Sales Forecast for Rossmann Stores

SUBMITTED BY:
GROUP A-8

Executive Summary:

a. Problem description:

Business Problem: Rossman is Germany's second largest drug store chain with more than 1000 stores across the country. Every month, the store manager needs to set targets for the sales team and design incentives for them. Currently, the managers set the targets based on their intuition of how much the sales are going to be in next month- which often leads to wrong target settings. Setting targets that are too high or unrealistic can lead to failure of the sales teams to meet the targets and therefore, loss of morale. On the other hand, setting targets that are very low will have costs in terms of lost opportunities/revenues.

Forecasting Problem: The goal is to enable the managers predict monthly sales. Our model will predict sales for the month of September'15 for 6 different stores of Rossman. August'15 is taken as a lag month (forecasting horizon=1).

b. Description of the data:

- The **"train"** data (training data) contains daily sales data of 1,115 stores from January 13 to July 15. The data also contains number of customers on a day, data on whether the store was open or closed on a day, whether it was a school holiday or state holiday, whether there was any promotion active on that day or not.
- The **"store"** data also contains details of the stores such as store type, assortment, competition (distance from competitor stores, number of months since the opening of the competitor store etc.), promotion, number of weeks since when a promotion denoted as "promo2" have been given.
- The **"test"** data contains the same columns as the "train" data except that it doesn't have the daily sales- which needs to be forecasted. This is basically for the month of Aug 15 (this is the lagged month for deployment purpose) and Sept 15.

c. Source of the data: Kaggle

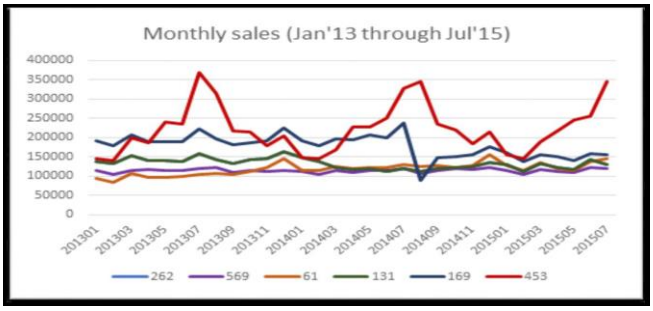
d. Key Characteristics:

- The timeseries data contains daily sales for 1,115 stores over 31 months (January 13 to July 15) which can be aggregated to monthly sales

- There are stores which are similar in attributes such as competition, assortment, and store size and therefore, each store can be representative of many other Rossman stores in the country.

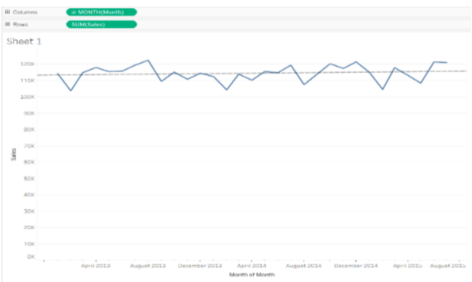
We have selected around 6 stores that are representative of the stores in the dataset and built forecasting models for them.

e. Charts:

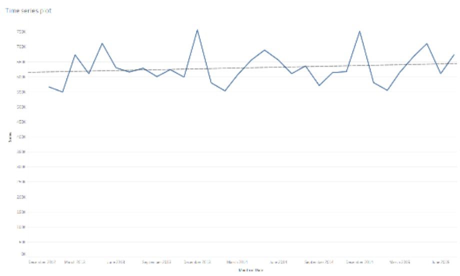


Some visible components:

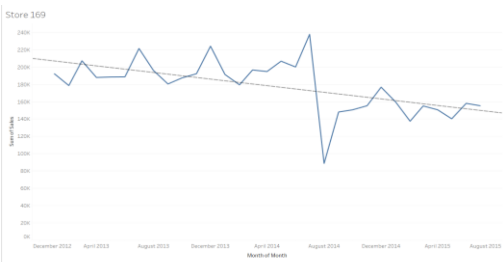
<u>Trend</u>: The following plots show the trend in the time series taken. A couple of time series do not have any trend, whereas one has downward linear trend and another has a quadratic trend.



