Analysis of Energia's Data set

Evan Ganson Saldanha

14/04/2019

Objective:

To discover the factors affecting the purchase of electricity by Energia on ISEM market and to optimise the buying by predicting the best market for every time period.

Exploring Dataset:

The data set consisted of 8,448 observation and 24 variable, of which 3 are charecter, 20 are numeric and one time variable. The observations under *Period.ending* were a time period which began from 30-09-2018 23:30 to 25-03-2019 23:00 which was in PosixCt format. The below table displays the first few rows of the whole data set:

Table 1: Energia's data-set

Period.Ending	Actual.WIND	DAM.Forecast.WIND	IDA1.Forecast.WIND	IDA2.Forecast.WIND	IDA3.Forecast.WIND	Actual.DEMAND
2018-09-30 23:30:00	1004	865	965	977	995	3427
2018-10-01 00:00:00	889	865	965	977	995	3232
2018-10-01 00:30:00	895	767	846	854	827	3150
2018-10-01 01:00:00	752	767	846	854	827	3085
2018-10-01 01:30:00	614	696	732	746	780	3001
2018-10-01 02:00:00	608	696	732	746	780	2947
2018-10-01 02:30:00	627	624	638	664	647	2887
2018-10-01 03:00:00	577	624	638	664	647	2854
2018-10-01 03:30:00	586	564	567	596	570	2807
2018-10-01 04:00:00	611	564	567	596	570	2809
2018-10-01 04:30:00	617	538	521	556	504	2813
2018-10-01 05:00:00	587	538	521	556	504	2838

Table 2: Energia's data-set

DAM.Forecast.DEMAND	IDA1.Forecast.DEMAND	IDA2.Forecast.DEMAND	IDA3.Forecast.DEMAND	DA.Price	IDA1.Price	IDA2.Price	IDA3.Price	BM.Price
3319	3299	3299	3296	71.267	67.630	NA	NA	80.07
3319	3299	3299	3296	71.267	54.318	NA	NA	80.81
3068	3050	3050	3048	67.212	52.210	NA	NA	54.57
3068	3050	3050	3048	67.212	49.547	NA	NA	36.75
2902	2876	2876	2885	60.500	39.000	NA	NA	33.68
2902	2876	2876	2885	60.500	39.000	NA	NA	-137.57
2813	2789	2789	2792	63.682	43.751	NA	NA	46.29
2813	2789	2789	2792	63.682	40.469	NA	NA	32.38
2756	2747	2747	2737	71.617	53.730	NA	NA	62.11
2756	2747	2747	2737	71.617	53.730	NA	NA	69.79
2776	2781	2781	2760	72.855	55.000	NA	NA	62.99
2776	2781	2781	2760	72.855	55.000	NA	NA	67.99

As the table 1 and 2 shows above, the data is for 4 different auction window on ISEM market namely: DAM, IDA1, IDA2 and IDA3. The DAM and IDA1 functions for 24 hours starting from 23:00:00 whereas the IDA2's acution period is for 12 hours staring 11:00:00 and IDA3's aution period for 6 hours starting from 17:00:00. All units are in MW and the prices in euros. The BM (Balancing market) prices are charged when Energia fails to buy required quantity of electricity.

Exploratory Data Analysis

Tidying the data:

Since the data was not tidy for visual representation, the complex variables were broken down to form few extra variables. The net demand calculated was the difference between the demand and wind. The column Period.Ending was further split into, *Time_hr* : time period from Period.Ending, *Time* : morning, afternoon, evening and night, *WeekDay* : 7 days of the week.

Table 3: Variables added to data-set after tidying

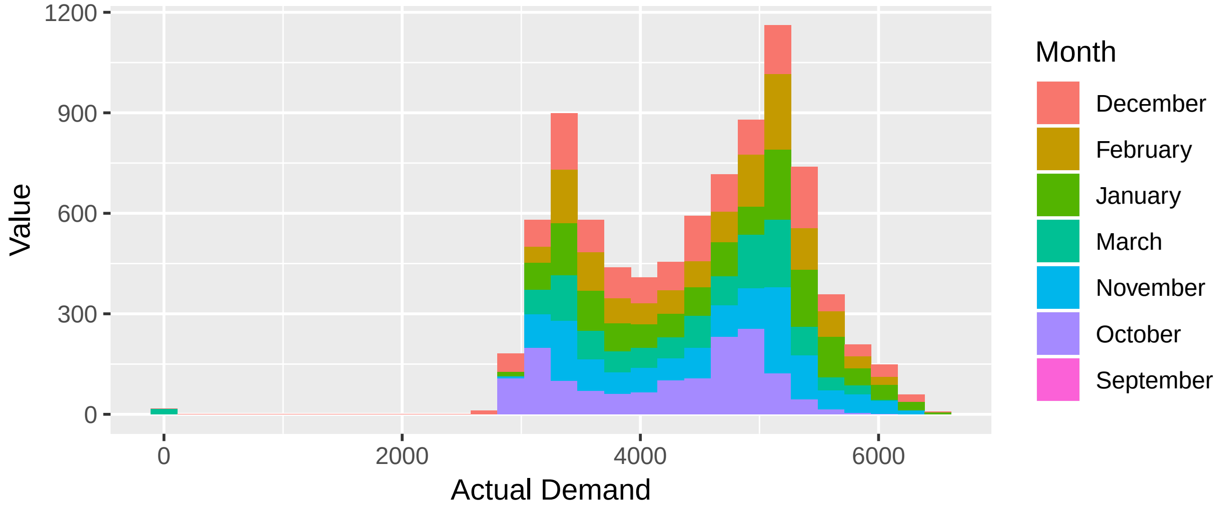
Actual_NetDemand	DAM_FC_NetDemand	IDA1_FC_NetDemand	IDA2_FC_NetDemand	IDA3_FC_NetDemand	Time_hr	Time	WeekDay
2423	2454	2334	2322	2301	23:30:00	night	Sunday
2343	2454	2334	2322	2301	00:00:00	night	Monday
2255	2301	2204	2196	2221	00:30:00	night	Monday
2333	2301	2204	2196	2221	01:00:00	night	Monday
2387	2206	2144	2130	2105	01:30:00	night	Monday
2339	2206	2144	2130	2105	02:00:00	night	Monday
2260	2189	2151	2125	2145	02:30:00	night	Monday
2277	2189	2151	2125	2145	03:00:00	night	Monday

Analysis is done on each market seperately, hence the whole data set is broken down into 4 individual data sets with respect to markets (DAM, IDA1, IDA2, IDA3). Every set consists the respective forecasted data, actual data, price and time factors.

