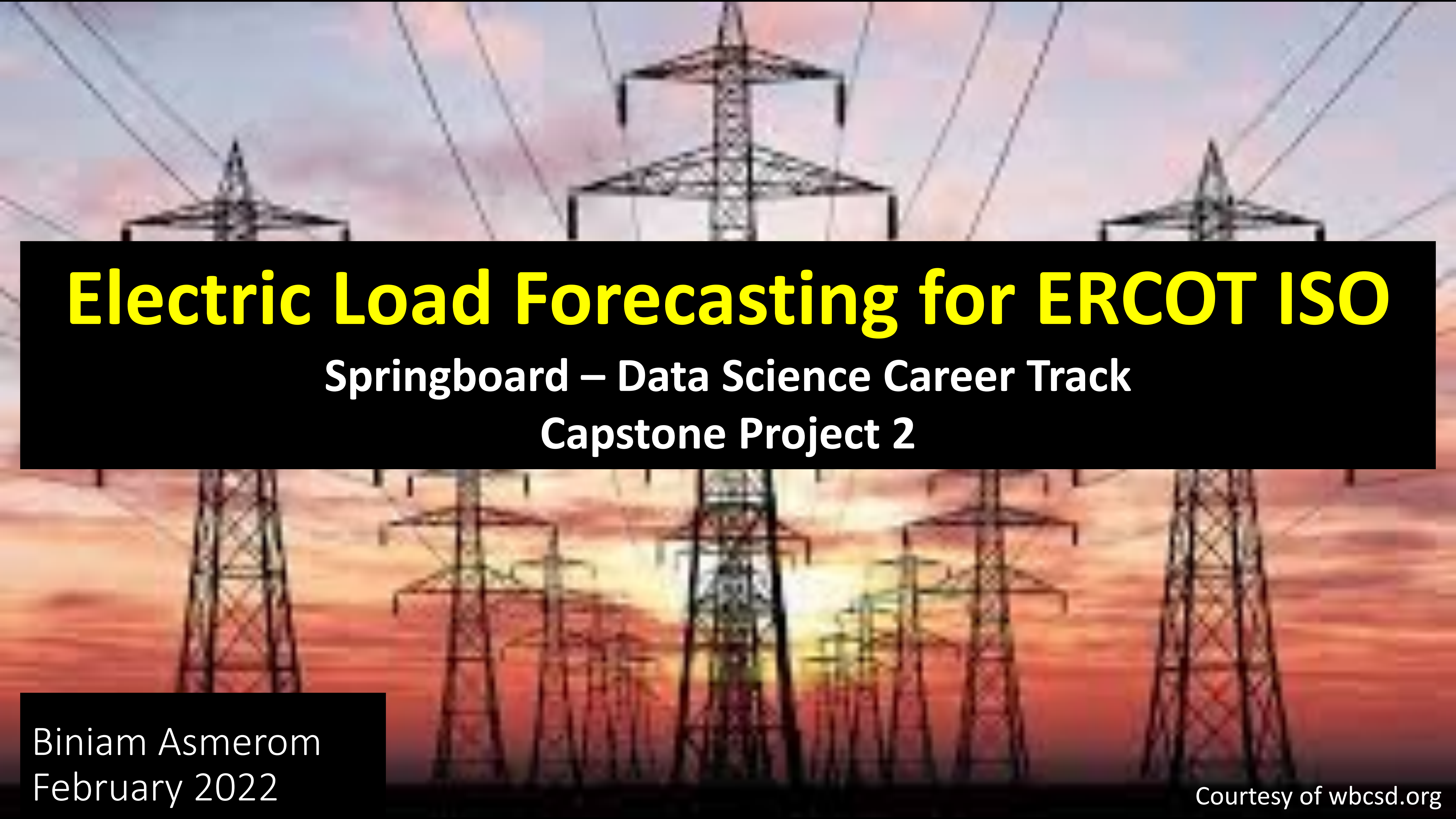
Electric Load Forecasting for ERCOT ISO

Springboard – Data Science Career Track

Capstone Project 2



Biniam Asmerom
February 2022

Courtesy of wbcsd.org

Business Problem

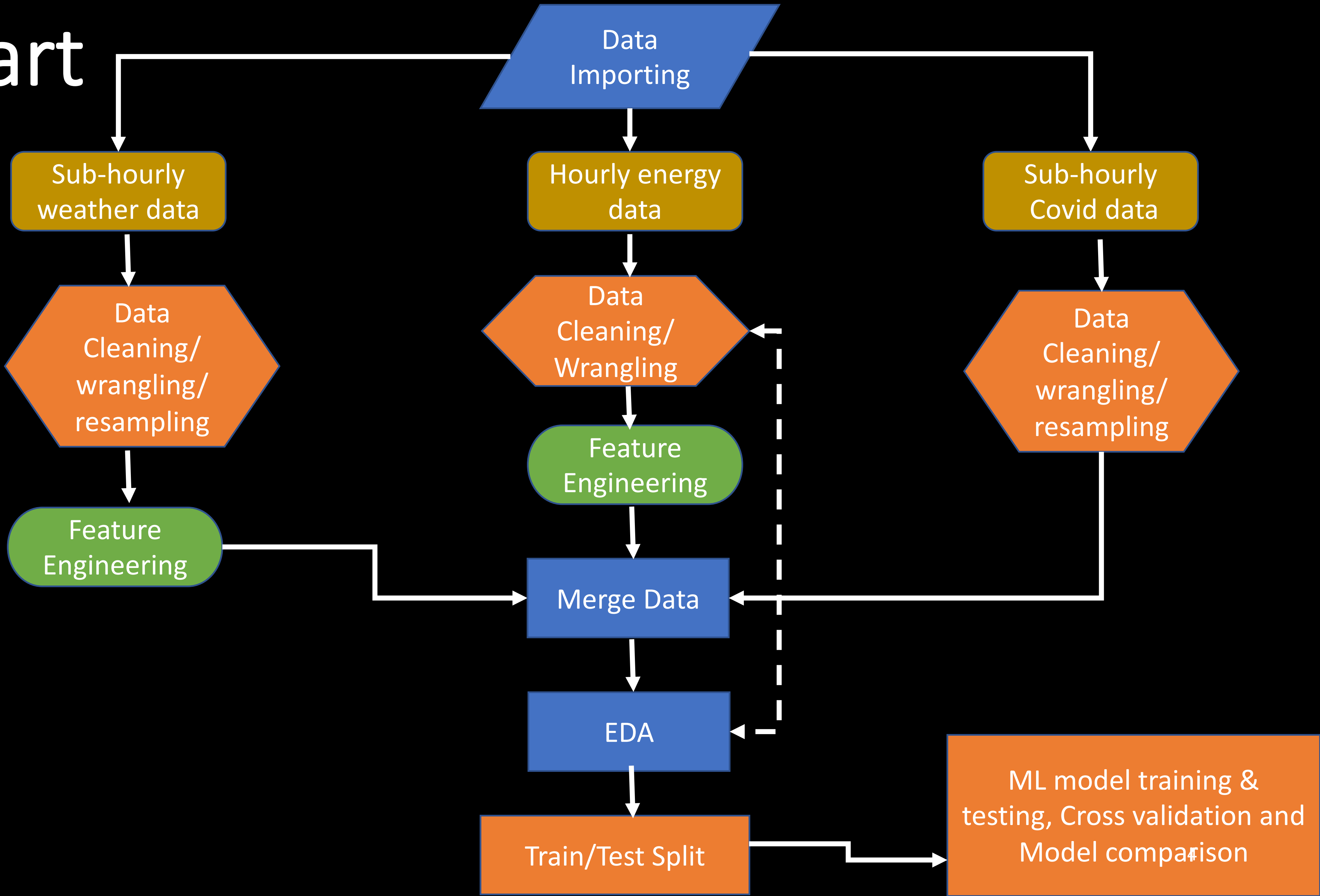
- Electricity can not be stored in large quantity, so demand and supply should be matched

- Utility companies need an accurate load forecasting methods to enable them to:
  - effectively plan their energy generation operations and balance the demand with appropriate supply.
  - Improve their day-to-day operations
  - Meet their customers' energy demand, and
  - Avoid grid failures or wastage of energy and costs of under or over cutting.