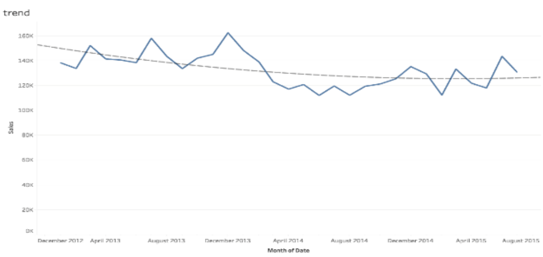
Store- 569



Store- 262



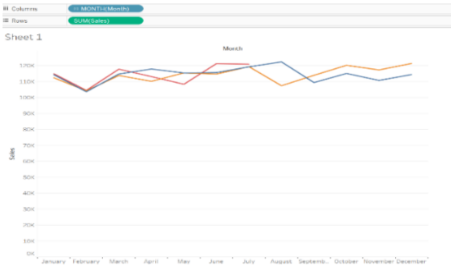
Store- 169



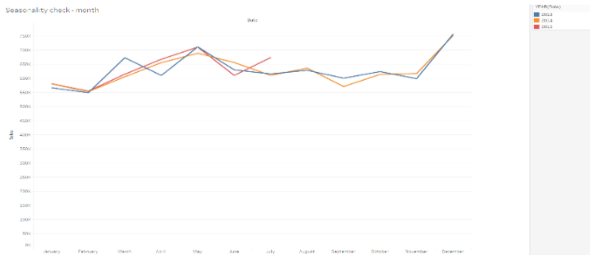
Store- 131

Seasonality:

Stores mentioned below have monthly seasonality.



Store- 569 (monthly seasonality)

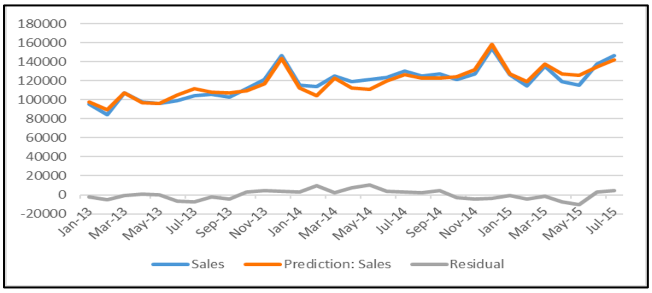


Store- 262 (monthly seasonality)

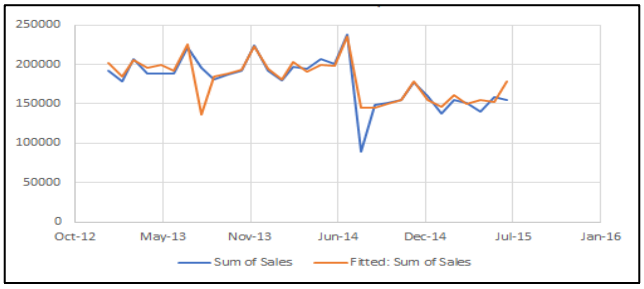
f. Final forecasting model and performance on meaningful performance metrics:

We’ve chosen six representative stores for six different models:

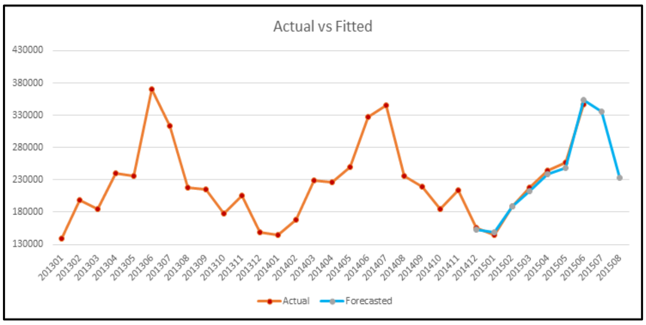
Store 61
Chosen Method: Linear Regression
(MAPE of chosen method: 5%, MAPE of Naïve: 10%)



