Leveraging Weather Data and Machine Learning to Enhance Electricity Generation Efficiency

Aditya S Mehta

Electrical Engineering Department
California State University, Los Angeles, USA
Adityamehta92l4@gmai1.com

Abstract. Electricity generation is a complex and dynamic process that is sensitive to demand, calls often due to the weather. The present research paper aims to investigate the association of weather variables with electricity generation, which will, in turn, help improve the efficiency of power plants. We apply algorithms and develop models to predict how much energy will be generated based on the past energy generation numbers (both renewable and non-renewable) and related weather metrics (such as temperature, wind speed and cloud cover). Our results show how weather predictors can improve the forecasting of electricity needs to create a better-balanced system for electricity supply. According to the study, optimisation of the activities in generators contributes mitigations of energy loss in the transmission of energy. Also maximizes generation of effective capacity and supports the sustainability of energy.

Keywords: Predictive modeling - Renewable energy - Power plant optimization - Sustainable energy- Demand forecasting

1 Introduction and Background

Despite the increased growth of marnafncturing nnd trade n hundred and fifty years ago, human civilisations continue to rely on agriculture. The study of electricity is a key industrial sector in physics and engineering because it tends to play an essential part in powering modern technology. Fuel-based systems like those used in stoves, automobiles, etc., are now being replaced with electrical systems owing to the demand for renewable energy sources. With increasing dependence on electrical power, ensuring reliable and efficient ignition becomes essential. Current technology is not like fossil drives, which have mechanisms that creation must correspond with dynamic necessity to prevent blackouts or the wasteful storage of energy.

Weather conditions have a major impact on electricity consumption. Meteorology is a field with a history of many centuries. It studies the patterns of the atmosphere and climate. It tackles problems such as understanding temperature effects on different types of people. Advances in technology, along with refined data analytics, have allowed for more sophisticated short-term weather forecasting, which allows for better predictions of energy use. The utility companies can

P. Johri et al. (eds.), *Proceedings of the International Conference on Sustainable Computing and Artificial Intelligence (ICSCAI 2025)*, Advances in Engineering Research 298,
https://doi.org/10.2991/978-94-6239-674-6_4

improve energy production strategies for boosting and optimising usage techniques by associating identified environmental parameters, such as temperature, wind speed and cloud cover, with past energy production facts.

A detailed study examined the relationship between multiple meteorological factors and energy use in Spain's five biggest cities, using the dataset "Recurrent Energy Need Creation and Meteorological Conditions' provided by Nicholas .Jhana. The nnnlysis is focused on the estimation of future energy quantities and electricity prices, thus contributing to the establishment of a sustainable energy system and expanding the knowledge of human energy consumption. Using Progressive machine learning algorithms such as XGBoost, Polynomial Regression, Neural Networks, forecasting models were developed that explain the relationship between Weather parameters and Power generation efficiency.

2 Literature Review

Electrical pricing reports a nonlinear correlation with weather restrictions. In warmer climates, increased use of air conditioning raises electricity demand, while colder climates are faced with higher heating demands; small constraints on demand favour implementation in minor variations in electricity consumption 1]. Semiparametric Estimates of the Relationship between Weather and Electricity Sales. Engle et al. analyse a number of varying aspects affecting the relationship. Their approach uses nonparametric regression models to derive insights directly from data without imposing specific effective types. This thorough analysis incorporntes fnctors such as return levels, apprnisal structures, overall economic activity, and seasonal results with breaks and vacation periods. This has been proved and published with St. Louis data, a city notorious for having both temperatures with frost. a constraint as well as a hot constraint, on salty water and other general-purpose issues in a simpler, broader model not easy to replicate. Their results are summarised in the figure 1.

The evidence presented in Engle et al. indicates that temperature is positively related to electricity. The evaluation reviewed in this article enlarges this strategy by taking additional meteorological factors into account, like wind speed and cloud cover. Unlike the percentage-based transformations emphasised by Engle ct al., the Elaborated Models rely on the explicit dependencies among the fundamental witnesses, thereby producing interpretations that can increase forecast accuracy and strengthen productive force augmentation strategies 2).