Project Objective

- Build a robust and accurate short-term forecasting model that predict hourly electric load for ERCOT ISO in Texas

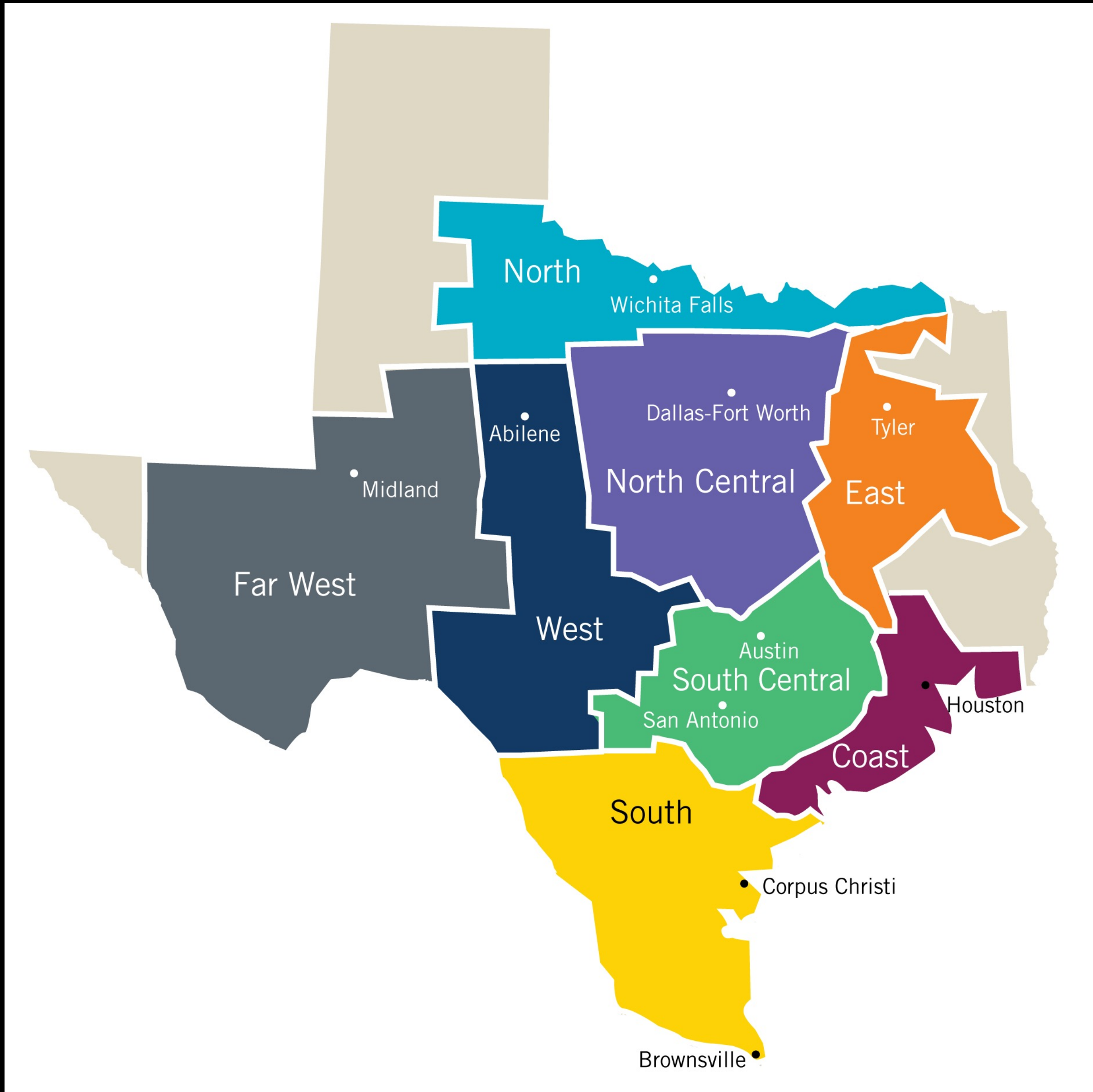
Flowchart



Data Importing
Sub-hourly weather data
Hourly energy data
Sub-hourly Covid data
Data Cleaning/ wrangling/ resampling
Data Cleaning/ Wrangling
Data Cleaning/ wrangling/ resampling
Feature Engineering
Feature Engineering
Merge Data
EDA
Train/Test Split
ML model training & testing, Cross validation and Model comparison

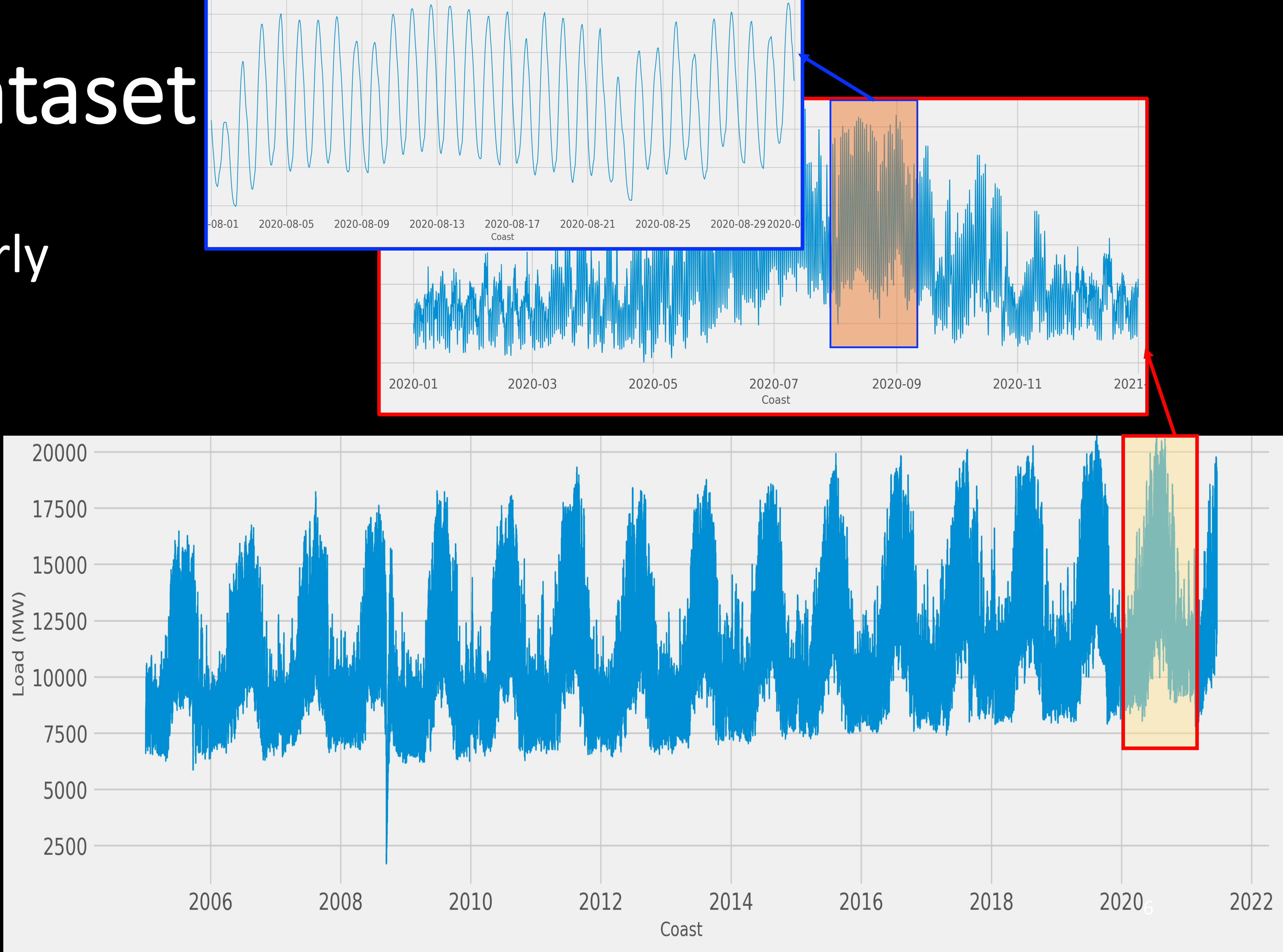
Data

- Three different datasets:
  - Energy load data
  - Weather data and
  - Covid-19 data



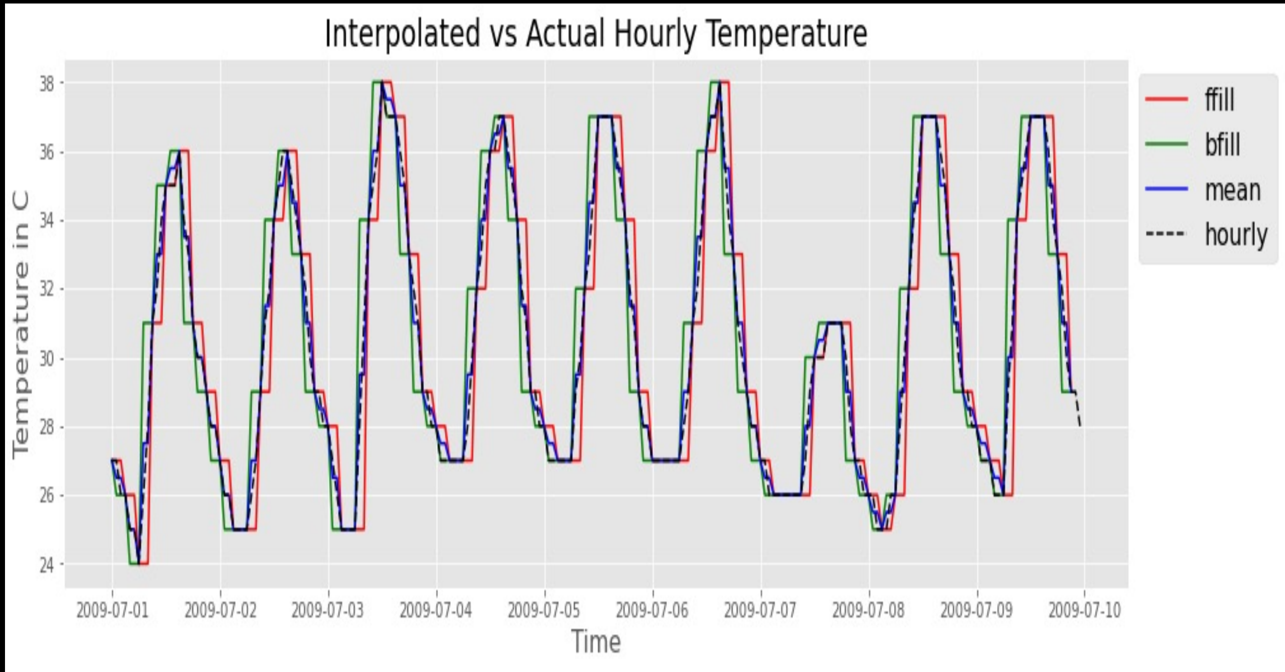
Energy dataset

- 16 years of hourly data for 8 Texas weather zones.



