# BEE 6310: Final Project Report

Electricity Price Spikes vs Renewable Grid

Harsh Meel, hm642

## Introduction:

My problem statement is to analyze Peaking of Electricity Prices in New York State, using the price, load and fuel mix data available from NYISO (New York Independent System Operator), in the context of transition to a renewable energy grid.

Under the New York State Climate Leadership and Community Protection Act (CLCPA) , the State of New York has set ambitious targets to eliminate 100% of the climate pollution caused by humans, calling for an 85% reduction in greenhouse gas emissions by 2050, with an interim target of 40% by 2030. However, an increase in reliability on renewable energy can lead to energy reliability issues and price hikes. A transition to a zero-carbon grid is necessary, but it cannot come at the expense of equity or access to reliable and cheap power. Price spikes hurt the vulnerable in society, who had little contribution to the carbon emission problem to begin with.

To understand this further, a regression between the grid attributes and the price spikes would let us find out the actual drivers of price spikes in a zero-carbon grid. A model fit will then help us to predict what could happen in the future in terms of frequency of price spikes, with a greater renewable energy mix in the grid.

This work is not a reproduction of any research but it sources data available on the NYGrid Model developed by Anderson Lab at CALS, Cornell University. (https://github.com/AndersonEnergyLab-Cornell/NYgrid)

## Data and Methods:

**Data Sources:**

1. Our Independent Variable is Electricity Price. NYGrid model collects NY state electricity prices for each hour from Jan 1st, 2016, to Dec 31st 2021, for 15 different sectors in the state.