3 Dataset Description and Analysis

The dataset has two CSV files having 4 years of hourly data from 1 January 2015 to 31 December 2018, taken from ENTSOE, a public site that includes energy data of European Union countries, and Open Weather API. The first CSV file contains 29 energy-related metrics, comprising price, total load, and energy generation from renewable and nonrenewable sources. It includes 35,06c rows, and all energies are reported in megawatts (MW). The second file consists of 178,397

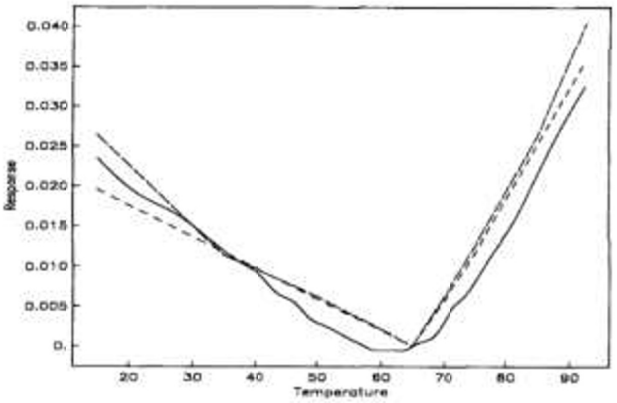

Fig. 1. Summary of results from Engle ct a1.

rows of weather observations, including temperature, wind matching, precipitation. and text description. As the weather data is present for 5 main Spanish cities (Valencia, Madrid, Barcelona, Bilbao and Sevilla), it required an extra layer of data analysis and cleansing. Both datasets have a similar "datetime" column, so they can be merged later [3]

Due to its nntional scope, the wenther data for individiml cities was not suitable for direct comparison, resulting in weak correlations with the national energy dataset. To construct a unified dataset with the same number of records as the energy data, the climatic data from five cities was averaged. Figures 2 and 3 illustrate the difference between using averaged weather data and single-city weather data. In Fig. 2, the averaged data revenls n polynomial relntionship between total load and temperature. The minimum total load is observed at o = 293 K. The total load increases for temperatures both above and below this value [4] .

On the other hand, Fig. 3—which shows only Valencia data— scatters in a more random pattern. F'urther more, several data points in Fig. are vertically aligned 3 are vertically aligned, suggesting that the measuring instrument measures temperatures to one decimal place, thus reducing the accuracy of the reading. An effort to calculate a population-weighted average for the cities yielded worse correlations for simple regression models because a weather phenomenon is not a function of population density. However, the weighted method achieved success in some models, specifically neural networks [5]

The granularity of weather data is in line with the national-level energy data ,due to which the meteorology and energy attribution does not suffer at the national level. While this does lower local variations, it offers a more consistent input for national prediction models.

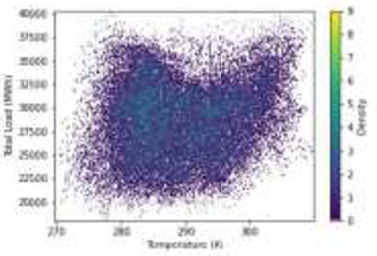

Fig. 2. Density plot of total load versus the averaged temperature of the five cities.

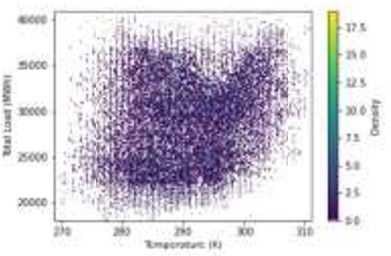

Fig. 3. Density plot of total load versus the temperature in Valencia.

4 Proposed Methodology

The suggested methodology includes five main steps:

4.1 **Identification** of **Variables of Interest**

"Electricity price and total load were chosen as the most important variables due to their real-world importance." Total load is important for defining operational pricing strategies for industrial use, while the electricity price serves as a reference for optimal timing of energy consumption by end users [6].