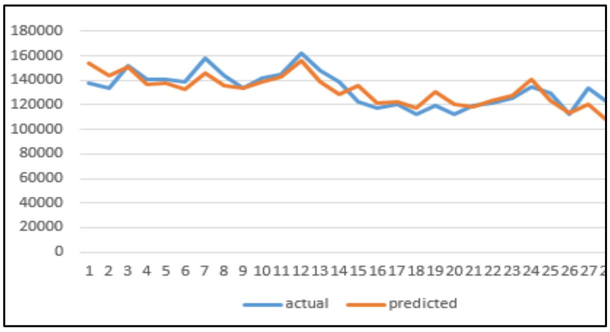
Store 169
Chosen Method: Holt-Winter’s Multiplicative
Alpha: 0.35 | Beta: 0.01 | Gama: 0.03
(MAPE of chosen method: 7.5%, MAPE of Naïve: 39.1%)



Store 453
Chosen Method: Linear Regression
(MAPE of chosen method: 2.03%, MAPE of Naïve: 5.12%)



Store 131
Chosen Method: Linear Regression
(MAPE of chosen method: 5.3%, MAPE of Naïve: 10%)

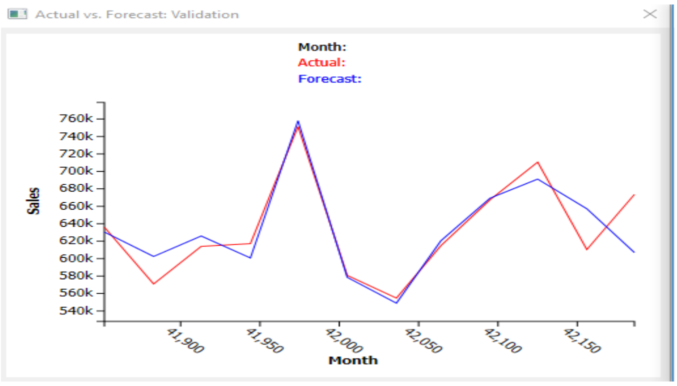


Store 262

Chosen Method: Forecasting using smoothing – Holt-Winter's with no trend

Alpha = 0.5 | Gamma = 0.75 | Beta = 0 (because, data has no trend)

(MAPE of chosen method: 2.89%, MAPE of Naïve: 2.93%)

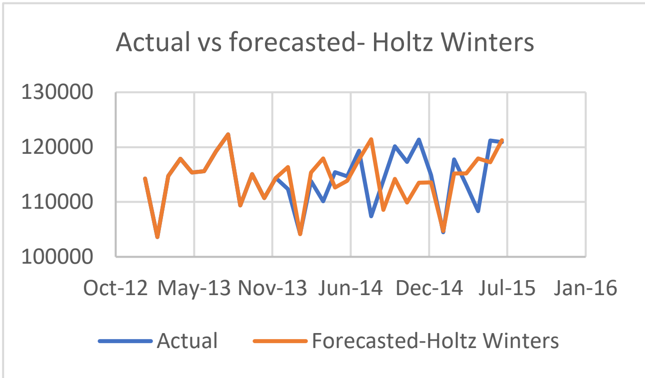


Store 569

Chosen Method: Holt-Winter's Multiplicative

Alpha = 0.368 | Gamma = 0.0009 | Beta = 0.763 (because, data has no trend)

(MAPE of chosen method: 4.5%, MAPE of Naïve: 4.8%)



g. Conclusion and Recommendation:

Since the stores are different in terms of assortments that they have, promotions that they give and their proximity from competitors- there is no one forecasting model best for all of them. Therefore, we've chosen six different stores that can together represent more than 1000+ stores and have built 6 different forecasting models for them.

