ASSIGNMENT SUBMISSION FORM

Course Name: Forecasting Analytics

Assignment Title: YourCabs - Demand by each car segment type

Submitted by: Group A6

(Student name or group name)

Group Member Name	PG ID
Ananya Guha	61410014
Devarishi Das	61410074
Nirman Sarkar	61410478
Pratyush Kumar	61410141
Shridhar S Iyer	61410529

Executive Summary:

YourCabs aggregates radio cabs from multiple taxi operators in Bangalore. For each model type, the supply shifts dynamically and depends on the number of drivers logged in. If cabs are unavailable, Yourcabs incurs cost on either upgrading the passenger or losing the sale. If several cabs do not have bookings, there is some loss of relationship with the supplier. Thus, **YourCabs needs to be able to predict the demand for each cab type at an hourly basis so it can reach out to taxi operators and plan the number of drivers logging into its system.**

Our forecasting problem predicts the demand for each cab type segment from November 15-22, 2013 at an hourly level. We first split the cab models into 5 segments – small cars, sedans, utility vehicles, premium cars, and buses. On plotting the data we see that the trend changes drastically from mid 2013 for small cars and sedans. Thus, only data for the last 6 months should be used to generate forecasts. Further, there is light seasonality at a weekly level and very strong seasonality at the hourly level for each segment. These seasonalities need be accounted for while making forecasts. We test the forecasting model over the first two weeks of November to evaluate and compare different models.

We first generate benchmark forecasts using a naive rule. We then create forecasting models using multiple linear regressions and the Holt Winter model. For Small cars, Sedans, and Utility vehicles, linear regression is a better predictor of cab demand over the validation period. We then generate forecasts for the forecast horizon using regression. However, the model has a tendency to under-predict the peak demand, and under-prediction cost is higher than over-prediction cost. Thus **we attach a factor of safety to the forecasts before making any business decisions.** We make suitable assumptions for the factors of safety for each of the popular segments. Premium cars and buses have seen very little demand over the past two years. Thus, it is not feasible to generate forecasts for their demand. Practically, these vehicles can be arranged specifically for any customer who asks for them.

In conclusion **we recommend using multiple linear regression including intraday and intraweek seasonalities for forecasting hourly demand for each cab segment**. The **forecasts can be generated at the start of the day for the entire day and rolled forward for the next day**. Our model does not take into account the location of the supply and demand. Nor does it factor in vehicles in transit that will become available for booking in the next hour. These aspects must be accounted for before deploying the forecasting model.

Technical Summary - Segment Wise Cab Demand Forecast for YourCabs

Business problem: A broad prediction of the demand of each cab type in the immediate future (1 week horizon). YourCabs offers a wide selection of taxis ranging in luxury and capacity. The cab supply is dynamically shifting and depends on the number of cab drivers logged in with YourCabs. The problem is ensuring availability of sufficient taxis of each type at any given time to meet the demand.

Forecasting problem: YourCabs needs to predict the demand for each cab type to ensure supply of cabs for the immediate succeeding hour at any given time. Demand for each segment is to be forecasted at an hourly level for one week

Assumptions: Some of the assumptions basis which we worked are as follows :

- YourCabs acts as an aggregator of radio-cabs from several operators
- Yourcabs is aware of the number of cabs available for booking at any given time
- Yourcabs can arrange for more cabs of any particular type within an hour
- Cabs can be redirected from any location in the city to any pick up point
- Cost of having too few cabs is the cost of upgrading a passenger or losing a sale
- Cost of having too many cabs standing by is loss of relationship with cab operator

Cab Portfolio Segmentation: The Assumption here is that vehicles that are similar to each other can be used interchangeably and hence can be clubbed together. This helped us generate aggregate demand for each type of segment to avoid extremely granular results.

1. Small car : Indica, I-10, Alto, Wagon R etc. (8 models)
2. Sedan: City, Civic, Logan etc. (31 models)
3. Utility: Scorpio, Bolero, Safari etc. (16 models)
4. Bus: Marcopolo, Swaraz Mazda, Volvo etc. (7 models)
5. Premium: Mercedes C class, BMW A8 etc. (17 models)

Macro level Segment-wise Visualization

Snapshot of Visualization			
	Small car	Sedan	Utilities
Level	There is a drastic jump in level from 2012 to 2013	There is again in jump in the level observed from 2012 to 2013	This segment has not seen any significant change in level
Trend	In the year 2013 there is a steep increasing trend	A linear increasing trend observed across the years	there is a marginal increasing trend
Seasonality	The demand seems to peak every August	The demand seems to peak every July	The demand seems to peak every August

Analysis: The first inspection report revealed that there is a definite increase in the demand for small cars first steadily since 2012 and then a drastically May2013 onwards. This is the reason why we could consider data over the last 6 months in our training set to get a more accurate picture. Sedans also showed a similar trend even though the actual numbers were much lesser compared to small cars. The utility vehicles have not shown much growth over the last couple of years. There were too few data points to make any forecast for Buses and Premium vehicles. YourCabs should consider moving out of this segment and focusing on the other more lucrative segments. For the Small Car segment light seasonality can be observed within the week- peaks on Fridays, however within a day there is a strong seasonality as demands seem to peak around 0800hrs and 1700hrs – presumably due to those being office hours for most. Moreover for the utility segment we observed that the demand peaks on the weekends at early hours and subsides during weekdays. This is possibly due to long/out of city trips planned during weekends. In general the seasonality appears additive rather than multiplicative for all the segments.

