead of rows. The data is now usable to be analysed for evaluation. This can be done in the excel using pivot tables or using TIBCO spotfire.

C. Sum across the stores to arrive at National Sales for every department as we want to forecast the sales for every department. At this level, if a particular department is not present in a store or if the sales are not present for a particular store-department level, then those records are avoided.

D. IsHolidaySales column gave the information whether the entire week was off or not. This boolean(TRUE/FALSE) data was converted to integer format (0/1) and was summed up for a particular month to determine the number of holiday weeks in a given month. It was observed that holidays on all the 45 stores were same in a given week.

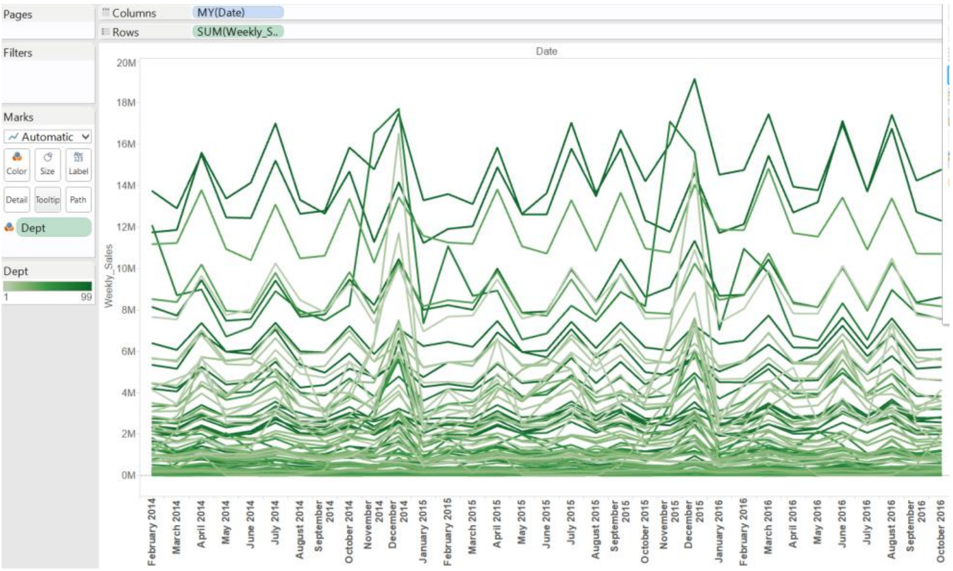
3.1.2)Department Selection: Post the data preparation, we followed the below mentioned steps for department selection:

1. We removed departments which had NULL or non-Numerical data in any one of its columns as these would skew our analysis.

2. We ranked the total sales of each department in descending order and chose a sample set of around 12-14 departments which could be considered. As these departments have the highest revenue, our client can focus bulk of their marketing budget towards these departments and realize increased sales.
3. Then we ran the random walk test by running the ARIMA model test on every shortlisted department to determine the value of the coefficient of AR1. If it is close to 1, then it is neglected and another department is analysed. We remove the data that shows a random walk as these datasets cannot be forecasted because the future prediction is independent of the historical data.
4. From the shortlisted departments, we have 33 months of data. We have split the data into training data with 24 months of data and validation data with 9 months. This has been done as we are forecasting for 2 months till the closing of the year.

3.1.3) Model Selection : Once the final 6 departments and their data were finalized, we first ran nonseasonal and seasonal naive forecasting to set the benchmark of MAPE that can be expected. As naive forecast involves the least computing to be done, hence it can act as a good benchmark to analyze other methods. Then we did smoothing techniques of moving average, exponential and double exponential techniques after removing seasonality from the data. Post that, Holts-Winter method was executed to understand the model fit for different values of trend, level and seasonality. Then we ran the multiple regression for additive and multiplicative seasonality to understand techniques which were work better than the Naive forecast. We then ran ACF to check autocorrelation in forecasted errors and introduced ARIMA model in the above models to ensure accurate forecasting. To ensure correct forecasting we calculated the 95% confidence interval of the forecasted data. We realised that if we over forecasted, then the client would spend more of their budget vis-a-vis others and that can lead to errors in optimal budget planning. We have then compared the forecast fit for each model by looking at the graph plot (Check Appendix) and the MAPE values to arrive at the best method. Below we have mentioned the MAPE values for both training and validation data sets for each of the chosen departments and plotted their residuals (in Appendix) to arrive at the best model for each department.