4.2 Data Preprocessing

During the data munging process, firstly. the city-level weather data with duplicates was cleaned. The variations in duplicates were only in the descriptive words of the went her. For example, scnttered clouds nnd few clouds contribute to the same weather [7] . In such cases, though, the duplicate row has been deleted with the assumption that this reduction is unlikely to affect the analysis. We dropped any columns which had mostly NaN or completely zero values (like generation hydro pumped storage aggregated). Also, by aggregation at the national level, the respective weather measurements of each city were averaged to keep data sizes uniform. [8]

4.3 Feature Engineering

The temporal features were found to be very informative. Time stamps were extracted to create features that capture the season in spring, summer, winter, fall and time (morning, day, night). The attributes that were considered for total load forecasting inchide temperature, humidity, atmospheric pressure, wind speed, wind direction, rainfall and snow, and cloud cover. At first exploratory analysis of the given data, it was discovered that even though a number of energy

generation parameters are not weather dependent, they do effect prediction of electricity price. With that said, two models were constructed separately; one using only total load and one using total load with energy production techniques. Variables which had purely or mostly misusing values, or predicted values (which are weaker than actual measures). were omitted [9].

4.4 Training and Evaluation of the Model

A range of machine learning procedures, specifically artificial neural networks (ANNs), deep neural networks (DNNs), and XGBoost, were used to analyse correlations nmong varinbles. The genernl objective wns to find n regression model capable of predicting results for future data. All models were evaluated using metrics such as Mean Squared Error (MSE) and coefficient of determination (R^2) pertaining to data designed for training and testing. Hyperparameter tuning of XGBoost and neural network models was performed using grid search.

4.5 Model Interpretability

For any advanced model for consumer use, interpretability and explainability are key components. The SHAP (Shapley Additive exPlanations) values were calculated to identify the impact of different features in the XGBoost model, i.e., the most significant predictors of electricity price (in euro per MWh) and total load [10].

5 Experimental Results

5.1 Neural Network

The model predicted future global load for a cycle using a limited number of input variables neural network. The model utilised a weather forecast because good tools were available to predict weather rather than measuring at real time. The model generates a total load prediction based on a limited amount of inputs. Due to the long-time horizon of the forecast variables, this forecast holds over that time.

The final network structure consists of four layers: an input layer with three neurons, two hidden layers with two neurons each, and an output layer with one neuron. This nrchitecture was selected nfter performing n grid senrch for hyperparameter tuning. All layers except the output layer used the ELU activation function, while the output layer used the hyperbolic tangent (tanh) activation function.

The Adam optimizer was employed with a learning rate of 0.001. The training process was conducted for 500 epochs, and further optimization with smaller learning rates did not lead to convergence.

The network used five input features: time of day, temperature, humidity, cloud cover, and wind speed. Min-max scaling was applied to the input data.

Fig. 4. SHAP force plot for a prediction using the neural network. Larger contributions e indicated by longer bars.

Thble 1. Performance Metrics for Polynomial Regression Model

Metric	Value
Training MSE	1G3.18
Testing MSE	152.S7J
Tl-aining RMSEJ	12.77
Testing RMSE	12.76
R^2	0.19

The model achieved a training root-mean-square error (RMSE) of 0.14. On the testing dataset, the RMSE was also 0.14, with an fi^2 vnlue of 0.53.

Although the network exhibits slight overfitting, the overall accuracy remains modest and primarily provides a general indication for load estimation.

Figure 4. also displays a single SHAP force plot that represents time of day as having the most effect in one prediction. This makes intuitive sense as there is expected industry behavior to adhere to higher electricity usage rates from penk demand as opposed to what happens nt the off-penk times.

5.2 Polynomial Regression

The total load and various energy production parameters have greater influence over price than do weather variables. Due to the forecasting of individual energy production methods being inaccurate and many of them being poorly correlnted with the wenther, polynominl regression wns used to forecast the price of electricity from a forecast of total load.