Forecasting period details

- Training data: 14th May 2013 to 31st Oct 2013 (6 months)
- Validation data: 1st Nov 2013 to 14th Nov 2013 (2 weeks)
- Forecast Horizon: 15th Nov 2013 to 22nd Nov 2013 (1 week)

Models considered

- Naïve forecast,
- Linear Regression

- Holt Winter method without trend

Seasonalities considered - ‘Day of the Week’ & ‘Hour of the day’.

Segment 1 – Small Cars

These are the highest demand vehicle comprising of 77% of the total demand. Following are the summary of each of the forecast used:

Naïve: the following table shows the different RMSEs observed for hourly, daily and weekly adjusted Naïve forecast:

	Training RMSE	Validation RMSE
Hourly Adjusted Naïve Forecast	3.618	3.742
Daily Adjusted Naïve Forecast	3.562	4.013
Weekly Adjusted Naïve Forecast	4.942	6.261

Based on the RMSE values the hourly adjusted prediction model fits best to the validation data. The Validation plot (Exhibit 1) is also indicative of the same phenomenon. However 1 hour would not be very helpful to help plan future capacity. Daily Adjusted Naïve forecast is therefore ideal trade-off considering accuracy and available reaction time.

Regression: The following linear regression model is used

$M_{0....}$ M_5 = Dummy variable for the days of the week

$N_{0....}$ N_{23} = Dummy variable for the hours of the day

$\beta_{0....}$ B_{30} =constant coefficients

Demand = $\beta_0 + \beta_1 M_{0+....} + \beta_6 M_5 + \beta_7 N_{0+...} + \beta_{30} N_{23}$

Holt Winter without trend: the model was not able to capture the trends and seasonalities as well as the linear regression

The regression plot is then compared with the naïve forecast and the visualization is displayed in Exhibit 2. The following is the RMSE comparison.

RMSE	**Naïve Forecast**	**Regression**
Training Period	3.56	**3.07**
Validation period	4.01	**3.69**

Thus basis RMSE Values over training period and validation period it can be concluded that linear regression is a better predictor of demand for small cars than Naïve forecast and Holt Winter method

Segment 2 - Sedan

These are the second highest demand vehicles comprising of 16% of the total demand. Similar to the small car segment, this segment is analysed with respect to different models.

Naïve: Again we find the daily adjusted plot ideal method for forecasting. The RMSE values and plots are shown in exhibit 3.

Regression: similar model used as described for the small car segment (additive linear)

Holt Winters: Again this method was not able to capture trend and seasonality

Finally regression model was chosen to determine the forecast and the plot is shown in exhibit 4.

Segment 3 Utilities: observation and conclusion were same for this segment as well and the plots are shown in the exhibit 5.

Factor of safety: we observed from the exhibits that the peak demands are not captured well in terms of their magnitudes by the chosen forecasting model (linear regression). Since the cost of under forecasting is typically greater than the cost of over forecasting, it makes business sense to have buffer capacity on demand. Therefore we recommend the following factors of safety of the respective segments. The values are chosen to match the forecast with actual peaks during validation period.

Actual peak demand/forecasted peak demand = FOS

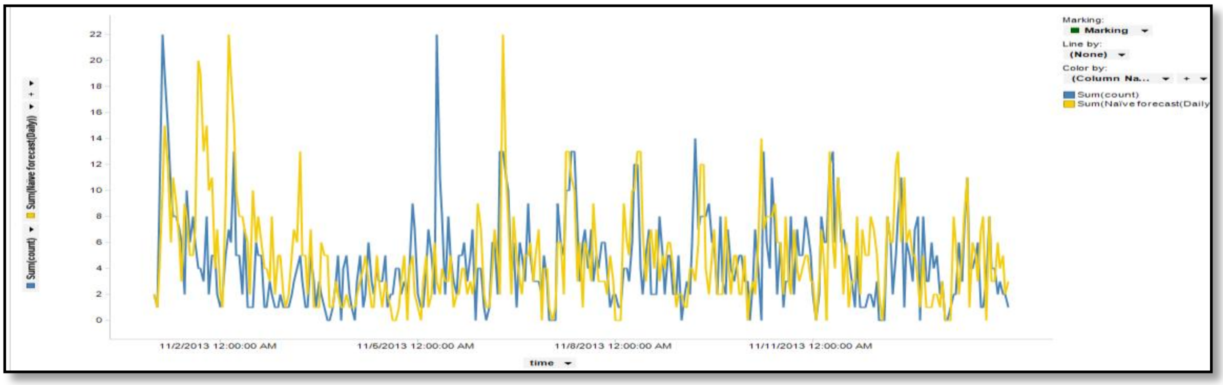
	Factor of safety
Small car	1.25
Sedan	1.30
Utility	1.50

