# REPORT

Ananya Palit

Student ID: 112077303

## TECHNOLOGIES USED:

I have used Python 3 and standard Python libraries like pandas (0.23) and scikit-learn (0.20) on jupyter notebook for this assignment (attached herewith).

## PREDICTION MODELS USED:

For this analysis, I have tried 4 different models, namely **Linear Regression, k-Nearest Neighbors Regression, Random Forest** and **Seasonal ARIMA(X)**.

- For Linear Regression, no special hyper-parameter was set.
- For k-Nearest Neighbors Regression, after tuning *no. of neighbors = 90* was found to give good results uniformly across all homes.
- For Random Forest, after tuning, *maximum depth of 20* was found appropriate.
- For Seasonal ARIMA, a grid-search algorithm was used to find the best combination of the p, d and q values. Then using the best values, predictions were made and evaluated.

## FEATURES FOR BASIC MODELS:

I have used the following sets of features for the basic models - Linear Regression, kNN Regression and Random Forest:

1) **Timeslots** – the Month of the year and Day of the month can capture information related to regularities in Weekly or Monthly consumption patterns. Likewise, hour of the day (and minute) can capture hourly regularities.
2) **Seasons** – Seasonal regularities cannot be captured in this dataset as only 1 year of data is available. Hence season-periodic patterns could not be captured. However seasonal variations may be. Hence I defined 'Seasons' as a feature – segregating the year into 4 seasonal periods: Fall (August-November), Winter (December-February), Spring (March-April) and Summer (May-July).
   Initially, I had tried 'Seasons' as a single parameter with each season represented by an id. But individual season-wise Boolean parameters proved more effective after training Linear Regression.
3) **Peak hours and Weekends** –
   a) Peak hours by definition have much higher consumption than other hours of the day. For each house, by plotting the data, I arrived at an optimum choice for peak-hours. Usually these were hours from morning to noon and hours from late-evening to night.
   b) Similarly, weekends have marked different consumption patterns.

## DATA PREPARATION FOR TRAINING AND EVALUATION:

First I loaded the data of all homes iteratively in a pandas DataFrame and performed initial diagnostics. The data was pretty clean and easy to parse. Hence, no data munging was needed in this assignment. Then, I plotted the data of each home, for one day, one week and the whole year to see trends. These plots are what gave me the ideas for what features I had to take.

Finally, as required by the assignment – I took an input of the particular timeslot from when we needed to predict for the next 96 slots. Then I created the Training-set as the whole history of data up to before this given timeslot, and the Cross-validation set as the data for the next 96 slots.

## PREDICTION ACCURACY

All the models were trained and cross-validated against a common random timeslot input – 2015-08-27 10:45:00. The following are the MAE errors of the different models for each home:

| | Linear Regression | k-NN Regression | Random Forest | Seasonal ARIMA |
|---|---|---|---|---|
| **Home 1** | 0.560928302 | 0.675412505 | 1.764305999 | 0.588551564 |
| **Home 2** | 0.677046726 | 0.672919518 | 0.876570942 | 0.654372660 |
| **Home 3** | 0.344733642 | 0.403340094 | 0.398481884 | 0.330747870 |
| **Home 4** | 0.618204081 | 0.692041276 | 0.686782515 | 0.604109599 |
| **Home 5** | 0.593880894 | 0.646595850 | 0.626689675 | 0.589354388 |
| **Home 6** | 0.378587093 | 0.526729013 | 0.204662532 | 0.144483044 |
| **Home 7** | 0.878811279 | 0.832264391 | 0.689789689 | 0.652135632 |
| **Home 8** | 0.562650402 | 0.563775234 | 0.574566950 | 0.654310762 |
| **Home 9** | 0.380513413 | 0.334748503 | 0.400984569 | 0.414796115 |
| **Home 10** | 0.301047690 | 0.363503397 | 0.378729999 | 0.388920689 |
| **Mean** | **0.529640352** | **0.567304214** | **0.660156475** | **0.502178232** |

## DIAGNOSTICS AND TRENDS

For diagnostics, I have mainly relied on plotting the prediction data from the trained models and comparing them with the actual values. Mostly the trend was **High-bias**. This was particularly high for some homes with abrupt irregular peaks as was visible from the plots. Naturally, this could not be avoided in general. However, all the models (except Linear Regression) usually mimicked the pattern of the consumption pretty well, with some lag.

## HYPER-PARAMETER TUNING:

As mentioned above in the Models section, several combinations were manually experimented for the kNN and Random-Forest Regression models and the optimum value was retained across all homes. For Seasonal ARIMA, these values were automatically imputed from grid-search.

## CONCLUSION

- In comparison to basic models, the Seasonal ARIMA seemed to generalize best.
- Most models suffer from High-bias due to unpredictable surges in consumption.

## CREDITS:

In this assignment, I have found help in several websites including the following:

1) https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b
2) https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
   https://machinelearningmastery.com/grid-search-arima-hyperparameters-with-python/
3) https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/