Appendix



Columns: MY(Date)

Rows: SUM(Weekly_S..

Marks: Automatic

Dept

Figure 1: Plot of monthly sales data across different departments

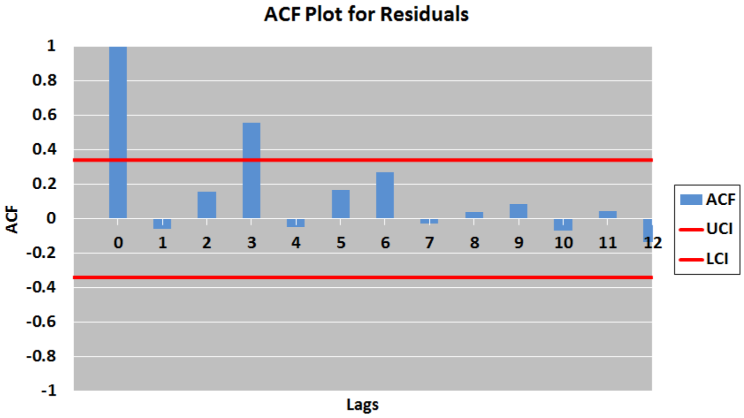


Figure 2: ACF Plot for Dept. 67 errors of Holt’s Winter Multiplicative Seasonality

Department monthly sales data plots:

