Electricity Consumption and Pricing Prediction Using Machine Learning

Ashima Arora
The University of Texas at Austin

ABSTRACT

The ability to accurately forecast electrical energy consumption values is an important factor for government and energy companies in planning for power production and maintenance. Apart from giving a competitive edge to electrical companies, it is also critical for improving the overall energy efficiency and reducing gas emissions in the environment. This research used an hourly energy generation, pricing and weather dataset from Kaggle containing energy load and pricing values from 2015 to 2019. Since electrical energy generation is influenced by many other factors, this paper aims to analyze the influence of weather information, as well as the generation power of other energy sources, in forecasting energy consumption and prices. It also compares the efficiency and precision of different machine learning models, such as random forests, gradient boosting, recurrent neural networks, and long short-term memory network models (LSTM), in forecasting energy load and prices.

1. Introduction

Electricity consumption is constantly increasing in all sectors of the economy, such as residential, industrial and commercial. Apart from being sensitive to weather conditions, its generation is also influenced by macroeconomic factors, such as generation of other sources of energy, such as wind, coal, and solar, etc. (U.S. Energy Information Administration, 2023). Due to the unpredictability involved in consumer behavior and climate conditions, accurately forecasting the energy demand is a difficult endeavor. Since it cannot be economically stored, the power transmission system constantly requires a balance between its production and consumption.

Today, various models have been applied to forecast electrical consumption and pricing, ranging from statistical and economic models such as supply and demand models to more complex and hybrid models. In this paper, we analyze the impact of weather information on the correctness of energy load and pricing forecasts, and then assess the use and efficiency of different machine learning models, such as random forests, gradient boosting, recurrent neural networks, and long short-term memory network (LSTM) models, in their ability to produce precise predictions for the energy demand as well as pricing. Precision in forecasting electrical demand and pricing reduces the risk of over or underestimating the revenues from production units, which in turn has significant implications for profits, market shares and ultimately shareholder value (NOVA School of Science and Technology & Energias de Portugal (EDP), 2022).

2. Research Background

Energy load forecasting focuses on generating insights into the electrical energy needed to meet future electricity demand. It has been categorized into three types: i) short-term, ii) medium-term, and (iii) long-term forecasting. Generally, short-term forecasting refers to forecasting intra-day and day-ahead demand, medium-term forecasting deals with predicting one week to several months ahead, and long-term forecasting looks one or more years ahead. (National Renewable Energy Laboratory & Lawrence Berkeley National Laboratory, 2023). Traditionally, electricity consumption and pricing models have been of 4 categories: i. Statistical, ii. Fundamental or Structural, iii. Machine Learning, and iv. Hybrid. Statistical methods use a mathematical combination of historical values and external factors, such as weather variables or consumption and production figures, to forecast the current values for energy consumption and pricing. These methods rely on linear regression to represent their target variable as a linear combination of the independent variables. Fundamental or Structural models aim to capture the physical and economic relationships present in the day-ahead market using fundamental factors such as fuel prices, wind generation or temperature values. Many of these are proprietary models used by the energy markets, whose details are often not disclosed (NOVA School of Science and Technology & Energias de Portugal (EDP), 2022).

Machine learning based methods use machine learning techniques to forecast the energy consumption and pricing values. Among the various machine learning methods, gradient boosting (particularly XGBoost) and neural networks (particularly RNNs) are of increasing interest. Hybrid models are extremely complex frameworks that combine different algorithms for feature selection and clustering data, as well as combining more than one forecasting model and using different optimization models to estimate the models or their hyper-parameters (Poggi et al., 2023). This paper aims to explore the machine learning based techniques used for energy load and price forecasting, and perform a comparative analysis of their ability to make correct predictions for the target variables of interest, i.e., energy consumption and price. The performance of these models was evaluated on four metrics: Mean Absolute Error (MAE), R-squared (R2) score, Root Mean Squared Log Error (RMSLE) and Root Mean Squared Error (RMSE).

3. Data and Variables

To forecast the total energy load consumption values using machine learning techniques, this paper uses the Hourly energy demand generation and weather data from Kaggle. The data is a combination of two datasets. The first dataset contains the electrical consumption, generation and pricing data collected over four years for Spain. It was obtained from ENTSO-E Transparency Platform, which contains a collection of electricity generation, transportation and consumption data information for the pan-European market. The pricing information was obtained from Red Electric España. The second dataset contains the weather data for Spain over the same time period. This was obtained from the OpenWeather Weather API.

The dataset contains hourly electrical consumption data and the corresponding consumption and pricing forecast values by the Transmission Service Operator (TSO). The energy dataset contains 29 columns containing information about the date and time of the recorded values, quantity of biomass, fossil fuel, coal, oil, peat, geothermal, hydro, sea, nuclear, solar, waste, wind and other generation values in megawatts. It also contains forecast values for the solar, offshore and onshore wind generation. Along with these, it has the forecast values of the electrical demand and pricing values (EUR/MWh), as well as the actual electrical demand and pricing values. The weather dataset contains the same date and time values as the energy dataset, and records the observed values for the minimum and maximum temperature (on Kelvin scale), pressure (in hPa), humidity (percentage), wind speed (m/s), wind direction, rain (mm), cloud (percent) and snow (mm).

