Energy Prediction Using the Building Data Genome 2 (BDGE) Data Set

https://github.com/mahishah19/Energy-Prediction-Using-BDG2-Data/tree/main

Authors: Sameera Boppana, Mahi Mitesh Shah, Inu Tenneti, Darwin Ye

March 6, 2024

Executive Summary

This project aims to enhance energy consumption estimation in educational buildings, leveraging a comprehensive dataset for precise analysis. Utilizing advanced machine learning models like Light Gradient Boosting (LGBM), Random Forest, and Neural Networks, the study addresses the complexities of energy prediction to foster sustainability and efficiency. Through data preprocessing and model optimization, significant predictive accuracy was achieved, particularly with the LGBM model. This research not only contributes to optimizing energy use in educational facilities but also sets a precedent for future sustainable practices and investments in the sector.

Introduction

The objective of our project is to develop a sophisticated approach for accurately estimating the total energy consumption, measured in kilowatt-hours (kWh), specifically within buildings designated for educational purposes. This initiative is paramount in light of the significant energy usage attributed to educational facilities, which constitutes a substantial part of the public sector's overall energy consumption. By addressing the unique energy dynamics prevalent in educational settings, our goal is to refine energy usage estimates significantly, enabling more informed decision-making concerning energy efficiency investments.

Educational buildings are not only centers for learning but also substantial energy consumers due to extensive operational hours and diverse functional areas. In this broader context, our project aligns with global sustainability goals, aiming to reduce greenhouse gas emissions and promote a healthier learning environment through improved energy management. Accurate prediction and efficient energy management in educational facilities can lead to notable cost reductions, enhanced educational quality, and a marked decrease in environmental impacts, thereby instilling sustainable habits within the academic community.

Data

We had 3 groups of datasets for a span of two year: 2016 & 2017

- Building Metadata - contained descriptive attributes of about 1636 buildings with details about what the building was primarily used for, how many floors it had, area, etc. The buildings were segregated by a site ID that would help group buildings in the same location together. (See Figure 1 to see distribution)
- Weather - a daily account of the prevailing weather conditions per site that had data like air temperature, cloud coverage, wind direction, etc. Comprehensive analysis was done for each attribute and compared for the two years to confirm weather conditions were comparable and had any seasonality (See Figure 2)
- Meter Readings - an hourly tab of how much energy was consumed (meter readings) for eight types of energy sources - electricity, chilled water, steam, hot water, gas, water, irrigation, solar. (See Figure 3 for how many readings we have per meter). Our analysis is particularly focused on the five most prevalent meter types: electricity, chilled water, steam, hot water, and gas, subsequently excluding steam from our analysis due to data inconsistency (Figure 3). Through this approach, we aim to contribute significantly to the transparency of energy consumption in educational buildings, serving as a model for sustainability and fiscal prudence within the educational sector.

Methodology

The dataset contains information regarding the building metadata, meter data and weather. For each, we first checked and corrected the data type, found the missing value of each feature, performed EDA for each of the categorical, discrete, continuous, and numeric data types, and inspected outliers and correlation.
[Please see github link for detailed EDA on the dataset]

Upon exploratory analysis of the distribution of the meter reading data, the energy was found to have the majority of the readings clustered near zero, indicating a highly right skewed distribution (Figure 4). There is an extremely low frequency of readings as the meter reading value increases, which suggests that high meter readings are rare or outliers. In order to correct for this, we will apply a logarithmic transformation of the meter reading data (Figure 5).

Since we are only interested in buildings utilized for educational purposes (~38% of the buildings in the dataset - Figure 1), the buildings were filtered by using the Primary Usage Category Column (~617 buildings) . Additionally, we will only be focussing on the top five

meter types: electricity, chilled water, steam, hot water, and gas. However, after further exploration of the data for both 2016 and 2017, we noticed that for steam, the majority of the readings were inconsistent even after the log-transformation was applied. Therefore, we will not be using the steam meter readings moving forward.

