aurum

AURUM MÜHENDİSLİK SİSTEMLERİ VE MİMARLIK DERGİSİ
AURUM JOURNAL OF ENGINEERING SYSTEMS AND ARCHITECTURE

Cilt 7, Sayı 1 | Yaz 2023
Volume 7, No 1 | Summer 2023, 99-108

XGBOOST ALGORITHM FOR ORECASTING ELECTRICITY CONSUMPTION OF GERMANY

Abdullahı Abdu IBRAHIM [1], Khalıd Mohamed Abdullah ELZARIDI [2*]
[1]Department of Electrical and Computer Engineering, Altınbaş University, İstanbul, Türkiye
abdullahi.ibrahim@altinbas.edu.tr (https://orcid.org/)
[2]Department of Electrical and Computer Engineering, Altınbaş University, İstanbul, Türkiye,
213721983@ogr.altinbas.edu.tr (https://orcid.org/ 0000-0002-0035-770X)

Received: 18.01.2023 Research Article
Accepted: 28.06.2023
*Corresponding author

Abstract

Stability requires energy demand prediction. We train and test 24-hour German load forecasting models. ENTSO-E Transparency Platform data covered European energy generation, transmission, and consumption. It uses German load data instead of PJM data for the eastern US, adds holidays and lag features to the XGB model, and benchmarks with a linear model and a random forest. Grid search CV refines the final XGB model. National load forecasting RMSE is 1740MW, which is suitable for the gradient boosting model. H-24 and H-48 lag is the most important for this job. Weekends and holidays help, but less. Regional holidays, average temperatures, and lag characteristics could improve the model (beyond H-48).

Keywords: Consumption, Electricity, Forecasting, Regression, XGBoost.

ALMANYA'NIN ELEKTRİK TÜKETİMİNİN TAHMİNİ İÇİN XGBOOST ALGORİTMASI

Öz

İstikrar, enerji talebi tahmini gerektirir. 24 saatlik Alman yük tahmin modellerini eğitiyor ve test ediyoruz. ENTSO-E Şeffaflık Platformu verileri Avrupa enerji üretimi, iletimi ve tüketimini kapsamaktadır. Doğu ABD için PJM verileri yerine Alman yük verilerini kullanır, XGB modeline tatiller ve gecikme özellikleri ekler ve doğrusal bir model ve rastgele bir orman ile karşılaştırma yapar. Şebeke arama CV'si nihai XGB modelini iyileştirir. Ulusal yük tahmini RMSE değeri 1740MW'tır ve bu değer gradyan artırma modeli için uygundur. H-24 ve H-48 gecikmesi bu iş için en önemlisidir. Hafta sonları ve tatiller yardımcı olur, ancak daha az. Bölgesel tatiller, ortalama sıcaklıklar ve gecikme özellikleri modeli geliştirebilir (H-48'in ötesinde).

Anahtar Kelimeler: Tüketim, Elektrik, Tahmin, Regresyon, XGBoost

1. Introduction

Maintenance on the electric power system is a never-ending task that requires constant communication and cooperation between generators and distribution substations in order to maintain a safe operating environment and consistently provide customers with electricity. This is necessary in order to keep the system in good working order. It is important to plan ahead in order to ensure that the production of electricity can keep up with the anticipated demand. When making these plans, it is important to take into account the patterns of production of renewable energy sources, the condition of power plants and the grid, and the significance of hydrothermal resources. In order to prevent damage to the grid, it is important to preserve the equilibrium that currently exists between energy generation and demand (Wood et al., 2020).

There are three distinct time frames that can be used when planning for the operation of a power system (Centro Nacional de Despacho - ETESA, 2020): the immediate, the intermediate, and the distant future. Each of these time frames places an emphasis on a different set of activities. In the shorter timeframe, which ranges from one day to one week, the operational and security aspects of the power system receive more attention than in the longer timeframe. Planning for the medium term considers the situation in terms of weeks to months, with the primary focus being on the management of production resources and the prevention of energy deficits through the use of existing power plants. As a result, the time frame for the long term is focused on the years and decades ahead of us, with the intention of defining the construction of new power plants or adjustments to the transmission system. These standards are subject to change depending on the location, but the overarching concept must always be maintained.