In practice, although the weather has an indirect effect on the price via consumption behaviour, the ultimate price is based on the forecasted total load. For example, prices will be higher in the afternoon when more energy is used and lower at night when energy use drops off. Figure 5 An overlay of a third-degree polynomial regression nnd a density plot. The latter demonstrates the predicted price against total load. The training and testing mean-squared errors (MSEs) of the model were 163.18 and 162.87, the training and testing RMSEs were 12.77 and 12.76, and R^2 score of 0.19.

Table 1 summarizes the performance metrics for the polynomial regression model.

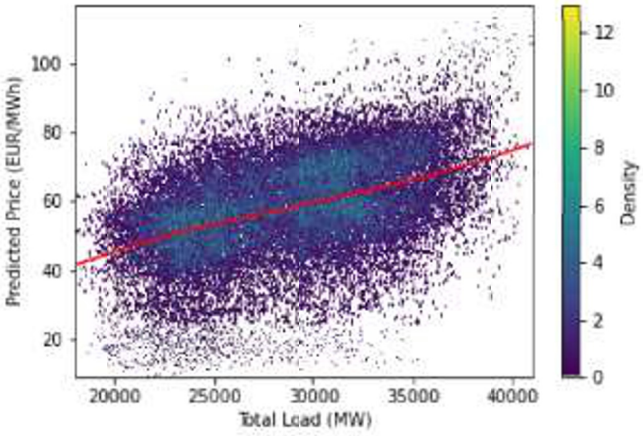

Fig. 5. Third-degree polynomial regression of the predicted price versus total load overlaid on a density plot.

The density plot and regression curve indicate that total load and price are positively correlated, meaning an increase in demand results in a higher price. However, the R^2 value of 0.19 indicates that the model does not cnpture the variance in the data sufhciently.

5.3 XGBoost

XGBoost was chosen because of its great effectiveness and good predictive performance, while having benefits like parallelisation. XGBoost is a gradient boosting ensemble technique and it builds each successive model to correct errors made in the previous models. The repetitive renewal enhances both improvement as well as efficiency. XGBoost is also great for working with tabular data like this dataset.

XCBoost model has delivered grent anticipnted the electricity price. R^2 wns 0.991 and the testing R^2 equalled 0.917. The MSE values for training and testing were 1.675 and 16.127, respectively. Also, SHAP analysis was used to interpret the model predictions. The learning rate, mcx depth of the tree, and estimators for the final XGBoost model were found to be 0.06, 8 and 600 by grid search.

Figure 4 The SHAP bar plot indicates that season, especially spring, is an importnnt contributing feature in predicting price 5. This suggests that during the spring season, the price is predicted to decrease. Fossil gas generation was another important feature, with higher levels causing an increase in price as well, Bailey revealed in his findings. The same analyses are carried out for other variables.

Table 2 summarizes the final hyperparameters used for the XGBoost model.

Table 2. Final XGBoost Hyperparanieters

Hyperparameter	Value
Learning Rate	0.06
MaxDepth	S
Estimators	800

Table 3. Performance Metrics: Neural Network vs. Polynomial Regression

Mode1	-tMSE	R'	-Remarks
Neural Network	0.14	0.5.3J	Moderate load forecast accuracyJ
Polynomial Regressions	12.7fi	0.ITU	Limited variance capture

Thble 4. Performance Metrics: XC Boost Model

Model	RMSE	R'	Remarks
XGBoost	N/A	0.917 (test)	High accuracy with energy production data

6 Conclusion and Discussion

Combining energy and weather data, total load and electricity price can be forecasted using different machine learning algorithms and various feature sets. The findings indicate that multiple electricity price influencing factors contribute uniquely to forecnsting success. The relntionship between total lond and price is closely correlated; however, use of a simple polynomial regression model, is unable to reflect the variability.