Recommendations:

- Linear regression with additive seasonality is the best predictor of Hourly demand
- ‘Hour of the day’ and ‘Day of the week’ seasonality must be taken into account
- Bookings data can be aggregated for cab model segments
- Forecasts can be generated at an hourly level for the entire day at the start of the day
- Forecasts can be rolled forward on a daily basis to include fresh data
- Since cost of under-prediction is higher than cost of over-prediction, a factor of safety must be used when using forecasts for capacity planning
- Multiplicative factor of safety is preferred to ensure capturing of peak demand
- Higher FOS for Utility and Sedans as cost of under-prediction will be higher
- Cabs in transit that will become available in the next hour must be taken into account when calculating supply

Appendix

Exhibit 1



Marking: Marking

Line by: (None)

Color by: (Column Na...

Exhibit 2

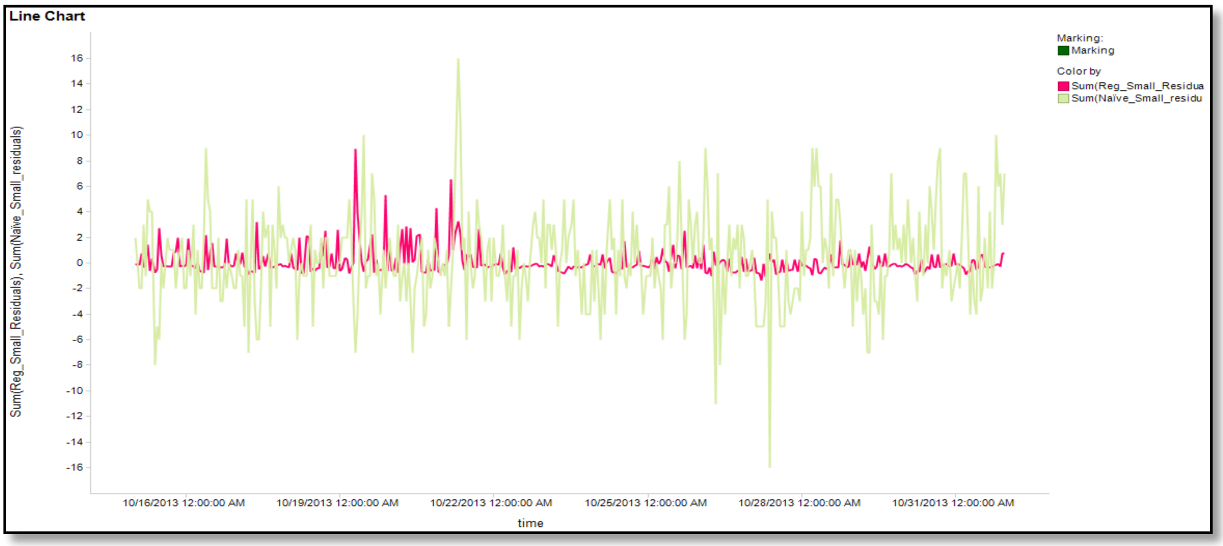
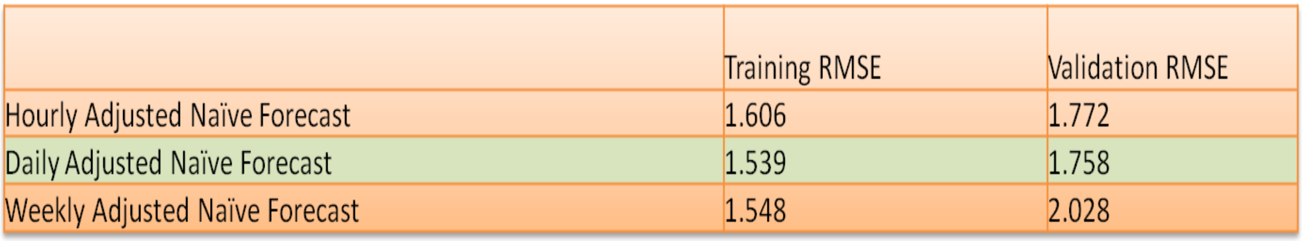


Exhibit 3



	Training RMSE	Validation RMSE
Hourly Adjusted Naïve Forecast	1.606	1.772
Daily Adjusted Naïve Forecast	1.539	1.758
Weekly Adjusted Naïve Forecast	1.548	2.028