The collection of data on the demand for electricity on a national or regional level is difficult to accomplish. Because the authors were familiar with Panama's electricity infrastructure and had access to publicly available data on the country's electricity load, they decided to use Panama's data in the development of their models. The National Dispatch Center (CND) of Panama is in charge of both the planning and the operation of the electrical grid throughout the country. The objective of demand forecasting with a tolerable margin of error, as outlined by CND methods (PSR NCP — Short term operation programming, 2023) is to anticipate and satisfy consumer demand while keeping expenses to a bare minimum. It is necessary to make short-term projections (for the upcoming week) in order to take into account any safety concerns regarding the operation of the electrical system. According to the short-term and mid-term methodologies CND is responsible for the weekly forecast planning that is performed. When it comes to short-term scheduling, CND makes use of optimization software that functions on an hourly basis (Forecaster | Hitachi Energy.2023). This optimization tool requires hourly updates on the demand prediction, the power plants, and the power grid in order to successfully solve the problem of the weekly least dispatch cost. At the moment, CND uses HITACHI ABB's Nostradamus (Centro Nacional de Despacho - ETESA, 2020) to estimate hourly demand. This information is then fed into the short-term optimization tool, which is then used to plan hourly dispatch (Madrid & Bosquez, 2017) for the following week.

This particular forecasting problem is referred to in the literature as short-term load forecasting (STLF), and more specifically, the STLF problem for the Panama power system, in which the forecasting horizon is one week, with hourly steps, for a total of 168 hours. The primary focus of this study is the prediction of the short-term load on the electrical system. An accurate STLF will assist in mitigating the additional planning uncertainty that has been brought about as a result of the unpredictable nature of power generation from renewable sources. After everything is said and done, it will determine the most accurate opportunity costs for reservoir-based hydroelectric facilities. Because of this, lowering the unit commitment production-transmission costs of the power system leads to an efficient dispatch of thermal power plants (Morales-España et al., 2013). It is possible to send in real time the combination of power plants that will produce the lowest possible operational costs.

The shifting patterns of power consumption and the emergence of new machine learning (ML) methodologies served as motivation for an endeavor to improve forecasting tools by equipping them with the most effective and dependable methods available for reducing the likelihood of errors. The present STLF models are going to be improved as a direct result of this study's findings. The models will be judged against Nostradamus' previous weekly forecasts for Panama's power grid in an effort to provide evidence that it is possible to improve the 168-hour STLF. This evaluation will take place in Panama. The dataset for this study consists of historical load, a comprehensive collection of weather factors, holidays, and historical load weekly prediction characteristics. These were included so that the suggested machine learning techniques could be evaluated and the aforementioned goals could be accomplished.

The Extreme Gradient Boosting Regressor (XGBoost) approach exhibited the highest performance when compared to the previous weekly predictions projected by an artificial neural network (ANN) tool. This was the case because XGBoost uses a gradient-based approach. Using this method also helped me gain a better understanding of the significance of characteristics.

In recent years, there has been an increase in the difficulty level of both maintaining and managing the nation's electrical infrastructure and market. In particular, it is the responsibility of aggregators and operators of transmission systems to guarantee that the ratio of output to demand is maintained in a consistent manner. The use of renewable sources of energy is becoming more widespread; but, because to the intermittent nature of these sources, it is becoming increasingly difficult to achieve an acceptable equilibrium. The prediction of the system's future energy requirements is absolutely necessary in order to maintain its stability. A number of different models for hourly German load forecasting with a lead time of 24 hours are put through their paces in a series of tests that we carry out to train and assess them. The ENTSO-E Transparency Platform served as the source for the collection of data on the generation, transmission, and consumption of energy all over Europe.