Weather dataset

- 13 years of sub-hourly data for 8 Texas weather zones.
- Interpolated to hourly level to match the energy data
  - forward fill, backward fill and mean interpolation methods
- One weather station per each zone are used to extract weather information

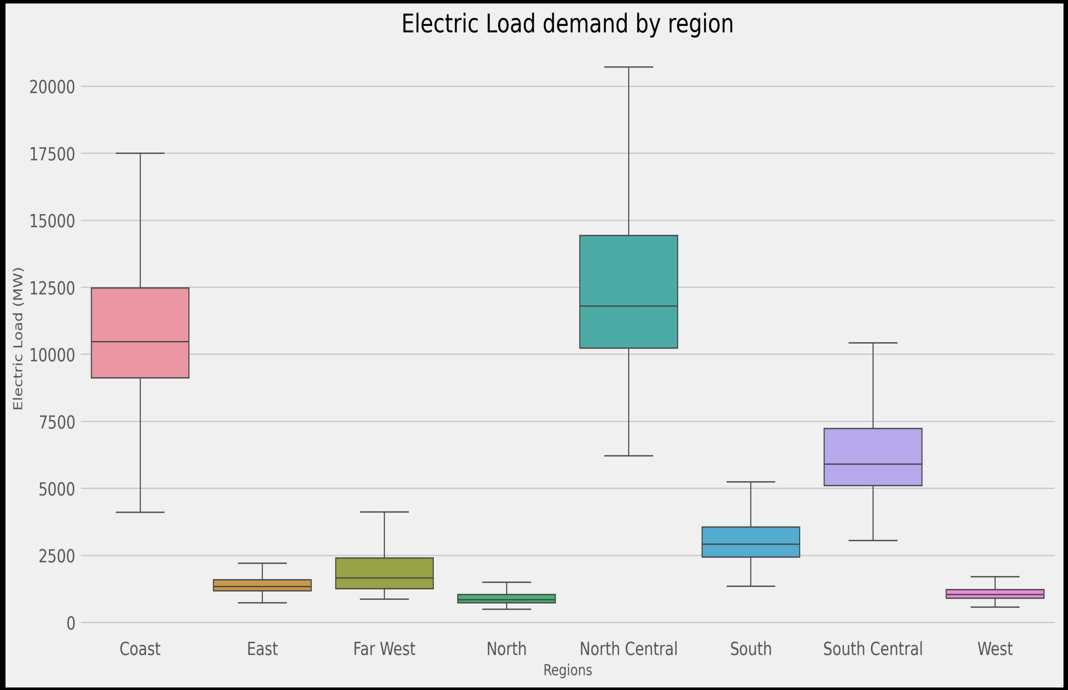


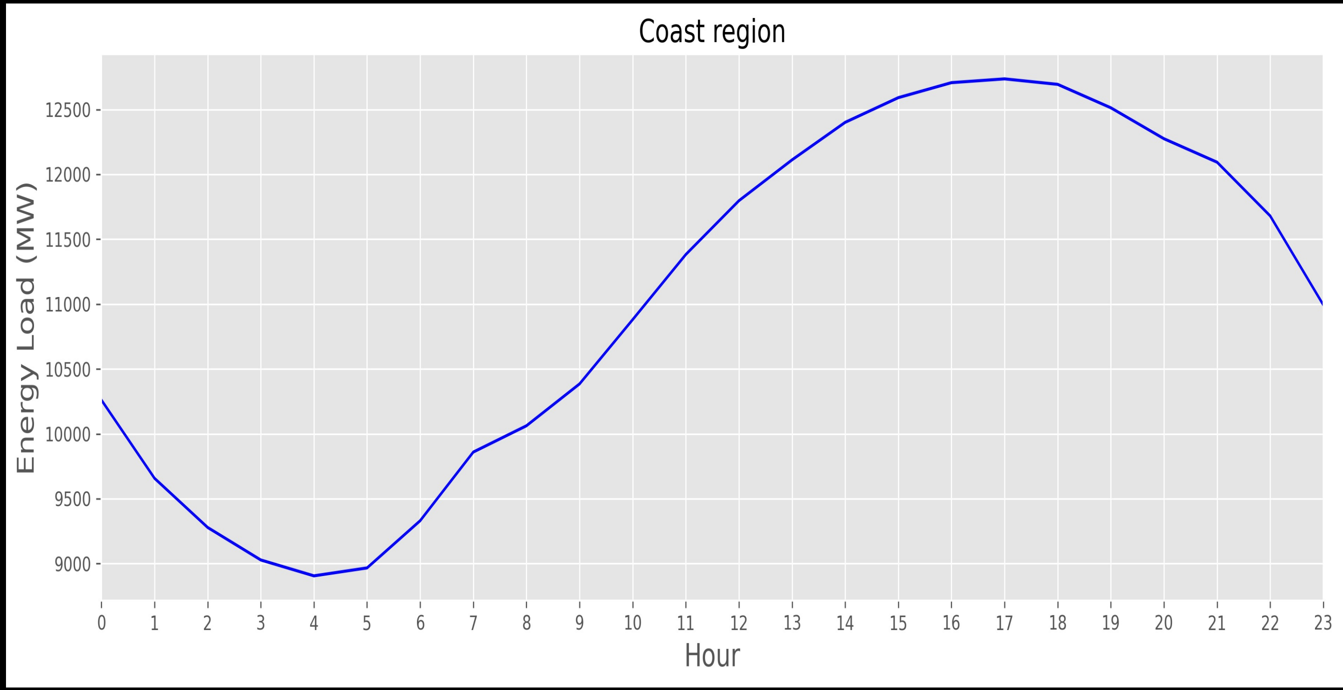
* Covid-19 data was covers only a small portion of the energy and weather data and it was not included in any of the modeling cases

Data Analysis

Total energy consumption by region

Average hourly energy consumption

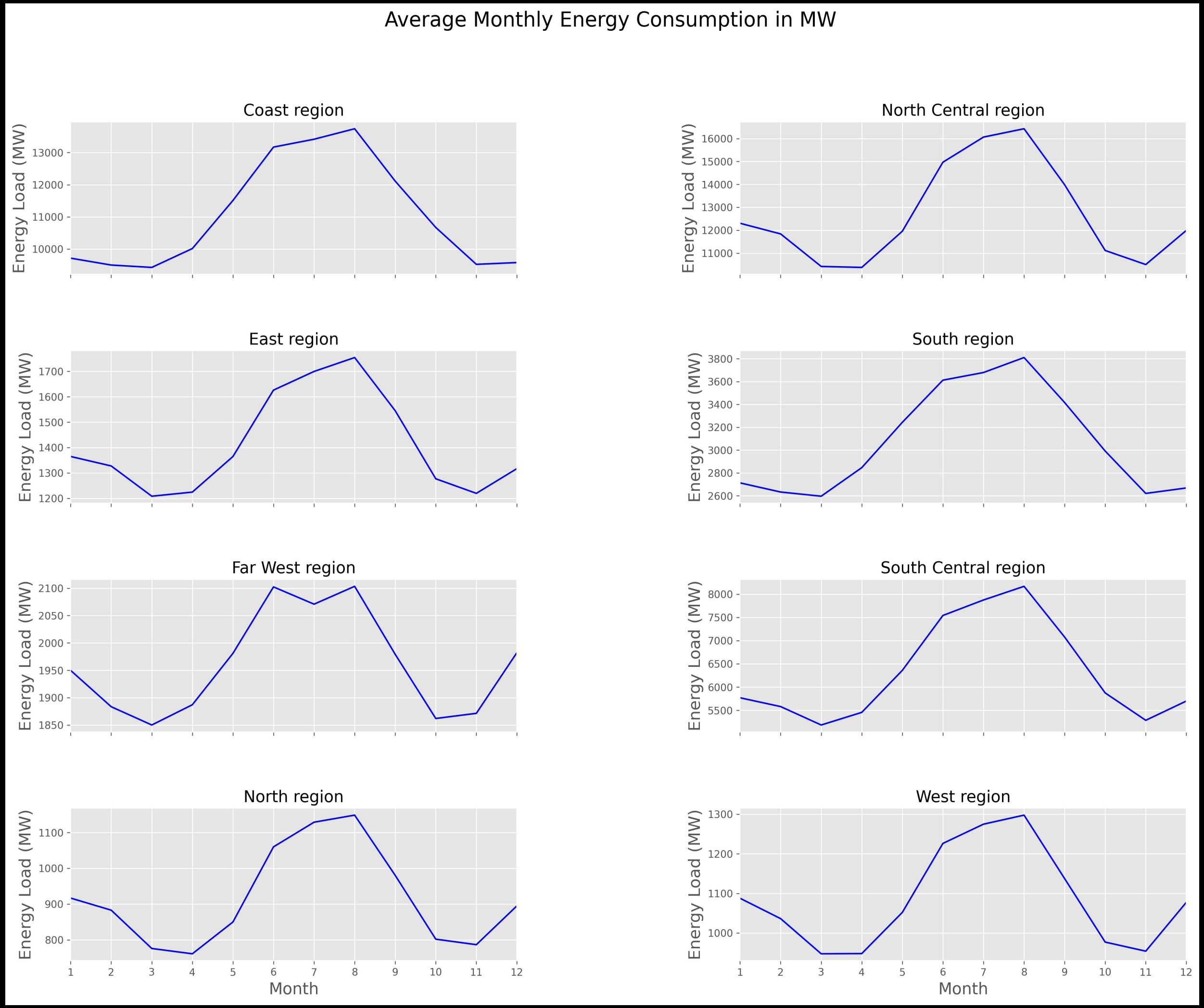




By far the North Central and the Coast zones account the majority of the electric load demand, followed by the South Central and South zones