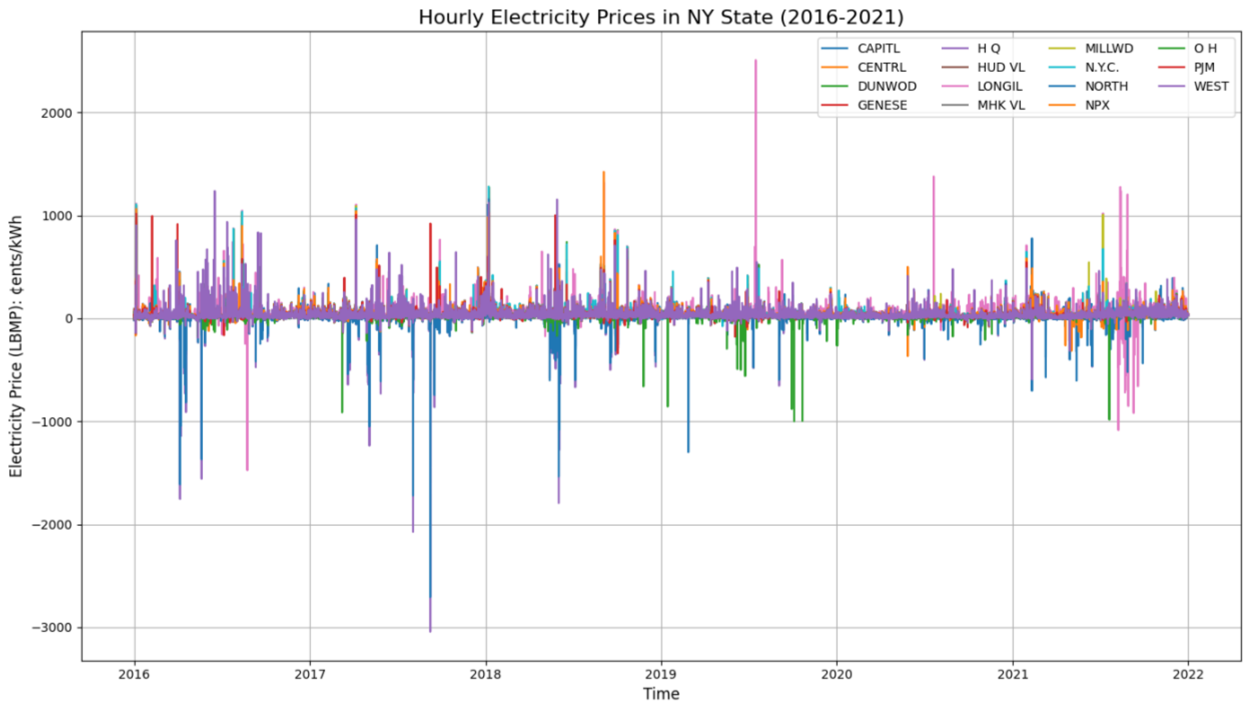


*Figure 1: Time series of electricity prices in NY state*

2. Our Dependent Variables are Electricity load and Electricity Generation fuel-mixture, both sources for each hour from Jan 1st, 2016, to Dec 31st, 2021. The load data is available for 12 different sectors from the 15 mentioned before and the generation fuel-mixture data is for the overall state. The 12 sectors are: 'CAPITL', 'WEST', 'N.Y.C.', 'MILLWD', 'NORTH', 'LONGIL', 'HUD VL', 'GENESE', 'DUNWOD', 'CENTRL', 'MHK VL', 'MILLWD'.

**Data Processing:**

**1)** First must change the hourly electricity prices into a binary variable: Price Spike – (True or False). This can be done by deciding a threshold for an increase in price. Two methods were tried:

**a)** Having a static threshold – where any time price is more than 100 cents /kWh, a price spike is registered. Time unit decided is per month

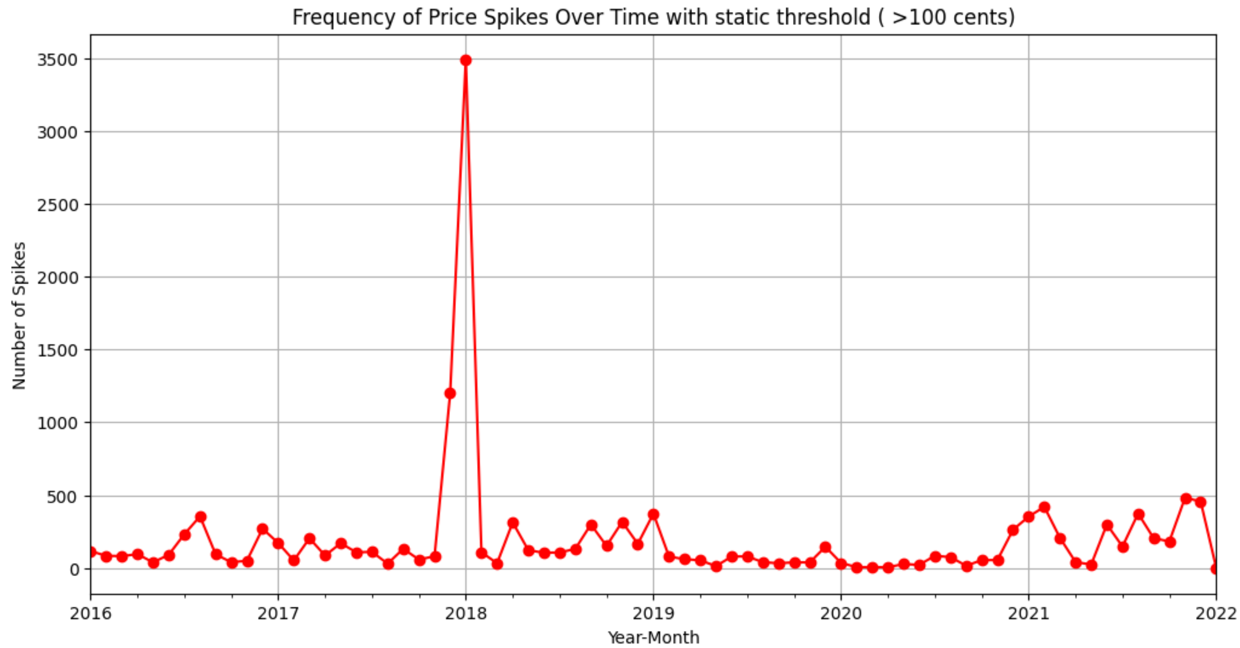


*Figure 2: Time Series of price hikes using static threshold*

**b)** Having a dynamic threshold based on temporal average – if price is 500% of the 7-day average around the current date a spike is recorded. Time unit decided is per month.
A interesting observation is the stability in Electricity prices during the Covid-19 lockdown.