2. XGboost Algorithm

XGBoost is an ensemble method for Machine Learning that is based on decision trees and uses a gradient boosting framework as its primary data structure. It is generally agreed that artificial neural networks are the most effective algorithms or frameworks for handling prediction problems that incorporate unstructured data (images, text, etc.). However, decision tree-based algorithms are only considered state-of-the-art for processing moderate to modest amounts of structured or tabular data at this time. The evolution of tree-based algorithms is seen in Figure 1, which may be found here.

Research conducted by professors at Washington State University led to the development of the XGBoost algorithm. The work that Tianqi Chen and Carlos Guestrin presented at the 2016 SIGKDD Conference generated a stir in the world of machine learning researchers (Chen & Guestrin, 2016). Since its creation, this algorithm has not only been acknowledged for winning several Kaggle competitions, but it has also been recognized for powering a variety of cutting-edge commercial applications. As a result, the XGBoost open source projects have a substantial community of data scientists, as seen by the over 350 contributors and around 3,600 modifications made on GitHub. In contrast to other algorithms of a similar kind, this one possesses the following characteristics:

- In a number of different settings: Helps solve a wide range of prediction problems, including those involving regression, classification, ranking, and the building of bespoke models.
- Mobility: Works perfectly on all of the most popular operating systems, including Windows, Linux, and Mac OS X.
- There is support for a wide variety of programming languages, including C++, Python, R, Java, Scala, and Julia, amongst many more.
- Compatible with Flink and Spark, in addition to other cloud-based ecosystems, and supported on clusters hosted by Amazon Web Services, Microsoft Azure, and Yarn.

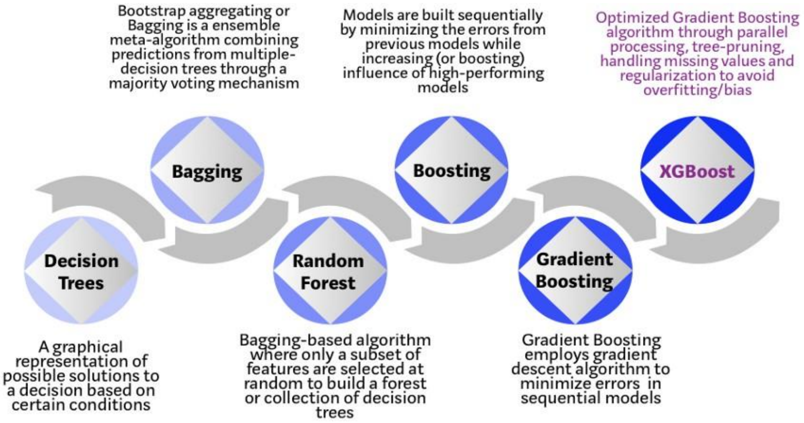
Bootstrap aggregating or Bagging is a ensemble meta-algorithm combining predictions from multiple-decision trees through a majority voting mechanism
Models are built sequentially by minimizing the errors from previous models while increasing (or boosting) influence of high-performing models
Optimized Gradient Boosting algorithm through parallel processing, tree-pruning, handling missing values and regularization to avoid overfitting/bias
Bagging
Boosting
XGBoost
Decision Trees
Random Forest
Gradient Boosting
A graphical representation of possible solutions to a decision based on certain conditions
Bagging-based algorithm where only a subset of features are selected at random to build a forest or collection of decision trees
Gradient Boosting employs gradient descent algorithm to minimize errors in sequential models

Figure 1: XGBoost Evolution

2.1. Build XGBoost

In its most basic form, decision trees are an algorithm that is both aesthetically appealing and reasonably clear to grasp. However, developing an intuitive knowledge of more complex tree-based algorithms can be more difficult. It's possible to think of each iteration of tree-based algorithms as a somewhat different take on the interviewing process.

Diagram of a Decision: Every hiring manager takes into account a variety of criteria when making a decision, including an applicant's degree of education, years of experience, and performance during the interview process. A decision tree, much like a real person in charge of recruiting, weighs applicants based on a variety of subjective criteria.