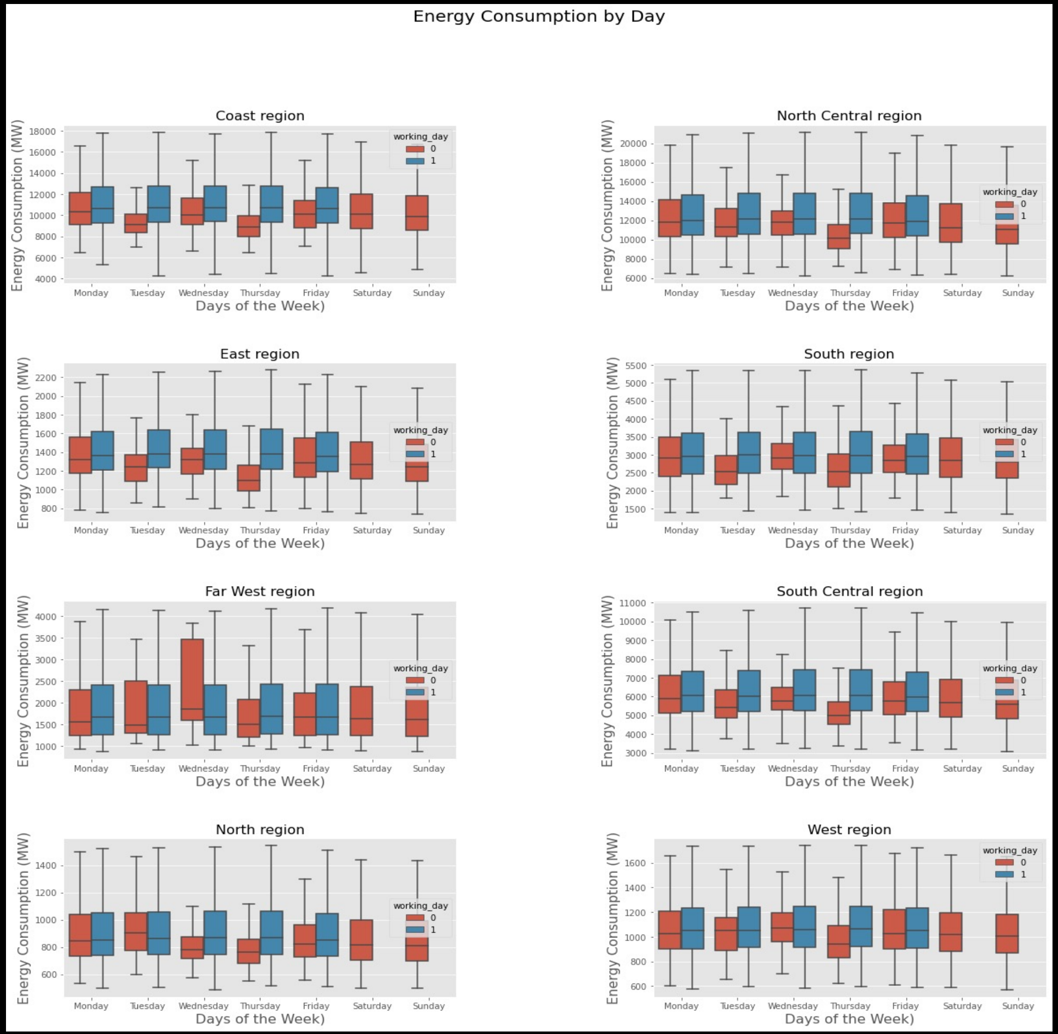
The load remains low over the night and then starts increasing as the region wakes up, and continues increasing during the office hours and peaks in the evening when everyone returns home and turns on the electrical appliances in their house

Average monthly energy consumption



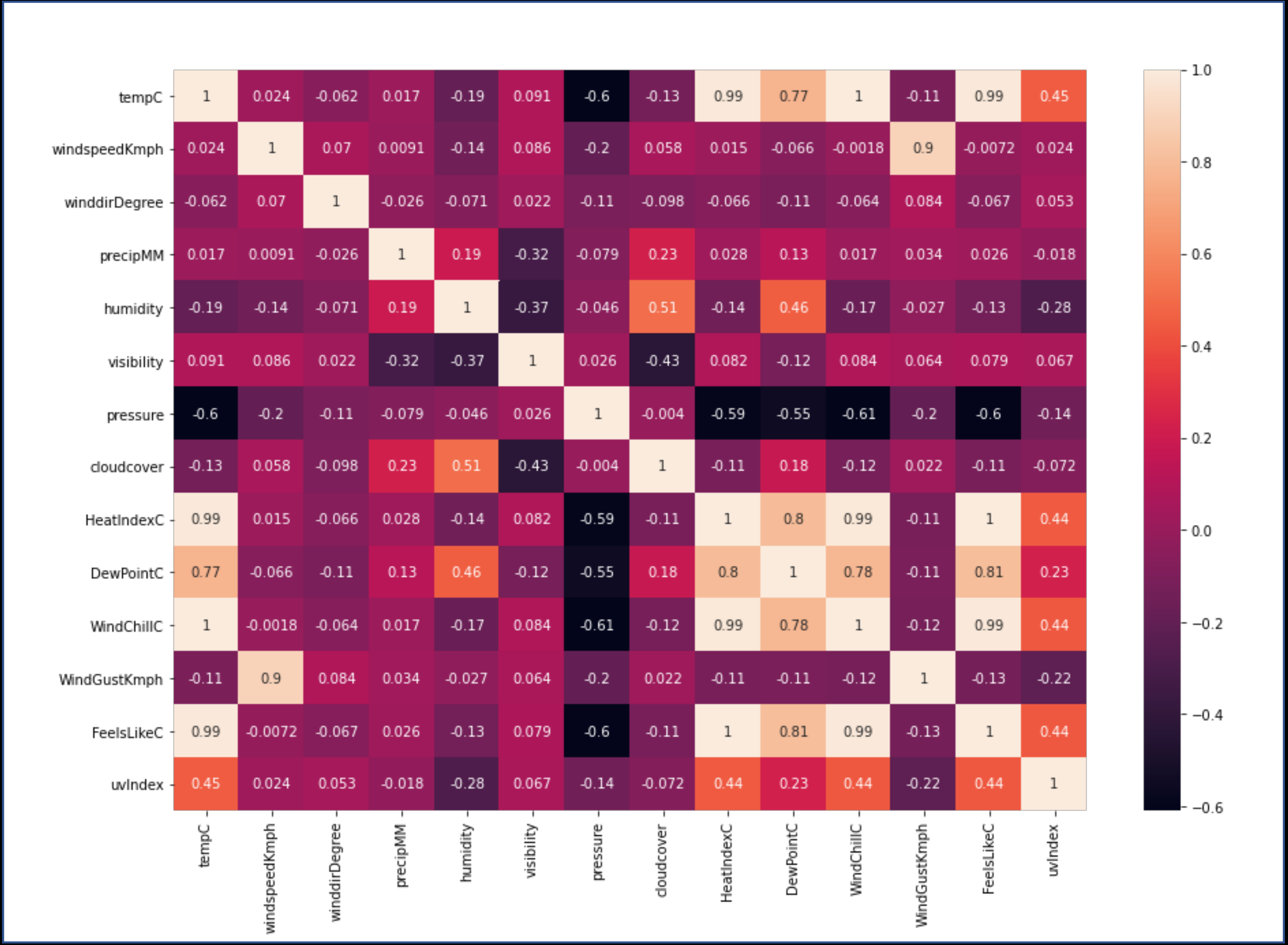
- All the regions have the highest demand during the summer month. This coincides with the hot summer of Texas.
- All regions, except the Coast and South regions, show mild increase in electric demand during the winter months of December up to February.
- The Coast and South regions have a mild or sometimes none-existent winter months (geographically they are the southern most regions), hence we don't see that much increase in electricity demand during the winter months.

Energy consumption by date type



- The median load consumption on working days remains fairly same from Monday to Friday and drops on the weekend.
- If a particular day is a holiday or non-working day the load consumption is much lower than if the same day was a working day.