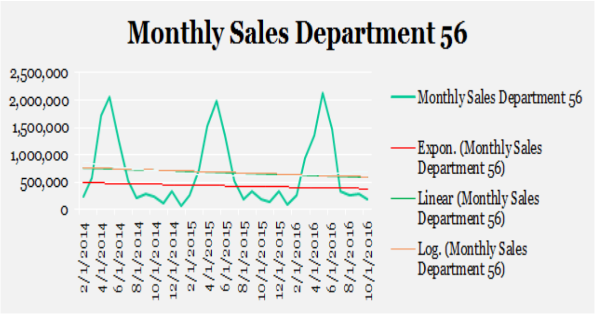


Figure: Dept. 56 plot with constant trend, annual seasonailty

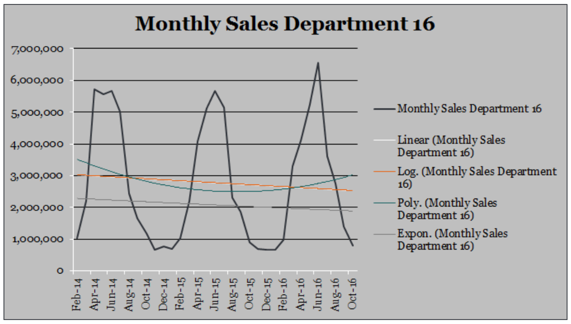


Figure: Dept. 16 plot with constant trend, annual seasonailty

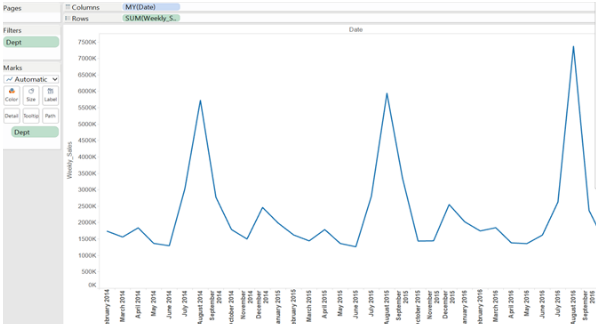


Figure: Dept. 3 plot with constant trend, annual multiplicative seasonailty

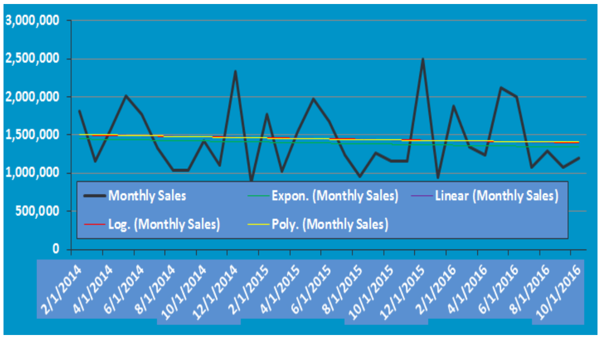
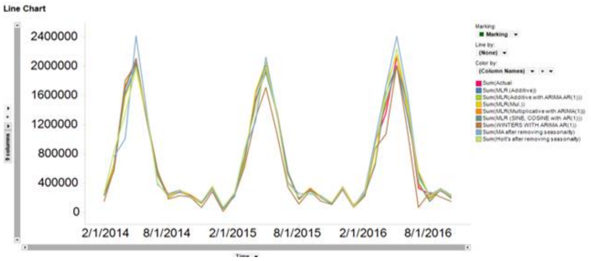
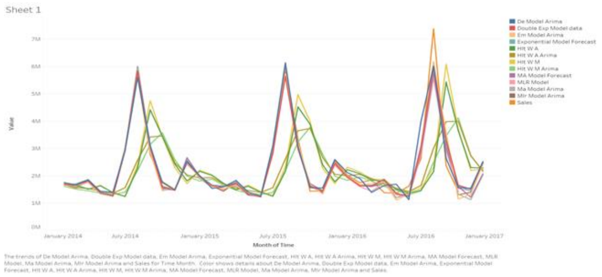


Figure: Dept. 67 plot with trend, to some extent annual seasonality (seen by overlapping yearly plot of sales data)

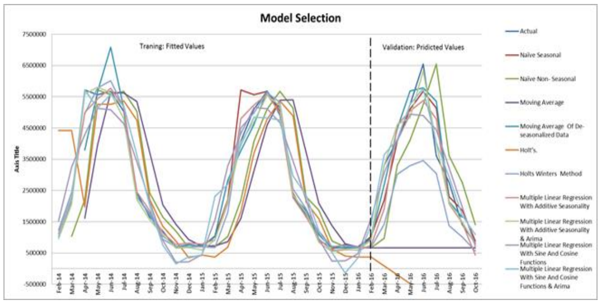


Actual vs Forecasts of all models Dept. 56

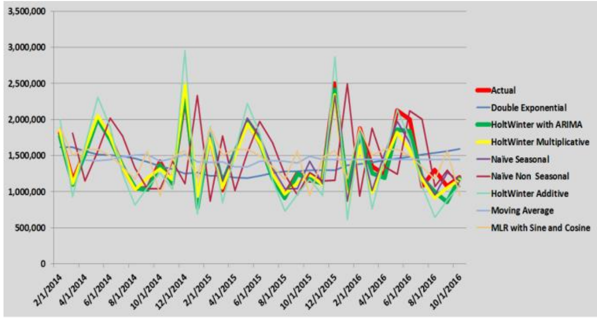


The trends of De Model Arima, Double Exp Model data, Em Model Arima, Exponential Model Forecast, Hlt W A, Hlt W A Arima, Hlt W M, Hlt W M Arima, MA Model Forecast, MLR Model, Ma Model Arima, Mlr Model Arima and Sales for Time Month. Color shows details about De Model Arima, Double Exp Model data, Em Model Arima, Exponential Model Forecast, Hlt W A, Hlt W A Arima, Hlt W M, Hlt W M Arima, MA Model Forecast, MLR Model, Ma Model Arima, Mlr Model Arima and Sales.

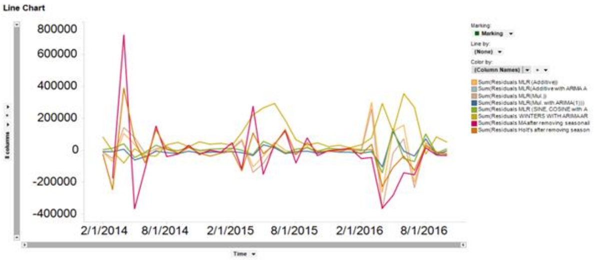
Actual vs Forecasts of all models Dept. 3



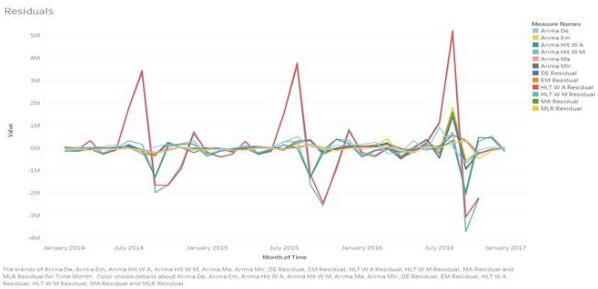
Actual vs Forecasts of all models Dept. 16