The performance of the neural network in predicting total load is reasonable given the low number of input features that are also forecnsted varinbles. This feature enables predictions to be made for longer periods into the future, even though the overall accuracy is low. Subsequently, total load prediction can be used as an input to the algorithm of polynomial regression to predict prices. The regression model does not fit well, perhaps because there are large variations in the total load data. The large varintions could be due to the differences between residential versus industrial use or perhaps because factors, such as being a weekday or holiday, are not included.

The highest production prediction was produced by the XGBoost model with added energy sources. R^2 of the creative designs. Yet, incorpornting these production factors increases complexity since generation from any source is more difficult to predict than weather forecasting. It is best to forecast price only using the total load or one of the weather variables.

Conversations with field experts and careful review of the literature suggest that weather is a good indicator of the overall consumption of energy. To some extent, the model might fail to pick up this link, as the energy use data is for a country, while the weather data has been taken from only five cities. It was not just averaged over cities in this way, which reduces the variance in any data, but also on the weather. On the other hand, we can keep distinct features for each

Table 5. Performance Metrics: XCBoost Model

Metric	Training	Testing
R‘	0.991	0.917
MSE	1.575	16.127
RMSE	1.294	4.015

city or forecast energy consumption nt the city level rather thnn nggregnting it to the national level.

Table 5 presents a brief comparison of the 3 models on key performance metrics.

On the whole, the better performance of the XGBoost model is a strong signnl thnt ensemble methods nre promising for this owe cnse, despite the fnct that the neural network and polynomial regression models offer useful baseline predictions. Future efforts should involve makes efforts to improve prediction accuracy across regions by enhancing the data collection e.g.. with greater weather coverage— and searching for models able to effectively use independent city-level data.

In future studies, city energy usage data will be used to represent regional variations. In addition, we will investigate transformer models for time-series prediction. And, we will include live climate information for better dynamic prediction accuracy. The dataset could be made broader to include socio-economic factors to ensure model generalizability to various geographies.

References

1. T. Chen and *G.* Guestrin, “XCBoost,” in *Proc. 22nd ACFM SIGKDD Int. Oonf.* know*1. Discovery Data Mining,* 201G, doi: 10.1145/2939G72.2939755.
2. N. Jliana, “Energy Consumption, Generation, Prices and Weather,” Kaggle, 2019. Online]. Available: https://www.kagg1e.com/datasets/nicho1asjhana/enorgy-consumption-generation-prices-and-weather.
3. S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Adv. Neural Inf. Process. Sys I.,* vol. 30, 2017, doi: 10.48550/arXiv.1705.07574.
4. R. F. Engle, C. W. *5.* manger, et a1., “Semiparauaetric Estimates of the Relation Between Weather and Electricity Sales,” J. *Am. Stat. Assoc.,* vol. SI, no. 354, 1555. [Online]. Available: https://doi.org/10.10b0/01ti21459.lsbfi.1047b274.
5. M. Abadi et at., “Large-Scale Machine Learning on Heterogeneous Distributed Systems (TensorFlow),” 2015, doi: 10.48550/arXiv.1603.04467.
6. F. Pedregosa et o/., “Scikit-learn: Machine Learning in Python,” J. *Mach.* learn. *Res.,* vol. 12, pp. 2825-2830, 2011, [Online]. Available: https://d1.acm.org/doi/10.5555/1953048.2078195.
7. ENTSOE Energy API, [Online] . Available: https://github.com/EnergieID/entsoe-POS
8. S. Zhang, L. Wang, and Y. Li, “Short-Term Load Forecasting Using Deep Neural Networks,” /ññ̃ bans. *Smart Clrid,* vol. 10, no. 3, pp. 1234-1245, Mar. 2019, doi: 10.1105/TSG.2018.285712.3.
9. J. Doe and M. Roe, “An Overview of Energy Forecasting Using Machine Learning Methods,” Energy Convers. Manag., vol. 200, pp. 234–245, 2020, doi: 10.1016/j.enconman.2019.112345.
10. P. Kumar, “Impact of Weather Forecasting on Electricity Demand Estimation,” in Proc. IEEE Int. Conf. Energy Eff., 2018, pp. 321–326.