Currently, the model predicts for only one month. It can be further modified to predict sales for 3 months.

Technical summary:

a. Data Preparation/issues:

- Daily Sales data for each store was rolled up to the monthly level
- Few stores had missing values for a few months, probably due to store renovation or moving of the store to a new location
- Identifying 6 stores that were a representative of the 1115 stores that the entire data set had. There were 4 types of store models and 6 combinations were identified based on unique combinations of store type and assortment type.
- 6 stores time series are chosen such that the store sales variance (both high and low) is captured

- There is a cost bearing to both over forecasting and under forecasting. Since there was no cost structure provided, we had to build a model that reduced the (absolute) error

b. Forecasting methods used:

With each series we **began with visualizing the data using Tableau**. Visualization was done to infer the components of each time series, i.e. what trend, seasonality, level, and noise it has. The component descriptions are mentioned in the previous section.

Post visualization, we **benchmarked each series with Naïve and seasonal Naïve forecast**. The **validation period** was chosen as the **maximum number between the forecasting horizon and seasonal number**.

Based on the components analyzed for each time series, we **ran the corresponding smoothing method(s)**, for example: for store #262, we had seasonality with no trend, so for this store we ran only Holt Winter's with no trend smoothing method. Time spent in visualizing and confirming the time series component helped us to narrow down the smoothing methods to be used for each series.

After smoothing was done, we modeled each time series data using **linear regression**. Again, knowing the components of each time series helped us to identify the predictor variables easily. One thing that we observed in one of the stores, i.e. store #169 was a sudden drop in the sales of the stores due to the invasion of a competitor store nearby. Usage of a **dummy variable to represent such competitor intervention** (pre and post competition) helped us model it well.

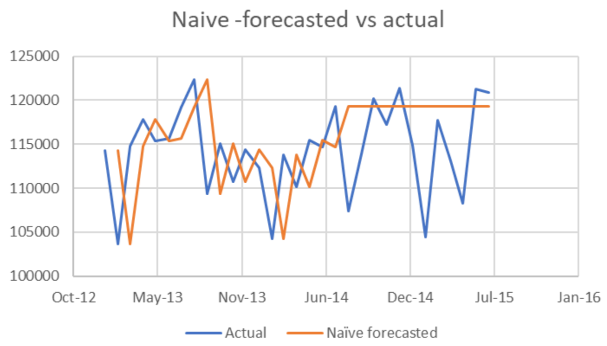
Eventually, all the methods were compared against benchmark, issues like overfitting and cost of under forecasting (lose money as sales reps will earn easily by beating the targets) and over forecasting (loss in morale of sales reps) were analyzed to select the final method for deployment.

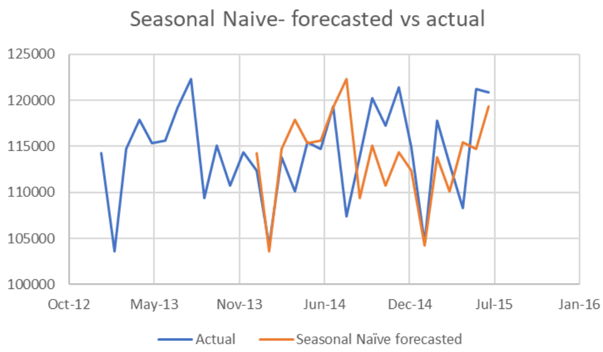
c. Performance compared to benchmarks

All the methods performed on each series are shown in the appendix with their performance compared against each other.