Heat Map : weather data



- High correlation b/n certain features.
- There is redundant information in the data, they can be excluded from input to the model

Feature Engineering

The following features were created

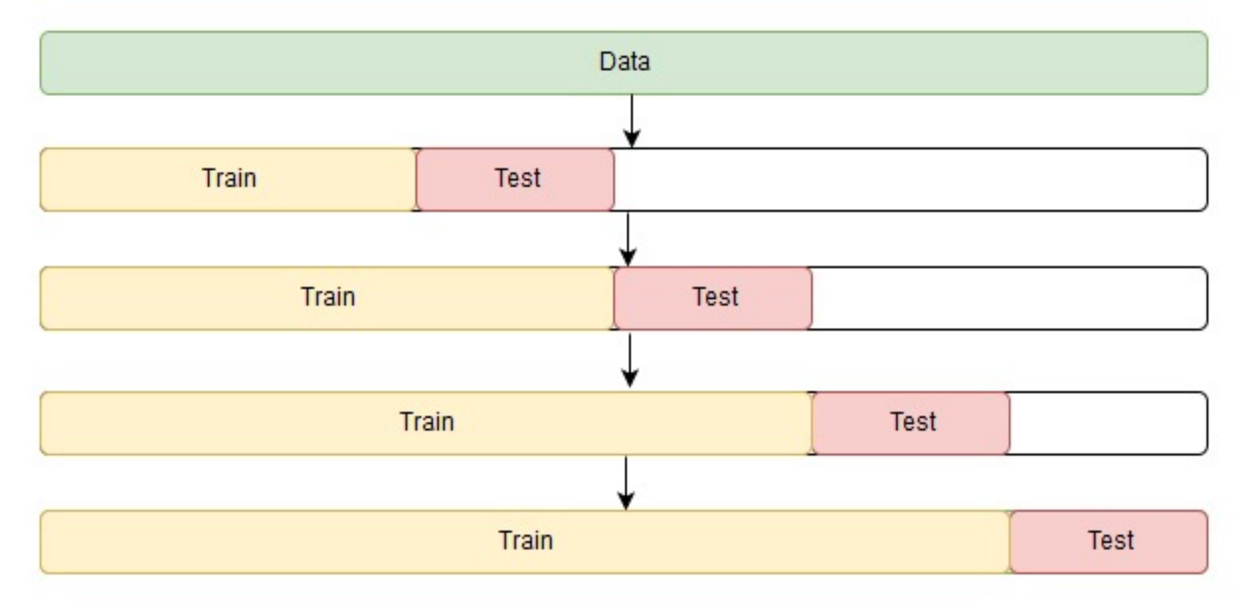
- Time related features
  - year, month, day, hour, weekday,
  - Holiday/non-holiday
  - Working/non-working day
- Lag and rolled features
- Fourier terms to account for the cyclic nature of a time series
- Apparent temperature that measures how people feel about the overall weather condition. It combines the effect of temperature, air pressure, humidity and wind speed

Modeling

Error Metrics and Cross validation (CV)

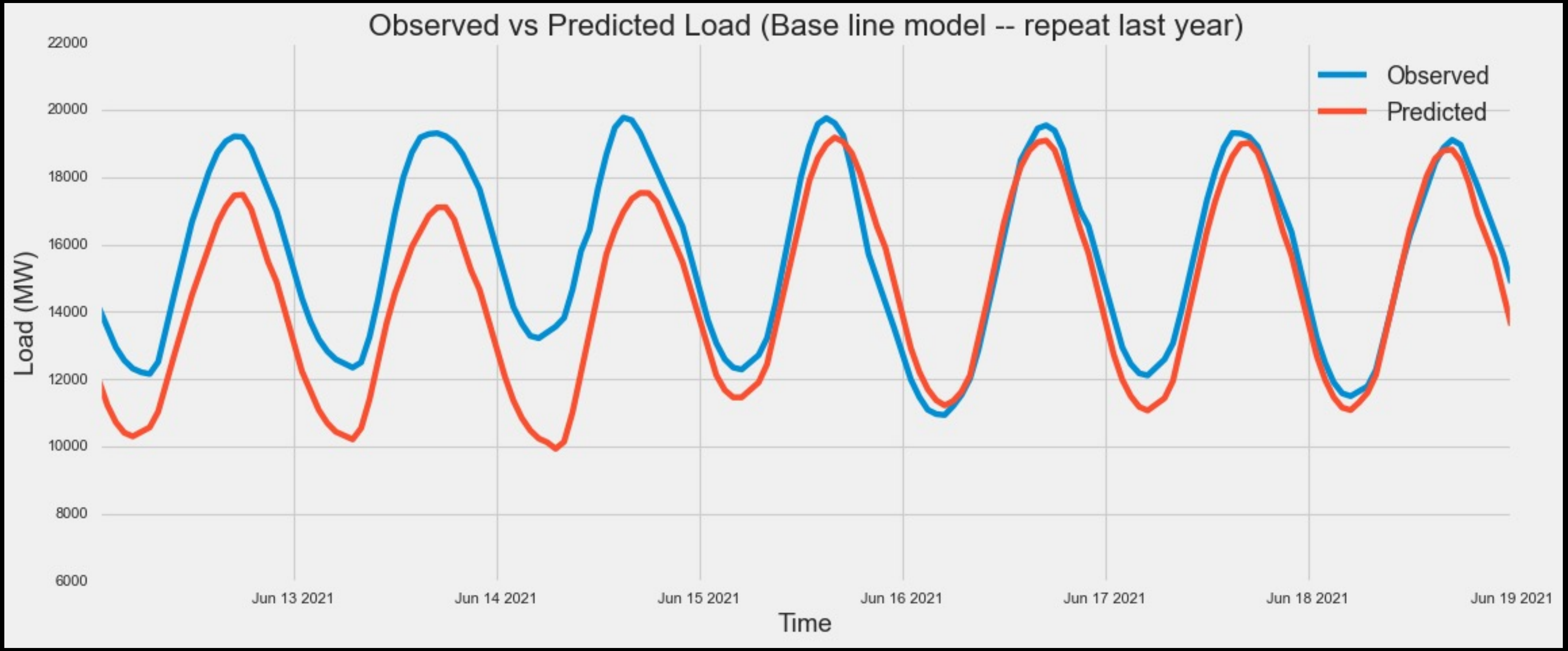
- R2 score
- MAE : mean absolute error
- RMSE : root mean square error
- MAPE : mean absolute percentage error