Exhibit 4

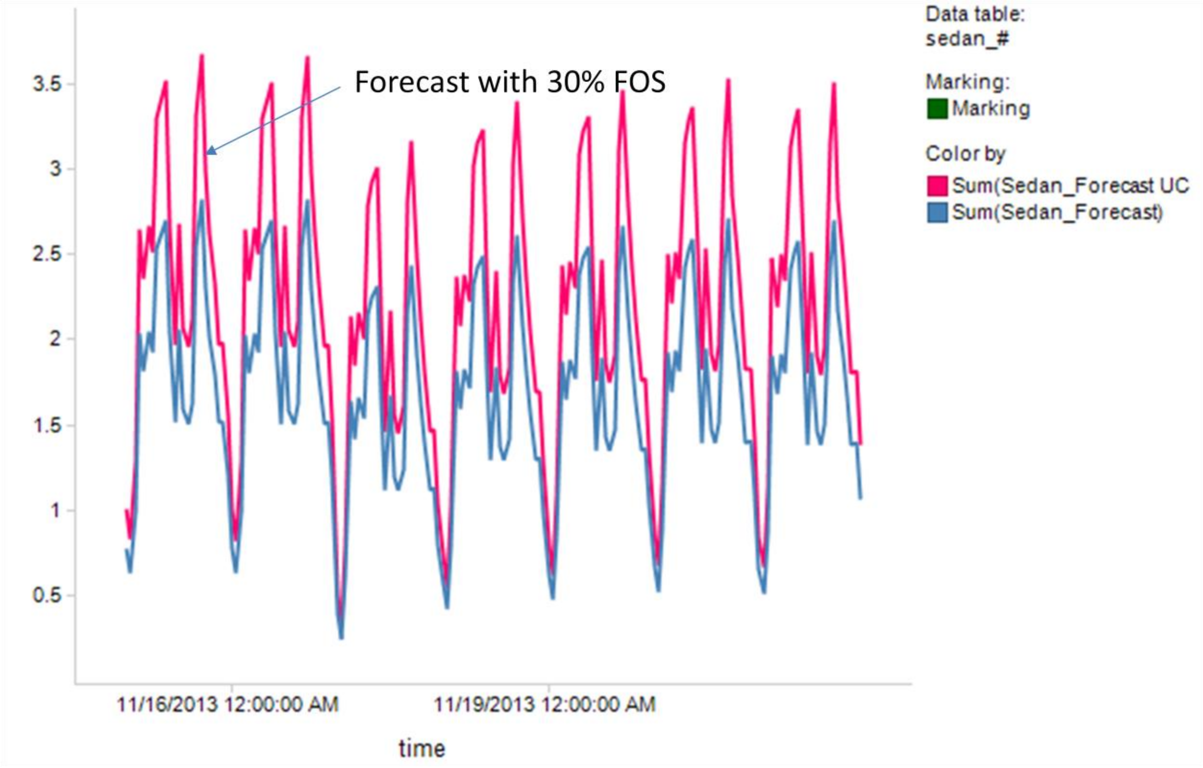
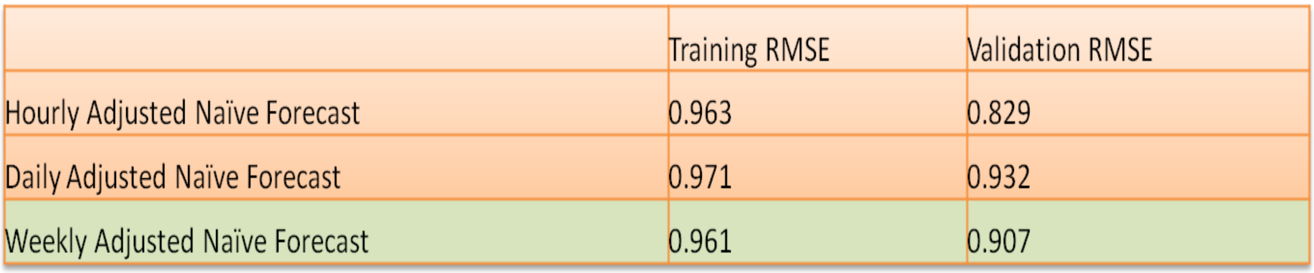
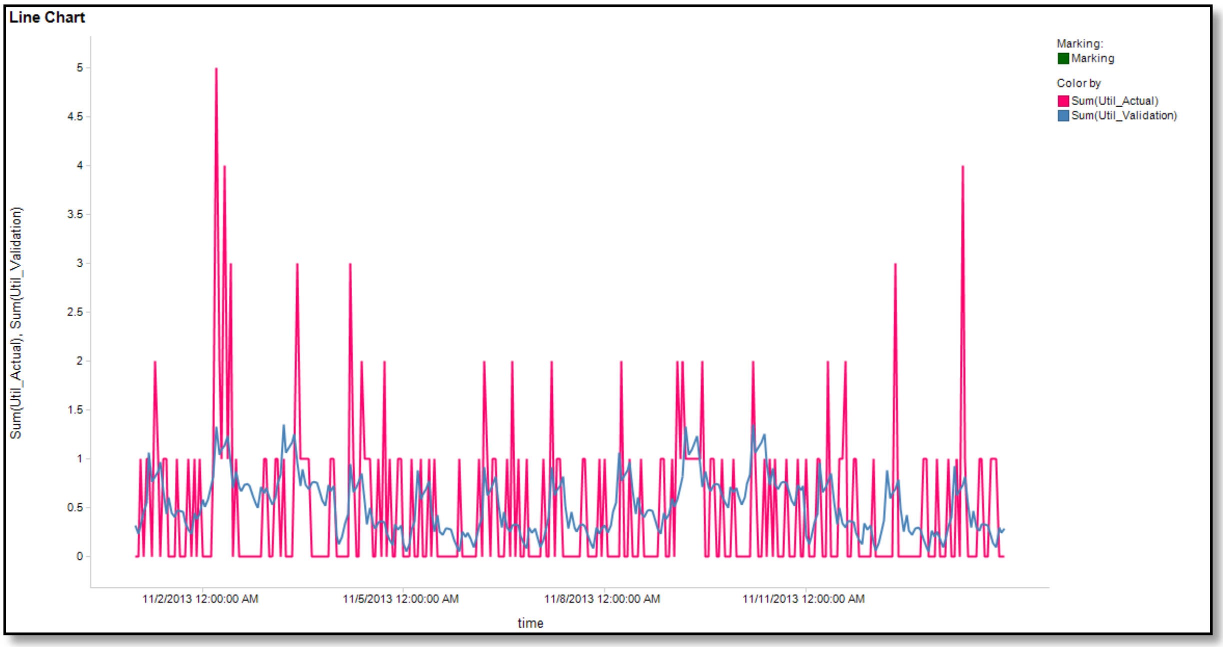


