**Week 3 - EDA and Modelling**

**Meeting notes 28 March 7-8pm**

\_\_\_\_\_\_  
Questions for Sonit

* Cleaned the data to remove null values after merge.
* Began EDA but the plotting takes too long, any tips?
* Vector Auto Regression Model is what we have decided to go with as we have explanatory variables. Is that the correct choice?
* Is there anything we need to watch out for when using VAR? (ARIMA, SARIMA, … Moving Average)
* Our plan:
  + Check for stationarity, otherwise convert to stationary data
  + Partial Auto correlation to find out how many lags we should use
  + Find correlation between the 2 variables, if P value is small then we pick that lag
  + Fit VAR model

\_\_\_\_\_\_\_\_

Notes:

* Sonit to share link to data source – Bureau of Meteorology
* Energy Load data
* App is also available for Energy load–
* Sonit suggests population projects – get how it will scale up, see if you have linear trend, for the next 20-30 years, this is how population will increase
* Temperature data projections are going to be considered (Kiara)
* Plug in the temperature data into the equation
* Build a single model that only takes into account the temperature data for the last 11 years
* Search photovoltaic solar energy
* Sonit to share the paper
* If you have a plot, please leave commentary underneath. What insights are we deriving from there?
* Do some error analysis. Why is one model doing worse than other?
* Given more expertise, how we can further improve the work?
* Whether it is sensible model, LSTM might be good for timeseries forecasting, GRUs,
* Econometric models could be option – Modelling from economics view point..
* Check the underlying assumptions of the modelling package we use and then see if our data and situations fits that model’s assumptions. Way to justify why we picked this model.
* Start EDA and finish that within 2 days.
* Start modelling.
* Title of project should be specific enough to show exactly what we are working on.
* Due to time constraints, use the data we have now. Collect data in parallel because any new data will need to be cleaned and used.
* Even if we use y=mx + c
* Data leakage is a problem in timeseries, so you cannot shuffle data. You just have to pick maybe 8 years for training, 3 years for validation.

Start taking 1 variable and get a model from there.

Ingest the data, fit the model, get R2, RMSE, RAMSE,

Create columns for month, days, week days, years, ….. We can then plot that. Use pandas. Focus on 1 year at a time, pick a year maybe.. i.e. shorter time span. Overlay 2 graphs together.

EDA – story telling

See why demand is high is some spots

Show correlation plot

Hourly energy consumption heatmap

Feature importance (correlation)

There should be justification for why we picked a particular model.