Time series CV



Base line model

Base line model : Repeat last year's load to current time

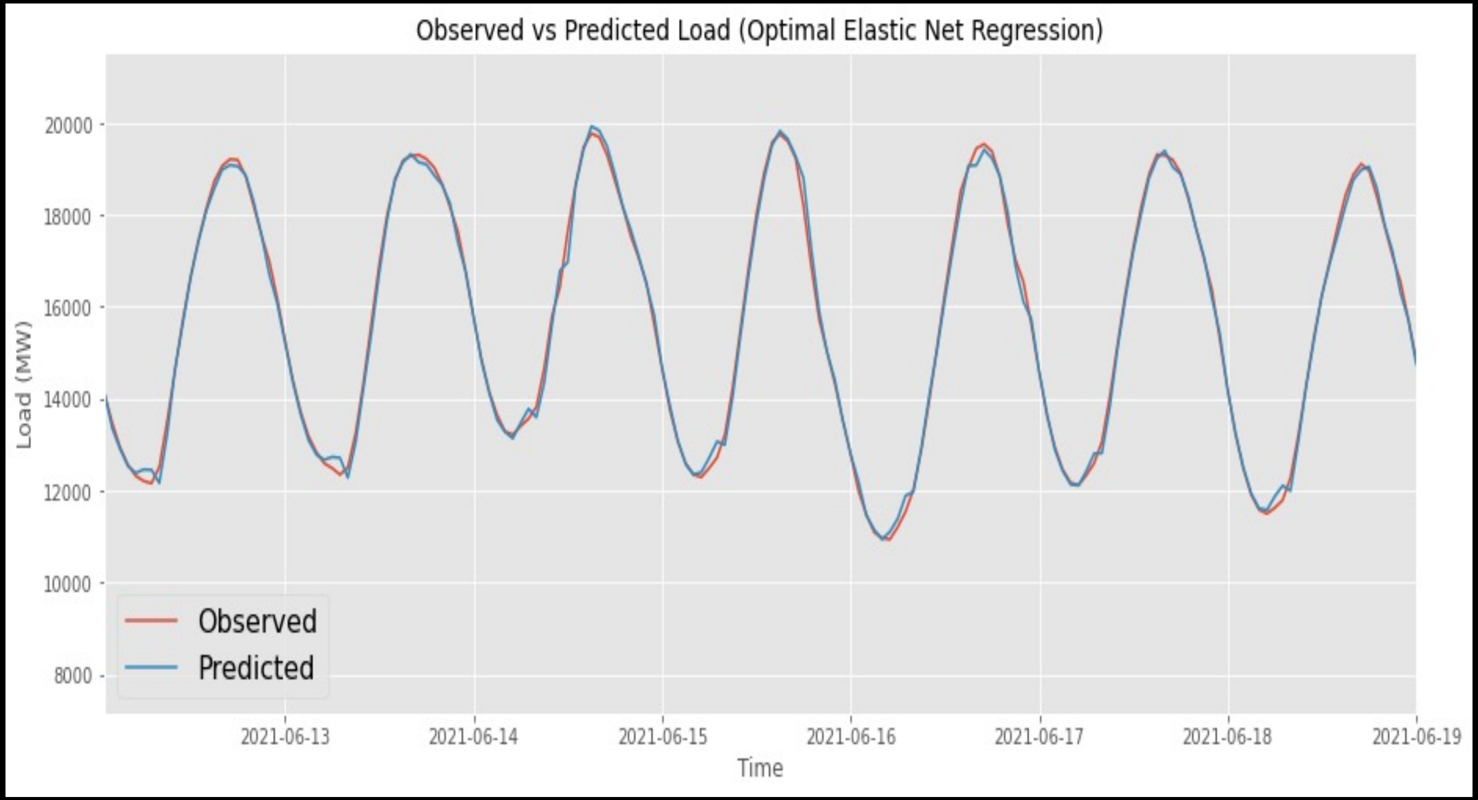


MAPE : 9.98%

RMSE : 1688.7Mw

Extended modeling

Elastic Net Regression



After hyper-param tunning

Before hyper-param. tunning

MAPE : 4.3%

RMSE : 667.6Mw

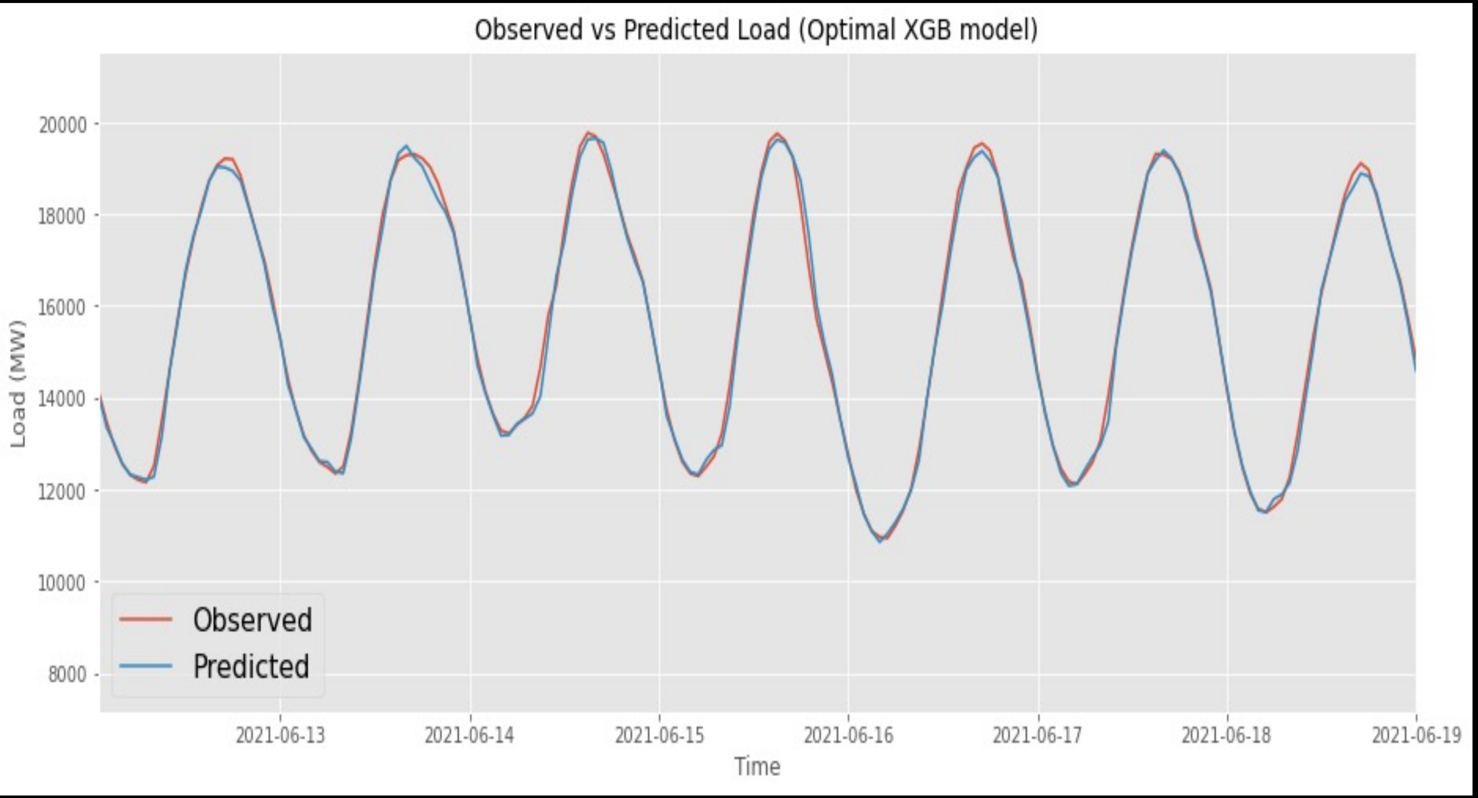
After hyper-param. tunning

MAPE : 0.87%

RMSE : 150.0Mw

Extended modeling

Extreme Gradient Boosting (XGB)



After hyper-param tunning

Before hyper-param. tunning

MAPE : 1.0%

RMSE : 179.8Mw

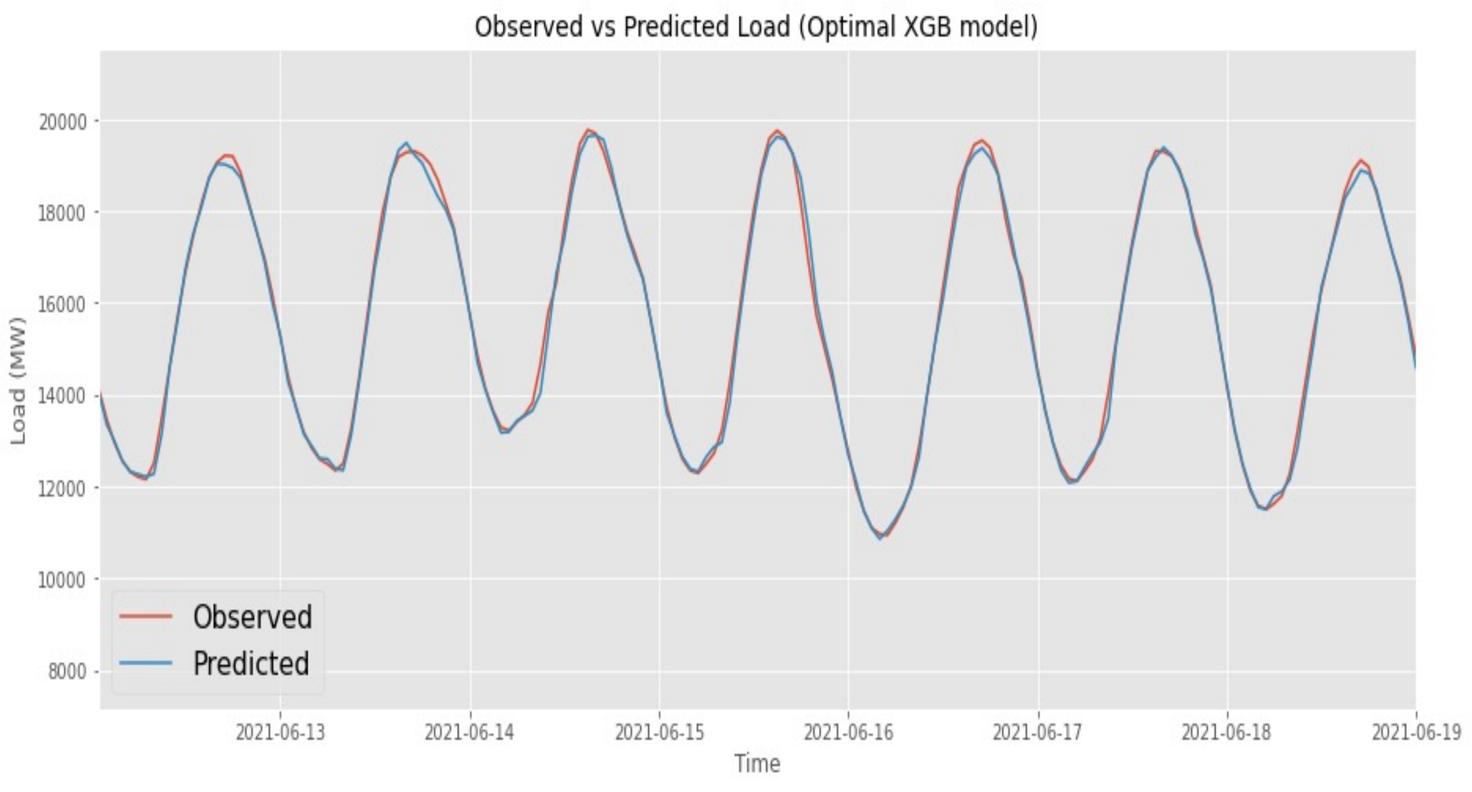
After hyper-param. tunning

MAPE : 0.89%

RMSE : 163.1Mw

Extended modeling

Light Gradient Boosting Method (LGBM)