Exhibit 5

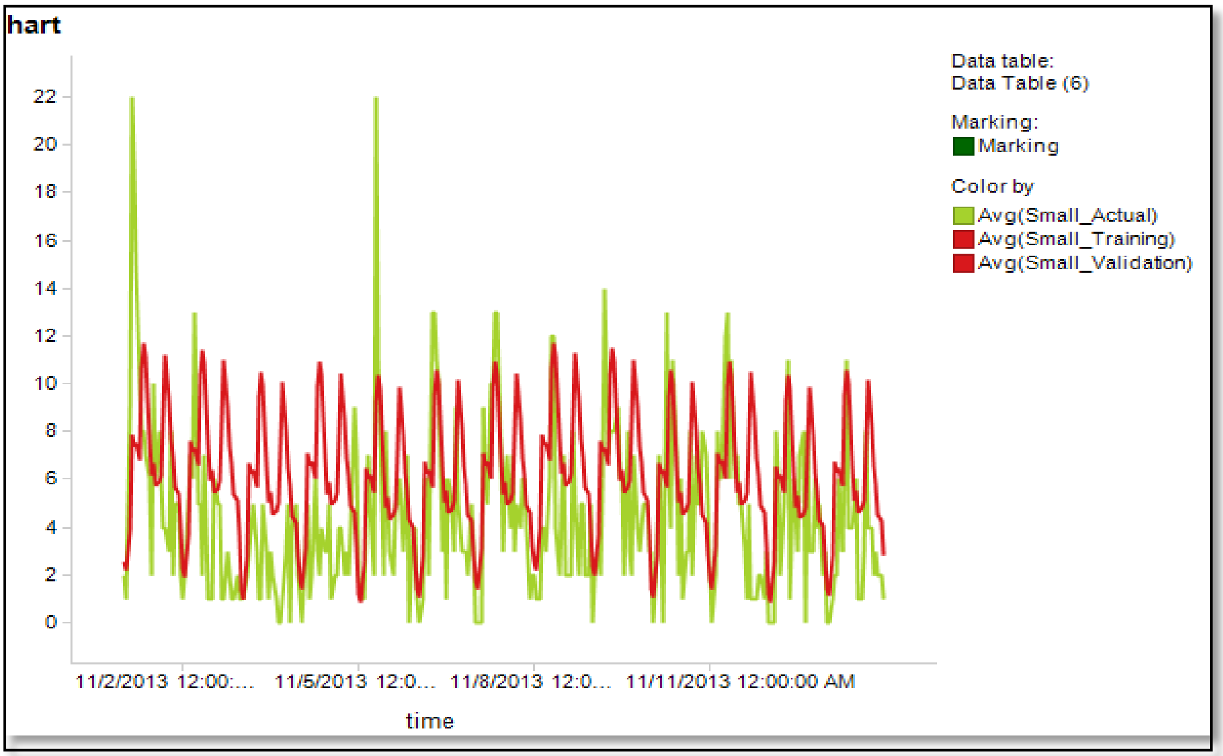


	Training RMSE	Validation RMSE
Hourly Adjusted Naïve Forecast	0.963	0.829
Daily Adjusted Naïve Forecast	0.971	0.932
Weekly Adjusted Naïve Forecast	0.961	0.907