3.1 Problem Formulation

Using the described dataset, the goal of this research is multi-fold:

1) Investigate the impact of weather elements on forecasting the total electrical energy consumption and pricing values, along with its influence on generation of other energy sources, such as solar or fossil gas, etc.
2) Investigate the performance of different machine learning techniques in accurately predicting the energy consumption and pricing values. We compare performance between Random Forests, Recurrent Neural Networks, LSTM, and Gradient Boosting techniques. The research aims to evaluate the performance of these techniques by investigating their impact on the overall Mean Absolute Error (MAE), R-squared (R2), Root Mean Squared Log Error (RMSLE) and Root Mean Squared Error (RMSE) scores.

3.2 Pre-Processing and Feature Extraction

3.2.1 Missing Values

The energy dataset contained several columns with missing values. Out of the 29 total, 2 columns had 100% missing values and 6 columns had 0% missing values. Out of the remaining 21 columns, 1 column had roughly 0.1% missing values while the remaining 20 columns had 0.05% missing data. For the 21 columns with 0.05 to 0.1% missing data, the missing values in these columns were imputed with a measure of the central tendency, i.e., by the mean values of each respective column. Data imputation for numeric features can help preserve relationships between any missing values and other features (Zhang, 2016). Of these 27 columns (21 columns that were imputed and 6 columns without any missing values) that were preserved, 6 columns were completely filled with zeros. Generally, data analysis techniques involving multiple variables (multivariate) require effective strategies for imputing or replacing the zero values in any columns of the sparse matrices generated from your data (Kiers et al., 2012). In this data, these columns with only zero values were swaying the outlier detection mechanism, as well as causing issues in using logarithmic operations. Hence, these 6 columns were dropped.

3.2.2 Outlier Detection

Detection and removal of outliers in the data was done using Z-scores. The Z-score method is used to represent items that are 'abnormal' with respect to the standard deviation and mean of the data (Venkata et al., 2019). The Z-Score Method measures how many standard deviations away a data point is from the mean of its data (Field, 2023). In this research, any rows in the dataset with a Z-score greater than or equal to 3 were considered to contain an outlier and were filtered out from the dataset.

3.2.3 Data Normalization

For this research, the data was normalized using min-max normalization, and then scaled to contain values between 0 and 100. It transforms each value by subtracting the minimum of the column from the value and then dividing by the range, i.e., the difference between the maximum and minimum value in the column (Aksu et al., 2019). The result was multiplied by 100 to scale the values to fit the desired range.

3.2.4 Correlation Analysis and Feature Extraction

One of the research questions was aimed at analyzing the impact of the weather data on the energy consumption, generation and pricing. The weather data contains the data of five cities in Spain. Some features in the weather dataset, such as 'weather_icon' or 'weather_id' were deemed irrelevant to our correlation analysis and hence were removed. Also, the weather data had several duplicate records for the same timestamp. To resolve this, the data from the latest timestamp was chosen. To analyze the correlation between our target variables and weather data, the data from each city was extracted and merged with the energy dataset on the common timestamp variable to create a merged dataset of weather and energy features.

After removing outliers and normalizing the merged dataset using the techniques explained above, a correlation analysis was performed. It was found that compared to the energy generation metrics, information about the weather elements was of little importance to the prediction of energy consumption and pricing. Out of the weather elements, the humidity and maximum temperature seemed to be the only factors with a relatively moderate absolute correlation value of 0.20 with the energy consumption values. For predicting the price, wind speed had a moderate impact, with a correlation value of -0.25.

When it comes to energy generation values, there seems to be an expected correlation between temperature and humidity on generation of solar energy, with correlation values of 0.37 and -0.38 respectively. Similarly, the wind speed correlated with generation of onshore wind with a correlation of 0.30. Pressure correlates with a score 0.23 to the generation of other renewable sources and waste generation. Other factors, such as wind degree, rain, snow and clouds, had no significant impact on the energy generation values. It is important to note that for the purpose of this analysis, data was filtered on the city of Madrid. It is, however, representative of the observations from other cities, as they also showed similar correlations. Since none of these correlation scores were greater than our threshold value of abs(0.4), weather elements were not used for the remainder of this research.

