We are given two datasets

1. Train.csv which contains multiple features along with target label.
2. Test.csv which contains all features from above except target label having NaN values

Columns given are*: [ 'Date', 'Is Working Day', 'Previous Hour Price', 'SDSH DAP',*

*'SDSH Volume', 'Sgn0 VolumeDir', 'P24HA Price', 'PDSH Price',*

*'PWSH Price', 'PWA Price', 'target' ]*

Train dataset has no missing values.

Index was set to “Date’ column

Used boxplots to check the quartile range of data for every column. From boxplot I can conclude that there are many outliers and outlier are important to this dataset, thus decided not to treat the outliers.

Used heatmap of seaborn to check the correlation of every feature. I used a threshold of 0.75 to filter out any feature.

{'SDSH DAP'} feature was removed from dataset.

Used distribution plot of seaborn to how the data is distributed. Found out most of the columns were skewed.

Used line plot to check which feature is most dependant on target label. ‘Sgn0 VolumeDir’ features didn’t align with target label, so it was removed.

Features were scaled via minmax scaler as values were with high range.

For ensemble method I used RandomForestRegressor with GriDSearchCV.

I found the most important features via this and removed two features ‘Hours’ and ‘Weekly hours’

After that With the new features trained an xgboost regressor which is again an ensemble model

For deep model I used neural network with 2 hidden layers for 400 epochs.

All models score were similar, but random forest was better.

All the results from 3 models are stored into 3 separate csv files.