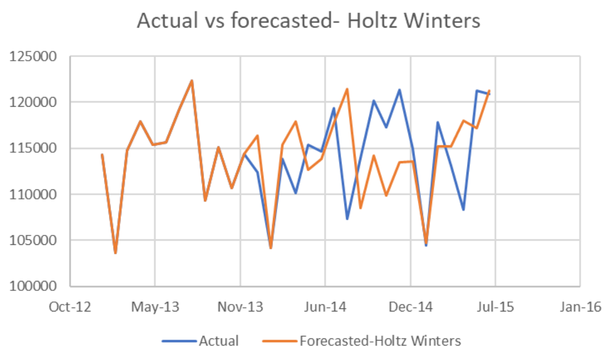
Appendix:

Store 569:



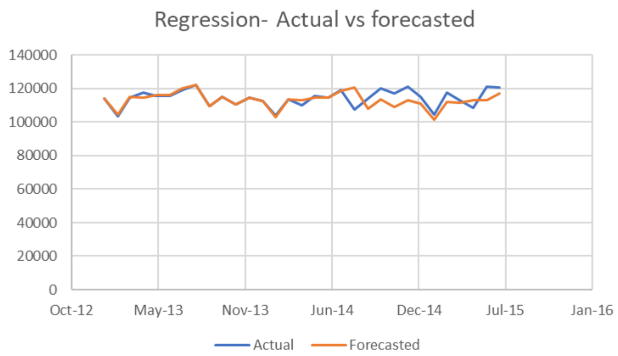


Holt Winter's multiplicative:



Model MAPE	Training	Validation
Naïve	4.7%	4.8%
Seasonal naïve	1.6%	4.6%
Holtz Winters	0.9%	4.5%
Linear Regression	0.6%	5.3%

Linear Regression:



Future forecasts using Holt Winter's multiplicative:

Aug'15: 115395.88 (this month will be lagged, and data has no use)

Sep'15: 113088.32 (this is the desired forecast)

Store 262:

Benchmarking using Naïve:

Month	Sales	Naïve	Residual - Naïve	MAPE - Naïve	Seasonal Naïve	Residual - Seasonal Naïve	MAPE - Seasonal Naïve
Jan-13	566482						
Feb-13	549174						
Mar-13	673085						
Apr-13	610222						
May-13	711428						
Jun-13	630094						
Jul-13	615851						
Aug-13	628556						
Sep-13	600554						
Oct-13	624051						
Nov-13	598850						
Dec-13	756285						
Jan-14	580046						
Feb-14	553094						
Mar-14	605661						
Apr-14	655468						
May-14	688993						
Jun-14	656008						
Jul-14	610550						
Aug-14	636065	610550	25515	0.040113825	628556	7509	0.011805397
Sep-14	570922	610550	-39628	0.069410532	600554	-29632	0.051902011
Oct-14	614107	610550	3557	0.00579215	624051	-9944	0.016192618
Nov-14	617180	610550	6630	0.010742409	598850	18330	0.029699601
Dec-14	751615	610550	141065	0.187682524	756285	-4670	0.006213287
Jan-15	580798	610550	-29752	0.051226072	580046	752	0.00129477
Feb-15	554704	610550	-55846	0.100677118	553094	1610	0.002902449
Mar-15	614812	610550	4262	0.0069322	605661	9151	0.014884225
Apr-15	667510	610550	56960	0.085332055	655468	12042	0.018040179
May-15	710640	610550	100090	0.140844872	688993	21647	0.030461274
Jun-15	610404	610550	-146	0.000239186	656008	-45604	0.074711175
Jul-15	673633	610550	63083	0.093645947	610550	63083	0.093645947
Level	**629575.5**			**0.066053241**			**0.029312745**