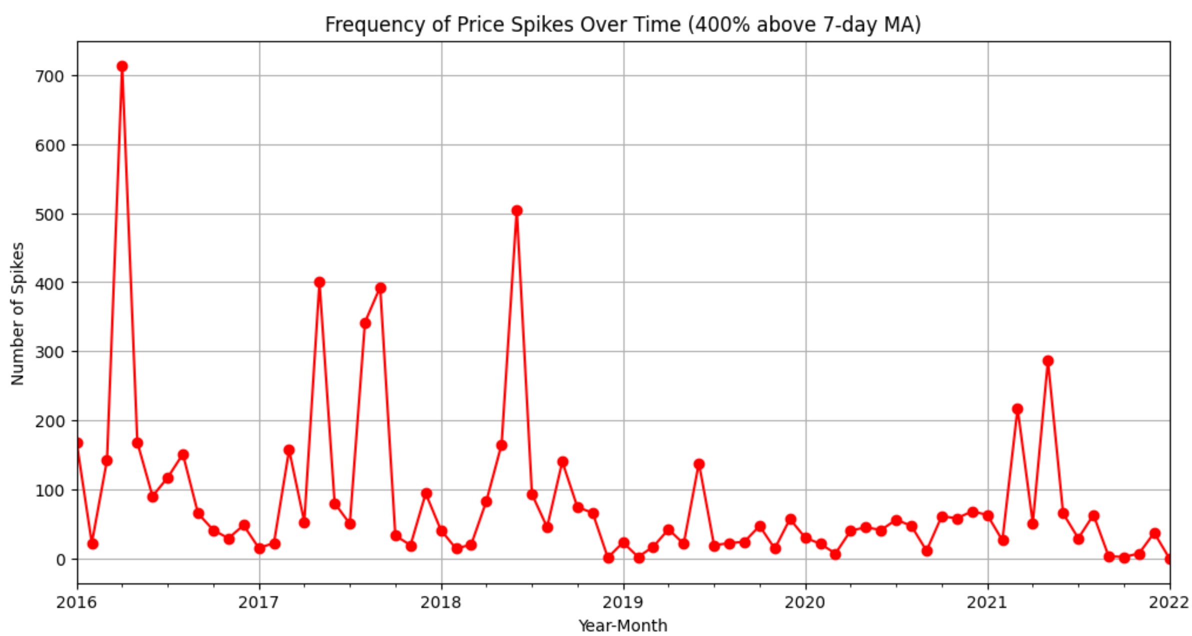


*Figure 3: Time Series of price hikes using dynamic threshold*

This dynamic threshold is chosen over the static one because:

- It is not affected by price increase or decrease due to seasonal loads and inflation in fuel prices.
- It is a better representation of abnormality or unexpectedness in budget, which is more aligned with our aim to predict reliability in electricity prices.

2) Bringing in the predictors now. Prices are governed by two things majorly: supply and demand. A peak can be result of a problem with either of them:
   a) A case of abnormal load - Thunderstorm or heatwave
   b) A case of abnormal problem with supply - loss of transmission, renewable energy not being productive due to weather

   Load data is in the units of MW/hr. for each sector.

   Generation Fuel Mix data is divided amongst seven types: 'Dual Fuel', 'Hydro', 'Natural Gas', 'Nuclear', 'Other Fossil Fuels', 'Other Renewables', 'Wind', each reported as GW/hr. Our interest is in the percentage mix of fuel type, rather than the actual generation. So, fuel mix data is normalized to provide ratio of fuel type for each hour. We assume this ratio remains equally accessible to each of the 12 sectors, thus the data is normalized. After some more changes to predictors the regression looks like:

   X:

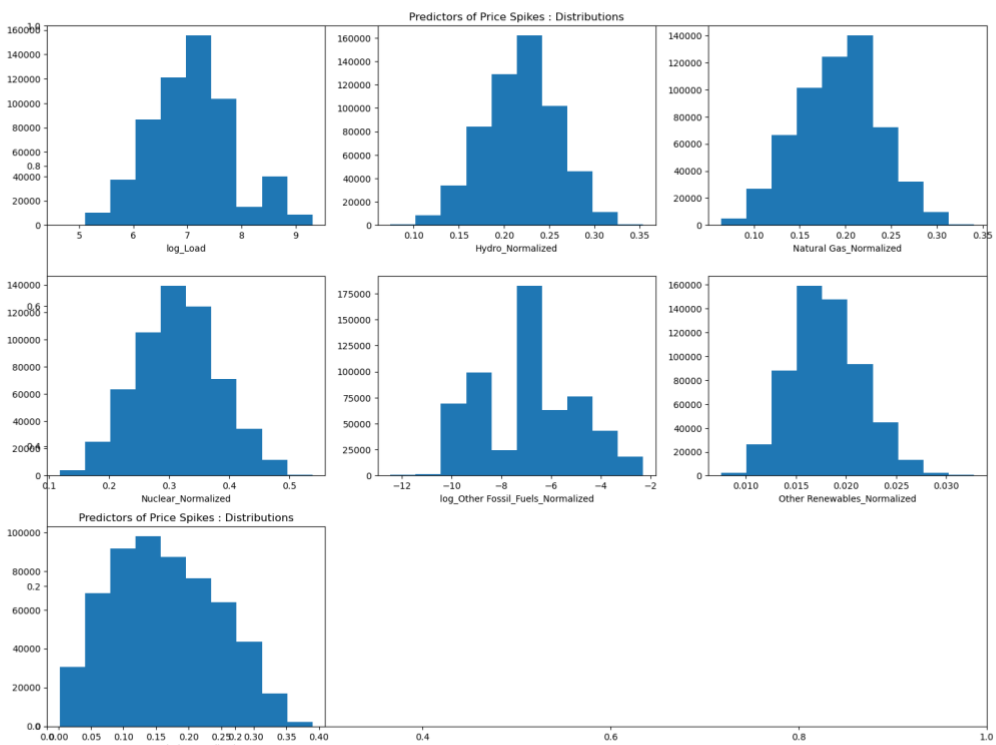


*Figure 4: Histograms of well distributed predictors*

Y: Price Spikes - Binary Value

**Collinearity Check:**

We would like to plot a logistic regression between these predictors and the price spikes. But first we would like to have a look at our data and ensure it doesn't have a collinearity problem - which is likely because a higher supply attracts higher demand.

i) A correlation matrix is plotted:

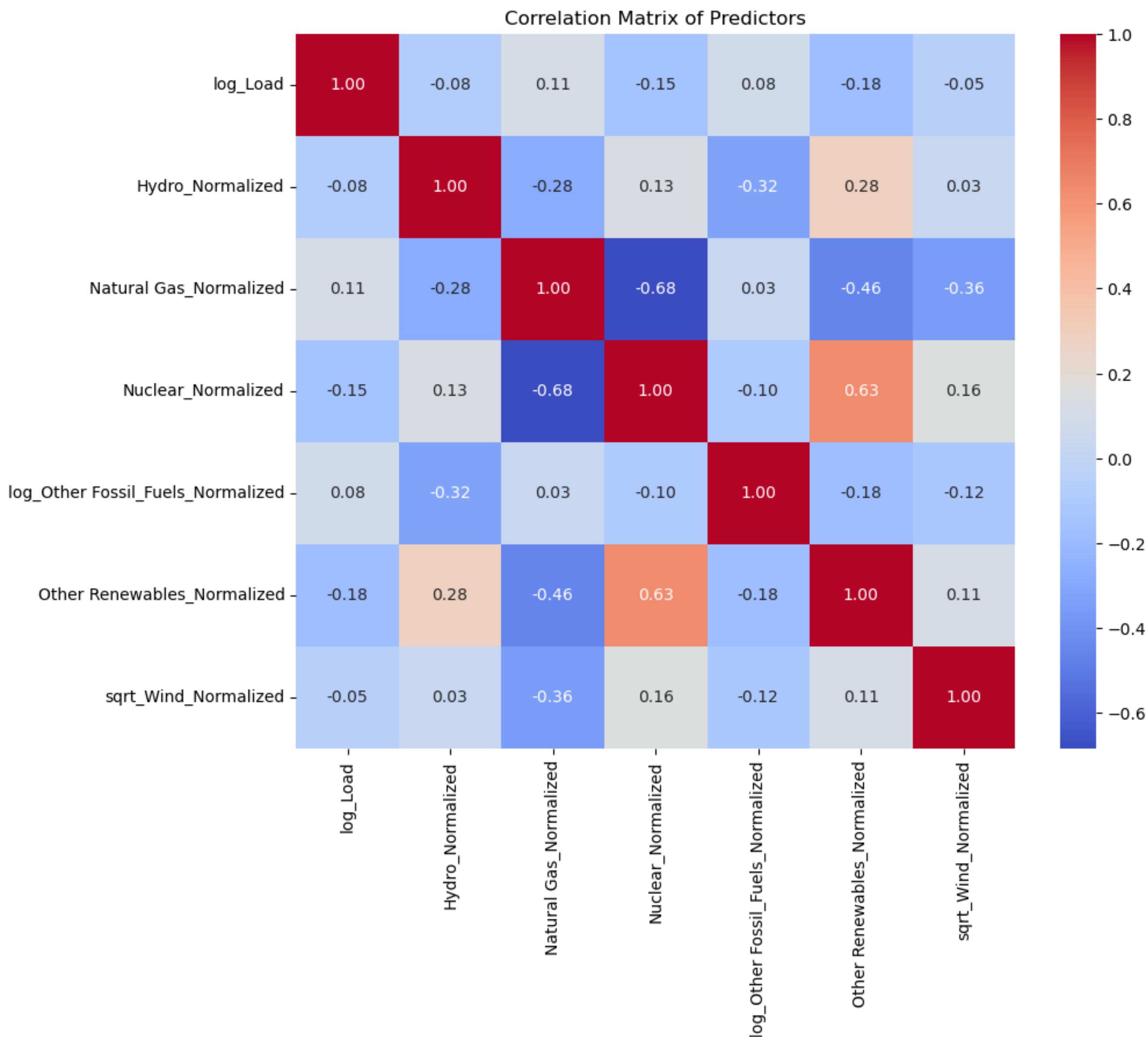


*Figure 5: Correlation Heat Map*

High correlation is observed between nuclear generation and other renewables generation: 0.63. Also, between hydro generation and other renewables generation: 0.28.

ii) VIF values of predictors are calculated. Except square root of normalized wind, every other predictor as a high VIF score.

| | feature | VIF |
|---|---|---|
| 0 | log_Load | 59.162816 |
| 1 | Hydro_Normalized | 32.630485 |
| 2 | Natural Gas_Normalized | 22.938991 |
| 3 | Nuclear_Normalized | 42.926726 |
| 4 | log_Other Fossil_Fuels_Normalized | 16.698786 |
| 5 | Other Renewables_Normalized | 48.788966 |
| 6 | sqrt_Wind_Normalized | 5.443627 |

## Results:

**Model Fitting:**

1) Linear Regression:

Because of high collinearity, Lasso and Ridge logistic regression are fitted on the dataset. Data is split in training and testing set and hyperparameter tuning is used to get the best alpha values for both regressions. The results are:

```
Best Ridge Alpha: {'alpha': 100}
Best Lasso Alpha: {'alpha': 0.001}
Ridge Regression Results:
Mean Squared Error: 0.007206623653120095
R^2 Score: 0.006736200984817309

Lasso Regression Results:
Mean Squared Error: 0.007219610170013645
R^2 Score: 0.004946314662659046

Lasso Coefficients:
log_Load                            -0.000893
Hydro_Normalized                     0.000000
Natural Gas_Normalized              -0.001593
Nuclear_Normalized                  -0.003916
log_Other Fossil_Fuels_Normalized    0.001504
Other Renewables_Normalized         -0.000099
sqrt_Wind_Normalized                -0.001083
```

With such low values of R^2, there doesn't seem to be a linear relationship between the predictors and price spikes.

2) Non-Linear Regression:

The dataset is fitted in a Decision Tree Classifier. Hyperparameter tuning is used to improve the model fit. Yielding the best parameters as:

Best parameters: {'criterion': 'entropy', 'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 10}

On fitting the best Decision Tree Classifier, following is the evaluation of test data:

**Precision: 0.41, Recall: 0.43, F1- score: 0.42**

With a confusion matrix of:

```
[[114385    518]
 [   482    364]]
```

F1 score of 0.42 is a low score but reflects much better fit than the linear relationship.

Further attempts are made to get a better fit using ensemble methods like Random Forest and Gradient Boosting. But no significant increase in F1 score on the test case is observed.

## Discussion and Conclusion:

1. No significant fit was found between the predictors and price spike: With an insignificant linear fit and a Decision Tree fit of F1 score 0.42, the predictors failed to provide any significant relation to price spikes. Thus, no predictions can be made about increase in price spikes relative to renewable generation with confidence.

2. Highly imbalanced data: The price spikes data was highly imbalanced with much of the instances being of no price hikes, so the model is struggling to correctly predict the minority class. Work can be done to address this problem.

3. High correlation between Renewable generation and Nuclear Power: A high correlation was observed between renewable energy generation and nuclear power generation. It was unexpected should be investigated further, to probably improve model fit.

4. Complexity of the problem: At the end predicting price spikes could be inherently difficult due to a variety of external factors (e.g., weather, market conditions). In such cases, a low F1 score might be the best achievable outcome with the available features.