Imagine for a moment that rather than their being a single interviewer, there is now a panel of interviewers, and each interviewer gets a vote. When making a final decision using either a bagging or bootstrap aggregation method, the responses from all of the interviewers are merged.

Random Forest: It works in a manner very similar to bagging algorithms, with the exception that it just selects a few features at random. That is to say, each potential employer will center their attention on the same collection of talents and experiences (e.g. a technical interview for testing programming skills and a behavioral interview for evaluating non-technical skills).

Boosting is an alternative strategy in which each interviewer modifies the criterion for evaluation depending on the remarks made by the interviewer who came before them in the line of questioning. The efficiency of the interview is "supercharged" as a result of the utilization of a method of evaluation that is more engaging.

For instance, organizations that provide strategy consulting services may employ case interviews as a technique of implementing gradient boosting in order to weed out persons who are not adequately competent.

Extreme Gradient Boosting, or XGBoost, is essentially gradient boosting, but on steroids. By improving the performance of both the software and the hardware, it produces greater results than earlier techniques while using less time and fewer resources (Figure 2).

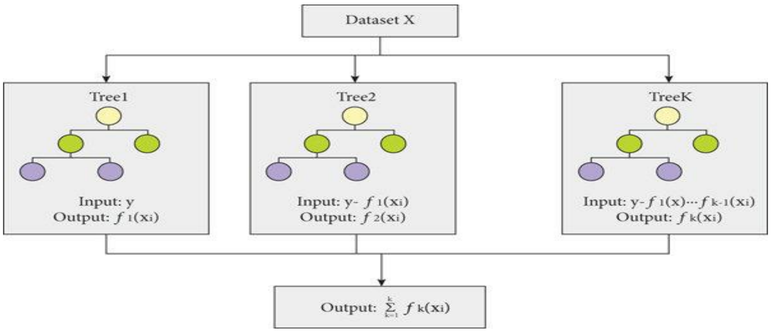

Figure 2: XGBoost Architecture

2.2. XGBoost Performans

Gradient Boosting Machines (GBMs), another form of ensemble tree technique, are quite like XGBoost in that they use the boosting weak learners (CARTs in general) via the gradient descent architecture. XGBoost, on the other hand, improves the underlying architecture of GBM by optimizing systems and implementing algorithmic enhancements (Figure 3).

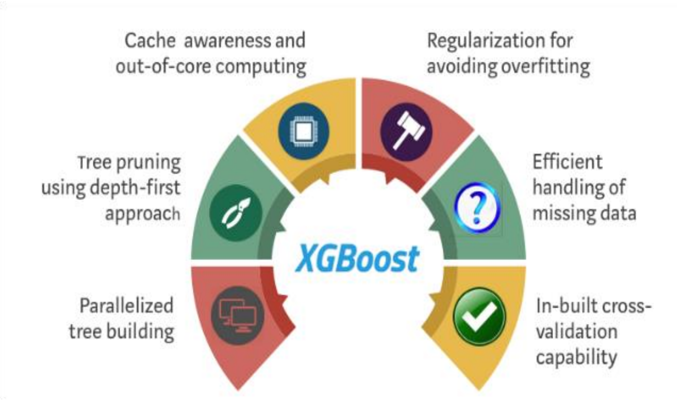
Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

Figure 3. XGBoost Performance

2.3. Enhancement of the Performance of the System

The sequential process of creating trees is something that XGBoost takes on thanks to its parallelized implementation. The fact that the two loops that are used to create base learners are interchangeable—the outer loop that enumerates the leaf nodes of a tree and the second inner loop that calculates the features—makes it simple to carry out the procedure. The processing requirements of the inner loop prevent the beginning of the outer loop until after the processing requirements of the inner loop have been satisfied, which limits the possibility for parallelization. Therefore, to improve run time, the initialization process, which involves conducting a global scan of all instances, is switched places with the sorting process, which makes use of parallel threads. This adjustment compensates for the additional work that is required by parallelization, which results in an increase in the performance of the algorithm.