Forecasting using smoothing - Holt Winters with no trend

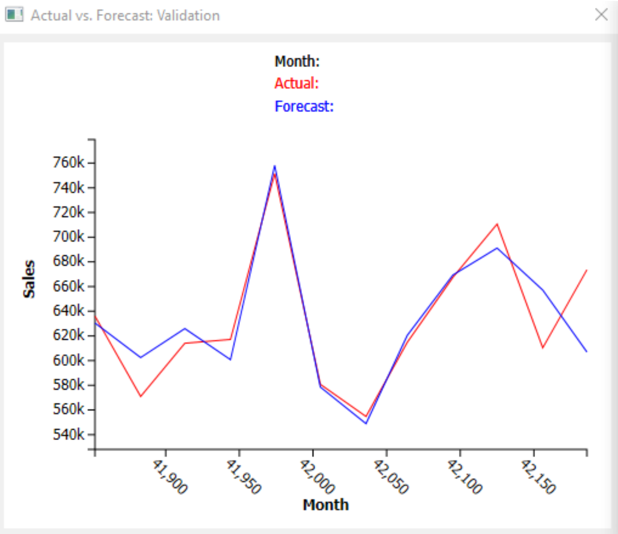
Alpha = 0.5 | Gamma = 0.75 | Beta = 0 (because, data has no trend)

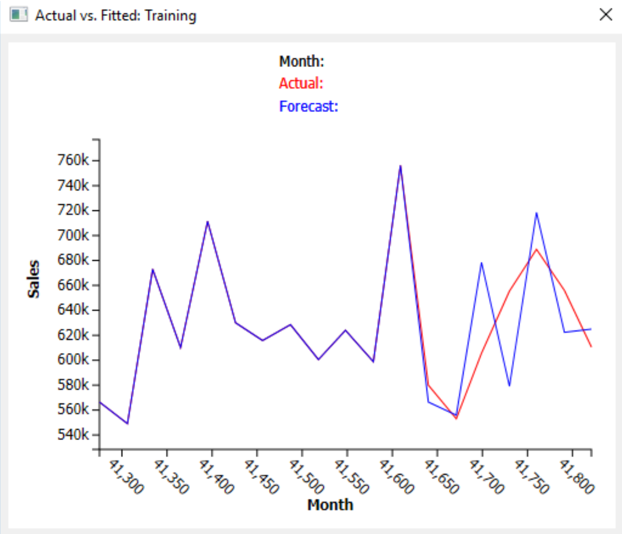
Error Measures: Training

Record ID	Value
SSE	1.35E+10
MSE	7.11E+08
MAPE	2.014649
MAD	12790.94
CFE	3823.438
MFE	201.2336
TSE	0.298918

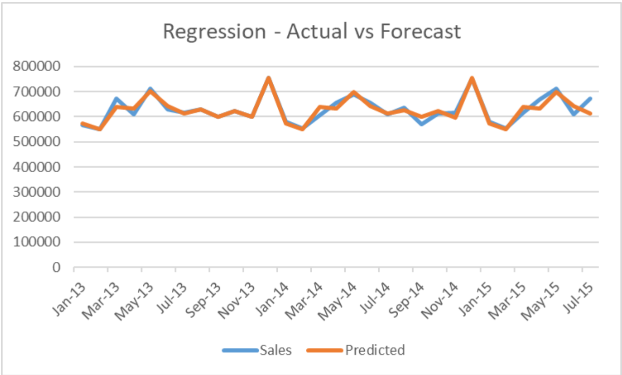
Error Measures: Validation

Record ID	Value
SSE	8567659354
MSE	713971612.9
MAPE	2.89864379
MAD	18363.52995
CFE	11949.79688
MFE	995.8164063
TSE	0.650735284





Linear Regression based forecasting (only seasonality):



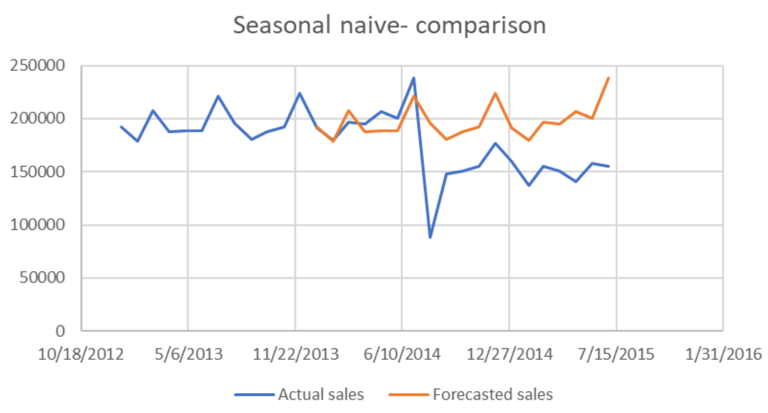
MAPE: Training: 1.53%; Validation: 3.25% (dropped because of overfitting)