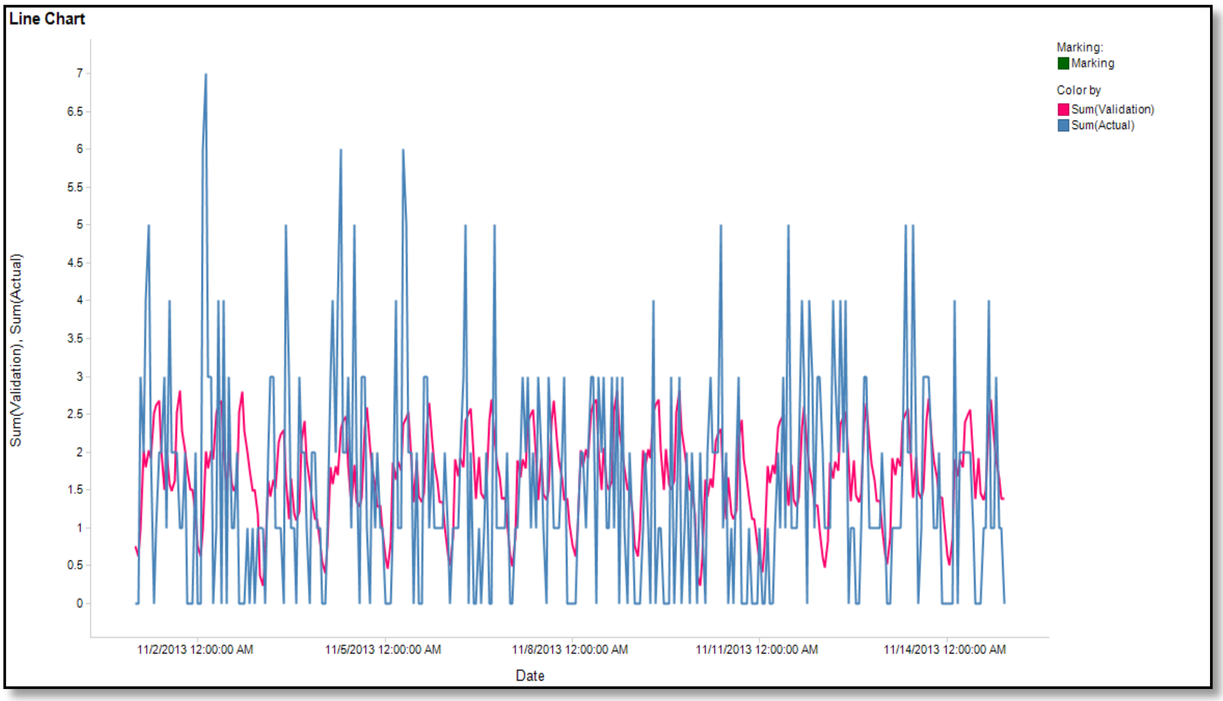


Graphs of the validation data for each of the segments for linear regression model