Actual vs Forecasts of all models Dept. 67

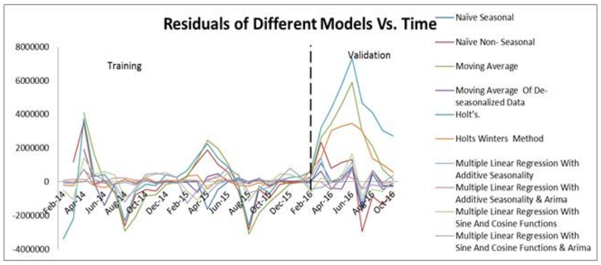


Residual plot of all models Dept. 56

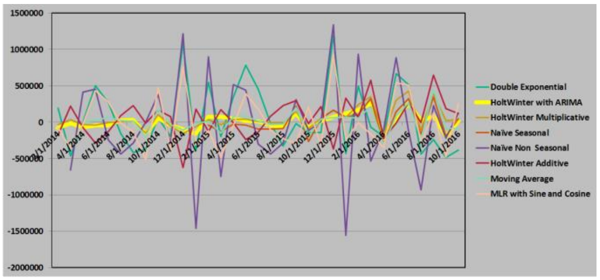


The trends of Arima De, Arima Em, Arima Hlt W A, Arima Hlt W M, Arima Ma, Arima Mlr, DE Residual, EM Residual, HLT W A Residual, HLT W M Residual, MA Residual and MLR Residual for Time Month. Color shows details about Arima De, Arima Em, Arima Hlt W A, Arima Hlt W M, Arima Ma, Arima Mlr, DE Residual, EM Residual, HLT W A Residual, HLT W M Residual, MA Residual and MLR Residual.