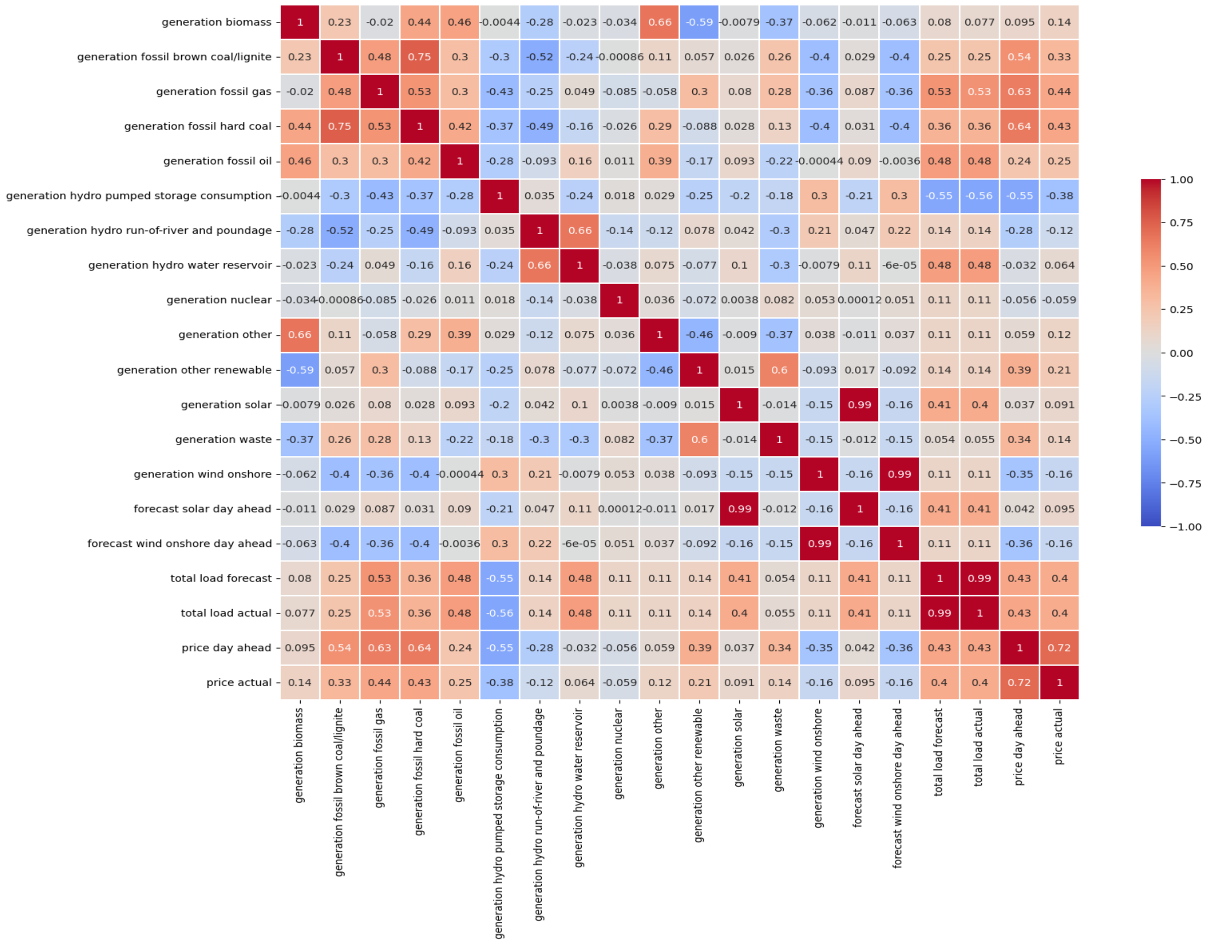

Figure 1: Feature Correlation Matrix

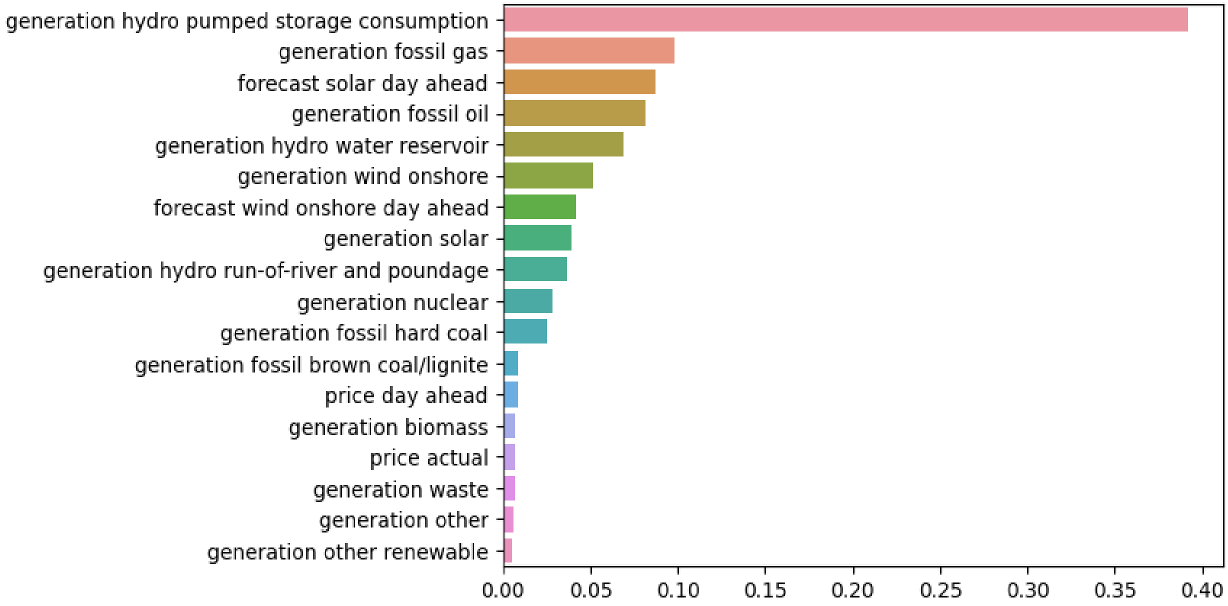

Figure 2: Feature Importance Scores for Electricity Consumption (Random Forest)

3.2.4.1 Forecasting Energy Consumption

Since the 'total load actual' variable is our target variable, the 'total load forecast' variable generated by the TSO is an explanatory variable in this regression model, i.e., it is highly predictive of the target variable, with the correlation score of 0.99 (Figure 1). Hence, this column was dropped from the set of independent variables (features). From the remaining 18 columns (after excluding the timestamp column which is used for indexing), only those features with a correlation score greater than abs(0.4) were selected. Pearson's correlation can effectively detect linear relationships; however, it is not reliable to identify nonlinear relationships between two variables (Deebani, 2020). Therefore, the feature importance values generated by a Random Forest (Figure 2) were also consulted. Features that received a feature importance score greater than 0.05 via the Random Forest selection method also met the selection threshold formed for the correlation matrix, with the exception of 'price day ahead', which contains the forecast price values that received a correlation coefficient of 0.43 but received a fairly low feature importance score.