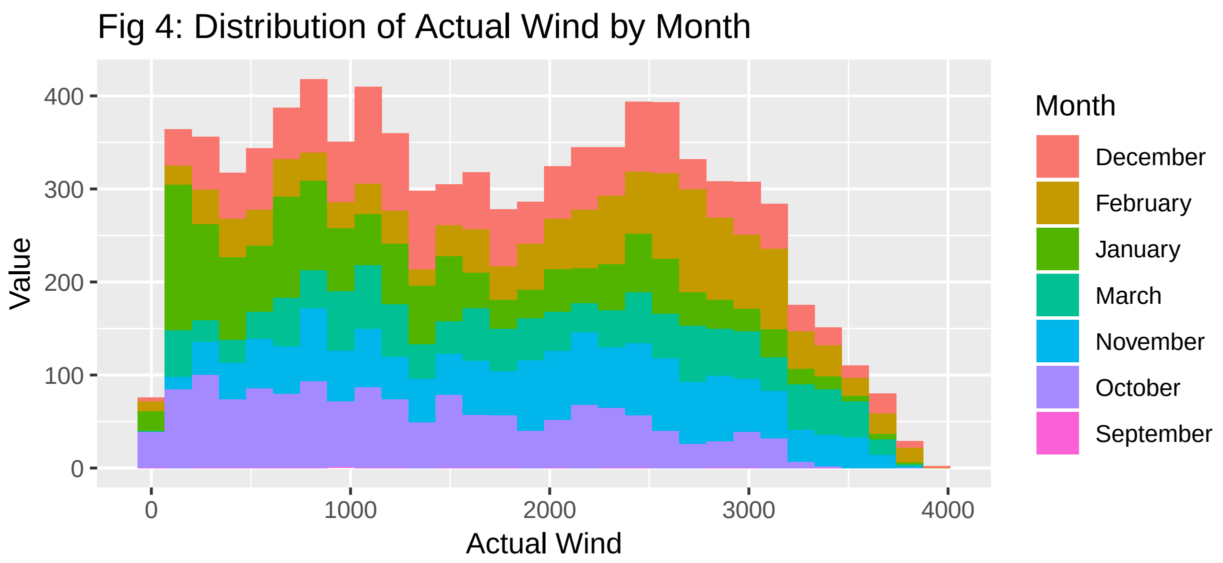
Visualization:

The two main factors on which the prices depend are the electricity produced by the wind and the demand of electricity by the customers. The below graph(Fig 3), respresents the demand of the customers of Energia on every month. The x axis represents the electricity in MW and the Y axis is the count of demand for that MW. The plot describes that, in the month of December the demand is highest of all other months followed by February. The cause of the demand rise may be due to the festive season in those two months.

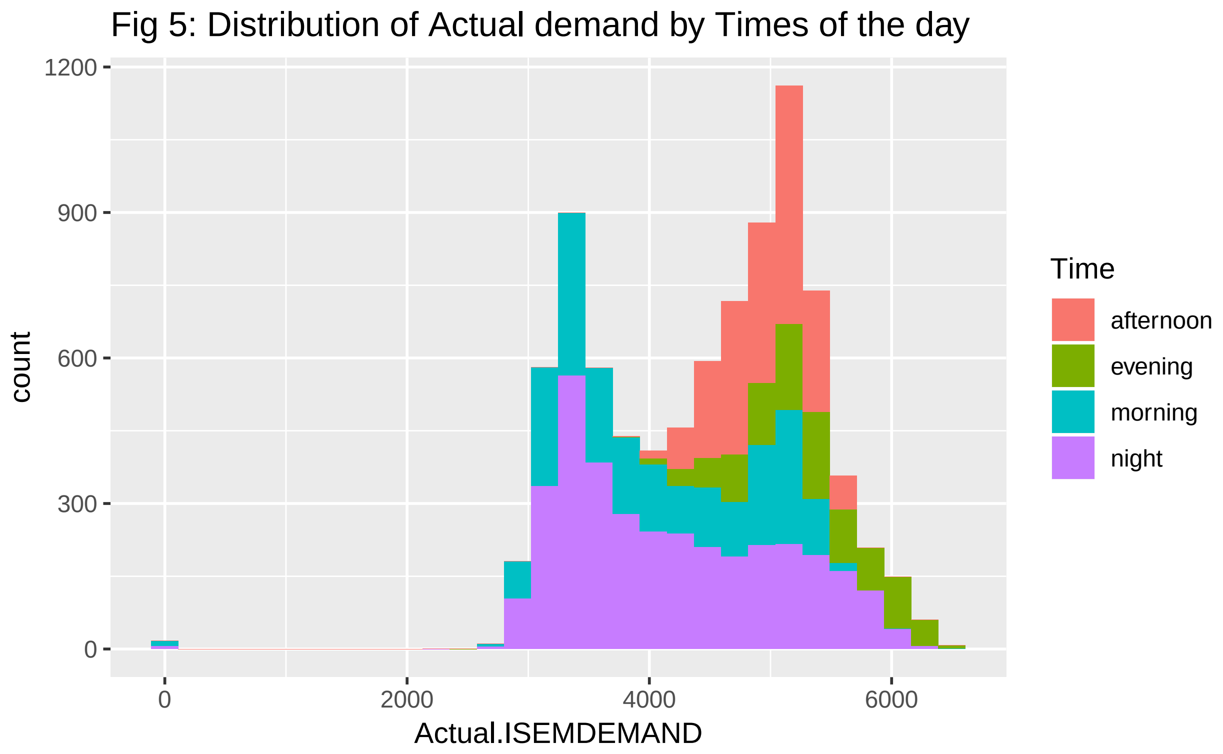
Fig 3: Distribution of Actual Demand by Month



The wind during December and February is also high which is interpreted from Fig 4, eventhough the electricity unit produced does not exceed 4000MW, the actual demand resides between 3000MW to 6500MW, which means that there is a high chance for buying the electricity from the ISEM market.

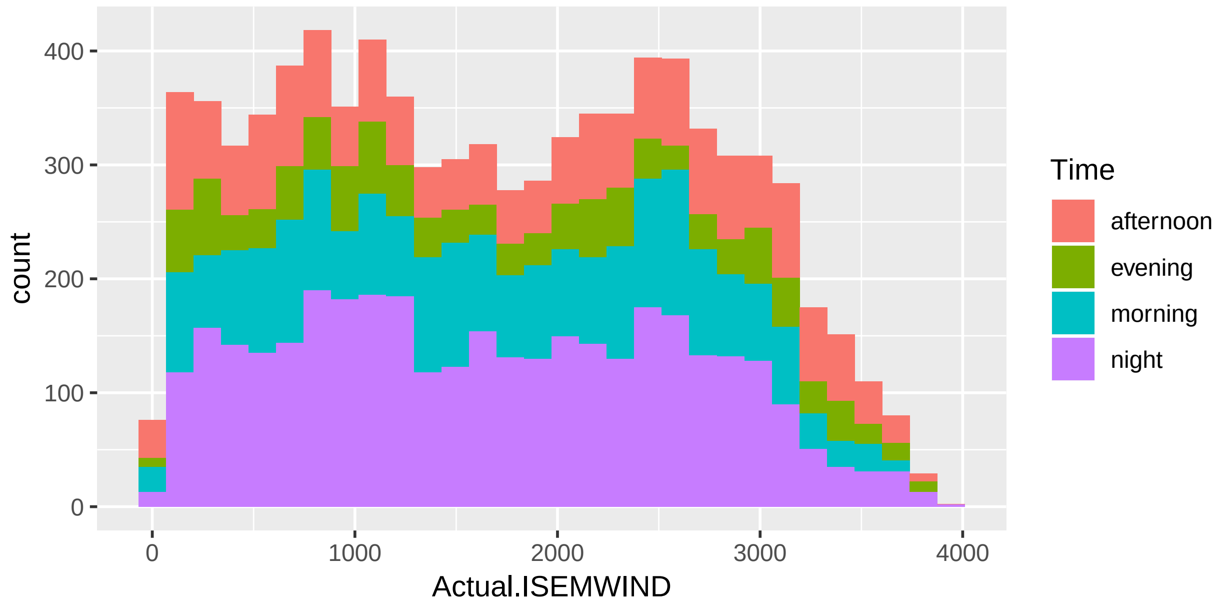


The time in hours and minutes, was then classified to 'Times of the day'. The plot below (Fig 5), shows that the demand in the afternoon is high compared to rest of the day time. The bars in blue which represent morning demand, seems to have high count, but of low electricity unit. But the graph places the demand of evening in the second as the power level is high even though the count is less. The afternoon rise must be due to the running of all the industries and other work environment technologies. Since major proportion of people rest during night, the demand is very low for electricity at nights.



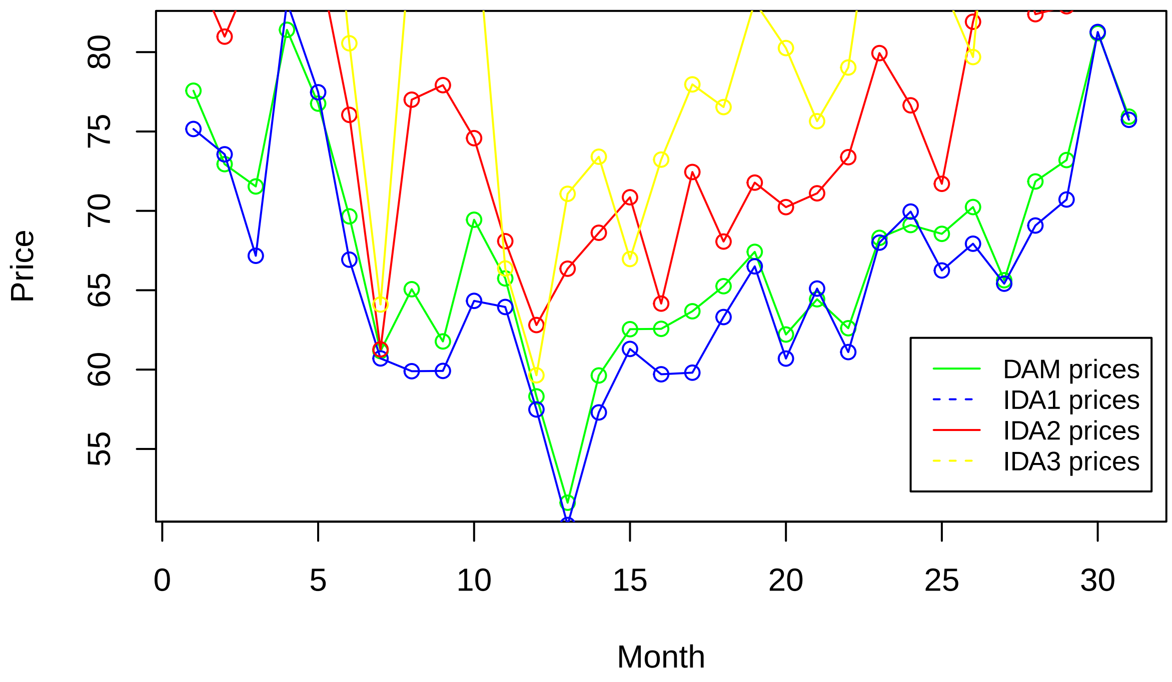
The production of energy by the generators using the wind is also very high during the afternoons. The nights are calmer and the count of electricity is not very high. The mornings have high count but have lower energy generation compared to evening. (Fig 6)

Fig 6: Distribution of Actual wind by Times of the day



Based on the prices of every market, lets find out which is more suitable for buying using the visualization technique. In the below graph (Fig 7), it clearly indicates that IDA1 is more optimal than DAM, IDA2, IDA3 prices.

Fig 7: Months vs Prices for all ISEM market



Modelling

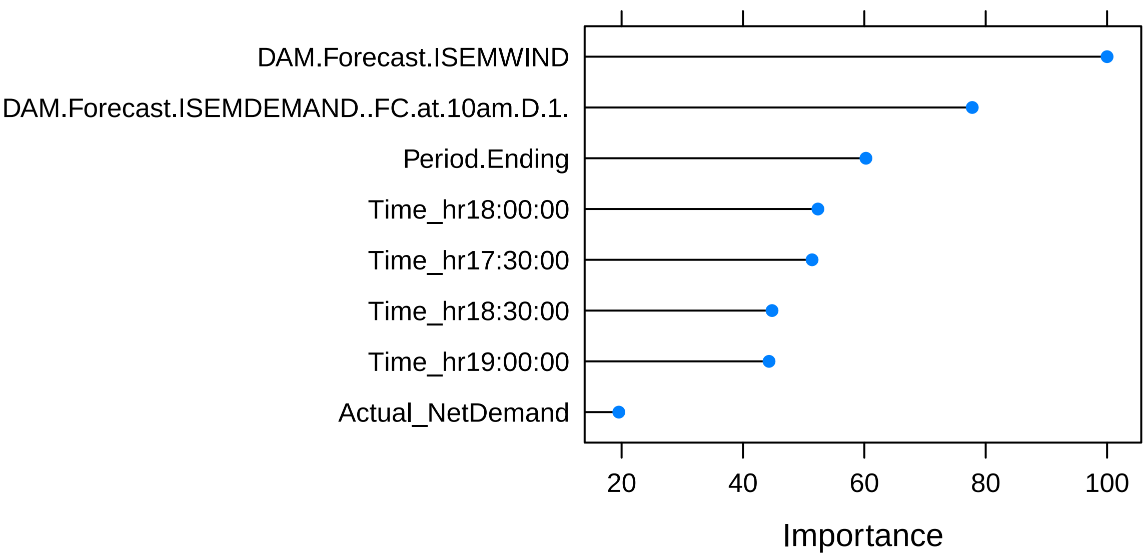
In CARET package of R, all models are trained using the `train()` function, while the `predict()` function is used for making predictions. The `trainControl()` function is used to create a set of configuration options known as a control object, which guides the train() function. These options allow for the management of model evaluation criteria such as the resampling strategy and the measure used for choosing the best model. Having the better understanding of the data by EDA and tidying the data for each market, the process of fitting a model takes place.

DAM market

Linear model for DAM:

The data is spilt into train and test data in which the train dataset consist of first 5 months(October,2018 to February,2019) data and test has the information of March, 2019 for prediction. Since this is a regression problem, Linear regression model is being fit to the train data. We note the R-squared values which is a statistical measure of how close the data are to the fitted regression line. The R-squared value for Linear model for DAM is 0.6357 and the important variables that influence this model are (see Fig 8):

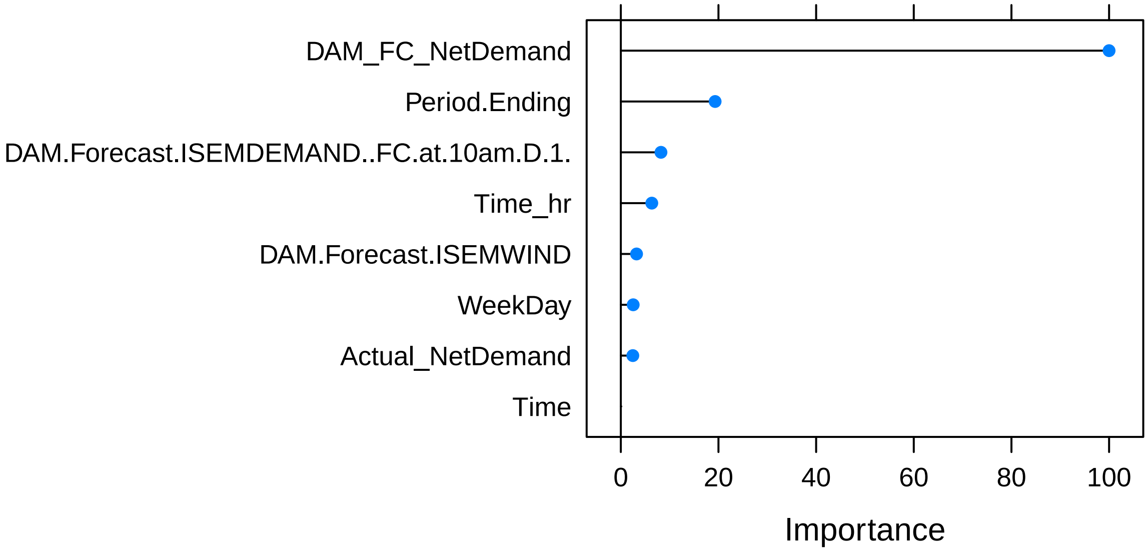
Fig 8: Important Variable of LM for DAM



Random Forest for DAM

Since the R-squared value of Linear model was not satisfactory, another regression model called Random forest is introduced on the training set. The R-squared value for this model is 0.821864 and the important variables that influence this model are (see Fig 9):

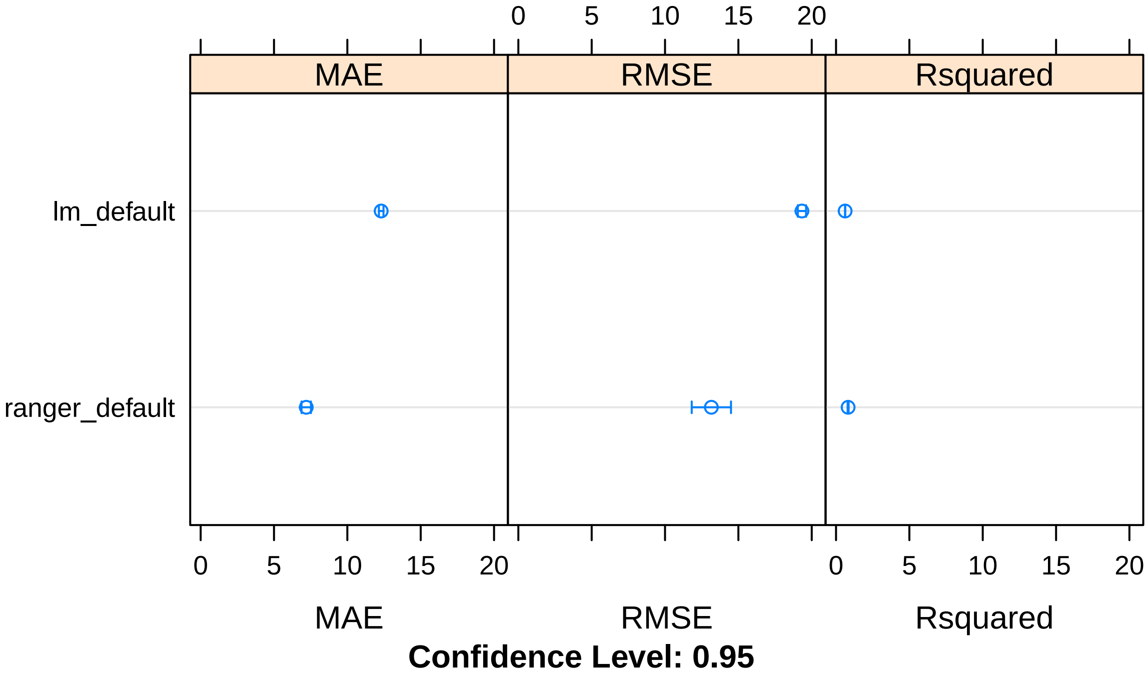
Fig 9: Important Variable of Random Forest for DAM



Comparision for DAM

Using the resampling method to compare the two models, the below dotplot (Fig 9) describes that the Random forest have better accuracy than the Linear model.

Fig 10 Dot plot: Resample for DAM

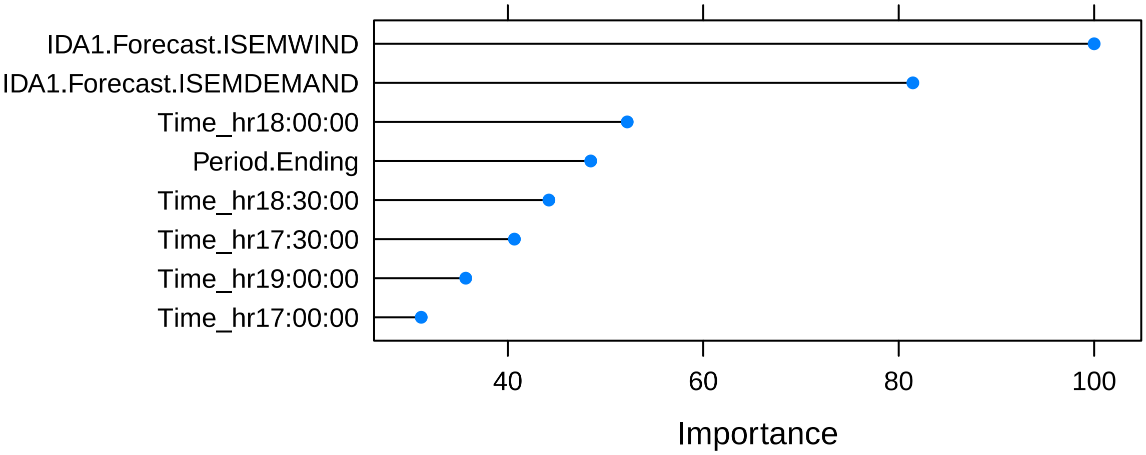


IDA1

Linear model for IDA1:

The IDA1 data set is spilt into train and test data in which the train dataset consist of first 5 months(October,2018 to February,2019) data and test has the data of March, 2019 for prediction. Linear regression models are being fit to the train data. We consider the R-squared values which is a statistical measure of how close the data are to the fitted regression line. The R-squared value for Linear model for IDA1 is 0.6108 and the important variables that influence this model are (see Fig 11):

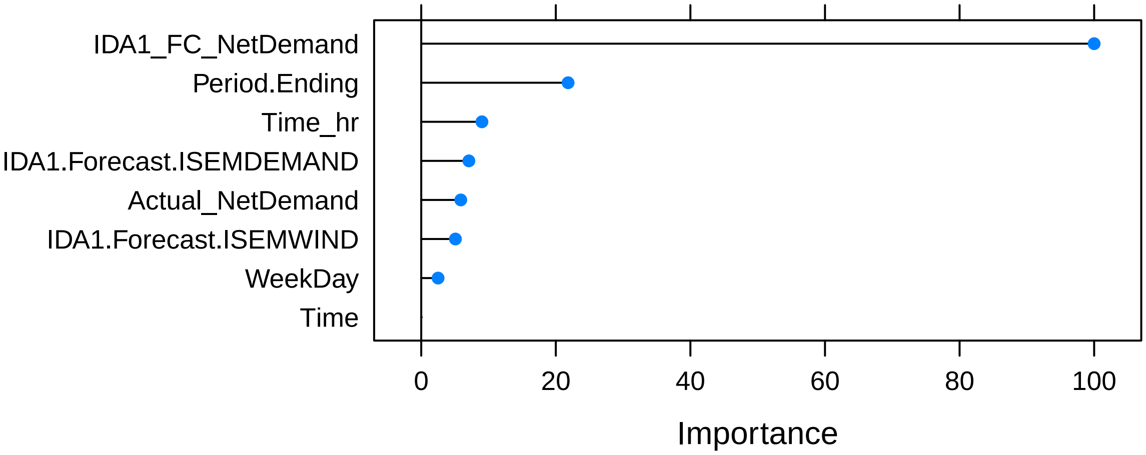
Fig 11: Important Variable of LM for IDA1



Random Forest for IDA1

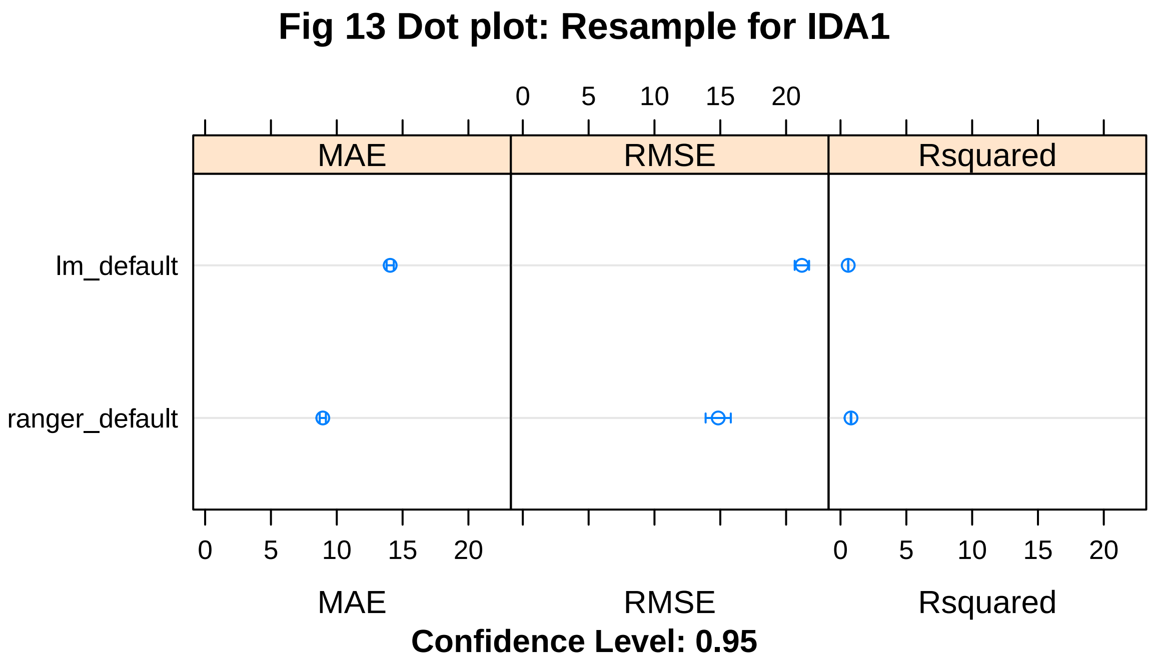
Since the R-squared value of Linear model was nearly half, Random forest is introduced on the training set to solve this regression problem. The R-squared value for this model is 0.7930251 and the important variables that influence this model are (see Fig 12):

Fig 12: Important Variable of Random Forest for IDA1



Comparision for IDA1

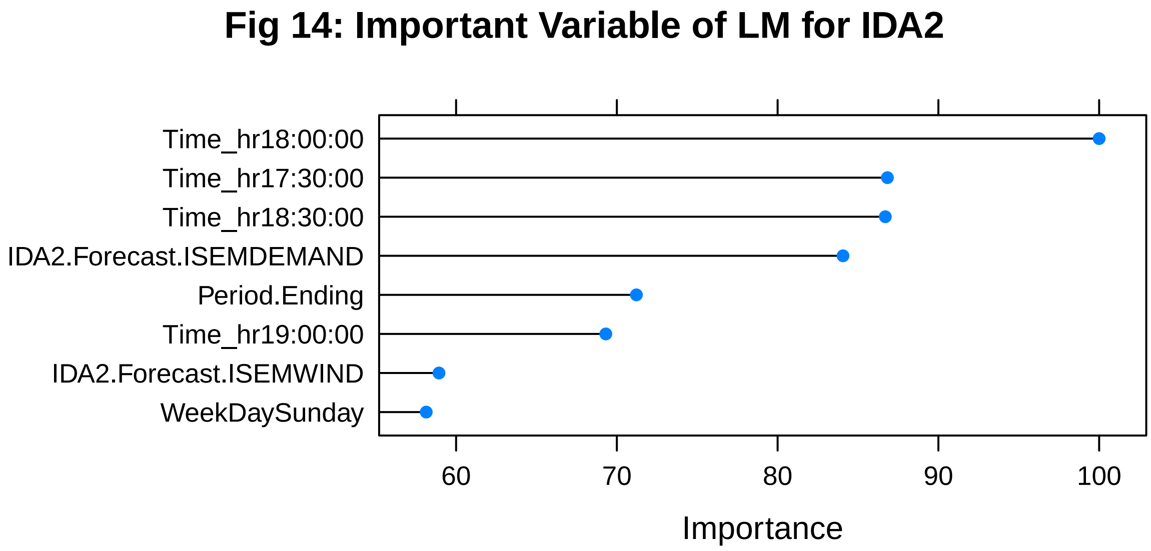
Using the resampling method to compare the two models applied on IDA1, the below dotplot (Fig 13) describes that the Random forest have better accuracy than the Linear model.



IDA2

Linear model for IDA2:

The train and test dataset of IDA2 consisted of first 5 months(October,2018 to February,2019) entries and the data of March, 2019 respectively. Linear regression models are being fit to the train. The R-squared value for Linear model for IDA2 is 0.5529. The variables that influence this model are as shown in the below plot (see Fig 14):



Random Forest for IDA2

The R-squared value for Random forest is 0.7210797 and is more accurate than the above Linear model. The important variables that impact this model are as shown below (see Fig 15):

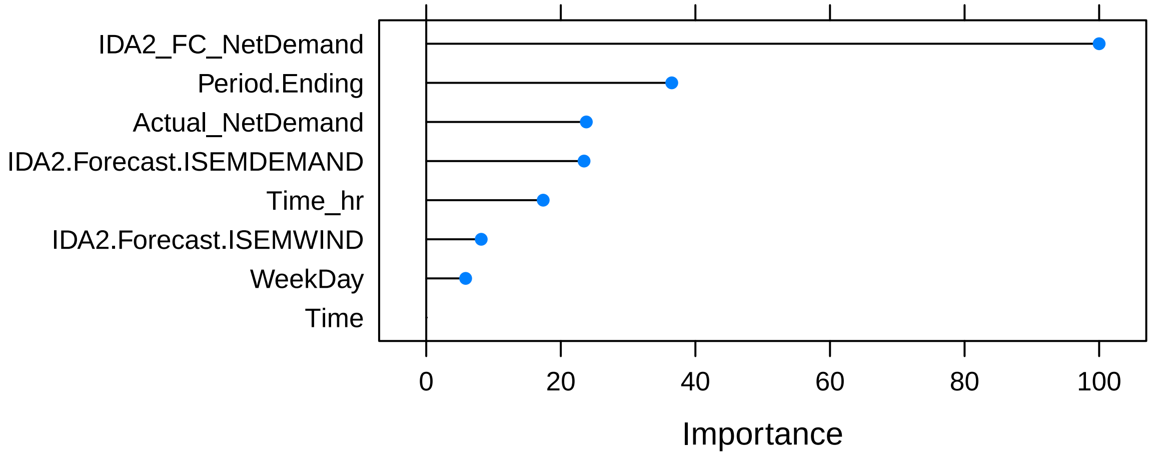


Fig 15: Important Variable of Random Forest for IDA2

Comparision for IDA2

The below dot plot(Fig 16), shows that the RMSE for linear model is highier compared to random forest, which clearly indicates that the random forest provides more accurate predictions compared to Linear model of IDA2

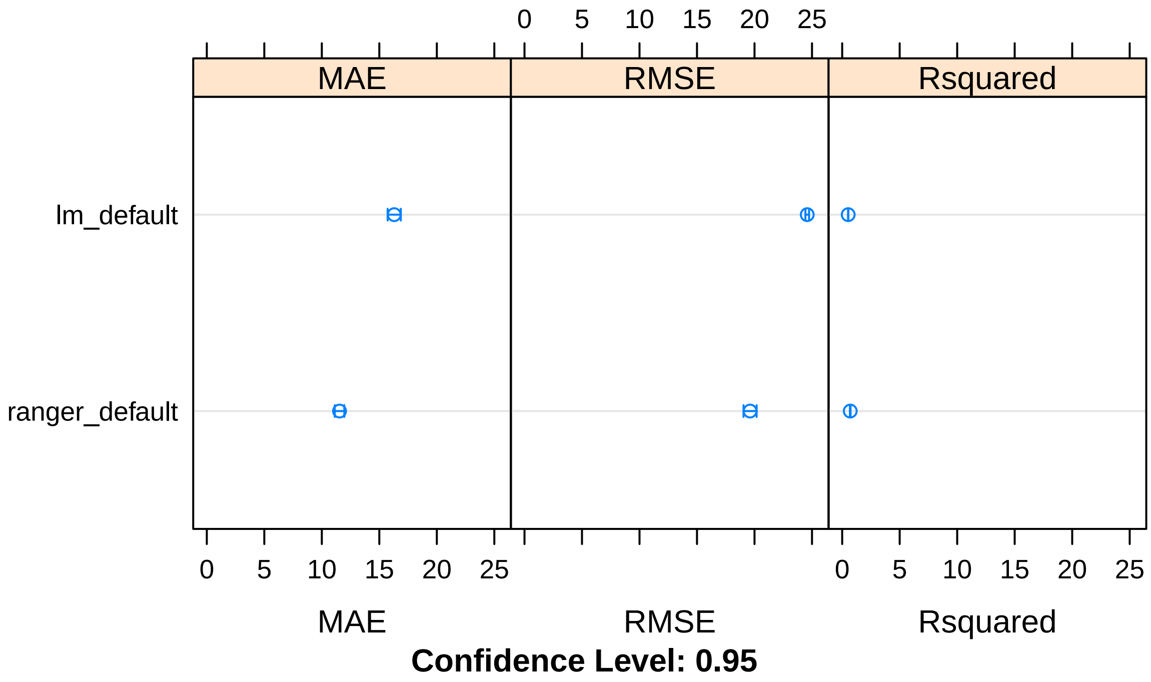


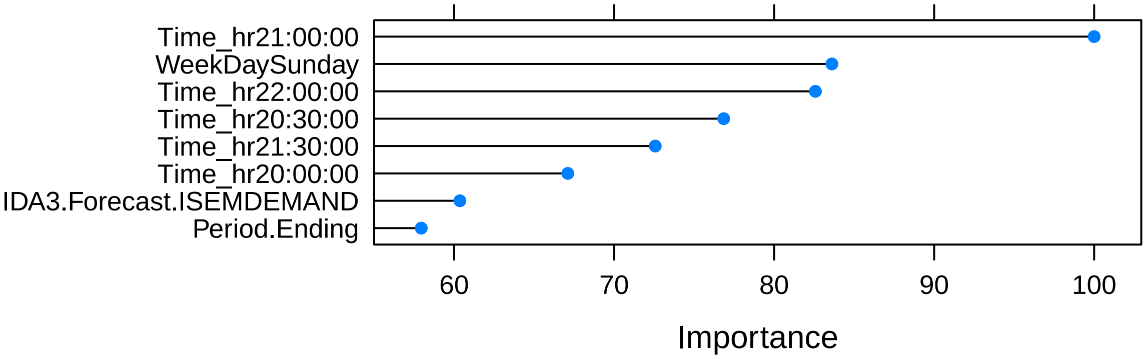
Fig 16 Dot plot: Resample for IDA2

IDA3

Linear model for IDA3:

Similar to the previous data partitions, the first 5 months(October,2018 to February,2019) entries from the IDA3 dataset are placed in the train dataset, leaving the last month data whihc is March, 2019 is copies to the test dataset. A Linear regression model is being fit to the train dataset which includes the 5months data and the model is being trained. The obtained R-squared value for this Linear model is 0.5388. The variables that influence this model are as shown in the below plot (see Fig 17):

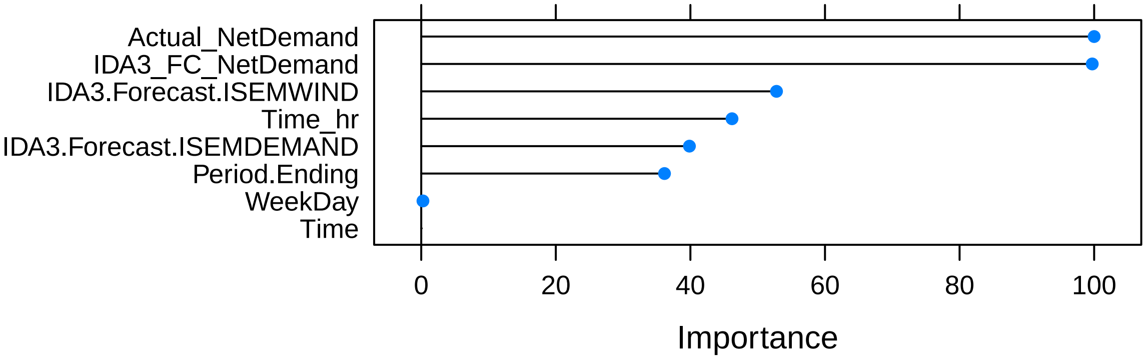
Fig 17: Important Variable of LM for IDA3



Random Forest for IDA3

From above Linear model for IDA3, the obtained R-squared is comparitively low when compared to the value for Random forest model which reads 0.7210797. Thus, Random forest shows more accurate predictions compared to the Linear model of IDA3. The variables that play a vital role in this model for prediction are as shown below (see Fig 18):

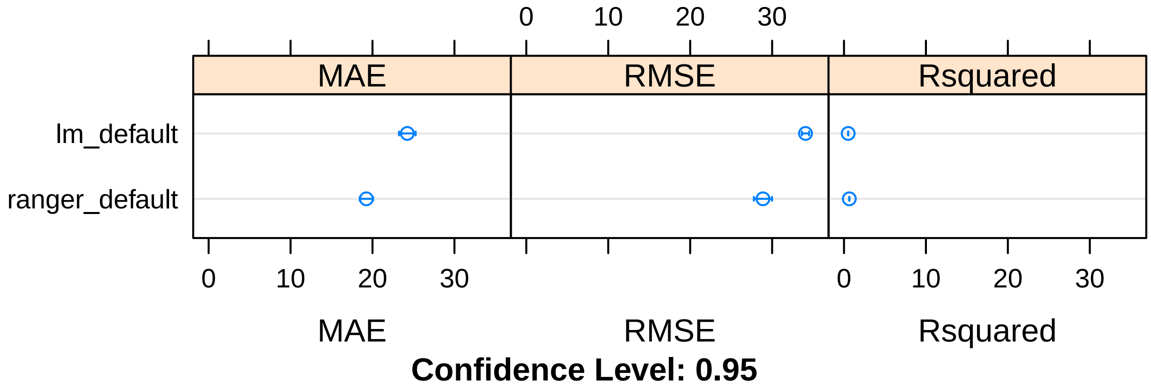
Fig 18: Important Variable of Random Forest for IDA3



Comparision for IDA3

On comparing the linear model and random forest by using resamples function in r, the RMSE value which should be as low as possible, seems to be in favour of Random forest, as shown in Fig 19, proving that Random forest gives more accurate predictions for the given data.

Fig 19 Dot plot: Resample for IDA3



Prediction using Test data:

Now that the best model has been decided, the test data is utilized as prediction data and given as input to the random forest model for every market. The output for all the market resulted in giving a promising result. This is graphically illustrated as shown in the below line graphs, where the actual output is represented by black line and the predicted output is shown in red line.

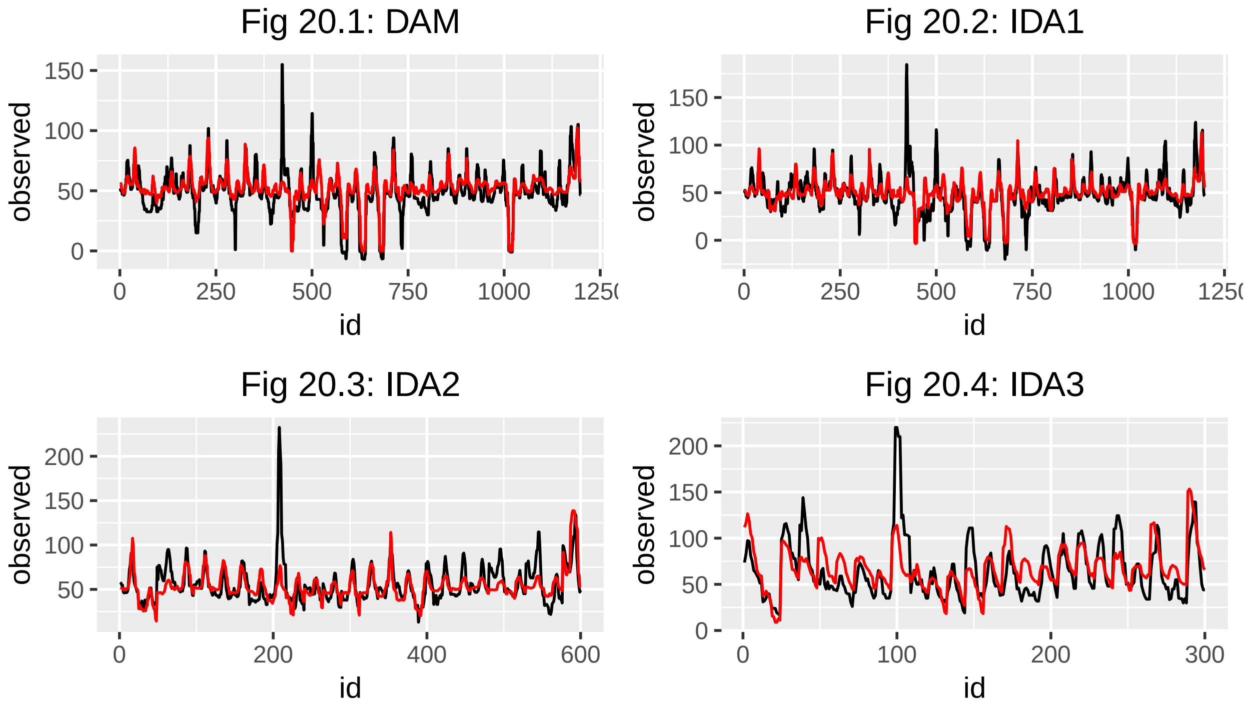


Fig 20.1: DAM

Fig 20.2: IDA1

Fig 20.3: IDA2

Fig 20.4: IDA3

Fig 20: Oberserved vs Predicted graph

Results and Findings

Using the Random forest model, the values, have been predicted for the month of March,2019. Now, lets see which price and market are better for buying the electricity optimally (see table 8). The above table shows

Table 4: Best market and price predicted for March, 2019

Period.Ending	DAM.Price	IDA1.Price	IDA2.Price	IDA3.Price	BestMarket	BestMarketPrice
2019-03-01 19:30:00	80.76665	91.64510	107.59680	104.69736	DAM.Price	80.76665
2019-03-01 20:00:00	79.34429	84.69227	97.84890	100.37107	DAM.Price	79.34429
2019-03-01 20:30:00	58.86669	61.40430	69.73666	85.63456	DAM.Price	58.86669
2019-03-01 21:00:00	58.68973	59.75264	66.05077	78.80936	DAM.Price	58.68973
2019-03-01 21:30:00	58.83024	53.48662	52.30540	66.11783	IDA2.Price	52.30540

the Best Market for every 30mins time slot, which Energia can bid to get the best price. On comaparing with the oberserved data from Energia for the month of March 2019, plotting the graph for a day (2019-03-03 10:00:00 to 2019-03-04 10:00:00), shows the optimal price predicted.

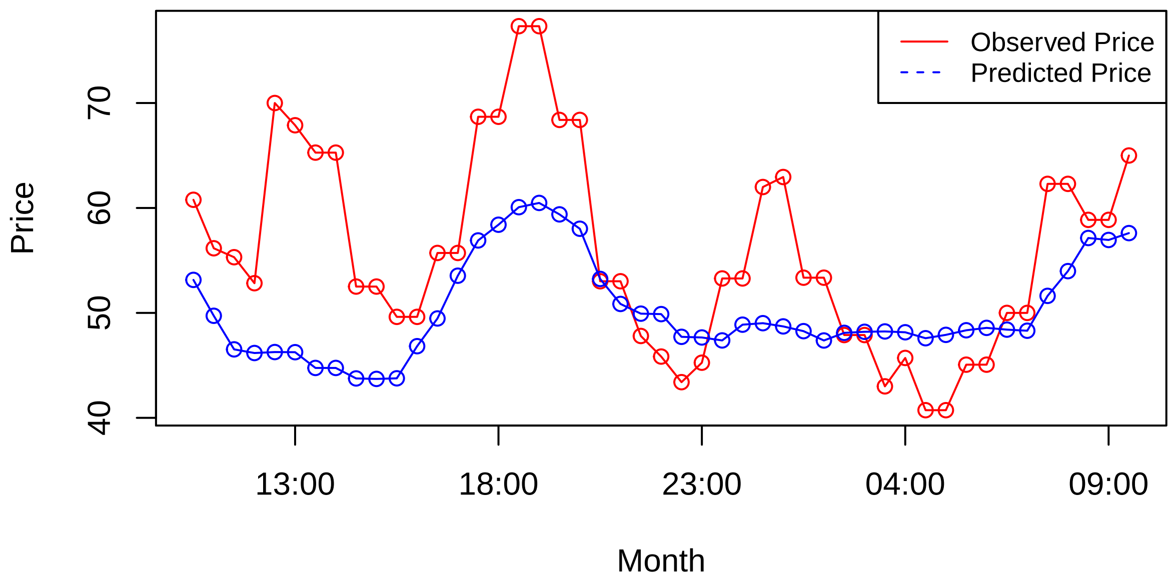


Fig 21: Observed price Vs Predicted price for a day

Conclusion

As conclusion, responding to the problem statement which Energia is facing: 1) Assume Energia have 100MW of electricity to buy in each half hour of the day. How can Energia optimise the purchases in ISEM? What factors drive this decision? *Solution:Energia can use the above model to predict the best market for every 30mins interval and optimise their purchase. The above graph (Fig 21) represents that Predicted price(in blue) is more cheaper than the forecasted price(in red) is the factor which drives this decision of choosing the above predicted best market. .*
2) Can Energia increase profitability by speculating across the markets by trading more actively? *Solution: Energia can increase profit by choosing the market based on the above prediction market and buy electricity for more less price as show in Fig 21.*