Residual plot of all models Dept. 3

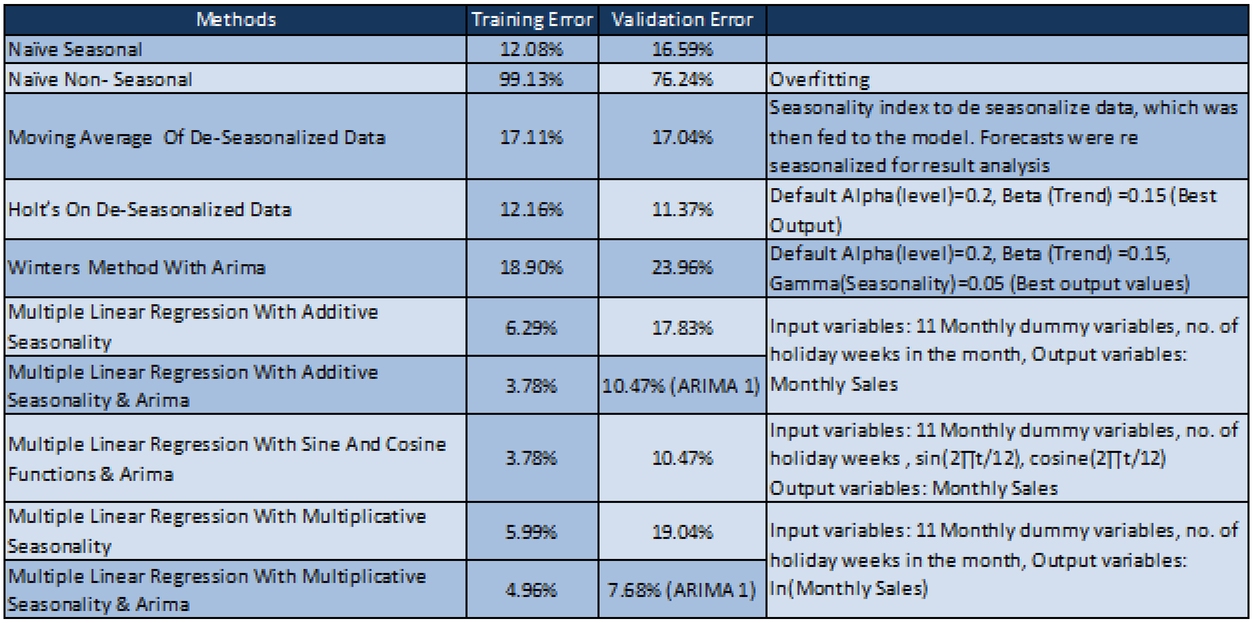


Residual plot of all models Dept. 16



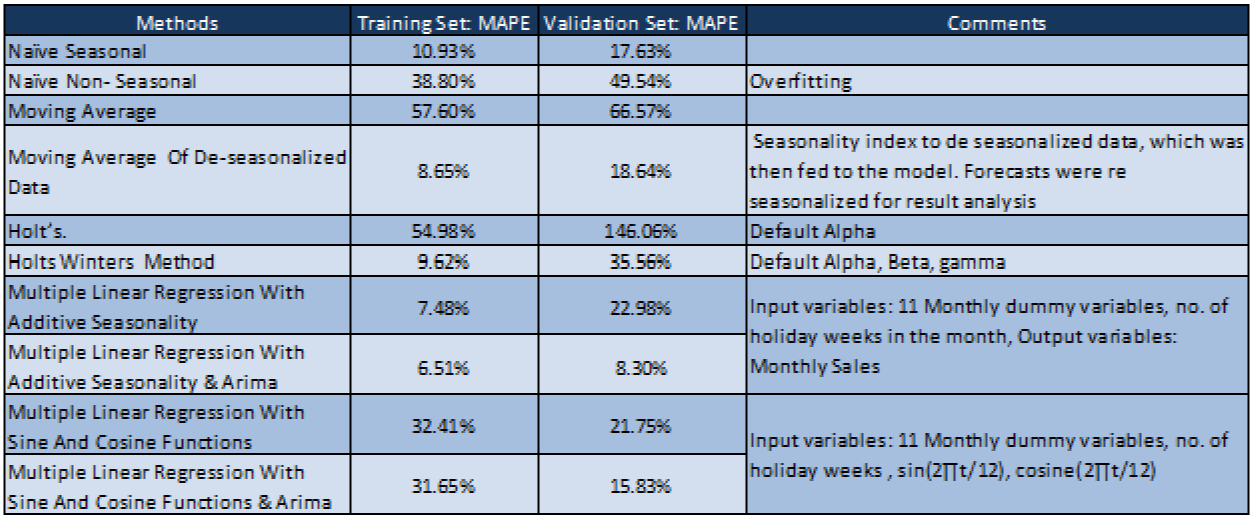
Residual plot of all models Dept. 67

Department #56 : Results of Different Methods



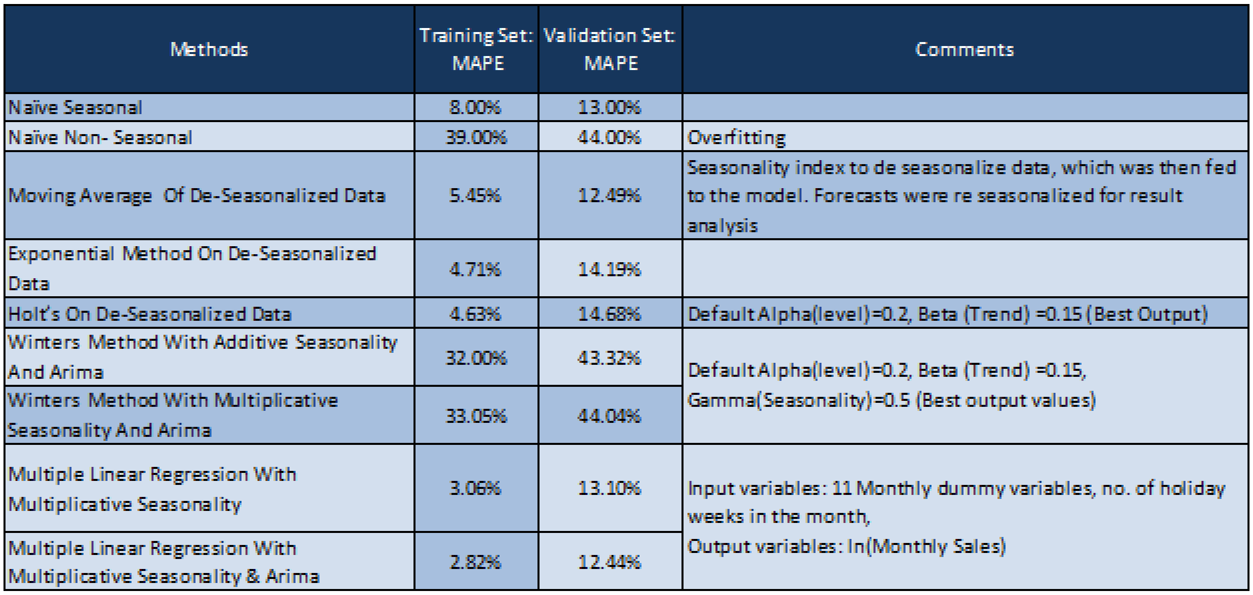
Methods	Training Error	Validation Error	
Naïve Seasonal	12.08%	16.59%	
Naïve Non- Seasonal	99.13%	76.24%	Overfitting
Moving Average Of De-Seasonalized Data	17.11%	17.04%	Seasonality index to de seasonalize data, which was then fed to the model. Forecasts were re seasonalized for result analysis
Holt's On De-Seasonalized Data	12.16%	11.37%	Default Alpha(level)=0.2, Beta (Trend) =0.15 (Best Output)
Winters Method With Arima	18.90%	23.96%	Default Alpha(level)=0.2, Beta (Trend) =0.15, Gamma(Seasonality)=0.05 (Best output values)
Multiple Linear Regression With Additive Seasonality	6.29%	17.83%	Input variables: 11 Monthly dummy variables, no. of holiday weeks in the month, Output variables: Monthly Sales
Multiple Linear Regression With Additive Seasonality & Arima	3.78%	10.47% (ARIMA 1)	
Multiple Linear Regression With Sine And Cosine Functions & Arima	3.78%	10.47%	Input variables: 11 Monthly dummy variables, no. of holiday weeks , sin(2∏t/12), cosine(2∏t/12) Output variables: Monthly Sales