Through this process, we extracted 6 numeric independent variables or features that will be used to predict our target variable. Using this technique, columns that would introduce multicollinearity were also avoided. Multicollinearity among independent variables can obscure the identification of important effects of such variables on the target variable, given the overlapping information shared between them. (Rahbar, 2016)

3.2.4.2 Forecasting Price

In this case, the target variable is ‘price actual’. Since the ‘price day ahead’ feature contains forecast pricing values that are highly predictive of the target variable, this column was dropped from the set of features. From the remaining 18 columns (after excluding the timestamp column used for indexing), there were only 4 features (Figure 1) with a correlation score greater than or equal to abs(0.4): ‘generation fossil gas’, ‘generation fossil hard coal’, ‘total load forecast’, and ‘total load actual’. The feature importance values generated by a Random Forest (Figure 3) were consulted to include features that received a feature importance score greater than 0.045. Through this process, 10 numeric independent variables or features were extracted that will be used to predict our target variable.

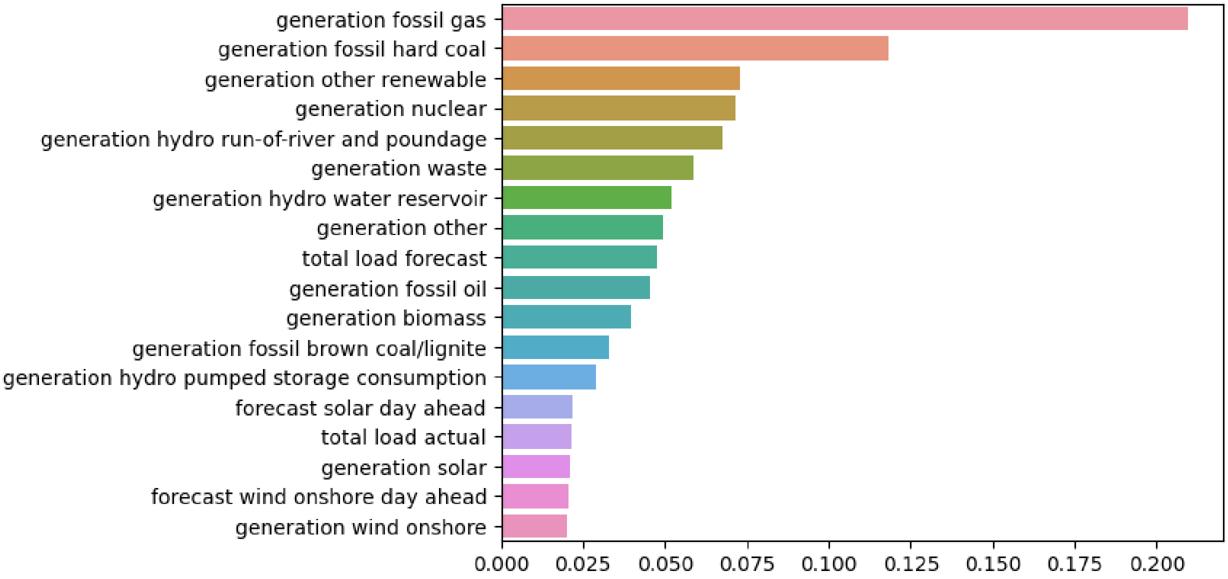

Figure 3: Feature Importance Scores for Electricity Price (Random Forest)

4. Research and Methods

4.1 Statistical Time Series Methods

Traditionally, statistical time series methods, specifically AutoRegressive Integrated Moving Average Model (ARIMA), have been used for time series forecasting tasks like forecasting energy consumption and pricing. ARIMA is a linear regression based forecasting model that combines the process of auto-regression (AR) and moving averages (MA) to build a combined time-series model (Schmid et al., 2023). The term AutoRegressive (AR) implies that it is a regression model with lagged values of the target variable until some value of the parameter p in the past. A stationary time series is beneficial for accurate time series forecasting, since it has statistical properties that are constant over time. Its properties don't depend on the time at which the series is observed. Integrated (I) defines a parameter d which controls the number of times the difference is taken into account before the original series becomes stationary. Moving Average (MA) is defined by the parameter q, which also uses a regression based model, but with lagged values of the past forecast errors (Bora, 2021). For this research, p was chosen to be 5; d was chosen to be 0 since both the energy load (Figure 5) and pricing data (Figure 6) is stationary; and q was chosen to be 1.

4.2 Machine Learning Methods

To understand the effectiveness of machine learning methods in predicting total electrical energy consumption and pricing values, we compare the performance of four different ML methods - Random Forest, Gradient Boosting, Recurrent neural networks, and LSTMs. The performance was evaluated by comparing the Mean Absolute Error, R-squared, and Mean Squared Error scores.