After hyper-param tunning

Before hyper-param. tunning

MAPE : 1.0%

RMSE : 184.4Mw

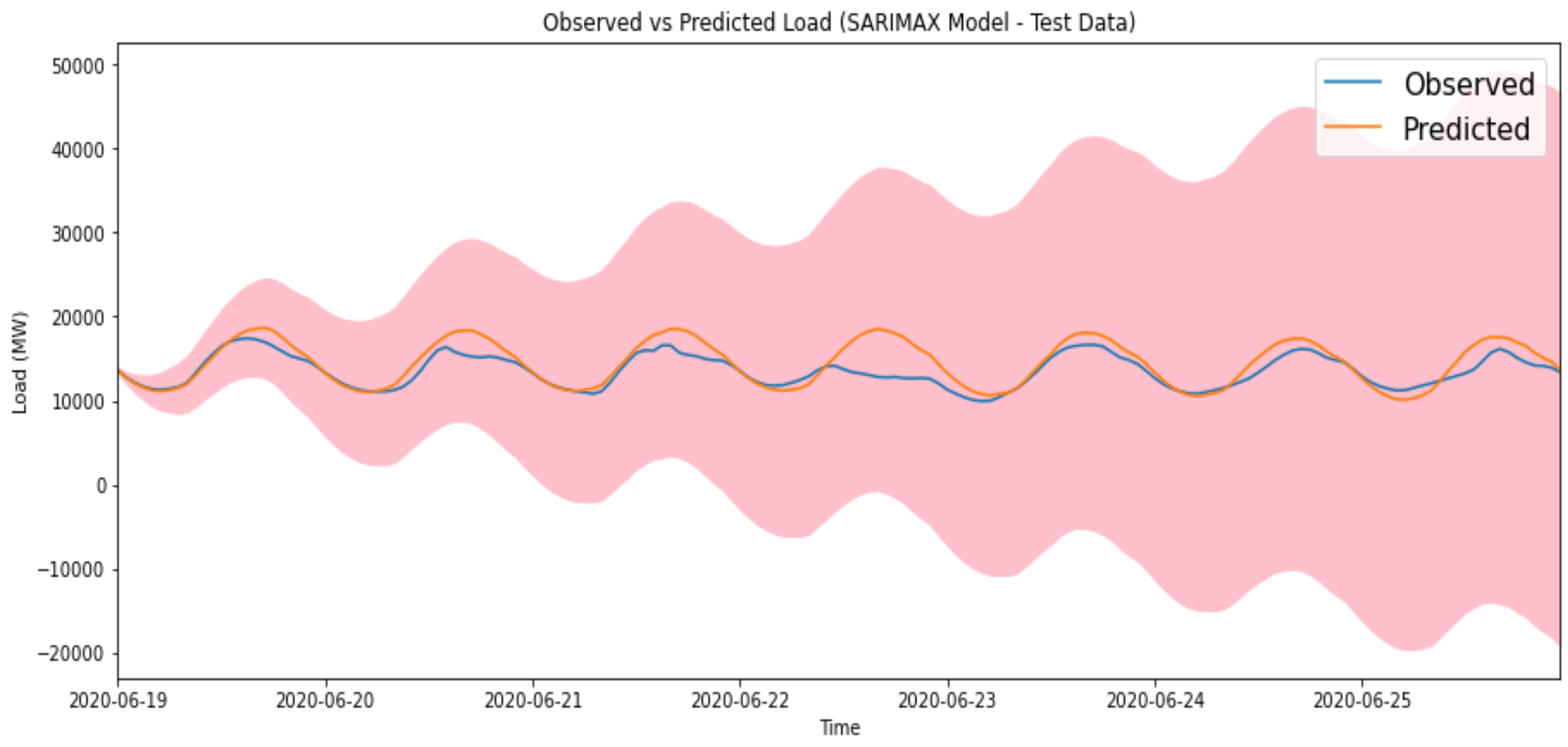
After hyper-param. tunning

MAPE : 0.86%

RMSE : 160.6Mw

Extended modeling

SARIMAX model (2,1,1)x(1,0,1,24)

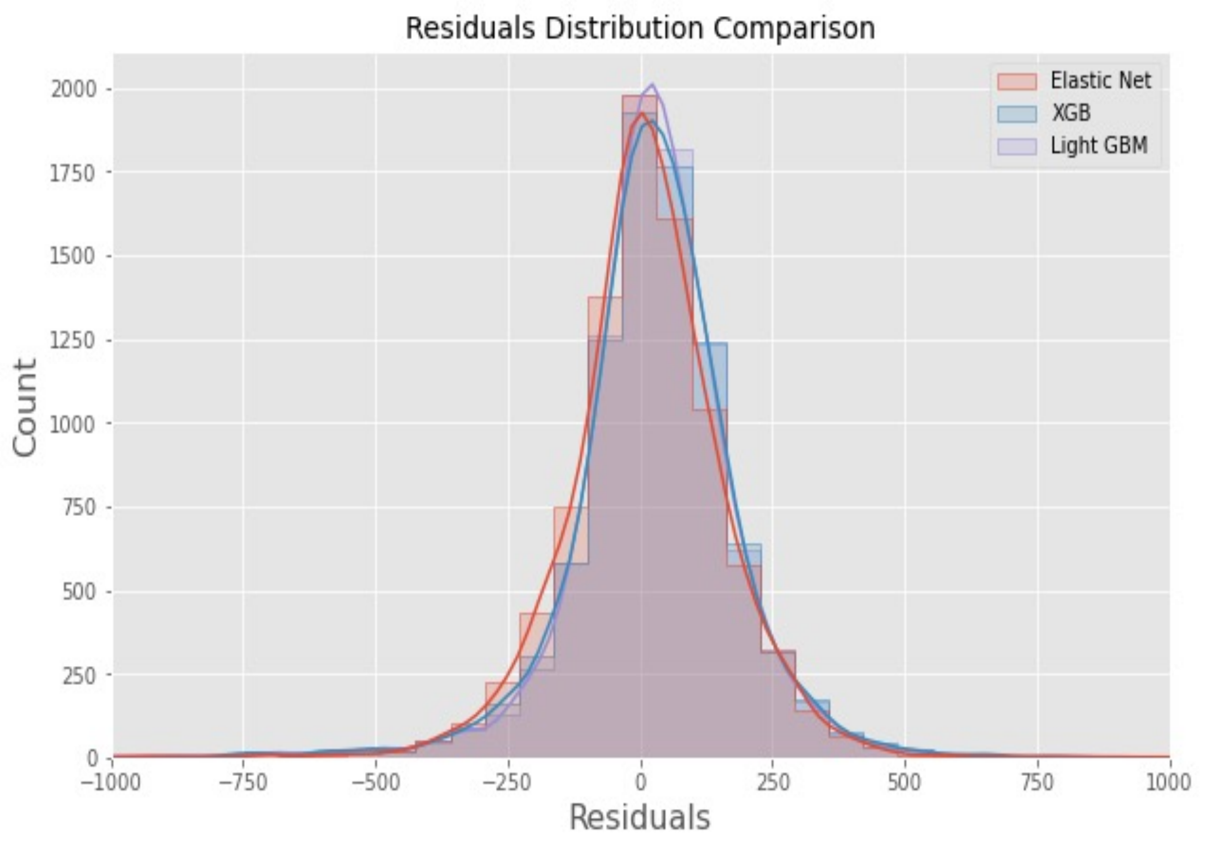


MAPE : 8.2%
RMSE : 1615.1Mw

Residual (Observed – Predicted)

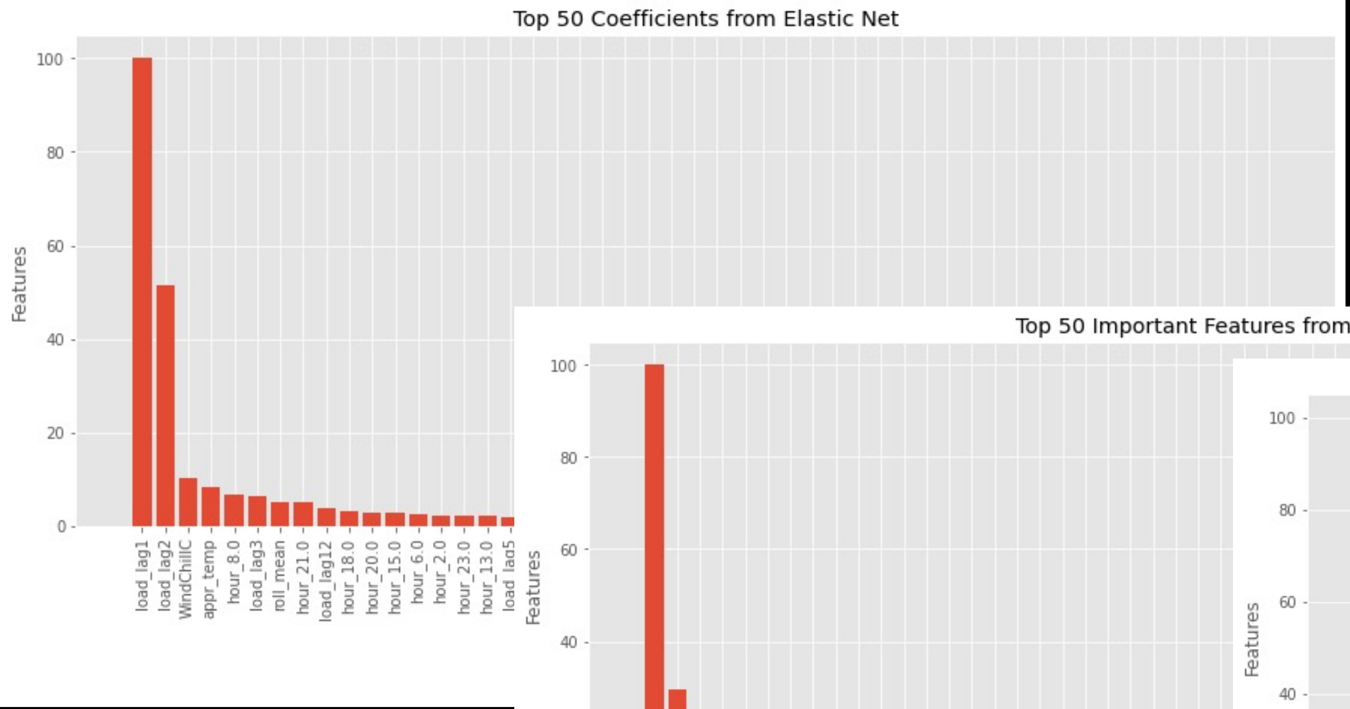
- The residuals distribution is a normal distribution with peak around zero.