Tree Pruning: The stopping criterion for tree splitting in the GBM framework has a greedy quality to it, and it is dependent on the negative loss criterion at the instant when the tree is split. Instead of beginning with the criterion, XGBoost uses the'max depth' parameter to select which trees should be removed first before moving on to the next. The use of this "depth-first" method results in a substantial increase in the effectiveness of the computer system.

This technique makes the most efficient use of the available hardware resources. Cache awareness enables us to do this by generating buffers on the fly within each thread to maintain a record of the gradient statistics. Out-of-core computation and other innovations help deal with enormous data frames that cannot be stored in RAM. These advances also optimize disk space.

Enhanced Algorithms: It prevents overfitting by penalizing more intricate models. This is accomplished using LASSO (L1) and Ridge (L2) regularization.

XGBoost "learns" the optimal missing value automatically based on training loss, and as a result, it quickly takes sparse features for inputs. This allows it to better handle the many types of sparsity patterns that can be found in the data.

XGBoost makes use of a technique called the distributed weighted Quantile Sketch to partition weighted datasets in the most effective way possible.

Because the method contains a cross-validation process at each iteration, there is no need to explicitly design this search or to determine the exact number of boosting rounds that are necessary in a single run.

3. Methodology

The ENTSO-E Transparency Platform offered the members of the study team information on the generation, transmission, and consumption of energy throughout Europe. This data set is compiled using contributions from several European TSOs. Europe to the Western Hemisphere The following countries and their consumption rates have been eliminated from consideration: Austria, Belgium, Switzerland, Denmark, Germany, Spain, France, United Kingdom, Italy, Ireland, Luxembourg, the Netherlands, Norway, Portugal, and Sweden. The time is presented in increments of either 15 minutes, 30 minutes, or 1 hour, although this can vary depending on where you are located. The amount of electricity consumed in megawatts (MW) is broken down for the specified time period, which begins in January 2015 and continues through August 2020. The procedure that was carried out is laid out in a flowchart that can be seen in Figure 4.

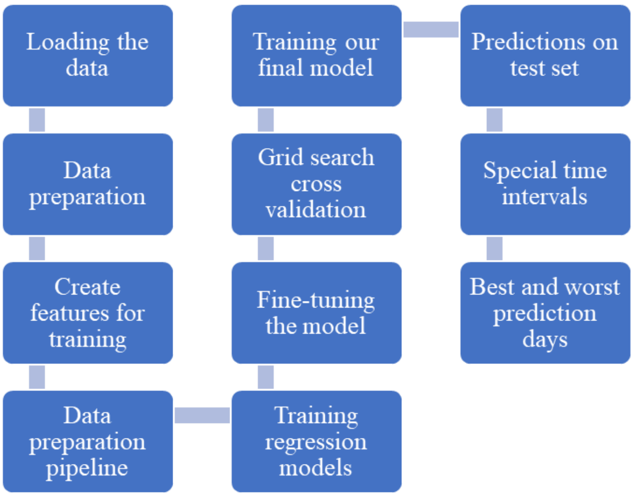
Loading the data
Data preparation
Create features for training
Data preparation pipeline
Training regression models
Fine-tuning the model
Grid search cross validation
Training our final model
Predictions on test set
Special time intervals
Best and worst prediction days

Figure 4. Our methodology

On the test set, the RMSE of the XGB model was calculated to be 1740MW, which is an acceptable value. The results of a comparison between the curves that were found in the test data and the curves that were predicted may be shown in Figure 5. When compared to the other time periods, the 15th of April 2019 through the 6th of May 2019 and the 16th of December 2019 through the 6th of January 2020 appear to be fairly out of the ordinary. This is probably connected to the times around holidays (Figure 6). Around the 21st of April and on the 2nd of May, performance begins to somewhat decline. The first of May is a vacation in Germany; thus, the model may be exploiting lag characteristics to estimate May 2nd volume using May 1st without adequately modifying scale, which would make May 2nd appear to be less than it actually is (as May 2nd is not a holiday). Aside from this one aspect, the model's forecasts of volumes are reasonably accurate. The model has a little tendency to underestimate consumption throughout the holiday season and on New Year's Eve (Figure 7). The days that had the worst performance according to the forecast are shown in Figure 8. The model has a considerable tendency to overestimate consumption on the 20th of June. This day is in fact a regional holiday in Germany known as Fronleichnam, which is not yet considered by our model. Although they are confined to a few states, these have an impact on the burden carried by the country. On the 23rd and 25th of January in the year 2020, the predictions were accurate (Figure 9).

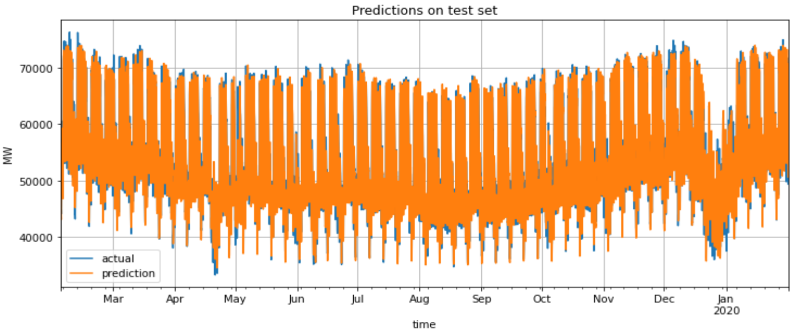

Figure 5. Comparing actual and predicted curves

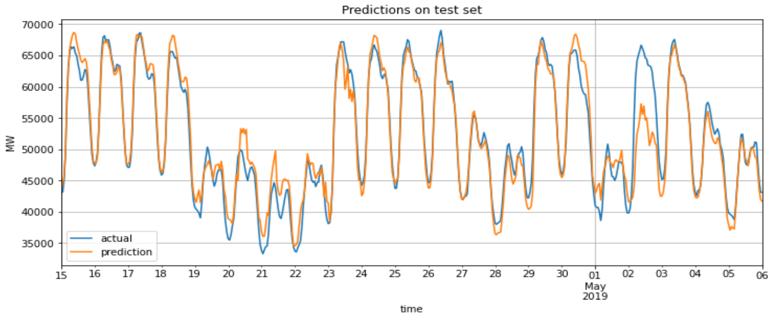

Figure 6. Comparing actual and predicted curves interval 1

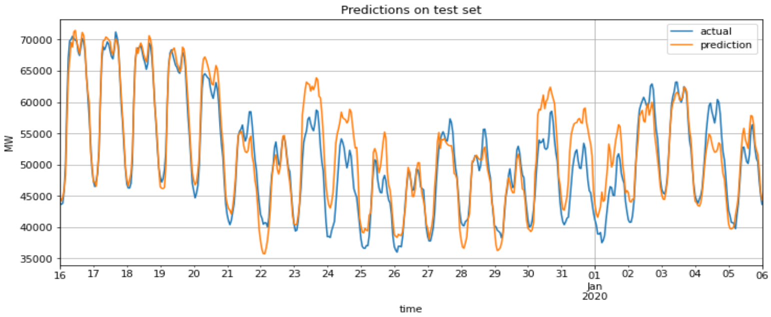

Figure 7: Comparing actual and predicted curves interval 2

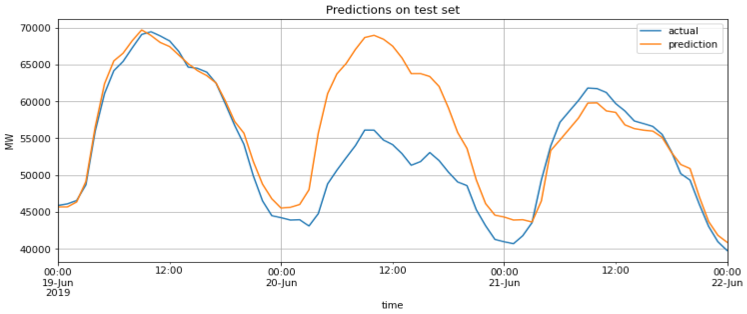

Figure 8. Labour Day's and end of year's periods results

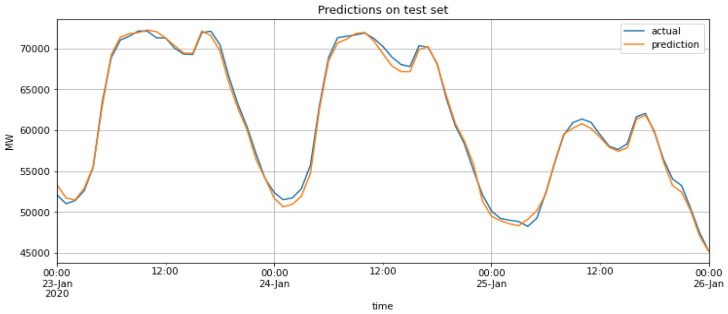

Figure 9. 23rd and 25th of January 2020 results

5. Conclusion

The prediction of the system's future energy requirements is absolutely necessary in order to maintain its stability. A number of different models for hourly German load forecasting with a lead time of 24 hours are put through their paces in a series of tests that we carry out to train and assess them. The ENTSO-E Transparency Platform served as the source for the collection of data on the generation, transmission, and consumption of energy all over Europe. It employs German load data rather than PJM data for the eastern region of the United States. Furthermore, it enhances the XGB model with characteristics such as holidays and lag features, and it employs a linear model and a random forest as benchmarks. Grid search CV is employed in the process of fine-tuning the XGB model. This is a respectable result for the gradient boosting model, as the RMSE for load forecasting at the national level comes in at 1740MW. The H-24 and H-48 lag characteristics are the ones that matter the most for this particular project. Although weekends and holidays have less of an impact overall, it is still beneficial to take advantage of them. There is a possibility that the model could be improved by taking into account additional lag characteristics in addition to factors such as regional holidays and average temperatures (beyond H-48).