Multiple Linear Regression With Multiplicative Seasonality	5.99%	19.04%	Input variables: 11 Monthly dummy variables, no. of holiday weeks in the month, Output variables: ln(Monthly Sales)
Multiple Linear Regression With Multiplicative Seasonality & Arima	4.96%	7.68% (ARIMA 1)	

Department #16: Results of Different Methods



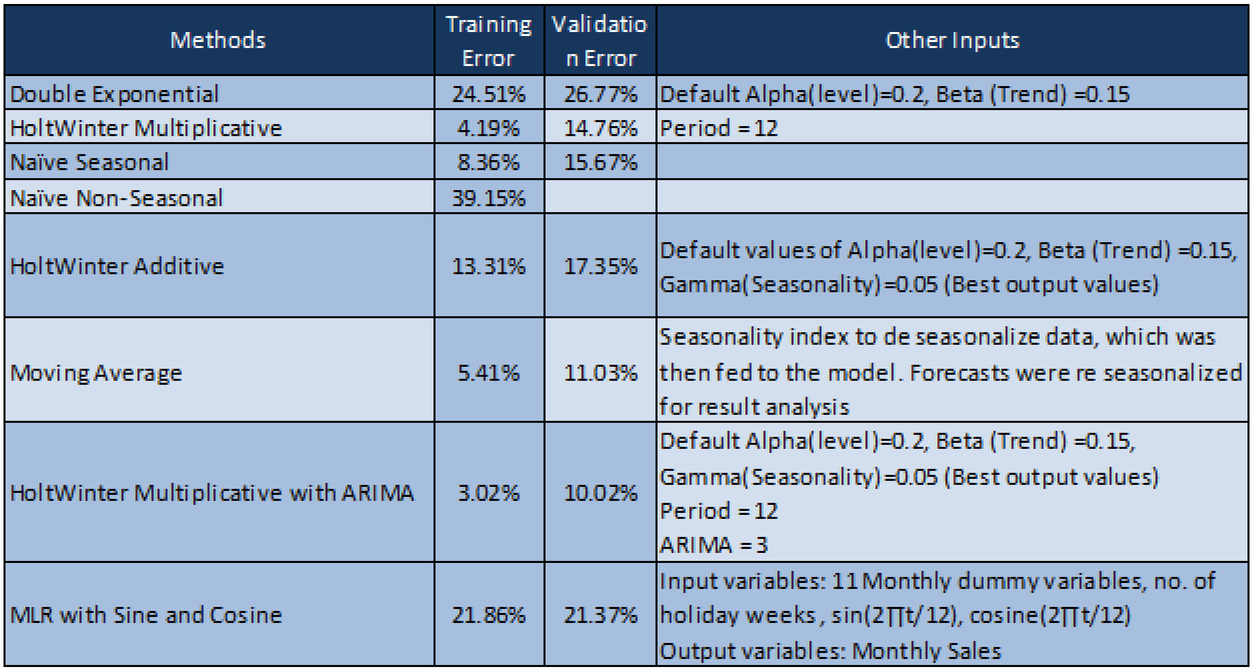
Methods	Training Set: MAPE	Validation Set: MAPE	Comments
Naïve Seasonal	10.93%	17.63%	
Naïve Non- Seasonal	38.80%	49.54%	Overfitting
Moving Average	57.60%	66.57%	
Moving Average Of De-seasonalized Data	8.65%	18.64%	Seasonality index to de seasonalized data, which was then fed to the model. Forecasts were re seasonalized for result analysis
Holt's.	54.98%	146.06%	Default Alpha
Holts Winters Method	9.62%	35.56%	Default Alpha, Beta, gamma
Multiple Linear Regression With Additive Seasonality	7.48%	22.98%	Input variables: 11 Monthly dummy variables, no. of holiday weeks in the month, Output variables: Monthly Sales
Multiple Linear Regression With Additive Seasonality & Arima	6.51%	8.30%	
Multiple Linear Regression With Sine And Cosine Functions	32.41%	21.75%	Input variables: 11 Monthly dummy variables, no. of holiday weeks , sin(2∏t/12), cosine(2∏t/12)
Multiple Linear Regression With Sine And Cosine Functions & Arima	31.65%	15.83%	