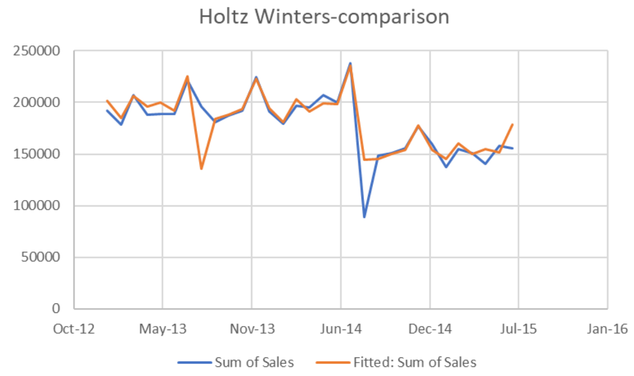
Future forecast using chosen method (Holt Winter's with no trend):

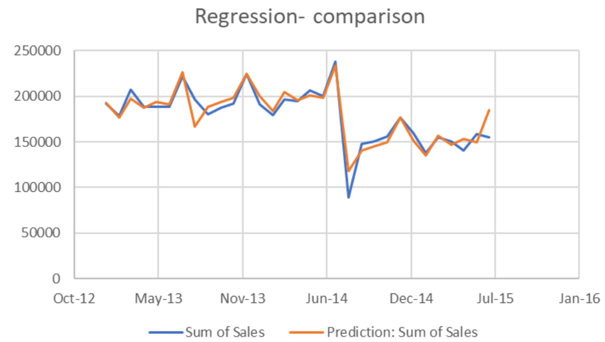
Aug'15: 660608.1368 (this month will be lagged, and data has no use)

Sep'15: 602022.5325 (this is the desired forecast)

Store 169:

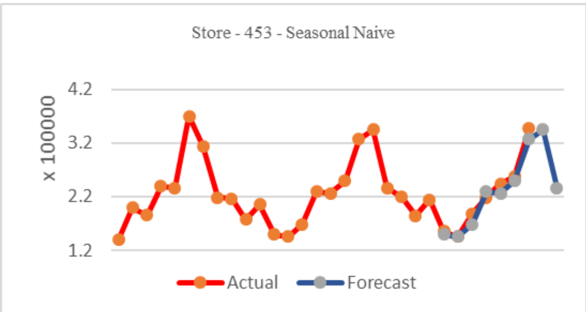






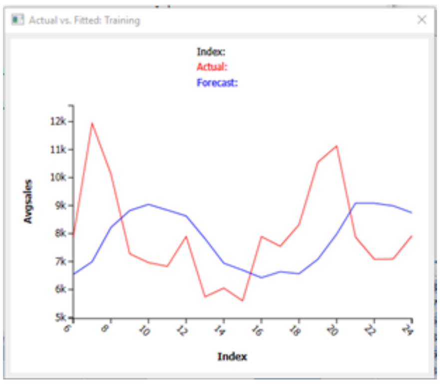
Method	Training MAPE	Validation MAPE
Holtz Winters	5.6%	7.5%
Seasonal naive	21.8%	39.1%
Regression	4%	9%

Store 453:



Store ID	Model	MAPE	Forecasted Sales (for Sept)
453	Linear Regression	2.03%	233093
453	Seasonal Naive	5.12%	236223
262	Holts Winter (with no trend)	2.89%	
262	Regression		

Smoothing:



Linear Regression: