1495
UNIVERSITY OF
ABERDEEN

Thesis on:

Predicting electricity prices and quantifying dependence on renewable energies

Submitted by

Manohar Babu Polamarasetti

MSc data Science (2023-2024)

Contents

Abstract

The integration of renewable energy sources into the power generation mix introduces significant volatility into electricity prices, owing to the inherent unpredictability of wind and solar outputs. This volatility poses challenges for energy suppliers and market participants in the UK, who are required to forecast electricity demand and supply accurately within each settlement period to minimize *imbalance* costs. In this context, our study aims to develop and assess predictive models that can accurately forecast electricity prices in the UK, considering the fluctuating share of renewable energies. Utilizing historical data on power generation mix and electricity prices from 1st January 2009 to 31st December 2023, we employed both statistical and machine learning techniques, including linear regression, random forest, and gradient boosting regressors, to construct our models. Our evaluation metrics—Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared (R^2)—facilitated a comprehensive performance assessment of each model. The results indicate that while renewable energies significantly impact price volatility, models that incorporate detailed generation data can improve price predictions. This study contributes to the ongoing efforts to enhance electricity market efficiency, offering insights that could aid energy suppliers in optimizing purchasing strategies and reducing operational risks amidst the growing share of renewable energies.

1.Introduction

1.1 Context and background of UK electricity pricing

In the United Kingdom, the process of setting electricity prices is intricate and involves several different market players and regulatory bodies. Elexon is a non-profit for organisation. Elexon settles contracts for the electricity industry. The electricity cannot be stored. So, that makes it intrinsically difficult to manage on the network. Balance on second-to-second basis is taken by the electricity system operator which is national grid. This is where Elexon comes into picture where all of this must be commercialized as an activity or contracted out and that's usually done in wholesale markets. Suppliers and generators meet at the marketplace called wholesale market. They decide how much electricity they are buying or selling in future that can be a day ahead, it can be intraday as in same day. And those contracts are then sent to Elexon. And Elexon checks whether they did what they said they would be buying or selling. In UK that is calculated for each *settlement period* which is for every 30-minute time frame. *Net imbalance volume* is the difference, within that given half-hour *settlement period,* between the contracted volumes for suppliers and generators and the actual metered electricity volumes. This has a direct influence on the balancing price of electricity. Imbalances can result in either a surplus or deficit of power and can be caused by generation problems, transportation challenges, or inaccurate predictions. To maintain the efficiency and balance of the electrical system, these discrepancies are calculated and a fee for the imbalances is assigned through the imbalance settlement process.

It is not possible to predict demand or generation and because electricity cannot be stored, the generation cannot be precalculated precisely, but it must exceed the demand on the network. The increase in demand, which again is becoming unpredictable because of the large amount of electrification happening in the country, many people are shifting towards electric vehicles. Many are shifting from gas-based heating to underfloor heating which is largely electric to do the right thing to the environment. The demand aspect is also being unpredicting, resulting in the creation of imbalance. Hence more balancing actions must be taken by the electricity system operators.

1.2 Why Predicting Electricity Prices Accurately Is Essential

The estimation of prices is essential for the energy markets efficiency and stability. It helps the providers to lower operational risks, optimise their purchasing models and provide customers with competitive pricing. It also helps the electricity system operator (national grid) to keep system balanced and to provide steady supply of electricity.

1.3 Impact of renewable energies on electricity prices

Price volatility increases when power market changes and integrates more renewable energy sources. There is a large part of intermittent generation playing a large part or the supplying the pool of generation from renewable energy sources. A large part of renewable generation happens from wind and solar. Accurate forecasts for wind and sun are not always reliable, especially in the UK. This leads to unpredictability in the production of electricity and thereby resulting in price swings. Particularly when demand is low or renewable output is high. Hence it is necessary to see how renewable energies effect the electricity prices so that it could help us in increasing sustainability and to reduce price volatility.

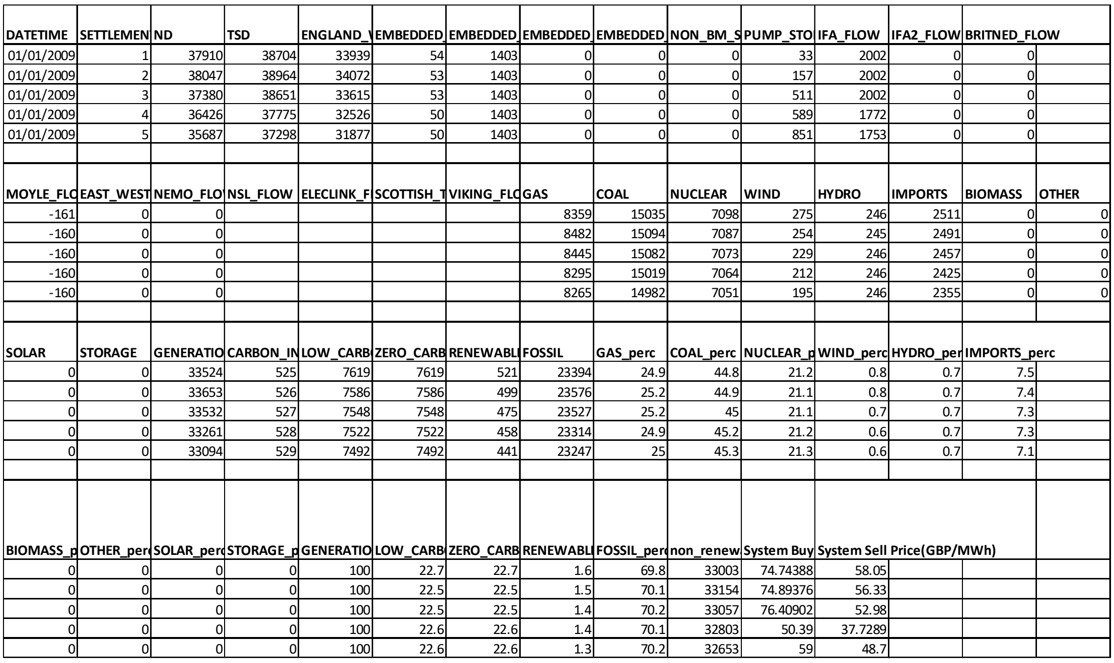
2. Data and exploratory data analysis

2.1 Data Preprocessing

Table 1 consists of data from 1st January 2009 to 31st December 2023. It includes DATETIME column with date along with every settlement period [30-min time frame], and the amount of energy generated by different energy sources in MWH and percentage of generation of each energy source out of total generation, for each settlement period of each day from historic generation mix, demand data and prices excel sheet.

Table 2 consists of data from 1st January 2019 to 31st December 2023. It contains all the columns from table 1, and along with them it also has Net imbalance volume which is the difference between total generation and total demand. The variables present in the tables are defined in the supplementary file (1. Variables Definition) because of the inclusion of sspsbpniv file.

Table 1 pre-processed data from 2009 to 2023.



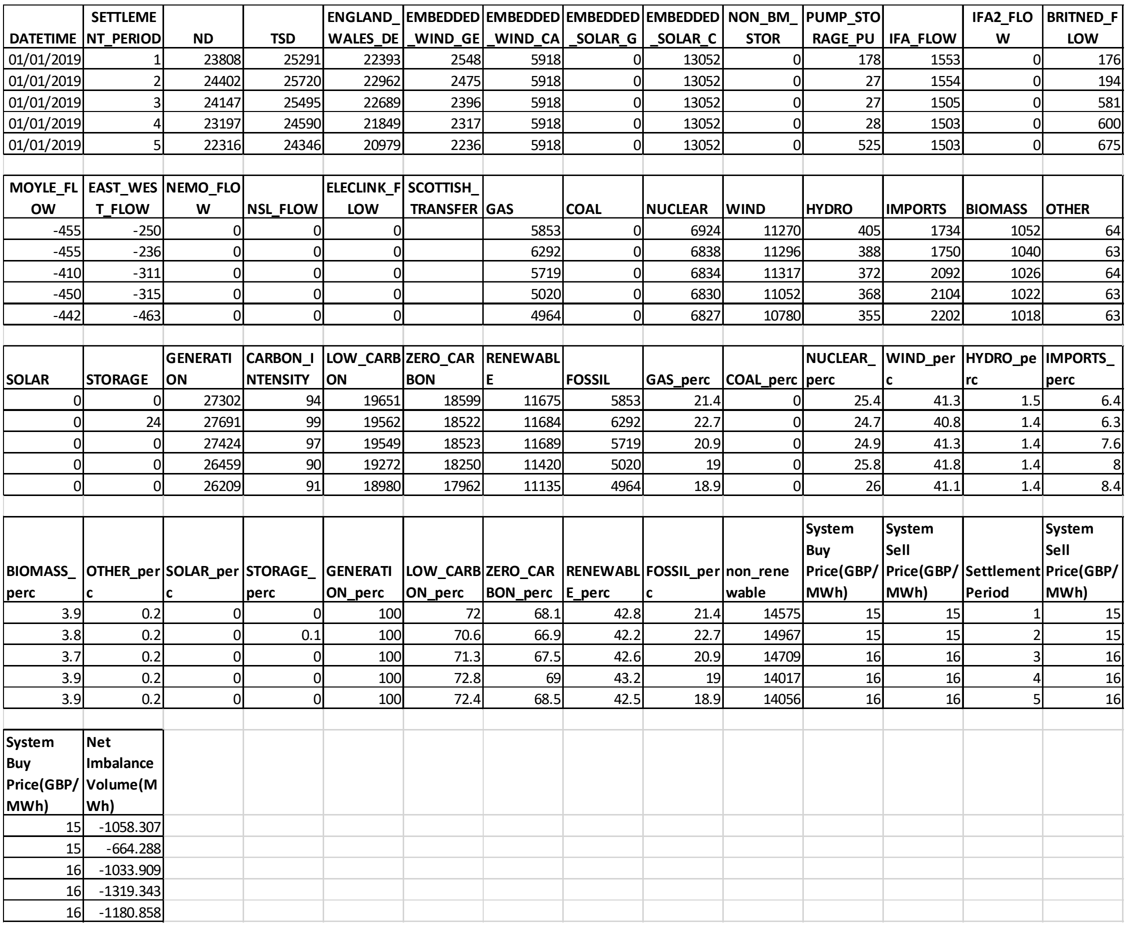
DATETIME	SETTLEMEN	ND	TSD	ENGLAND_\	EMBEDDED_	EMBEDDED_	EMBEDDED_	EMBEDDED_	NON_BM_S	PUMP_STO	IFA_FLOW	IFA2_FLOW	BRITNED_FLOW	
01/01/2009	1	37910	38704	33939	54	1403	0	0	0	33	2002	0	0	
01/01/2009	2	38047	38964	34072	53	1403	0	0	0	157	2002	0	0	
01/01/2009	3	37380	38651	33615	53	1403	0	0	0	511	2002	0	0	
01/01/2009	4	36426	37775	32526	50	1403	0	0	0	589	1772	0	0	
01/01/2009	5	35687	37298	31877	50	1403	0	0	0	851	1753	0	0	

MOYLE_FLC	EAST_WEST	NEMO_FLO	NSL_FLOW	ELECLINK_F	SCOTTISH_T	VIKING_FLC	GAS	COAL	NUCLEAR	WIND	HYDRO	IMPORTS	BIOMASS	OTHER
-161	0	0					8359	15035	7098	275	246	2511	0	0
-160	0	0					8482	15094	7087	254	245	2491	0	0
-160	0	0					8445	15082	7073	229	246	2457	0	0
-160	0	0					8295	15019	7064	212	246	2425	0	0
-160	0	0					8265	14982	7051	195	246	2355	0	0

SOLAR	STORAGE	GENERATIO	CARBON_IN	LOW_CARB	ZERO_CARB	RENEWABLI	FOSSIL	GAS_perc	COAL_perc	NUCLEAR_p	WIND_perc	HYDRO_per	IMPORTS_perc	
0	0	33524	525	7619	7619	521	23394	24.9	44.8	21.2	0.8	0.7	7.5	
0	0	33653	526	7586	7586	499	23576	25.2	44.9	21.1	0.8	0.7	7.4	
0	0	33532	527	7548	7548	475	23527	25.2	45	21.1	0.7	0.7	7.3	
0	0	33261	528	7522	7522	458	23314	24.9	45.2	21.2	0.6	0.7	7.3	
0	0	33094	529	7492	7492	441	23247	25	45.3	21.3	0.6	0.7	7.1	

BIOMASS_p	OTHER_per	SOLAR_per	STORAGE_p	GENERATIO	LOW_CARB	ZERO_CARB	RENEWABLI	FOSSIL_per	non_renew	System Buy	System Sell Price(GBP/MWh)			
0	0	0	0	100	22.7	22.7	1.6	69.8	33003	74.74388	58.05			
0	0	0	0	100	22.5	22.5	1.5	70.1	33154	74.89376	56.33			
0	0	0	0	100	22.5	22.5	1.4	70.2	33057	76.40902	52.98			
0	0	0	0	100	22.6	22.6	1.4	70.1	32803	50.39	37.7289			
0	0	0	0	100	22.6	22.6	1.3	70.2	32653	59	48.7			

Table 2 pre-processed data from 2019 to 2023 along with net imbalance volume.



DATETIME	SETTLEMENT_PERIOD	ND	TSD	ENGLAND_WALES_DE	EMBEDDED_WIND_GE	EMBEDDED_WIND_CA	EMBEDDED_SOLAR_G	EMBEDDED_SOLAR_C	NON_BM_STOR	PUMP_STORAGE_PU	IFA_FLOW	IFA2_FLOW	BRITNED_FLOW
01/01/2019	1	23808	25291	22393	2548	5918	0	13052	0	178	1553	0	176
01/01/2019	2	24402	25720	22962	2475	5918	0	13052	0	27	1554	0	194
01/01/2019	3	24147	25495	22689	2396	5918	0	13052	0	27	1505	0	581
01/01/2019	4	23197	24590	21849	2317	5918	0	13052	0	28	1503	0	600
01/01/2019	5	22316	24346	20979	2236	5918	0	13052	0	525	1503	0	675

MOYLE_FLOW	EAST_WEST_FLOW	NEMO_FLOW	NSL_FLOW	ELECLINK_FLOW	SCOTTISH_TRANSFER	GAS	COAL	NUCLEAR	WIND	HYDRO	IMPORTS	BIOMASS	OTHER
-455	-250	0	0	0		5853	0	6924	11270	405	1734	1052	64
-455	-236	0	0	0		6292	0	6838	11296	388	1750	1040	63
-410	-311	0	0	0		5719	0	6834	11317	372	2092	1026	64
-450	-315	0	0	0		5020	0	6830	11052	368	2104	1022	63
-442	-463	0	0	0		4964	0	6827	10780	355	2202	1018	63

SOLAR	STORAGE	GENERATION	CARBON_INTENSITY	LOW_CARBON	ZERO_CARBON	RENEWABLE	FOSSIL	GAS_perc	COAL_perc	NUCLEAR_perc	WIND_perc	HYDRO_perc	IMPORTS_perc
0	0	27302	94	19651	18599	11675	5853	21.4	0	25.4	41.3	1.5	6.4
0	24	27691	99	19562	18522	11684	6292	22.7	0	24.7	40.8	1.4	6.3
0	0	27424	97	19549	18523	11689	5719	20.9	0	24.9	41.3	1.4	7.6
0	0	26459	90	19272	18250	11420	5020	19	0	25.8	41.8	1.4	8
0	0	26209	91	18980	17962	11135	4964	18.9	0	26	41.1	1.4	8.4

BIOMASS_perc	OTHER_perc	SOLAR_perc	STORAGE_perc	GENERATION_perc	LOW_CARBON_perc	ZERO_CARBON_perc	RENEWABLE_perc	FOSSIL_perc	non_renewable	System Buy Price(GBP/MWh)	System Sell Price(GBP/MWh)	Settlement Period	System Sell Price(GBP/MWh)
3.9	0.2	0	0	100	72	68.1	42.8	21.4	14575	15	15	1	15
3.8	0.2	0	0.1	100	70.6	66.9	42.2	22.7	14967	15	15	2	15
3.7	0.2	0	0	100	71.3	67.5	42.6	20.9	14709	16	16	3	16
3.9	0.2	0	0	100	72.8	69	43.2	19	14017	16	16	4	16
3.9	0.2	0	0	100	72.4	68.5	42.5	18.9	14056	16	16	5	16

System Buy Price(GBP/MWh)	Net Imbalance Volume(MWh)
15	-1058.307
15	-664.288
16	-1033.909
16	-1319.343
16	-1180.858

The code used for combining different data sets to make them suitable for exploratory analysis are placed in in the supplementary file (2.1 combining historic generation mix and demand data)

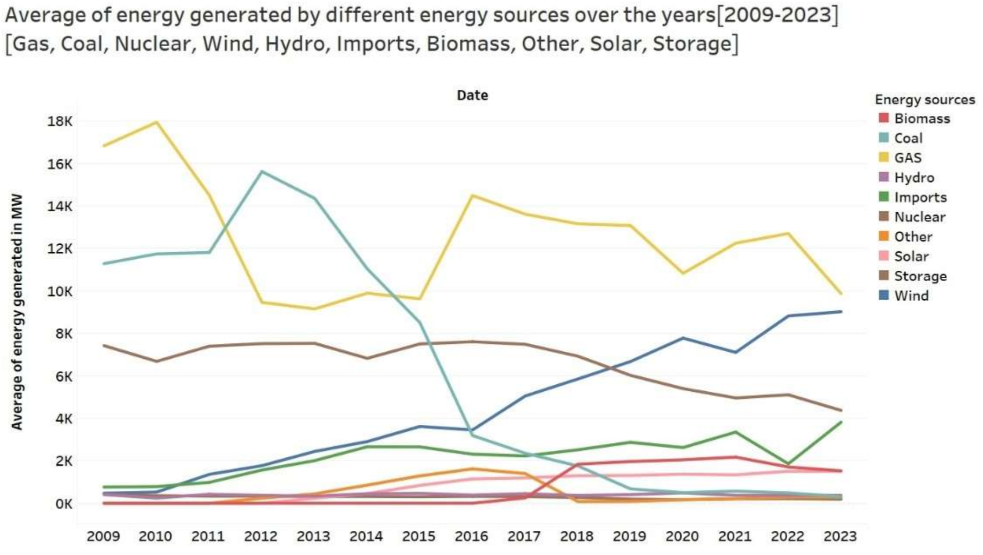
These processed datasets are used for exploratory analysis and building prediction models.

2.2 Tools

The tools used are python and tableau.

2.3 Exploratory Data Analysis of Renewable Energy Impact on Electricity Prices using tableau.

Figure 1 shows us the trend in the average energy generated by different energy sources for each year ranging from 2009 to 2023. That represents the average annual production of energy in MWH from each energy source. For instance, the average quantity of energy produced by petrol in 2009, and so forth. From the plot we could see that over the years gas and coal used to contribute more to the total generation. However, after 2015 the coal declined drastically. This is because of the *large combustion plant directive (LCPD)* and the *industrial Emissions directive (IED)* which required coal plants to either install pollution control equipment or to opt out and shut down the plants. Most coal plants increased their production before they could discontinue their operations. Gas remained to be the largest contributor in the total generation of electricity in UK.



Average of energy generated by different energy sources over the years[2009-2023]
[Gas, Coal, Nuclear, Wind, Hydro, Imports, Biomass, Other, Solar, Storage]

Figure 1 Average of energy generated by different energy sources over the years (2009-2023)

Figure 2 presents a comparison between the average amount of renewable energy generated and the total amount generated between 2009 and 2023. People are becoming more interested in adopting sustainable lifestyles, which can be explained by the rise in usage of renewable energy sources.

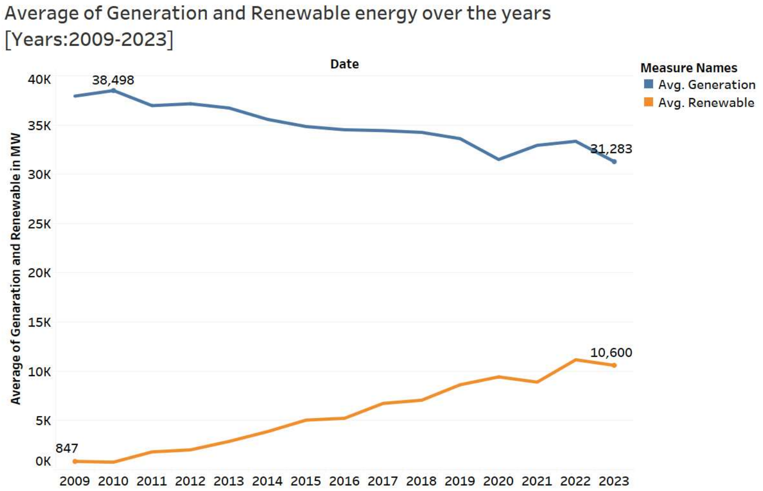


Figure 2 Average of generation and renewable energy (2009-2023)

The average generation of renewable energy sources—wind, solar, hydro, and biomass—from 2009 to 2023 is depicted in Figure 3. Although biomass generation began in 2017, it grew to 4.9% of total power by 2023 because, in contrast to solar energy, biomass can always be converted into biofuel and used to generate electricity. A further factor is the UK government's investment in the conversion of coal-fired power facilities to biomass. Because of the UK's flat terrain and low annual rainfall, there is less potential and resource availability to produce hydropower, which results in lower volumes of hydro energy being created. Since solar energy is, as far as we know, produced by the sun, it can only be produced during the day. Because of that there are irregularities in the generation resulting in high generation during summer months of the year.

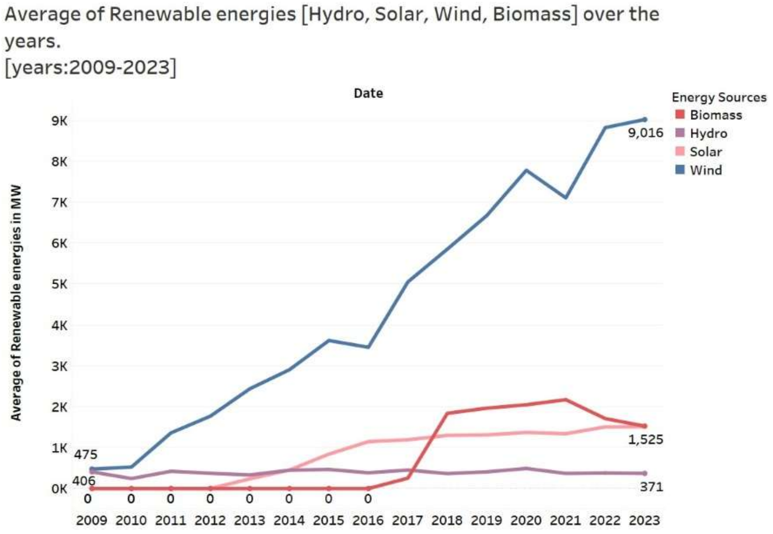


Figure 3 average generation of renewable energy sources [wind, solar, hydro, and biomass] (2009 to 2023)

2.4 The Correlational Weave between Energy and Economics

Analytically creating correlation matrices sheds light on the relationships between the economic counterparts of energy production sources and electricity pricing. The matrix's complex colour gradients convey a narrative of complex variable interaction that go against conventional thinking in the field of energy economics by displaying both positive and negative correlations. The codes for the below generated correlation heat maps are included in the supplementary file (2.2.3. correlation heat map (2009 to 2023) and 2.2.4 correlation heat map (2019 to 2023))

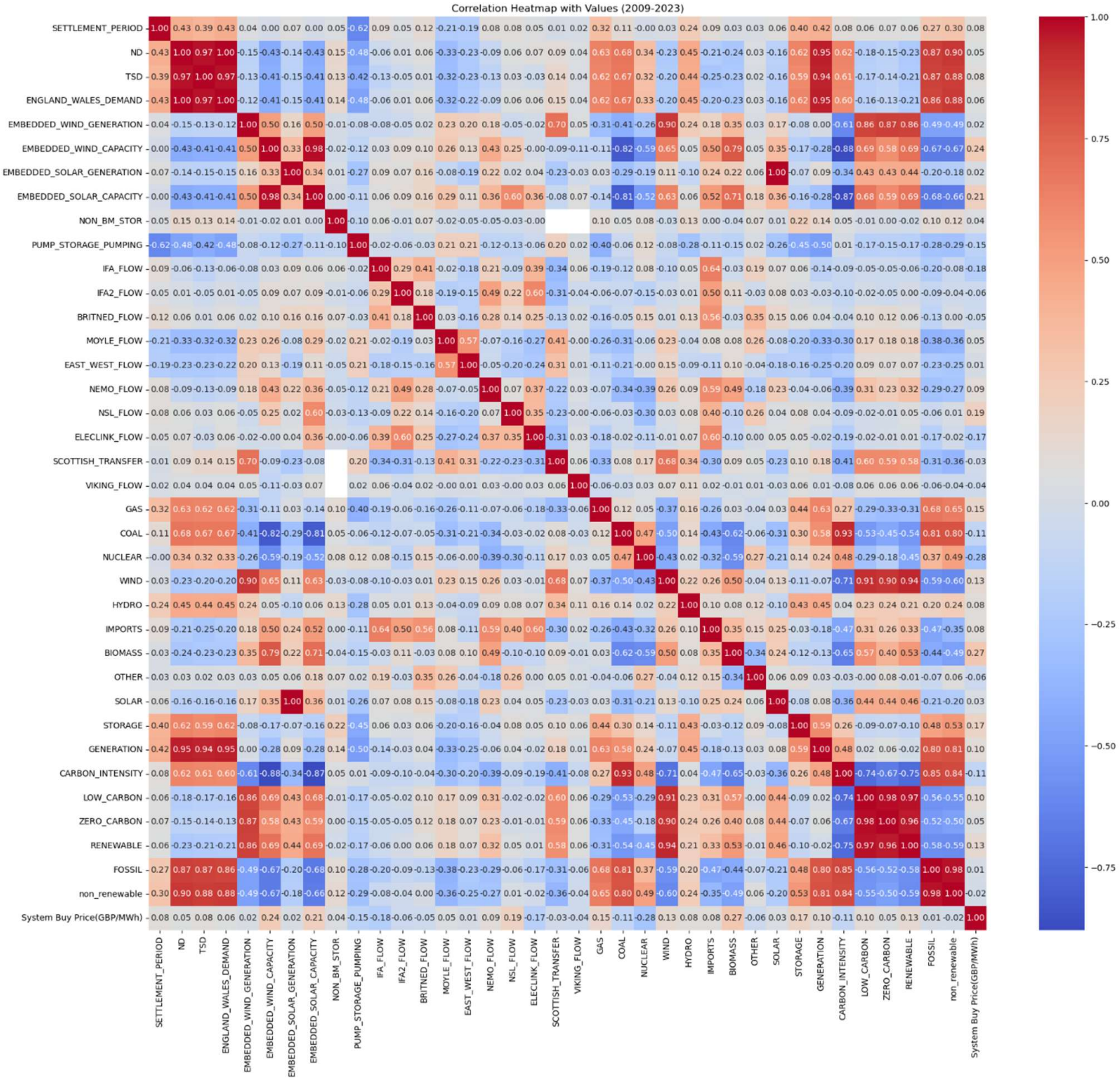


Figure 4 correlation heatmap (2009 to 2023)

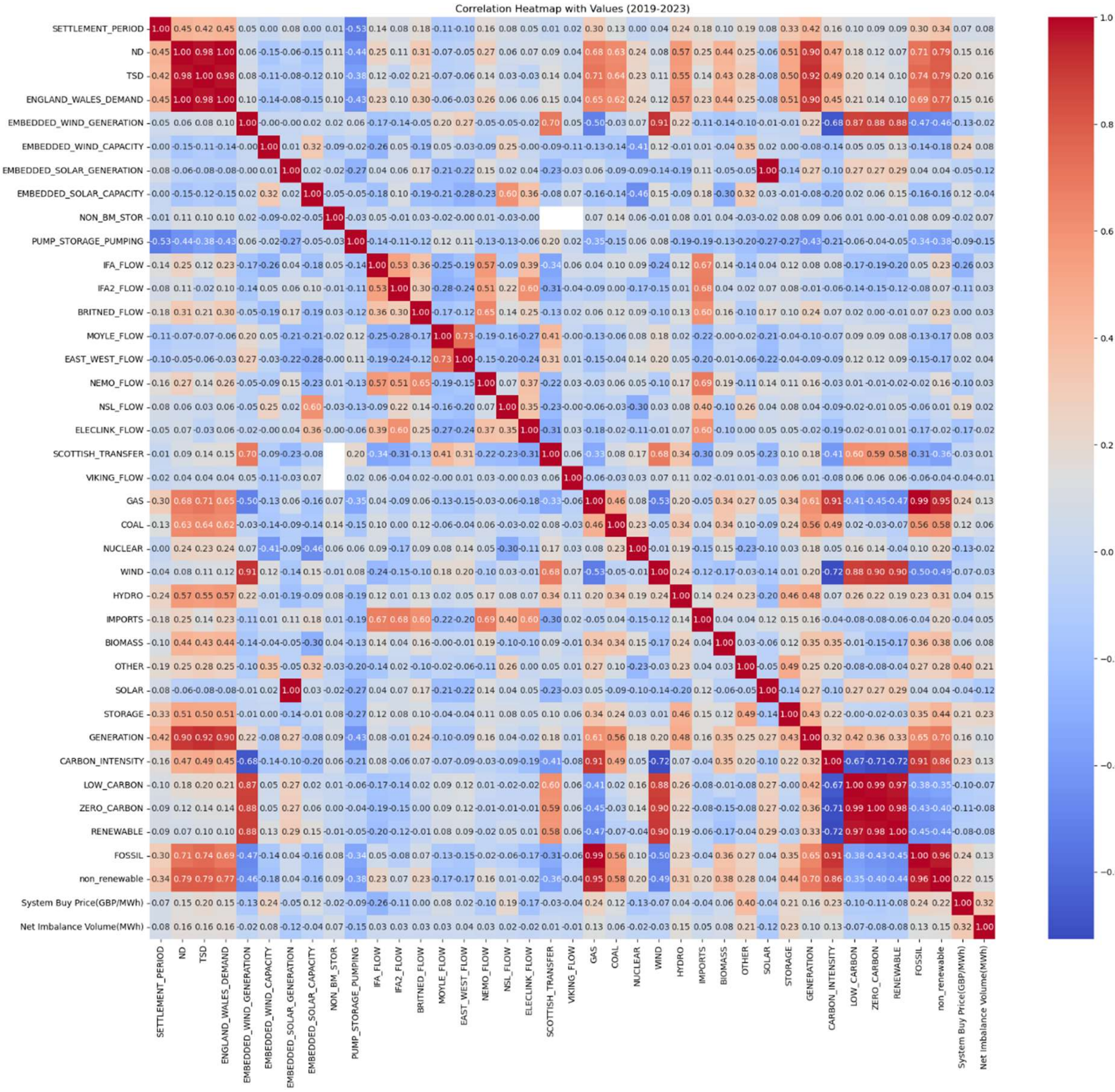


Figure 5 correlation heat map (2019 to 2023)

Figure 4 show the correlation matrix heatmap between different energy sources, flows(interconnectors), Embedded generations, embedded capacities and the system buy price during the years 2009 and 2023. Moderate correlation is shown between EMBEDDED_WIND_CAPACITY (0.24) and EMBEDDED_SOLAR_CAPACITY (0.21).

Figure 5 show the correlation matrix heatmap between different energy sources, flows(interconnectors), Embedded generations, embedded capacities along with net imbalance volume and the system buy price during the years 2019 and 2023. The System Buy Price may tend to rise in accordance with an increase in fossil generation, as suggested by a positive correlation with a variable such as "FOSSIL" (0.24). The negative correlation between "EMBEDDED_WIND_GENERATION" and higher wind generation (-0.13) implies that wind power may be linked to a drop in the system buy price. The strong and positive connection with "Net Imbalance Volume (MWh)" (0.32) suggests that more substantial imbalances in the energy system may result in higher system buy prices.

3 Evaluation metrices

3.1 Mean Absolute Error (MAE):

This is the average of the absolute errors between predicted and actual values. It gives an idea of how big the errors are on average.

MAE = $\frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$

MAE = Mean absolute error
y_i= Prediction
x_i= True Value
N= Total number of data points

3.2 Mean Squared Error (MSE):

This is the average of the squares of the errors. It penalizes larger errors more than MAE does.

MSE =$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$

MSE= Mean Squared error
n= Number of data points
$Y_i =$ Observed Values
$\widehat{Y}_i$= Predicted values

3.3 Root Mean Squared Deviation (RMSD):

Root Mean Squared Deviation also known as Root Mean squared error. This is the square root of the MSE.

RMSD= $\sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$

RMSD=Root mean square deviation
i= variable i
N= Number of non-missing data points
x_i= actual observations time series
$\hat{x}_i$= Estimated time series

3.4 R-Squared (R^2):

This is the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides a measure of how well the model captures the variation in the data.

$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

R^2: The coefficient of determination

y_i: Actual values of the variable being predicted.
$\hat{y_i}$: The predicted values generated by regression model
$\bar{y}$: The mean value of the actual values (y_i)
n: The number of observations or data points

R^2 ranges from 0 to 1, with both 0 and 1 included. The higher the R^2 value tells us that there is better fit between the model and the data.

3.5 Feature importance (Random Forest regressor):

Finds out which feature is mostly used by random forest in taking decision (most influenced feature in decision trees)

$$n_i = \frac{N_t}{N}[impurity - \left(\frac{N_{t(right)}}{N_t} * right\ impurity\right) - \left(\frac{N_{t(left)}}{N_t} * left\ impurity\right)]$$

$$f_{i_k} = \frac{\sum_{j\varepsilon\ node\ split\ on\ feature\ k} n_i}{\sum_{j\ \varepsilon\ all\ nodes} n_i}$$

N_t= number of rows that particular node has

n_i= the total number of rows present in data

Impurity is gini index value.

Gini index=$1 - [(P_+)^2 - (P_-)^2]$

Where P+ is the probability of a positive class and P_ is the probability of a negative class.

$N_{t(righ\ )}$= number of nodes in right node

$N_{t(left)}$= number of nodes in left node

4.Methods and Results:

ND, TSD, ENGLAND_WALES_DEMAND, NON_BM_STOR, GAS, STORAGE, GENERATION, FOSSIL, non_renewable, PUMP_STORAGE_PUMPING, MOYLE_FLOW, EAST_WEST_FLOW, NUCLEAR, OTHER, SOLAR, BIOMASS, WIND, HYDRO, RENEWABLE, are the independent variables and System buy price is the dependant variable that is taken into consideration for building models. This is due to the correlation done during exploratory analysis showing correlation with buy price, as well as to observe the impact of renewable energy sources on buy price.

The correlation heaps maps are included in supplementary file. (2.2.5 correlation heat map (2009 to 2015) and 2.2.6 correlation heat map (2014 to 2018))

4.1 Linear Regression model

Linear regression model tries to fit model into a mathematical representation which helps to predict a response y (dependant variable) with one or more predictor(independent) variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \varepsilon$$

y is the variable that is getting predicted (System Buy price).

x_1, x_2,…… x_n are the independent variables used to predict y.

β_0 is the y-intercept when all the predicted value of y when all the independent variables are 0.

β_1, β_2,…, β_n are the coefficients of independent variables representing the change in the dependent variable y [System Buy price] for one unit change in in the respective x variable, holding all other predictors constant.

ε is the error term, it is the difference between observed and predicted values of y.

4.1.1 Linear Regression [Years trained 2009 to 2015 and predicted January 2016]

The code for this is included in supplementary file (3.1.1).

Linear regression equation:
y = 29.15859 - 0.00108 * GAS + 0.24775 * WIND + 0.23828 * HYDRO + 0.24881 * SOLAR + 0.00000 *
BIOMASS + 0.03147 * STORAGE + 0.00301 * ND + 0.00010 * TSD - 0.00019 * ENGLAND_WALES_DEM
AND + 0.01174 * MOYLE_FLOW - 0.00127 * EAST_WEST_FLOW - 0.36881 * non_renewable + 0.0011
8 * FOSSIL + 0.36602 * GENERATION - 0.00526 * EMBEDDED_WIND_GENERATION - 0.61551 * RENE
WABLE + 0.07779 * NON_BM_STOR + 0.00482 * PUMP_STORAGE_PUMPING

Evaluation metrices:
MAE: 14.346757145776504
MSE: 402.49715983949795
RMSE: 20.06233186445429
R²: 0.17748118969032822

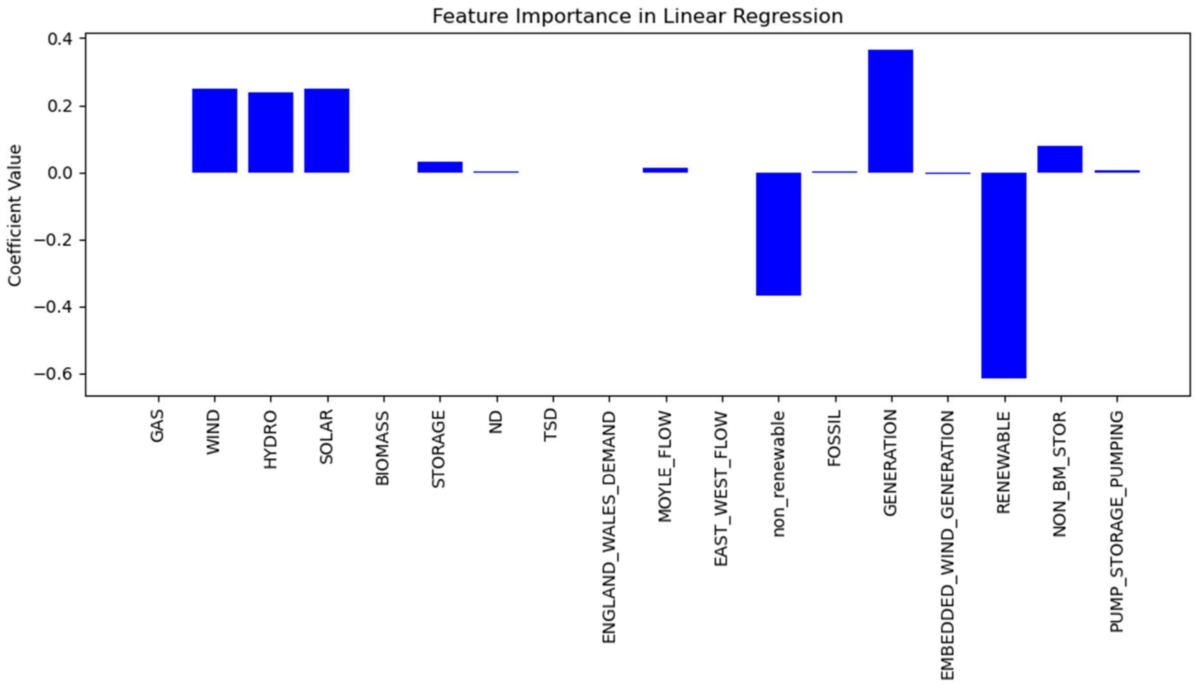


Figure 6 Linear regression: Feature importance while predicting January 2016

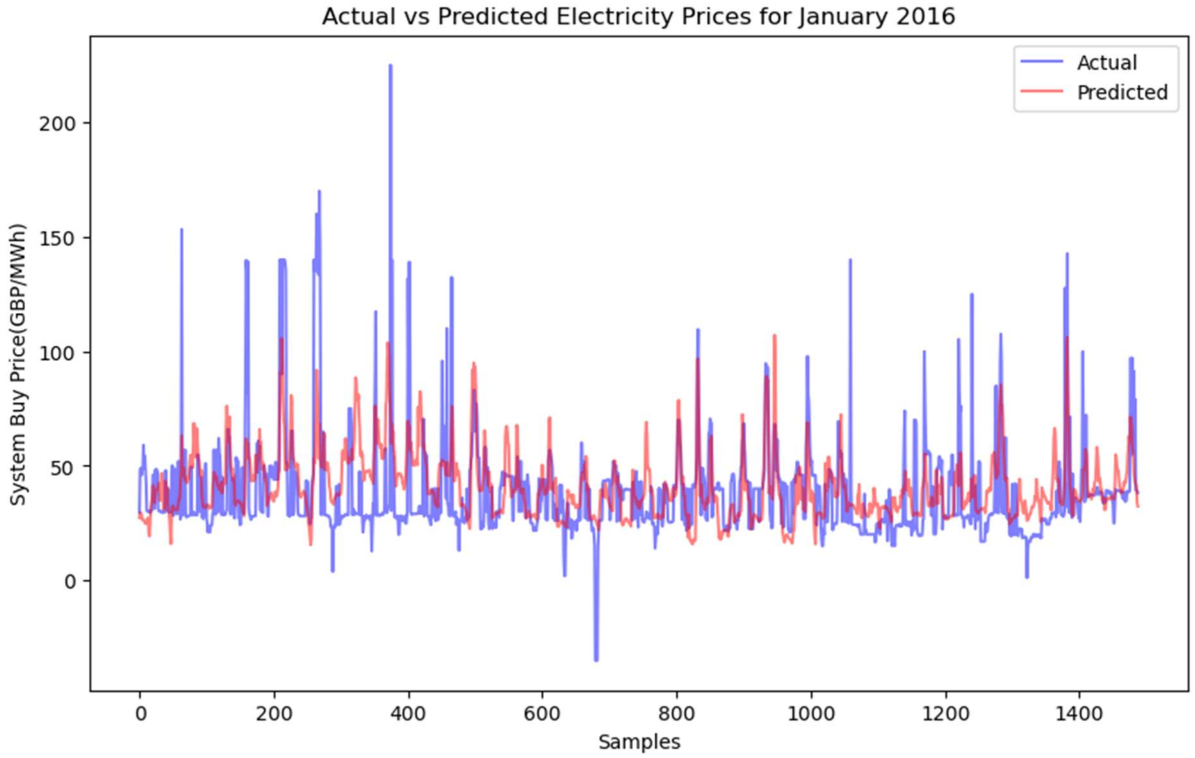


Figure 7 Linear regression actual vs predicted for January 2016

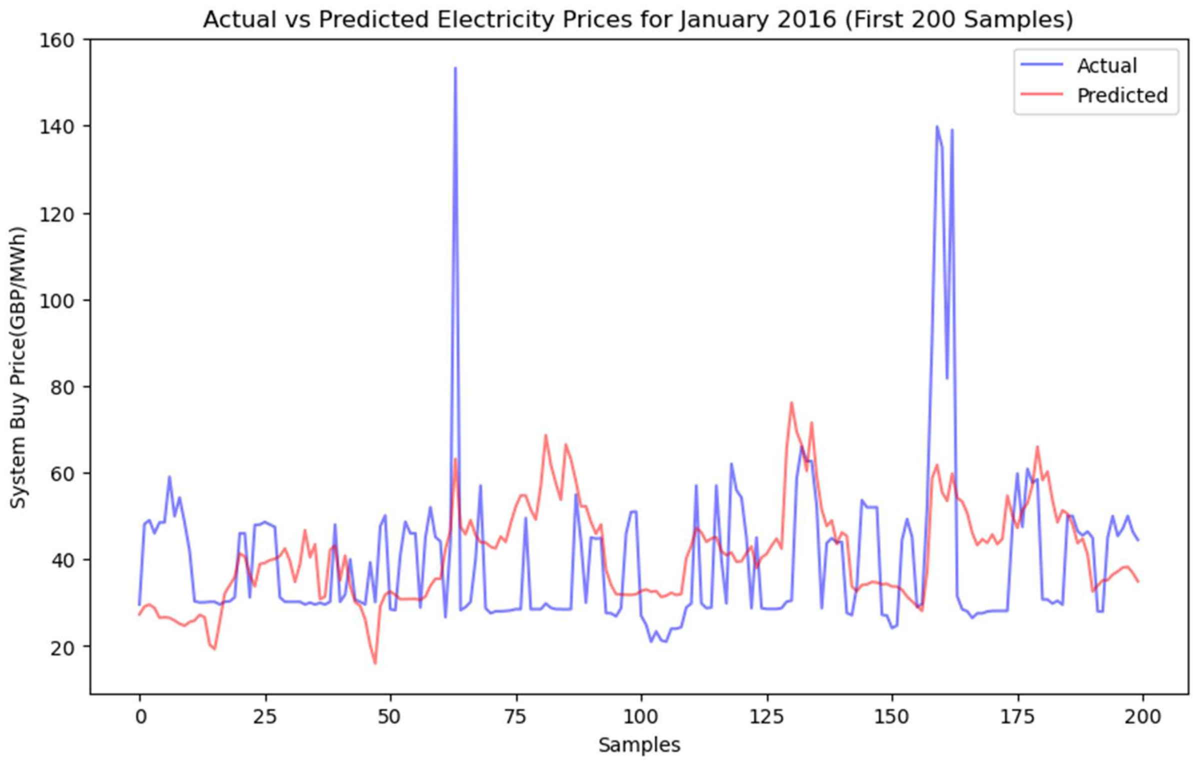


Figure 8 Linear regression actual vs predicted for January 2016(200 samples)

From figure 6 we could see that there is no effect of biomass on the model. This is because there was no biomass generation during this period.

This model suggests that the dependent variable's variation can be predicted by the independent factors to the extent of 17.7%.

4.1.2 Linear Regression [Years trained 2014 to 2018 and predicted January 2019]

The code for this is included in supplementary file (3.2.1).

Linear regression equation:
y = 42.93923 + 0.00481 * ND - 0.00611 * TSD - 0.00080 * ENGLAND_WALES_DEMAND + 0.08570 * NON_BM_STOR - 0.00047 * GAS + 0.02923 * STORAGE + 0.78920 * GENERATION + 0.00312 * FOSSIL - 0.78923 * non_renewable + 0.00477 * PUMP_STORAGE_PUMPING + 0.01008 * MOYLE_FLOW - 0.00152 * EAST_WEST_FLOW + 0.00096 * NUCLEAR + 0.00008 * OTHER + 0.78381 * SOLAR - 0.77927 * BIOMASS + 0.78717 * WIND + 0.78672 * HYDRO - 1.57492 * RENEWABLE

Evaluation Matrices:
MAE: 16.535552377209964,
MSE: 446.63336118562006,
RMSE: 21.133702022731846,
R²: 0.2617454571123552

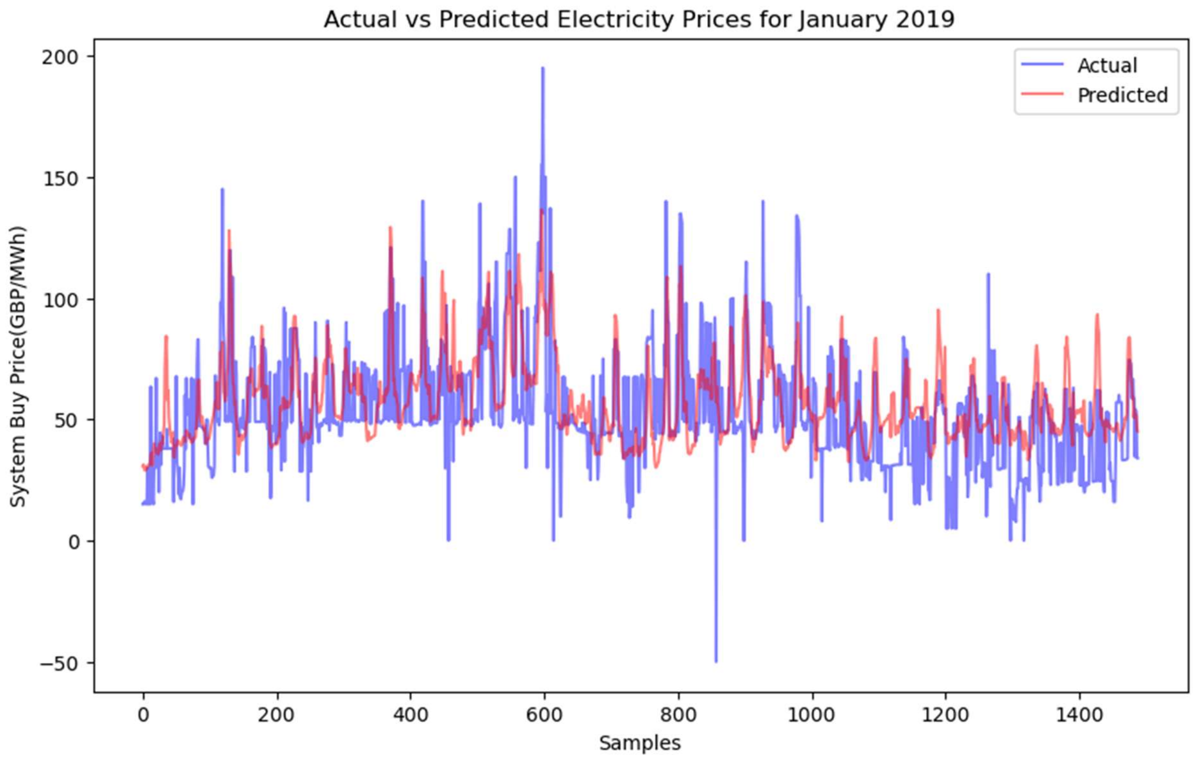


Figure 9 Linear regression actual vs predicted for January 2019

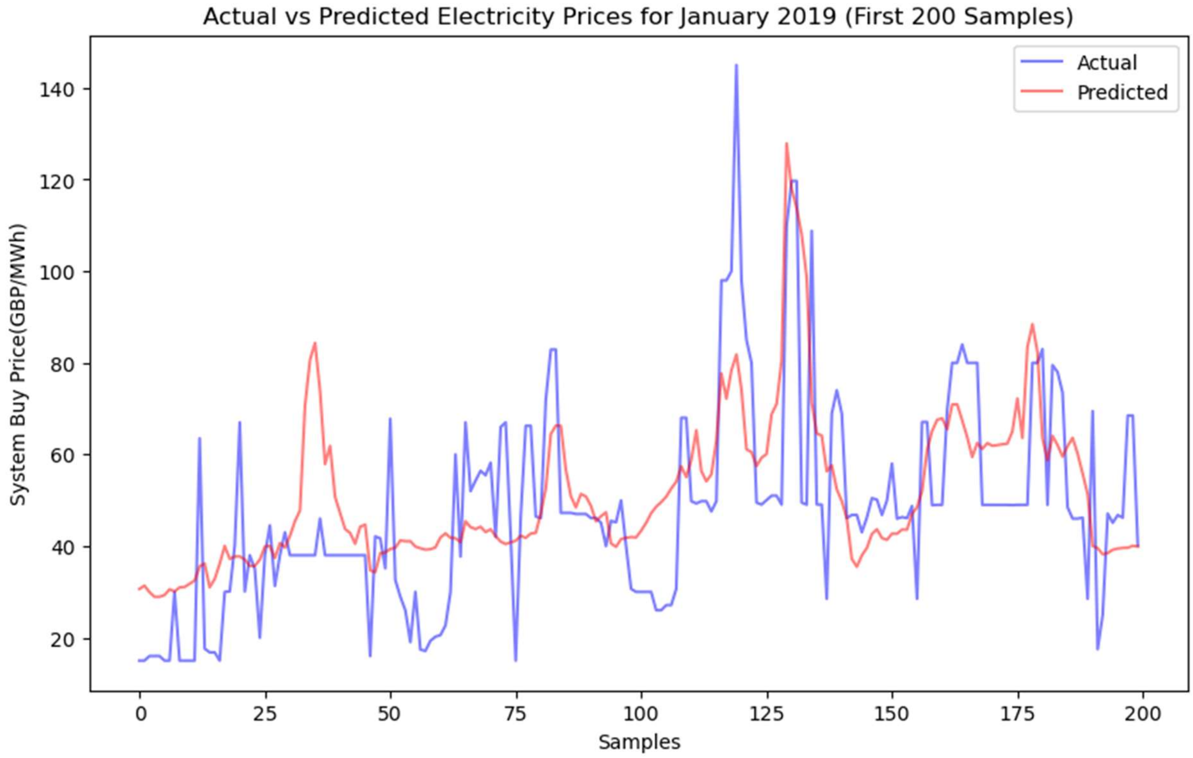


Figure 10 Linear regression actual vs predicted for January 2019(200 samples)

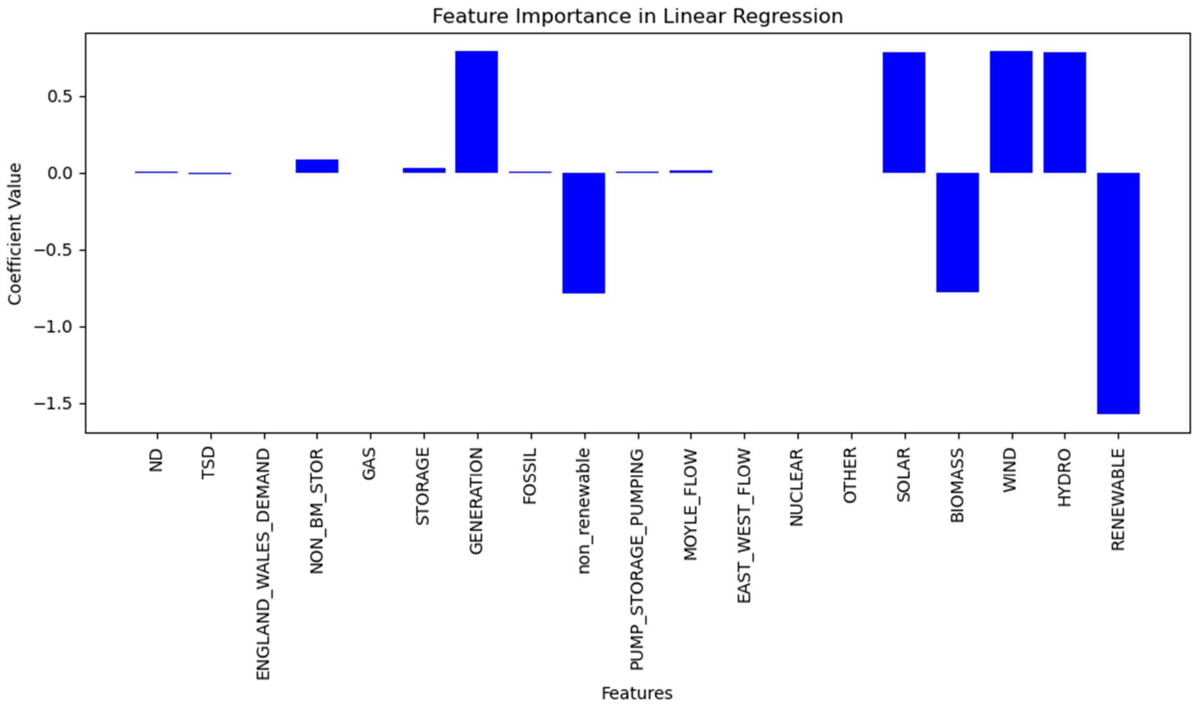


Figure 11 linear regression: feature importance while predicting January 2019

This model suggests that the dependent variable's variation can be predicted by the independent factors to the extent of 26.1%.

When the model is trained with all the data points that is for every settlement period and every da y between the years 2014 and 2018 and predicted January 2019. The model fitted better than the previous one with R²: 0.26. The introduction of Biomass during this period resulted in influencing the model's prediction. Unlike Solar, Biomass can be generated at any time of the day, government start ed to invest more on Biomass generation by converting Coal based stations into biomass stations.

The possible reasons for relatively low performance linear regression model are that, the model assu mes a linear relationship between dependant and independent variables which is not the case here. Multicollinearity problems arises because the independent variables are correlated among themselv es considering the trends, seasonality relations.

4.2 Random Forest model

A random forest is an ensemble learning method that constructs multiple decision trees during training. For regression, it returns the average prediction from all the trees. A random forest model grows many decision trees based on different subsets of the dataset. Each tree is trained on a random sample of data, a method known as bootstrap aggregating, or bagging. When splitting a node during the construction of a tree, the split is chosen from a random subset of features, rather than all features. It considers the predictions made by all the decision tress and give us the final predicted value.

Why Random Forest model

- No overfitting
- High accuracy
- Estimates missing data

4.2.1 Random Forest model [Years trained from 2009 to 2015 and tested January 2016]

The independent variables used are GAS, COAL, WIND, HYDRO, SOLAR, BIOMASS, STORAGE, ND, TSD, ENGLAND_WALES_DEMAND, MOYLE_FLOW, EAST_WEST_FLOW, non_renewable, FOSSIL, GENERATION, EMBEDDED_WIND_GENERATION, RENEWABLE, NON_BM_STOR, PUMP_STORAGE_PUMPING and dependent variable is System Buy price. Similar years used while doing the linear regression model are considered. The model is trained from years 2009 to 2015 and tried to predict January 2016.

Evaluation metrices:
MAE: 14.756, MSE: 429.901, RMSE: 20.734, R²: 0.121

This model suggests that the dependent variable's variation can be predicted by the independent factors to the extent of 12.1%.

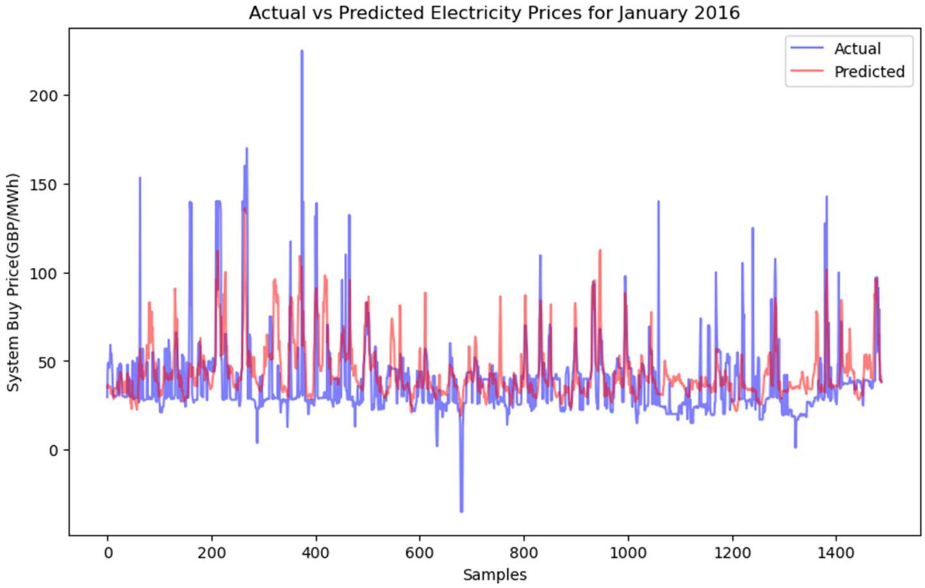


Figure 12 Random Forest: trained from 2009 to 2015 and predicted January 2016

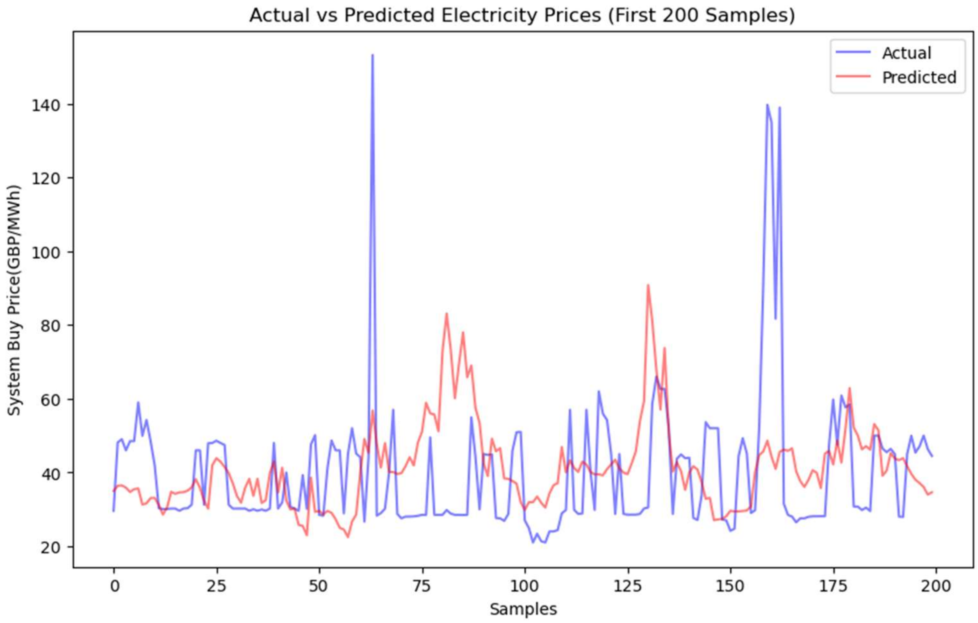


Figure 13 Random Forest: trained from 2009 to 2015 and predicted January 2016(first 200 samples)

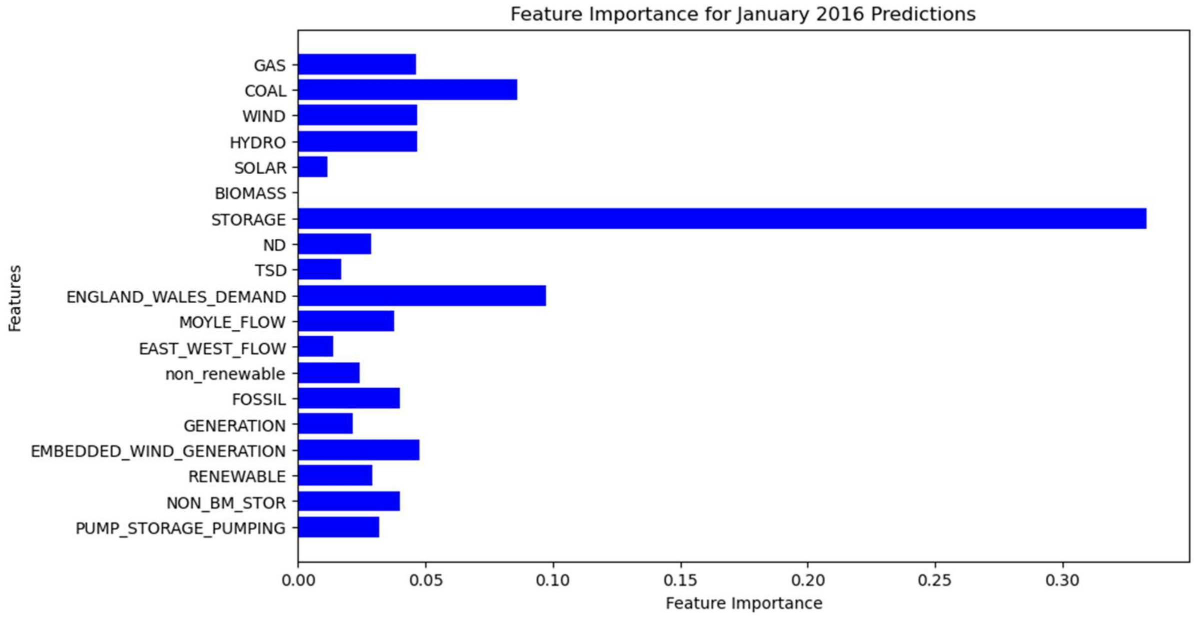


Figure 14 Random Forest feature importance (trained from 2009 to 2015 and predicted January 2016)

Figure 15 represents single decision tree from random forest model. It shows how the model predicts using various input features. The goal of each node is to minimise error by representing a decision point that divides the data according to specific thresholds.

The code for this is included in supplementary file (4.1).

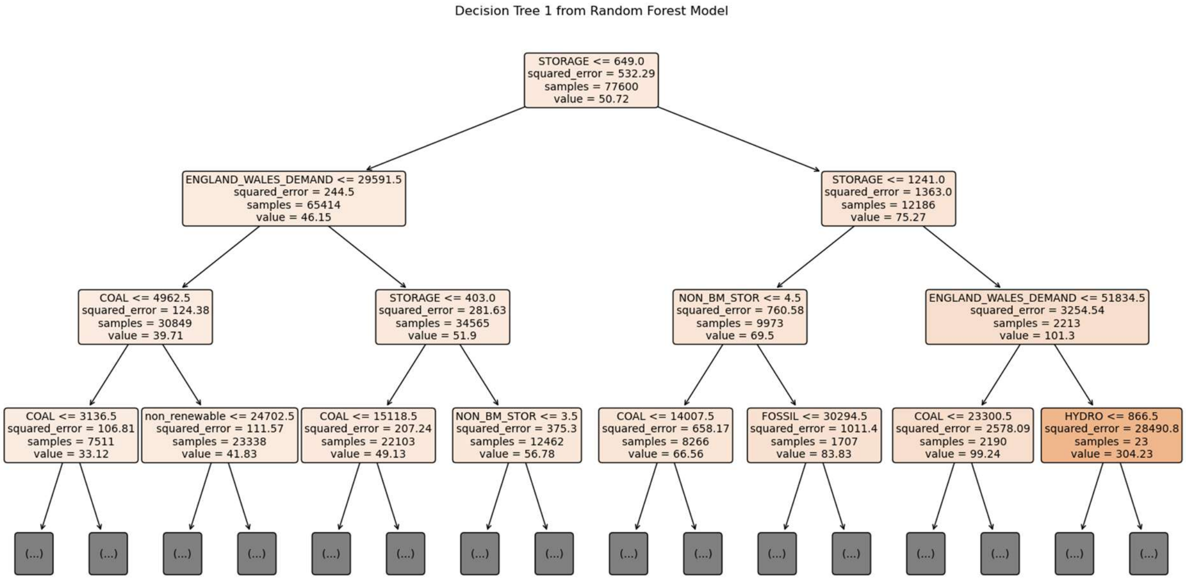


Figure 15 Random Forest (one of the Decision trees among many others)

4.2.2 Random Forest model [Years trained from 2014 to 2018 and tested January 2019]

When the model is trained from 2014 to 2018 and tested for January 2019 the following results are obtained.

The code for this model is included in supplementary file (4.2).

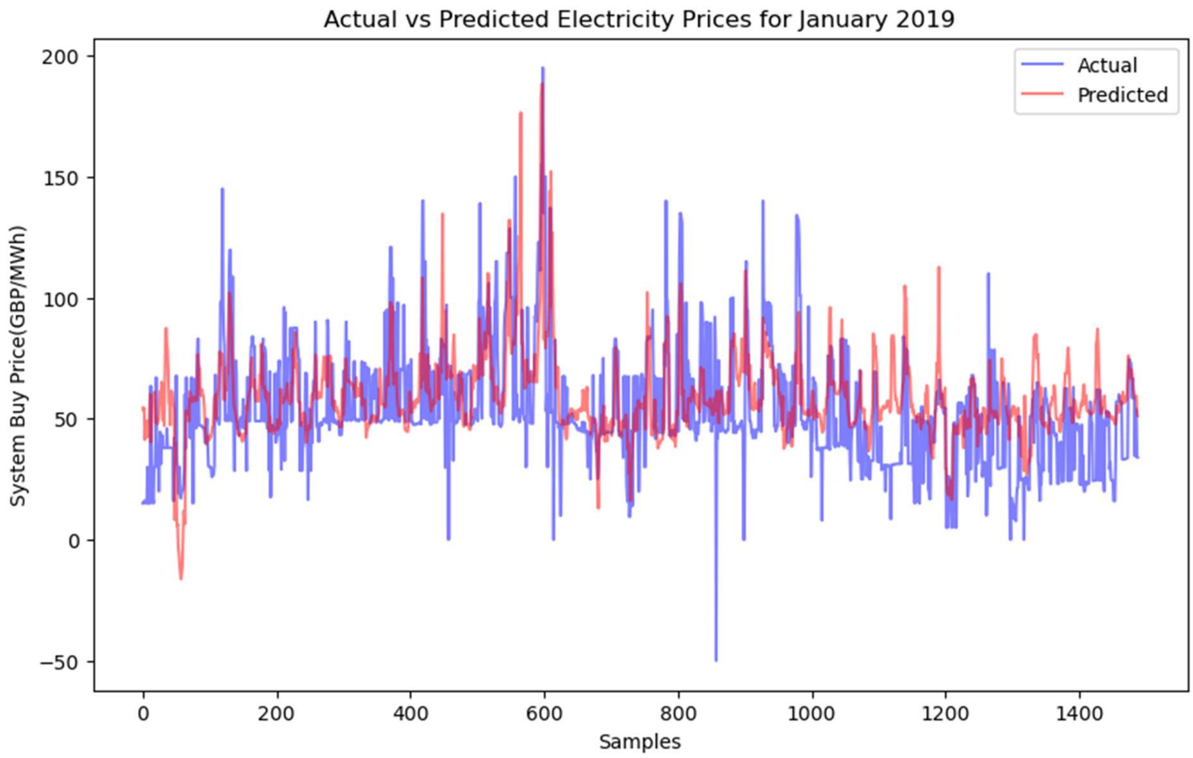


Figure 16 Random Forest: trained from 2014 to 2018 and predicted January 2019

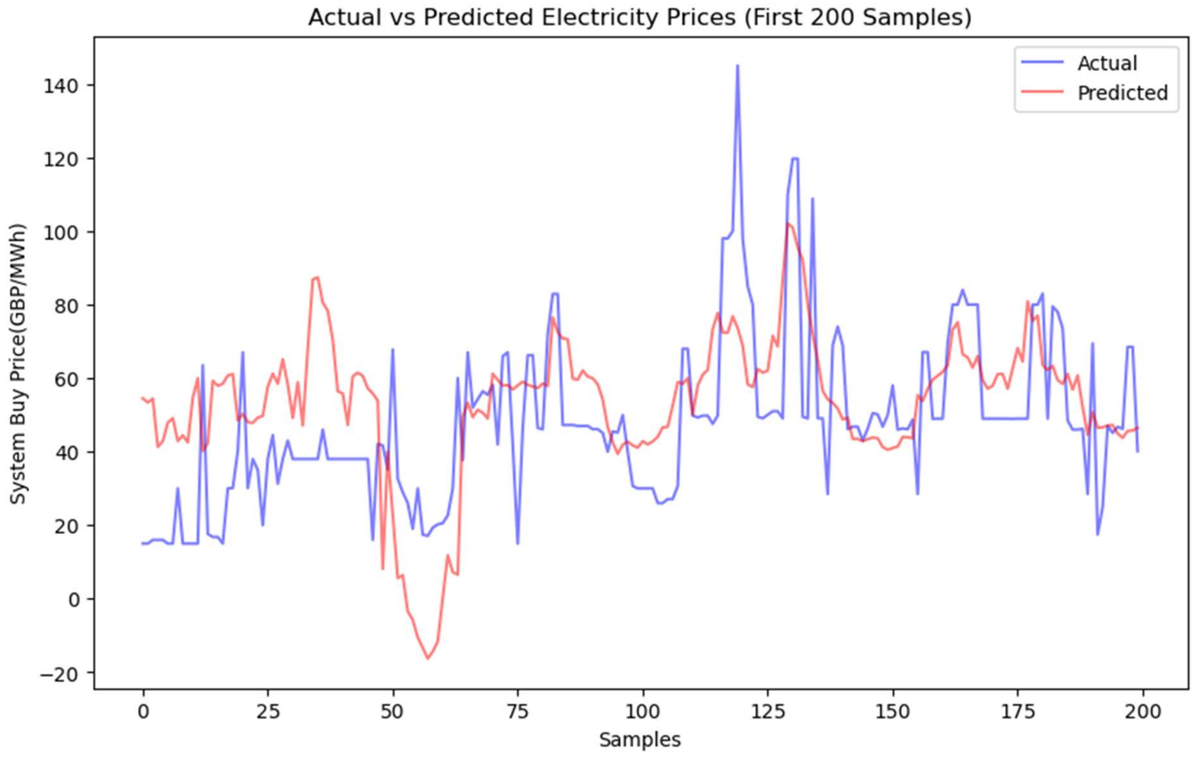


Figure 17 Random Forest: trained from 2014 to 2018 and predicted January 2019(first 200 samples)

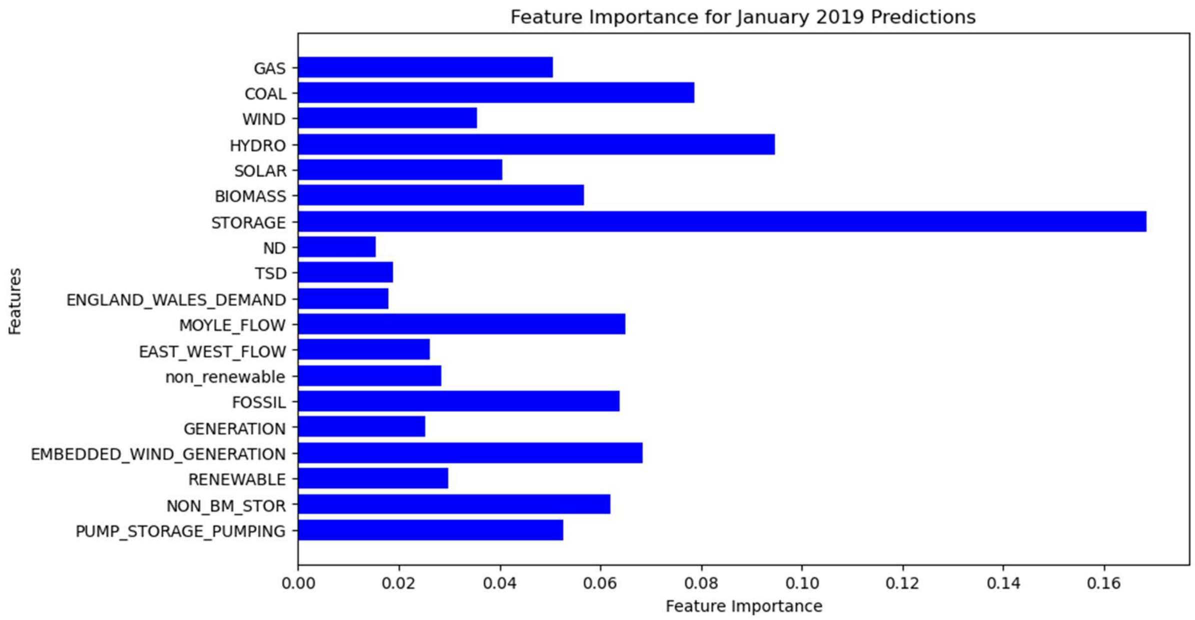


Figure 18 Random Forest feature importance (trained from 2014 to 2018 and predicted January 2019)

Evaluation metrices: MAE: 17.749, MSE: 516.312, RMSE: 22.723, R²: 0.147
This model suggests that the dependent variable's variation can be predicted by the independent factors to the extent of 14.7%.

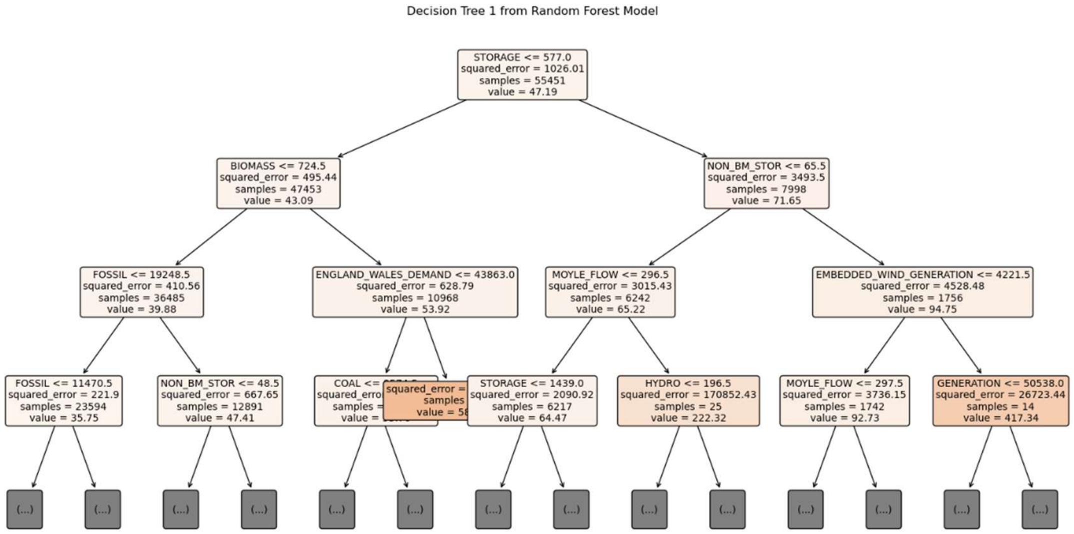


Figure 19 Random Forest (one of the Decision trees among many others) while predicting January 2019

4.2.3 Random Forest model [Trained 70 percent and predicted 30 percent for each year] (2010-2023)

The code for this model is included in supplementary file (4.3).

Table 3 Random Forest evaluation matrices when 70 percent of data is trained and predicted 30 percent for each year (2010 to2023)

Year	MAE	MSE	RMSE	R^2	Top 3 Feature Importances
2010	11.73	480.46	21.92	0.43	STORAGE, COAL, WIND
2011	13.41	262	16.19	0.11	STORAGE, COAL, EMBEDDED_WIND_GENERATION
2012	10.14	219.13	14.8	0.52	STORAGE, non_renewable, HYDRO
2013	11.79	309.72	17.6	0.18	non_renewable, STORAGE, EAST_WEST_FLOW
2014	9.86	333.62	18.27	0.36	STORAGE, TSD, HYDRO
2015	13.26	373.1	19.32	0.22	STORAGE, GAS, NON_BM_STOR
2016	24.04	4788.07	69.2	0.1	NON_BM_STOR, STORAGE, GAS
2017	15.51	434.96	20.86	0.15	STORAGE, SOLAR, PUMP_STORAGE_PUMPING
2018	18.5	568.93	23.85	0.04	COAL, WIND, EMBEDDED_WIND_GENERATION
2019	17.78	488.2	22.1	0.08	STORAGE, non_renewable, GAS
2020	26.33	2167.97	46.56	-0.2	non_renewable, STORAGE, TSD
2021	129.2	45583.6	213.5	-0.9	non_renewable, EAST_WEST_FLOW, STORAGE
2022	95.61	19931	141.2	0.04	COAL, non_renewable, WIND
2023	51.47	4642.6	68.14	-0.5	TSD, COAL, FOSSIL

The effectiveness of Random Forest models in forecasting energy prices is shown in Table 3, which covers the years 2010 through 2023. Notable metrics include MAE, MSE, RMSE, and R^2, as well as the best predictive features for each year. Although negative R^2 values in some years indicate areas for model development or the impact of external factors not included in the model, it also highlights the significance of features like "STORAGE", "COAL", and "WIND".

4.2.4 Random Forest model [Trained 70 percent and predicted 30 percent for each year] along with net imbalance volume (2019-2023)

The code for this model is included in supplementary file (4.4).

Table 4 Random Forest evaluation matrices when 70 percent of data is trained and predicted 30 percent for each year along with net imbalance volume (2019 to 2023)

Year	MAE	MSE	RMSE	R^2	Top 3 Feature Importances
2019	10.76	205.84	14.35	0.61	Net Imbalance Volume (MWh), FOSSIL, non_renewable
2020	17.76	1403.85	37.47	0.23	Net Imbalance Volume (MWh), non_renewable, TSD

2021	128.7	44312.3	210.5	-0.9	non_renewable, STORAGE, Net Imbalance Volume (MWh)
2022	62.28	13057.1	114.3	0.37	Net Imbalance Volume (MWh), COAL, non_renewable
2023	45.38	4113.38	64.14	-0.3	Net Imbalance Volume (MWh), ND, FOSSIL

Together with FOSSIL and COAL, the Table 4 shows how important Net Imbalance Volume (MWh) is for projecting energy costs, especially in 2019 and 2020 and 2023. Variable R^2 values over time demonstrate the model's varying accuracy, with substantial success in 2019 and difficulties in 2021, suggesting the need for model improvements or the impact of certain market situations.

4.3 Gradient Boosting

Gradient boosting is an ensemble strategy that combines predictions from multiple models to increase overall forecast accuracy.
It operates in a step-by-step additive fashion, with each subsequent model fixing the mistakes of the preceding ones.

Important elements of enhancing gradients:
Loss function: Calculates the variation between expected and observed values.
Weak learners: straightforward models (decision trees, for example) that make few assumptions about the data.
Additive model: An ensemble of weak learners is the ultimate prediction.

Regression, classification, and multi-class issues can all be solved with gradient boosting.

4.3.1Gradient boosting regressor model [Years trained from 2009 to 2015 and tested January 2016]

The code for this model is included in supplementary file (4.5).

Evaluation metrices: MAE: 14.820319038381186, MSE: 412.5278216528103, RMSE: 20.310780921786595, R²: 0.1569831369224669

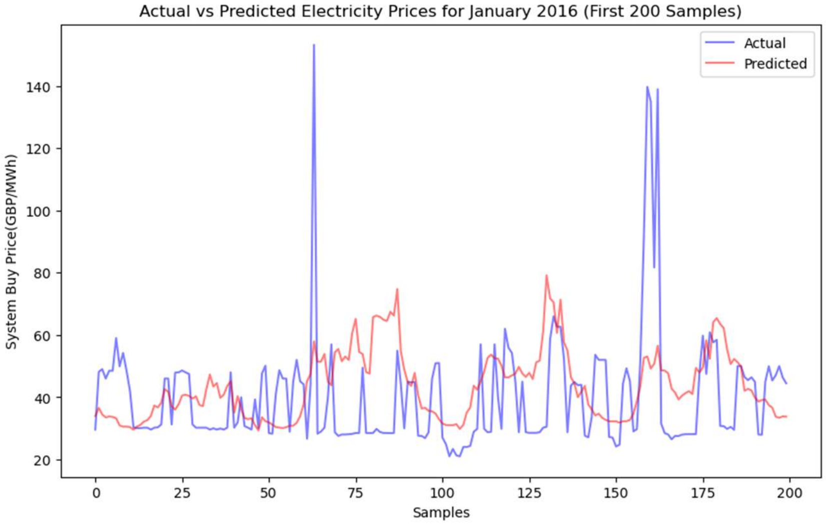


Figure 20 Gradient boosting trained from 2009 to 2015 and tested January 2016(first 200 samples)

4.3.1Gradient boosting regressor model [Years trained from 2014 to 2018 and tested January 2019]

The code for this model is included in supplementary file (4.6).

Evaluation metrics: MAE: 17.483657101928973, MSE: 517.4618773420171, RMSE: 22.74778840551356, R²: 0.14467074133285107

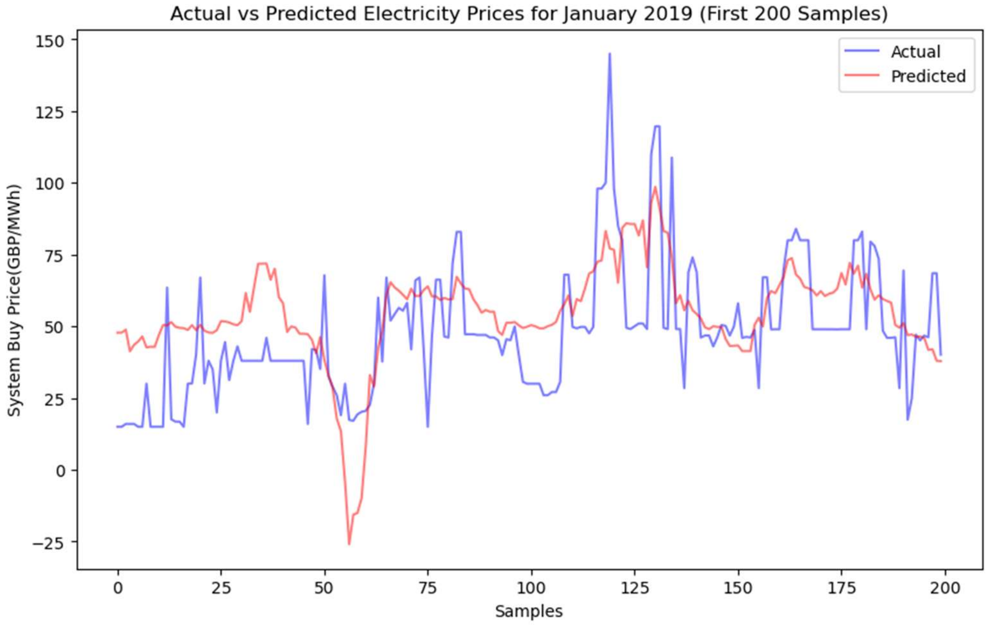


Figure 21 Gradient boosting trained from 2014 to 2018 and tested January 2019(first 200 samples)

5. Seasonal prediction using Random Forest model, Gradient boosting regressor and Linear regression.

The new analysis, which looks at the various seasons of 2015, demonstrates how well three different regression models—Linear Regression, Random Forest Regression, and Gradient Boosting Regressor—perform in projecting power prices in a nuanced manner. For Spring, Summer, Autumn, and Winter, each model's performance metrics—Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2)—are carefully put together to enable a thorough comparison.

The code for these models is included in supplementary file (5).

Gradient boosting regressor:

Season	**Mean Absolute Error (MAE)**	**Mean Squared Error (MSE)**	**Root Mean Squared Error (RMSE)**	**R-squared (R²)**
Spring	9.1	208.94	14.45	0.204
Summer	10.12	203.73	14.27	0.148

Autumn	11.09	259.82	16.12	0.158
Winter	10.46	215.92	14.69	0.34

Random forest Regressor:

Season	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R-squared (R^2)
Spring	8.86	210.52	14.51	0.198
Summer	9.52	198.84	14.1	0.169
Autumn	10.63	231.82	15.23	0.249
Winter	11.1	236.42	15.38	0.278

Linear regression:

Season	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R-squared (R^2)
Spring	10.83	252.79	15.9	0.037
Summer	9.67	195.42	13.98	0.183
Autumn	9.76	182.57	13.51	0.409
Winter	10.35	210.68	14.51	0.356

The models of Gradient Boosting and Random Forest exhibit noteworthy uniformity throughout the seasons, highlighting their resilience and dependability in identifying intricate patterns present in the data. Conversely, while performing inconsistently in the spring and summer, linear regression significantly improves in the autumn and winter. The R^2 values, which indicate the percentage of the dependent variable's variation that can be predicted from the independent variables, show how much the improvement has improved. Because seasonal changes in Autumn and Winter more closely match the linear assumptions of the model, higher R^2 values in these seasons indicate that Linear Regression fits the data better throughout these seasons.

6.Observations and Discussion

This thesis tries to close the gap between the difficulty of precisely forecasting power costs in the UK and the uncertain nature of renewable energy sources. To traverse the complicated dynamics of power pricing in the context of rising renewable energy integration, we used models such as gradient boosting, random forest, and linear regression. Although every model had distinct advantages and disadvantages, our research adds to the larger conversation on energy economics, especially when it comes to predictive modelling and the effect of renewable energy on market stability.

Previous studies have looked at this problem from several angles. Deep learning networks may be used to estimate electricity prices, according to Zhang, Cheema, and Srinivasan (2018) [2]. This suggests that sophisticated machine learning methods may be able to identify the intricate patterns present in energy markets. By contrasting machine learning techniques with conventional statistical models, our study broadens this viewpoint and highlights the distinct capabilities of each technique over a range of time scales.

Because renewable energy sources like solar and wind power are intermittent, it has been observed that their incorporation into the electricity markets has increased price volatility. The authors of Aitor Ciarreta, Blanca Martinez, and Shahriyar Nasirov (2022) [3] stressed the need of having reliable forecasting techniques that can adjust quickly to changes in demand and generation capacity. Their findings are consistent with our investigation of seasonal fluctuations in prediction accuracy, which shows how seasonal variations can have a big impact on model performance.

Lago, Marcjasz, De Schutter, and Weron (2021) [4] offered a thorough analysis of cutting-edge algorithms for projecting electricity prices for the next day, highlighting significant issues and areas in need of development. Their suggestion to use ensemble techniques and external factors to increase accuracy is in line with our methodology, especially when it comes to our usage of gradient boosting and random forest models, which combine several forecasts to increase reliability.

Our research supports the notion put forth by a number of writers [1-4] regarding the significance of comprehending the financial effects of integrating renewable energy. Because renewable energy sources fluctuate, new forecasting models that can handle the inherent uncertainty are needed to maintain market stability and efficiency. Our results highlight the complexity of forecasting energy prices and the crucial role of renewable energies in influencing market dynamics, with varied degrees of prediction accuracy across models and seasons.

The investigation of how to forecast UK electricity prices in the face of fluctuation in renewable energy produces important new information for the field of energy economics. Low R-squared values were the result of the linear regression model's struggles with the non-linear intricacies of the data, particularly when taking new energy sources and regulatory changes into account. On the other hand, the random forest model was adaptable in handling extensive datasets and robust against overfitting; its forecast accuracy varied yearly because of external market factors including policy modifications and technological breakthroughs. The unstable dataset and the changing energy situation posed additional difficulties for the gradient boosting regressor, which sought to improve accuracy by fixing earlier mistakes. The significance of taking temporal dynamics into account was highlighted by seasonal analysis, which also suggested that models customised for could increase accuracy.

To further improve forecasting accuracy and market stability, future research should keep investigating the incorporation of external variables, the creation of more complex modelling tools, and the use of real-time data analysis.

7.References:

[1] https://www.elexon.com/about-elexon/

[2] W. Zhang, F. Cheema, and D. Srinivasan, "Forecasting of Electricity Prices Using Deep Learning Ne tworks," 2018 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), Kota Kinaba lu, Malaysia, 2018, pp. 451-456, doi: 10.1109/APPEEC.2018.8566313.

[3] Aitor Ciarreta, Blanca Martinez, Shahriyar Nasirov, "Forecasting electricity prices using bid data," I nternational Journal of Forecasting, 2022, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2022. 05.011.

[4] Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, Rafał Weron, "Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices, and an open-access benchmark," Appli

ed Energy, Volume 293, 2021, 116983, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2021.116983.

[5] https://www.elexonportal.co.uk/category/view/175?cachebust=yvxb0b7syw

[6] https://www.nationalgrideso.com/data-portal/historic-generation-mix/historic_gb_generation_mix

[7] https://www.nationalgrideso.com/data-portal/historic-demand-data?page=1

[8] https://www.tableau.com/trial/what-is-tableau

[9] M. Shahidehpour, H. Yamin, Z. Li, "Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management," IEEE Press, Wiley-Interscience, 2002.

[10] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," International Journal of Forecasting, Volume 30, Issue 4, 2014, Pages 1030-1081, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2014.08.008.

[11] S. Fan, R. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," IEEE Transactions on Power Systems, Volume 27, Issue 1, 2012, Pages 134-141, doi: 10.1109/TPWRS.2011.2159218.

[12] P. Pinson, H. Madsen, "Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models," Journal of Forecasting, Volume 31, Issue 4, 2012, Pages 281-313, https://doi.org/10.1002/for.1237.

[13] K. Zhou, S. Yang, Z. Shao, "Electricity price forecasting with confidence-interval estimation through an extended ARIMA approach," IEEE Transactions on Power Systems, Volume 31, Issue 1, 2016, Pages 782-791, doi: 10.1109/TPWRS.2015.2409119.

[14] L. G. Papageorgiou, E. S. Fraga, "Optimisation of electricity market bidding strategies for power producers," International Journal of Electrical Power & Energy Systems, Volume 29, Issue 3, 2007, Pages 230-238, ISSN 0142-0615, https://doi.org/10.1016/j.ijepes.2006.08.005.

[15] N. Amjady, F. Keynia, "Day-ahead price forecasting of electricity markets by a new fuzzy neural network with market condition analysis," IEEE Transactions on Power Systems, Volume 25, Issue 4, 2010, Pages 1784-1794, doi: 10.1109/TPWRS.2010.2045883.