4.2.1 Linear Regression

As a baseline, a linear regression model was used, since it is a simpler model in comparison to the more complex ensemble based models. Multiple variable linear regression assumes that the target variable Y depends on its independent variables X_1, X_2,…, X_n linearly. Its equation can be written as

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

Where Y is 'total load actual' for forecasting energy consumption and 'price actual' for forecasting energy pricing, X_1 through $X_n$represent the independent variables or features, and β_1 through $\beta_n$represent the slope of the equations for the variable of X (Mustapa et al., 2020).

4.2.2 Random Forest Regression

Random Forest is an ensemble based learning algorithm that utilizes the technique of bagging. It trains an ensemble of decision trees parallelly, with bootstrapping followed by aggregation. Bootstrapping trains several individual decision trees in parallel on different subsets of the training dataset, using various different subsets of the available features (Misra et al., 2019). The results of the predictions are then aggregated to get the overall prediction at the end. For this research, the number of trees in the random forest (n_estimators) was 100, while all the other parameters, such as maximum tree depth or number of leaf nodes, were all set to their default values provided by the scikit-learn library modules.

4.2.3 Gradient Boosting

Gradient Boosting is another ensemble based method, which combines multiple relatively simple and weak models into a stronger ensemble. Boosting methods like gradient boosting iteratively add a new model to the existing ensemble (Natekin & Knoll, 2013). This new model is trained with respect to the error of the whole ensemble learnt until that iteration. The new decision tree added aims to minimize the loss function (Touzani et al., 2018). For the purpose of this research, 6 numeric features were used to predict the total energy load, and 10 were used to predict energy pricing. The number of trees in the gradient boosting regressor (n_estimators) was chosen to be 100, while all the other parameters, such as shrinkage rate, learning rate (0.1) and maximum depth, were kept as defaults provided by the scikit-learn library.

4.2.4 Recurrent Neural Network

Recurrent Neural Networks (RNNs) are artificial neural networks with loops inside it. Neural networks have been of interest for forecasting energy load values and pricing as an alternative to auto-regressive moving average (ARIMA) models, because they can model non-linear time series information (Connor et al., 1994). The loops allow information to be passed across different timesteps. For this research, the RNN used 3 recurrent layers, each with 40 activation units and a tanh activation function. Between two recurrent layers, there was a dropout layer with a dropout rate of 0.15. This is followed by a dense output layer. For training, the Adam optimizer was used with a learning rate of 0.001 for 10 epochs on a batch size of 32. The model was trained to minimize the Mean Squared Error for its loss function.

For training an RNN model, the 'time' variable providing timestamp data for each hour (Figure 3) was used to index into our target variable 'total load actual' or 'price actual' for forecasting energy consumption or pricing respectively. A sequence length of 10 was used to generate sequences containing energy demand or pricing values of the last 10 hours (intra-day) so we could utilize this historic data to forecast current values. This model is no longer using only regression-based methods which model the correlation between the dependent and independent variables while relying on the reliability of those independent variables (Abdallah et al., 2022). Rather, it is doing a time-series analysis to predict energy demand or pricing values. Time-series analysis involves statistically analyzing a target variable that is generated repeatedly at regular intervals (i.e., it is sequential data) over a large number of observations. Such an analysis can show how the target variable changes over time, and can be used to forecast its future values based on some historical data (Velicer & Fava, 2003). RNNs are capable of handling varying lengths of sequential data and temporal patterns within such data effectively (Buhl, 2023).

