Air Liquide : Future Ready Data Challenge

Gayan Bandara : Monash University, Australia

July 3, 2018

Problem Statement

Predicting monthly energy consumption demand of Air Liquide customers using explanatory variables such as *Market domain, Time stamp, Type of Gas, Sum of Sales.*

My Approach

I used time series forecasting approach along with machine learning techniques to predict the corresponding monthly energy consumption demand of each customer.

Why Time series Forecasting?

Predicting energy consumption demand of a customer is a highly uncertain use case. Therefore providing only the point predictions (only one value) is not adequate as there is no way of telling how accurate the predictions/forecasts are. This is mainly because in the energy consumption sector, customer satisfaction is highly important (You don't want your customers, experiencing shortage of energy). These shortages can be occurred if the predictions are point forecasts (You need to know the upper/lower bounds with the prediction intervals and the amount of buffer you need to preallocate for uncertainties).

As a result, providing prediction intervals along with the point forecasts are imperative in this domain. However, traditional state-of-the –art machine learning algorithms are constrained at providing such prediction intervals (i.e Gradient boosting, Deep learning, SVM). On the other hand, statistical time series forecasting are rich at capturing the uncertainty. Providing the prediction intervals helps us to clear how much uncertainty is associated with each forecast/prediction.

Figure 1 shows such forecasting plot generated for Customer ID 1071. In this plot, you can see the 95% prediction interval (Grey color) calculated on top of the point forecasts (Blue color). Therefore, incorporating prediction intervals for prediction (energy forecasting in particular) is quite useful [1][2]

You can access all the forecasting plots generated by each forecasting algorithm (ETS, ARIMA, and ES namely), including the prediction intervals. As submission files only requested the point predictions, i only submitted the corresponding point forecasts.

Challenges for Time series forecasting

During the initial explanatory data analysis, i discovered the given dataset is not tailor made for time series forecasting, as there were incomplete records of each customer (i.e Not every customer had records from 2015 Nov- 2016 May). This is not a limitation of the dataset. But to make it more ideal for time series forecasting (before applying forecasting algorithms) . Therefore to impute these missing records, i used lasso regression, which accounts for the exogenous variables (Market domain, Time stamp, Type of Gas, Sum of Sales) available in the dataset.

After imputing the incomplete energy consumption records of each customer, i used state-of-art forecasting techniques to forecasts the expected energy demand.

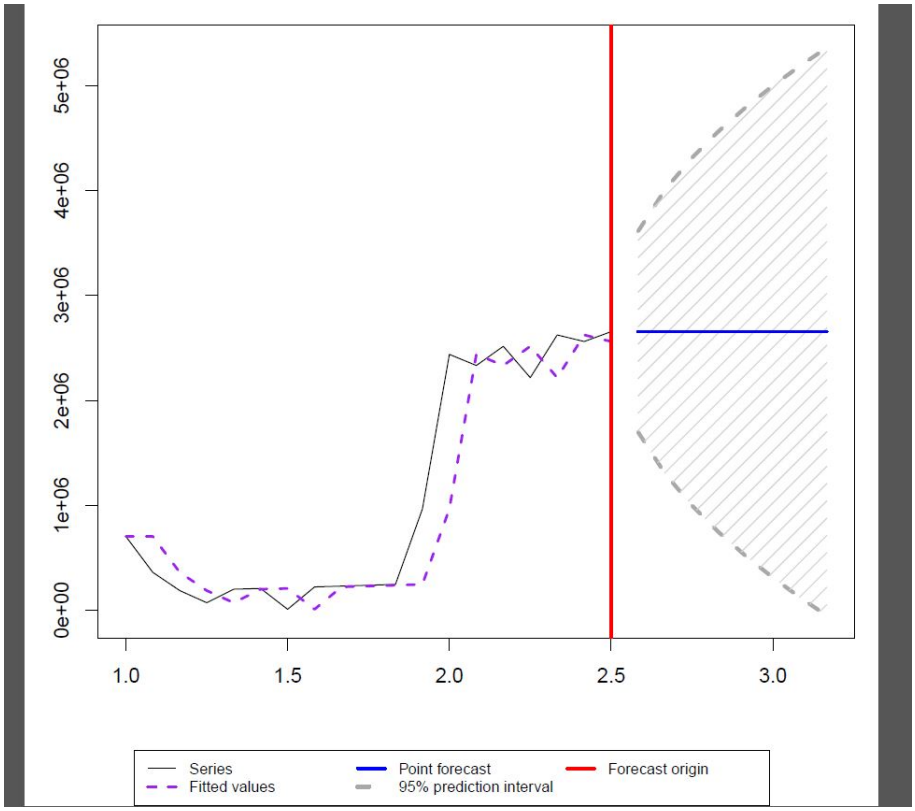

Figure 1: Prediction Intervals for Customer ID 1071

Tools and Packages

I used RStudio/ R Language as my primary scripting language. For imputation purposes i used lasso regression imputation technique, which is available as a R Package (**imputeR**) [3]. For the forecasting purposes, i used the ETS (), ARIMA () algorithms, which are available from the (**forecast**) R Package [4] and ES() algorithm available from the (**smooth**) R Package [5]. All these algorithms are considered as state-of-art techniques in the forecasting domain (ES() algorithm is developed to mitigate the limitations of traditonal ETS() algorithm)

Overall Procedure

Below is the overall summary of my approach, Please look into the ReadMe file and the corresponding script file for the detail implementations.

1. Convert *train dataset IM challenge.csv* to *train-dataset.csv* : This replaces the missing values values of Zip codes by "0" (so we can consider as another category as there are customer IDs with "0" zip codes)Also removes the record of 1143, as the energy consumption is zero (doesn't add anything for learning), also this customer isn't in the test set.
2. Adding a new feature (combined) using existing timestamp feature.
3. Generate the incomplete records of each customer.
4. Feature normalization and factorization
5. Imputing incomplete (N/A) values of Energy consumption using lasso regression.
6. Prepares the raw test.csv for the forecasting (Facilitate to match with the appropriate forecast in the forecasting scripts)
7. Forecasting using ETS() function from forecast() package.

8. Forecasting using Auto.ARIMA() function from forecast() package.
9. Forecasting using ES() function from Smooth() package.
10. Combining the forecasts using weighting (ensembling)

Acknowledgement

I bestow my thank to Air Liquide for the efforts of organizing such a successful data challenge event and providing us (students) a common platform practically implement and evaluate our knowledge.

Personally for me, this was a great challenge. Hope to see more challenges from your company in future.

References

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5565406/
[2] https://ieeexplore.ieee.org/document/7325539/
[3] https://cran.r-project.org/web/packages/imputeR/imputeR.pdf
[4] https://cran.r-project.org/web/packages/forecast/forecast.pdf
[5] https://cran.r-project.org/web/packages/smooth/smooth.pdf