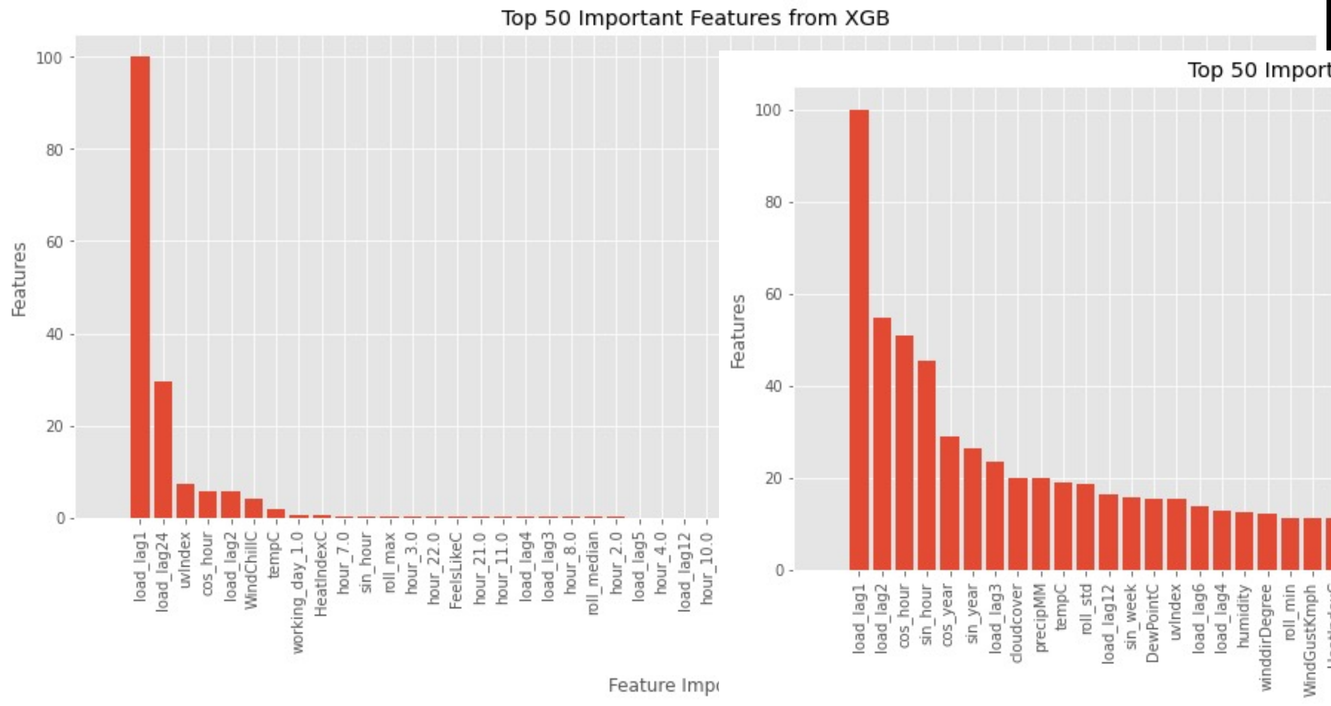
- The distribution is narrow meaning the magnitude of the errors is small.

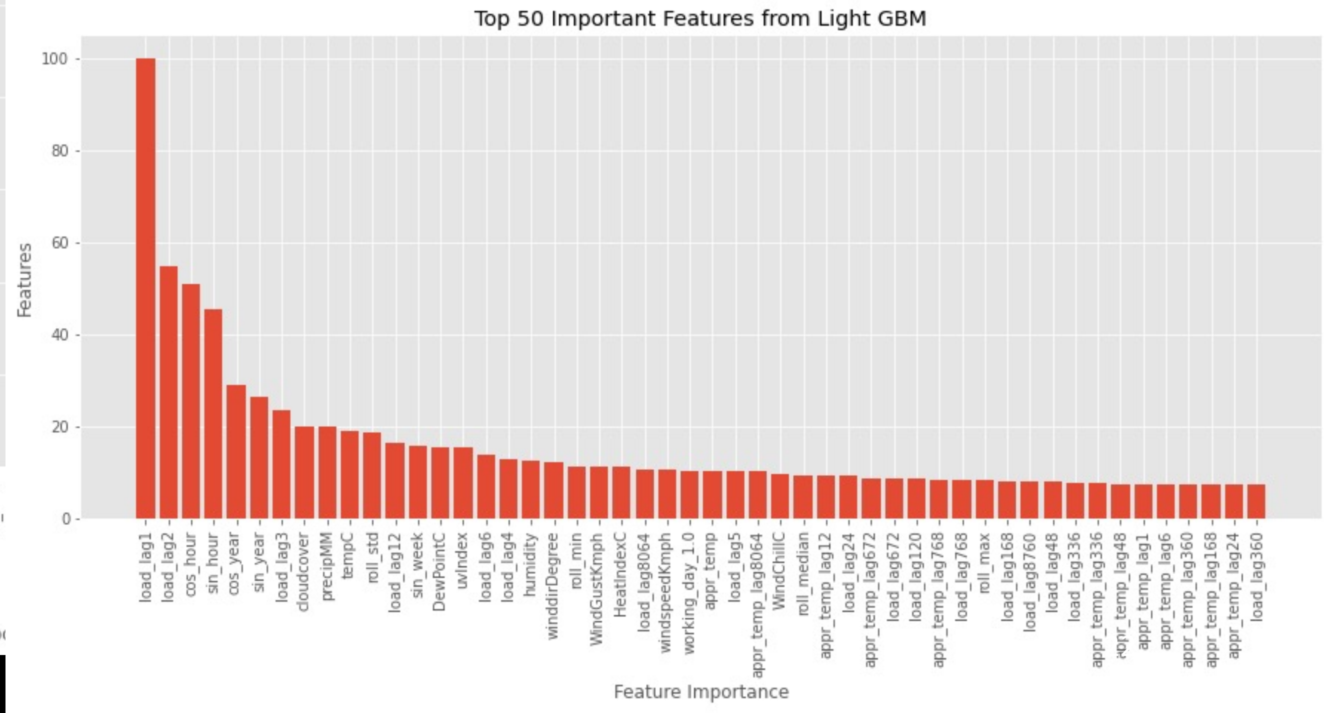


Feature Importance

- Load lag features especially load lag 1hour are the most dominant features
- Temperature, windchill, Uvindex and the Fourier terms are also among the top 10 features







Model Comparison

Models		R2 score		MAE (Mw)		RMSE (Mw)		MAPE (%)	
		Train	Test	Train	Test	Train	Test	Train	Test
Elastic Net	Default	0.951	0.932	452.45	532.542	584.389	677.566	4.034	4.294
	tunned	0.997	0.997	104.88	109.895	144.43	151.037	0.889	0.877
XGBOOST	Default	0.999	0.995	68.987	128.178	90.889	179.815	0.586	1.007
	tunned	1.000	0.996	6.345	114.341	8.662	163.150	0.056	0.891
Light GBM	Default	0.998	0.995	90.036	133.139	124.1	184.427	0.752	1.045
	tunned	0.993	0.996	53.613	111.014	72.233	160.581	0.454	0.866
SARIMAX		0.451	0.278	1705.9	1128.52	2058.26	1615.07	11.397	8.225
Base Line Model		Nan	0.576	Nan	1260.75	Nan	1688.67	Nan	9.982

Model Comparison

- Elastic net, XGB and Light GBM with hyper-parameter tunning gave excellent performance
- XGB takes longer time compared to LGBM and Elastic Net
- The SARIMAX model performance is bad is only slightly better than the base line model
- Among the three models, Light GBM has slight advantage and was selected as our final model.

Conclusion

- we built a very robust and accurate forecasting LGBM model utilizing the hourly interval data that resulted in MAPE of 0.86%

- Some observation and take away from the project
  - Knowledge of previous few hours load consumption is very crucial for short-term forecasting.
  - Electric load usage is highly correlated to temperature. Therefore, accurate weather forecast is needed to get a good load forecasting.
  - Energy consumption is non-stationary. It has trend as well as multiple seasonality (daily, yearly). This is probably one of the main reasons why the SARIMAX model performed poorly.

Future Works

- Build forecasting models using Long Short Term Memory (LSTM) and Facebooks' Prophet methods

Recommendations

- To have accurate short-term forecasting (with MAPE < 1%), you need to have knowledge of recent past electric load consumption. With todays' availability of smart meters, last hour load data can be readily available. We recommend to use lag features as short as possible.

- Weather information particularly temperature has direct impact on the forecasting. The accuracy of the weather information we feed to the model is of great importance. There are several API that provide weather forecasting. We recommend to use reliable weather API's to get the weather forecast information.

References

[1] https://engineering.electrical-equipment.org/electrical-distribution/electric-load-forecasting-advantages-challenges.html
[2] https://people.eng.unimelb.edu.au/smonazam/publications/Fahiman2019IJCNN.pdf
[3] https://medium.com/analytics-vidhya/hyperparameter-tuning-hyperopt-bayesian-optimization-for-xgboost-and-neural-network-8aedf278a1c9
[4] https://analyticsindiamag.com/complete-guide-to-sarimax-in-python-for-time-series-modeling/