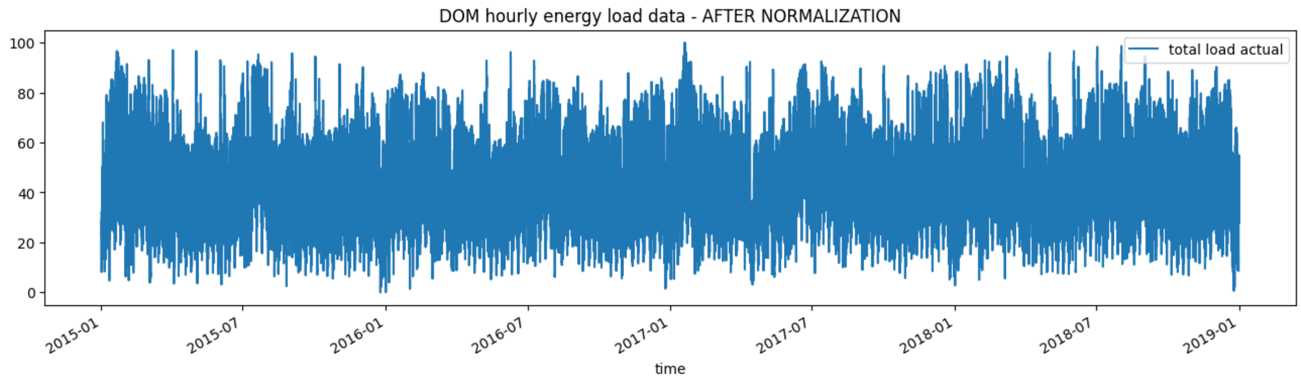

Figure 5: Hourly electrical energy load data

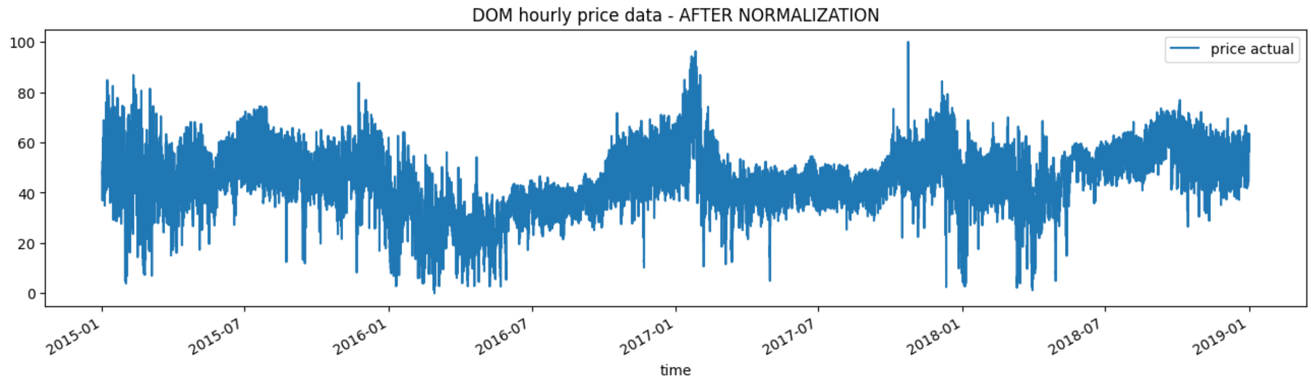

Figure 6: Hourly electrical price data

4.2.5 LSTM

Traditionally, LSTM models have shown a better capacity compared to RNN models in handling noise generated due to temporal variations in data (Cai et al., 2019). The LSTM model used a total of 3 LSTM layers, each with 40 activation units and a tanh activation function. A dropout layer with a dropout rate of 0.15 was used between each LSTM layer to minimize the effect of overfitting. These layers are followed by a dense output layer. The LSTM network increased its number of total parameters by approximately 74.5% in comparison to the RNN model. For training, the Adam optimizer was used with a learning rate of 0.001 for 10 epochs on a batch size of 32. The model was trained to minimize the Mean Squared Error for its loss function.

Similar to the RNN model, the ‘time’ variable providing timestamp data for each hour (Figure 3) was used to index into our target variable ‘total load actual’ or ‘price actual’ for forecasting energy consumption or pricing respectively. However, the LSTM model provided comparative results with only a sequence length of 5, as opposed to the RNN model that used a sequence length of 10. It is important to note that using a sequence length of 10 for the LSTM model generates a negative R2 score, indicating that the model performs worse than a simple mean based model (A Data Driven Life, 2017). This may have been because longer sequences introduce more parameters and increase the model's complexity. If the model is too complex in comparison to the amount of data available, it might not generalize well (Peng & Nagata, 2020). Hence, the sequences of this LSTM model contained the energy consumption or pricing values of the last 5 hours in order to forecast the current values.

5. Results

The table below (Table 1) shows the performance of the time series statistical model ARIMA for forecasting energy consumption and price values. ARIMA was evaluated on Mean Absolute Error (MAE), R-squared (R2) score, Root Mean Squared Log Error (RMSLE) and Root Mean Squared Error (RMSE). Each of the metrics capture a slightly different aspect of the correctness of the model's predictions. MAE takes the average of the distance between the target and predicted data (absolute error). This avoids positive and negative errors from canceling each other out. It keeps the units of measurement of the data itself and gives all the errors the same weight (Karaderili, 2022). Mean Squared Error (MSE) measures the square of the average (squared average) difference between the predicted and actual values. RMSE takes the square root of MSE. Due to the squaring, outliers have a worse impact on RMSE compared to the other error metrics (like RMSLE). Also, models making larger errors are heavily penalized. But it is generally considered an interpretable metric (Kumar, 2023).

RMSLE is preferred when the target variable has values on a larger scale, and one cares more about the relative error between the predicted and actual values, rather than absolute error values. Due to the use of logarithmic values, RMSLE is more robust compared to RMSE (Daniel, 2022). An R-squared value indicates how well the model predicts the outcome of the dependent variable. It indicates the percentage of variance in the dependent variable that can be explained by the independent variables of the regression model (Abba, 2023). Generally, one prefers lower error values and higher values for R2-score.

Target Variable	MAE	R2-Score	RMSLE	RMSE
Total load actual	2.820	0.950	0.1145	4.25
Price actual	1.940	0.960	0.0830	4.25

Table 1: Model Performance - ARIMA model

The table below (Table 2) shows the performance metrics of all four machine learning algorithms described in section 4 above that were used to forecast energy consumption values. Each algorithm's performance was evaluated on the above described metrics: MAE, R2 score, RMSLE and RMSE.

Algorithm	MAE	R2-Score	RMSLE	RMSE
Linear Regression (Baseline)	8.778	0.696	0.300	10.970
Random Forest Regressor	**6.180**	**0.833**	**0.209**	**8.133**
Gradient Boosting Regressor	7.548	0.769	0.254	9.572
Recurrent Neural Network (RNN)	6.758	0.722	0.658	10.430
LSTM	6.636	0.728	0.652	10.293

Table 2: Model Performance - Electricity Consumption Prediction

The table below (Table 3) contains the performance metrics of all 4 algorithms described above in section 4 that were used to forecast electrical energy prices. The algorithms were evaluated on MAE, R2 score, RMSLE and RMSE scores.