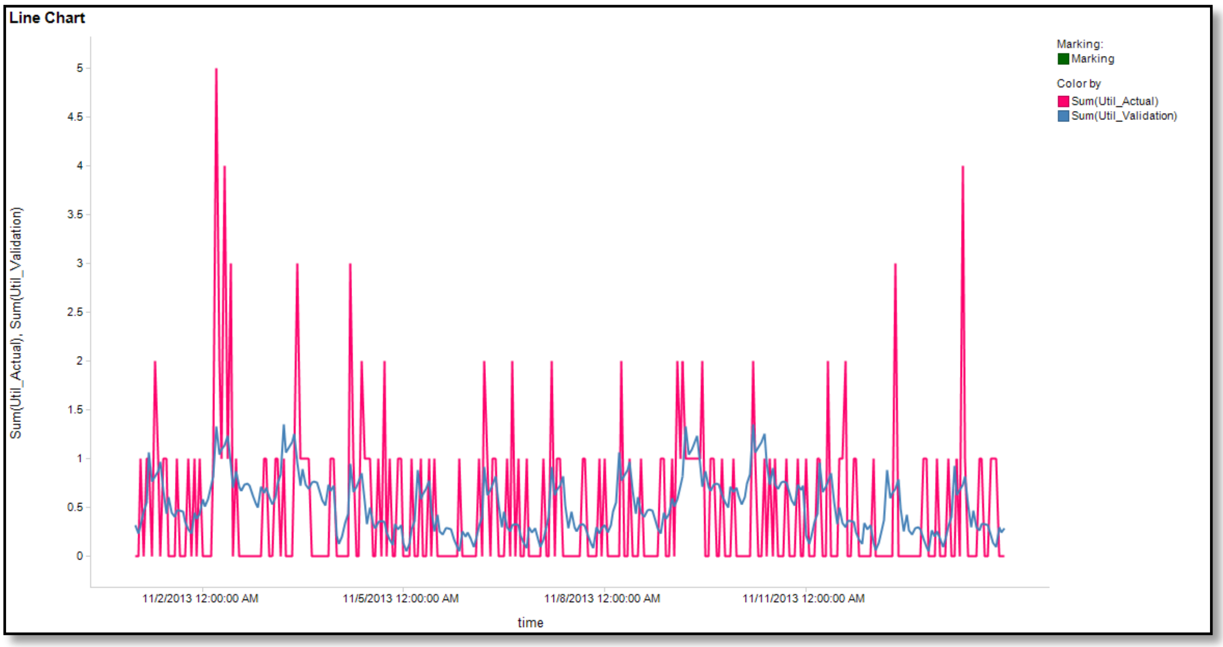
Small



Sedan

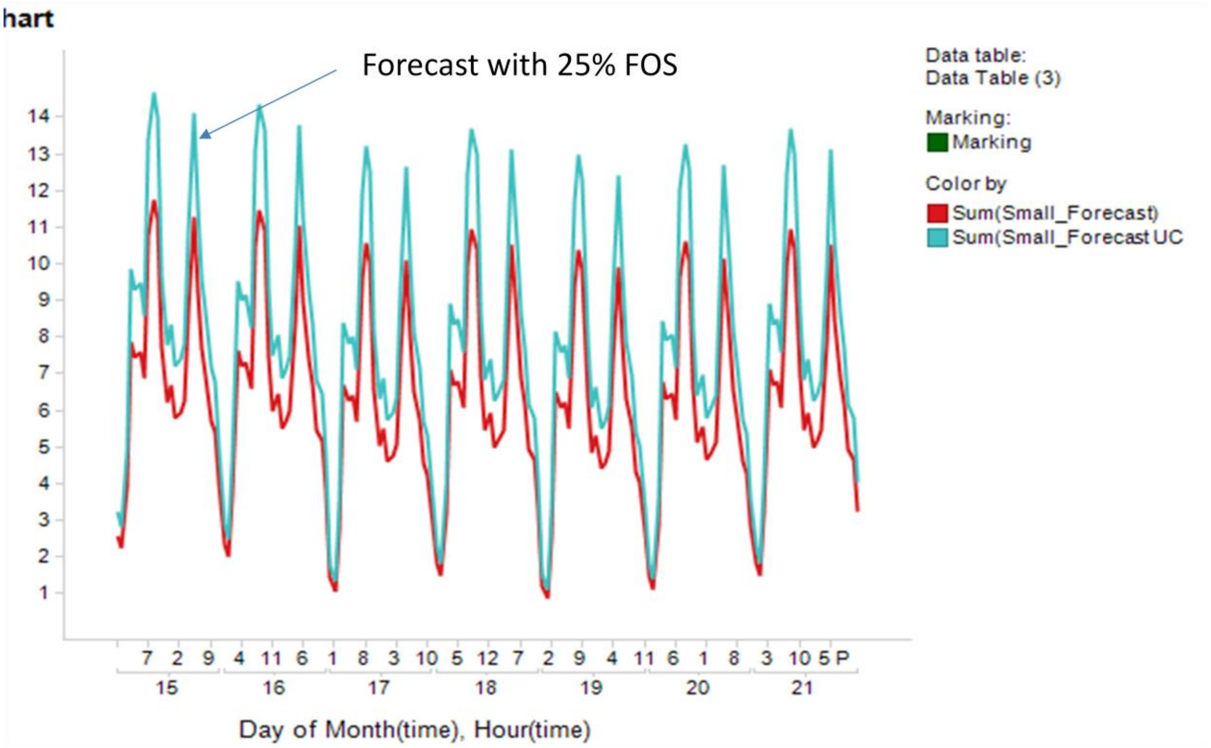


Utility

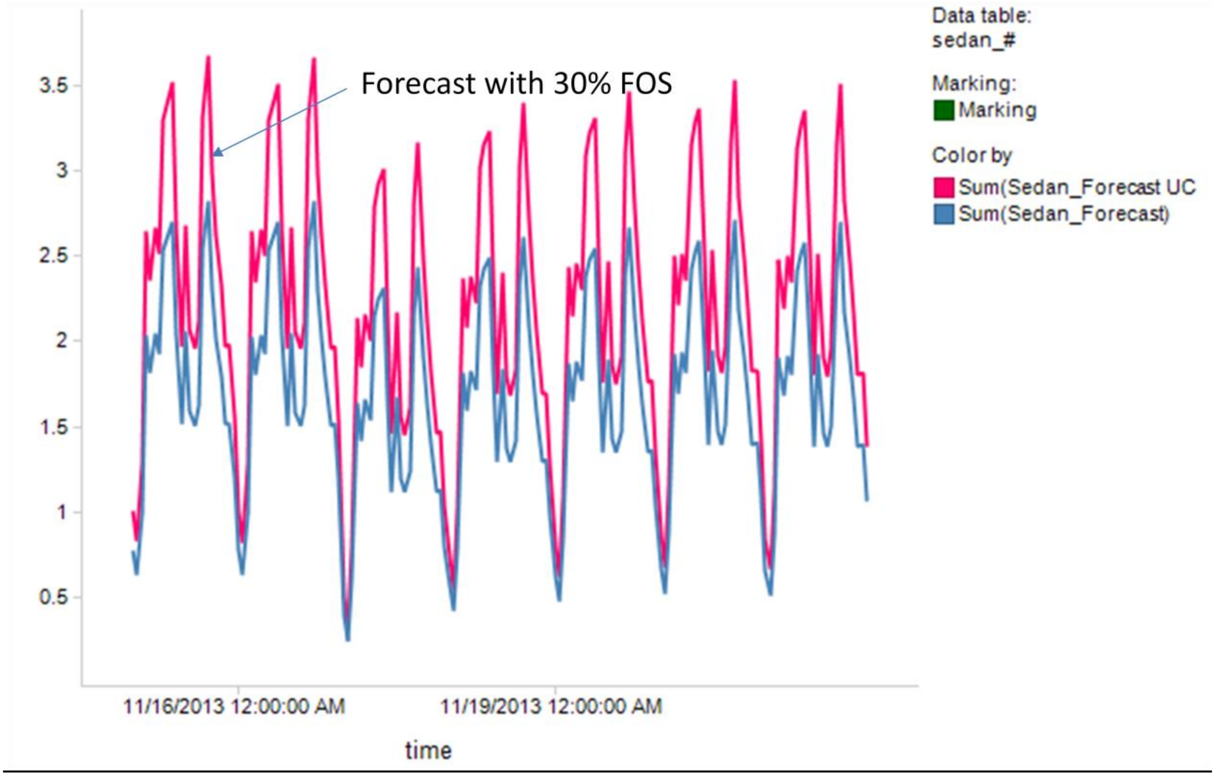


Graphs of the forecasts for each of the segments

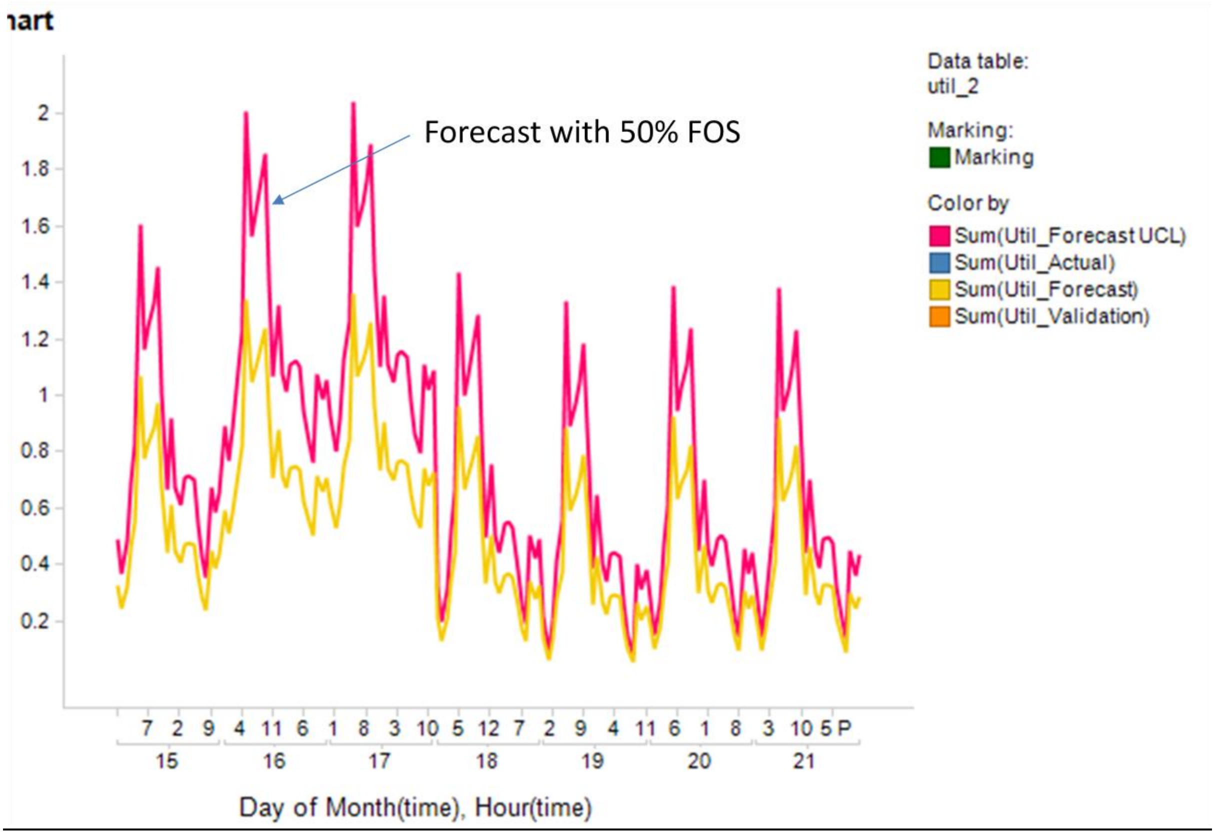
Small



Sedan


Data table:
sedan_#
Marking:
🟩 Marking
Color by
🟥 Sum(Sedan_Forecast UC
🟦 Sum(Sedan_Forecast)
Forecast with 30% FOS
3.5
3
2.5
2
1.5
1
0.5
11/16/2013 12:00:00 AM
11/19/2013 12:00:00 AM
time

<u>Utility</u>


hart
Data table:
util_2
Marking:
🟩 Marking
Color by
🟥 Sum(Util_Forecast UCL)
🟦 Sum(Util_Actual)
🟨 Sum(Util_Forecast)
🟧 Sum(Util_Validation)
Forecast with 50% FOS
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
7 2 9 4 11 6 1 8 3 10 5 12 7 2 9 4 11 6 1 8 3 10 5 P
15 16 17 18 19 20 21
Day of Month(time), Hour(time)