Department #3: Results of Different Methods



Methods	Training Set: MAPE	Validation Set: MAPE	Comments
Naïve Seasonal	8.00%	13.00%	
Naïve Non- Seasonal	39.00%	44.00%	Overfitting
Moving Average Of De-Seasonalized Data	5.45%	12.49%	Seasonality index to de seasonalize data, which was then fed to the model. Forecasts were re seasonalized for result analysis
Exponential Method On De-Seasonalized Data	4.71%	14.19%	
Holt's On De-Seasonalized Data	4.63%	14.68%	Default Alpha(level)=0.2, Beta (Trend) =0.15 (Best Output)
Winters Method With Additive Seasonality And Arima	32.00%	43.32%	Default Alpha(level)=0.2, Beta (Trend) =0.15, Gamma(Seasonality)=0.5 (Best output values)
Winters Method With Multiplicative Seasonality And Arima	33.05%	44.04%	
Multiple Linear Regression With Multiplicative Seasonality	3.06%	13.10%	Input variables: 11 Monthly dummy variables, no. of holiday weeks in the month, Output variables: ln(Monthly Sales)
Multiple Linear Regression With Multiplicative Seasonality & Arima	2.82%	12.44%	

Department #67: Results of Different Methods



Methods	Training Error	Validation Error	Other Inputs
Double Exponential	24.51%	26.77%	Default Alpha(level)=0.2, Beta (Trend) =0.15
HoltWinter Multiplicative	4.19%	14.76%	Period = 12
Naïve Seasonal	8.36%	15.67%	
Naïve Non-Seasonal	39.15%		
HoltWinter Additive	13.31%	17.35%	Default values of Alpha(level)=0.2, Beta (Trend) =0.15, Gamma(Seasonality)=0.05 (Best output values)
Moving Average	5.41%	11.03%	Seasonality index to de seasonalize data, which was then fed to the model. Forecasts were re seasonalized for result analysis

HoltWinter Multiplicative with ARIMA	3.02%	10.02%	Default Alpha(level)=0.2, Beta (Trend) =0.15,
			Gamma(Seasonality)=0.05 (Best output values)
			Period = 12
			ARIMA = 3

MLR with Sine and Cosine	21.86%	21.37%	Input variables: 11 Monthly dummy variables, no. of holiday weeks, sin(2∏t/12), cosine(2∏t/12) Output variables: Monthly Sales