References

Centro Nacional de Despacho - ETESA. (2020). Home - Centro Nacional de Despacho - ETESA. Retrieved from https://www.cnd.com.pa/ (accessed Jan. 11, 2023).

Centro Nacional de Despacho - ETESA. (2020). Metodologías de Detalle - Centro Nacional de Despacho - ETESA. Retrieved from https://cnd.com.pa/index.php/acerca/documentos/normas/981-metodologias-de-detalle-2 (accessed Jan. 11, 2023).

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Retrieved from Mar. 2016, doi: 10.1145/2939672.2939785

Forecaster | Hitachi Energy. (2023). Retrieved from https://www.hitachienergy.com/products-and-solutions/energy-portfolio-management/trading-and-risk-management/forecaster (accessed Jan. 11, 2023).

Madid, E. A., & Bosquez, L. V. (2017). Impacto de la entrada de la generación eólica y fotovoltaica en Panamá. I+D Tecnológico, 13(1), 71–82. Retrieved from https://revistas.utp.ac.pa/index.php/id-tecnologico/article/view/1440/html

Morales-España, G., Latorre, J. M., & Ramos, A. (2013). Tight and compact MILP formulation of start-up and shut-down ramping in unit commitment. IEEE Transactions on Power Systems, 28(2), 1288–1296. doi: 10.1109/TPWRS.2012.2222938

PSR | NCP — Short term operation programming. (2023). Retrieved from https://www.psr-inc.com/softwares-en/?current=p4034 (accessed Jan. 11, 2023).

Wood, A. J., Wollenberg, B. F., & Sheblé, G. B. (2013). Power generation, operation, and control. John Wiley & Sons. Retrieved from https://books.google.com.tr/books?hl=en&lr=&id=JDVmAgAAQBAJ&oi=fnd&pg=PA17&ots=CSPViJj2h3&sig=5HS1-xR4-Wl2maC8Vth4iB4H_7I&redir_esc=y#v=onepage&q&f=false (accessed Jan. 11, 2023).