Algorithm	MAE	R2-Score	RMSLE	RMSE
Linear Regression (Baseline)	9.686	0.326	0.342	12.609
Random Forest Regressor	4.538	0.808	0.208	6.730
Gradient Boosting Regressor	7.933	0.515	0.298	10.691
Recurrent Neural Network (RNN)	4.060	0.781	0.467	6.218
LSTM	**3.927**	**0.804**	**0.477**	**5.881**

Table 3: Model Performance - Electricity Price Prediction

For forecasting energy consumption values, Random Forest Regressor performs better than other machine learning models, such as gradient boosting, RNNs and LSTMs. It has the lowest error values for MAE, RMSLE and RMSE. It has the highest R-squared score, which indicates the best fit among these models. It is interesting to observe that random forest regressor performs better than more complex models, such as gradient boosting regressor and LSTMs. This may be due to the fact that other models were too complex for the amount of data available, and hence random forest was able to generalize better overall. However, it fails to outperform the ARIMA model (Table 1), which is one of the industry benchmark models for forecasting energy consumption values.

In the case of forecasting price for electrical energy, LSTM generally performed better than other machine learning models. It showed lower MAE and RMSE and higher R2 scores in comparison. However, it showed almost double the RMSLE score compared to random forest regressor. It is important to note that RMSLE penalizes underestimation more than overestimation (Gok, 2018). Underestimating the price estimates could have a bad financial impact on energy companies, by causing budget downfalls and making it challenging for them to remain competitive in their industry. Random Forest, on the other hand, generates quite similar performance metrics to LSTM in terms of MAE, R2 and RMSE, and a lower RMSLE score in comparison owing to the fact that more complex models might be fitting to the noise in the data. However, similar to the energy consumption forecast, the pricing forecast fails to compete with the performance of the ARIMA model (Table 1). This may be due to the fact that this research predicts the target variable from a sequence of 5 (LSTM) or 10 (RNN) hours; hence it falls in the category of short-term forecasting. ARIMA models are usually better fitted for short-term modeling in comparison to models like LSTMs, which are better suited for long-term forecasting. (Chatterjee, 2020)

6. Discussion & Conclusion

The research aimed to analyze the factors that influence the prediction of electrical energy demand (or consumption), as well as its price. It is first demonstrated in the correlation analysis that weather features such as temperature, humidity, pressure, wind speed or rain, etc. that are otherwise used widely in existing research to forecast electrical energy consumption and pricing appeared to have lesser influence in comparison to metrics associated with the generation of other energy sources, such as oil, coal or fossil fuel, among others.

Then, the research evaluated the performance of different machine learning models in forecasting energy consumption and pricing. There are four different methods that are evaluated: two of them are ensemble based regression models, including random forests and gradient boosting regressors; the other two models, which include recurrent neural networks and long short term memory networks (LSTM) perform a time series analysis. Among these four models, random forest performs competitively better than others to forecast energy consumption values whereas LSTM network model performs better to forecast energy pricing. The performance of the machine learning models was compared with the statistical time series ARIMA model that is widely used in the industry, and showed significantly lower error in its predictions compared to all machine learning models used.

This research has its own limitations. The data used for this research has been collected over a period of four years. Data collected over additional years may help get a better picture of the shocks or unusual events due to climate, shortages, or geopolitical events, etc., that might also influence energy prices and consumption. Also, a deeper analysis of why an ARIMA model outperformed the machine learning models could be utilized to achieve a nuanced understanding of the factors that influence electricity consumption and pricing values.

References

Abba, V. (2023, March 28). *What is R Squared? R2 Value Meaning and Definition*. freeCodeCamp. Retrieved November 29, 2023, from https://www.freecodecamp.org/news/what-is-r-squared-r2-value-meaning-and-definition/

Abdallah, M., Talib, M. A., Hosny, M., Waraga, O. A., Nasir, Q., & Arshad, M. A. (2022). Forecasting highly fluctuating electricity load using machine learning models based on multimillion observations. *Advanced Engineering Informatics*, *53*. https://doi.org/10.1016/j.aei.2022.101707

Aksu, G., Güzeller, C. O., & Eser, M. T. (2019). The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model. *International Journal of Assessment Tools in Education*, *6*(2), 170-192. https://dx.doi.org/10.21449/ijate.479404

Bora, N. (2021, November 8). *Understanding ARIMA Models for Machine Learning*. Capital One. Retrieved November 18, 2023, from https://www.capitalone.com/tech/machine-learning/understanding-arima-models/

Buhl, N. (2023). *Recurrent Neural Networks (RNNs) for Time Series Predictions*. Encord. Retrieved November 29, 2023, from https://encord.com/blog/time-series-predictions-with-recurrent-neural-networks/

Cai, R., Li, S., Tian, J., & Ren, L. (2019). Short-term Load Forecasting Based on Electricity Price in LSTM in Power Grid. *IOP Conference Series: Materials Science and Engineering*, *569*(4). 10.1088/1757-899X/569/4/042046

Chatterjee, S. (2020, April 28). *ARIMA/SARIMA vs LSTM with Ensemble learning Insights for Time Series Data*. Predict the future. Retrieved November 29, 2023, from https://techairesearch.com/arima-sarima-vs-lstm-with-ensemble-learning-insights-for-time-series-data/

Connor, J.T., Martin, R.D., & Atlas, L.E. (1994). Recurrent neural networks and robust time series prediction. *IEEE Transactions on Neural Networks*, *5*(2), 240-254. https://doi.org/10.1109/72.279188

Daniel, S. (2022). *Difference between RMSE and RMSLE*. Data Science Blog. Retrieved November 21, 2023, from https://www.datascienceland.com/blog/difference-between-rmse-and-rmsle-656/

A Data Driven Life. (2017, January 28). *What Is R Squared And Negative R Squared*. Fairly Nerdy. Retrieved November 18, 2023, from http://www.fairlynerdy.com/what-is-r-squared/

Deebani, W. (2020, April 13). *Association Factor for Identifying Linear and Nonlinear Correlations in Noisy Conditions*. NCBI. Retrieved November 3, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7516922/

Field, S. (2023, September 10). *Brushing out outliers from your dataset — The Z-Score Method*. Medium. Retrieved November 18, 2023, from https://medium.com/pythoneers/brushing-out-outliers-from-your-dataset-the-z-score-method-f46dd58e5091

Gok, H. (2018, October 8). *Metrics - MSE, R^2, RMSLE*. Data to Wisdom. Retrieved November 19, 2023, from https://hrngok.github.io/posts/metrics/

Karaderili, S. (2022, March 10). *My Notes on MAE vs MSE Error Metrics [graphic]*. HackerNoon. Retrieved November 21, 2023, from https://hackernoon.com/my-notes-on-mae-vs-mse-error-metrics

Kiers, H. A.L., Martín-Fernández, J.A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (Eds.). (2000). Zero Replacement in Compositional Data Sets. In *Data Analysis, Classification, and Related Methods* (pp. 155-160). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-59789-3_25

Kumar, A. (2023, November 17). *Mean Squared Error or R-Squared - Which one to use?* Analytics Yogi. Retrieved November 21, 2023, from https://vitalflux.com/mean-square-error-r-squared-which-one-to-use/

Misra, S., Li, H., & He, J. (2019). *Machine Learning for Subsurface Characterization*. Elsevier Science.

Mustapa, R., Mustapa, D., Nordin, M., Nordin, I., Hamizah, A., & Zabidi, A. (2020, 01). Energy consumption prediction through linear and non-linear baseline energy model. *Indonesian Journal of Electrical Engineering and Computer Science*, *17*, 102. 10.11591/ijeecs.v17.i1.pp102-109

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*(21). https://doi.org/10.3389/fnbot.2013.00021

National Renewable Energy Laboratory & Lawrence Berkeley National Laboratory. (2023, April 1). *Best Practices in Electricity Load Modeling and Forecasting for Long-Term Power System Planning*. NREL. Retrieved November 17, 2023, from https://www.nrel.gov/docs/fy23osti/81897.pdf

NOVA School of Science and Technology & Energias de Portugal (EDP). (2022). Electricity Spot Price Forecast by Modelling Supply and Demand Curve. *Computational and Applied Mathematics*, *10*(12). https://doi.org/10.3390/math10122012

Peng, Y., & Nagata, M. H. (2020). An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. *Chaos, solitons, and fractals*, *139*(110055). https://doi.org/10.1016/j.chaos.2020.110055

Poggi, A., Persio, L. D., & Ehrhardt, M. (2023). Electricity Price Forecasting via Statistical and Deep Learning Approaches: The German Case. *AppliedMath*, *3*(2), 316-342. https://doi.org/10.3390/appliedmath3020018

Rahbar, M. H. (2016, March 7). *Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies*. NCBI. Retrieved November 3, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4888898/

Schmid, L., Roidl, M., & Pauly, M. (2023, March 13). *[2303.07139] Comparing statistical and machine learning methods for time series forecasting in data-driven logistics -- A simulation study*. arXiv. Retrieved November 18, 2023, from https://doi.org/10.48550/arXiv.2303.07139

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533-1543. https://doi.org/10.1016/j.enbuild.2017.11.039

U.S. Energy Information Administration. (2023). *Short-Term Energy Outlook*. Short-Term Energy Outlook - U.S. Energy Information Administration (EIA). Retrieved November 17, 2023, from https://www.eia.gov/outlooks/steo/report/elec_coal_renew.php

Velicer, W. F., & Fava, J. L. (2003). Time Series Analysis for Psychological Research. *Handbook of Psychology*, *2*, 581-606. https://doi.org/10.1002/0471264385.wei0223

Venkata, A. P., Ch., A., Murty, P. S.R. C., & Ch., S. K. (2019, November). Detecting Outliers in High Dimensional Data Sets Using Z-Score Methodology. *International Journal of*

Innovative Technology and Exploring Engineering (IJITEE), *9*(1), 48-53. http://doi.org/10.35940/ijitee.A3910.119119

Zhang, Z. (2016, January 11). *Missing data imputation: focusing on single imputation - Zhang*. Annals of Translational Medicine. Retrieved November 1, 2023, from https://doi.org/10.3978/j.issn.2305-5839.2015.12.38