After EDA, we consolidated the data, transformed the timestamp into date and aggregated hourly weather to daily. Then we combined data frames and polished them for later splitting. Since it is time series data, we split the whole 2016 year and 2017 year into training and validation sets.

With the completion of the EDA portion and train test split of the data, we can now preprocess the data. We start by categorizing the variables into numerical and categorical types based on the dataset's characteristics. The numerical variables include measurements which are pertinent to energy consumption predictions. The categorical variables encompass attributes like timezone, season, and site specific information.

We then construct a preprocessing pipeline using the ColumnTransformer from Scikit-Learn, applying MinMaxScaler to the numerical features to normalize them between zero and one. This standardization is essential for optimizing the performance of our machine learning algorithms. For the categorical features, we employ the OneHotEncoder, which translates these features into a binary matrix, facilitating the use of these categorical attributes in our predictive models without introducing ordinal assumptions.

Upon establishing our preprocessing pipeline, we fit this setup into our training data to learn the necessary scaling and encoding parameters. Subsequently, we apply these transformations to both the training and test datasets to maintain consistency and ensure that our models are not biased by differing scales or feature representations between datasets.

We convert the processed datasets back into pandas DataFrames, maintaining clear and interpretable feature names post-transformation, which aids in subsequent analysis and model explanation. This step is crucial for ensuring data integrity and facilitating the interpretability of our preprocessing results.

Finally, we address the skewness in our target variable by applying a logarithmic transformation (log1p). This transformation normalizes the distribution of our target variable, which can improve the predictive performance and stability of our regression models. The transformed target data then becomes ready for use in training and evaluating our machine learning models, laying a solid foundation for accurate and robust energy consumption forecasting.

The machine learning models we applied to our preprocessed data are Light Gradient-Boosting Machine (LGBM), Random Forest, and Neural Network (Nnet). When choosing between Neural Networks, Random Forest, and LightGBM (LGBM) for predicting energy consumption, each model offers distinct advantages. Neural Networks excel in modeling complex, nonlinear relationships within large datasets and are capable of automatic feature extraction, making them suitable for detailed energy usage patterns. Random Forests are user-friendly, offering robust performance with minimal tuning. They are known for their interpretability, allowing an easier understanding of feature importance, and are adept at handling outliers and non-linear data. Finally, LGBM stands out for its efficiency and scalability, ideal for large datasets. It provides fast training speeds and achieves high accuracy with relatively straightforward parameter tuning, catering well to large-scale energy prediction tasks with an emphasis on performance.

We also conducted an essential hyperparameter tuning stage for each chosen model This process optimizes model performance tailored to our dataset.

For LGBM, we fine-tuned elements like the number of leaves and learning rate to balance model complexity and prevent overfitting. In Random Forest, we adjusted the number of trees and their depth to enhance accuracy and generalization. For the Neural Network, variations in the number of hidden layers, neurons, and learning rates were explored to mitigate overfitting while maintaining stability.

We employed grid search for this tuning, aiming to identify optimal settings that improve our models' predictive accuracy. This tuning ensures our models are accurately adjusted for effective energy consumption forecasting in educational buildings. Then using scoring metrics like negative MSE and R-squared, the best parameters for each model and meter were selected.

Results

Having now fit our models and identified how models will be scored, we will delve into the performance of our predictive models. Below are the R-squared values obtained from our testing and training phases, followed by a detailed examination of the predictions made by the Light

Gradient Boosting Machine model, which emerged as our chosen approach for forecasting future energy consumption in educational buildings.

Meter	Neural Network Test R-Squared	Random Forest Test R-Squared	**Light Gradient Boosting Test R-Squared**
Electricity	0.206	0.853	**0.818**
Chilled Water	0.185	0.613	**0.726**
Hot Water	0.368	0.562	**0.601**
Gas	0.169	0.740	**0.726**

In the context of this problem, where our goal is to accurately estimate the total energy consumption in educational buildings, a test R-squared value of 0.726 indicates that approximately 72.6% of the variance in the actual energy consumption can be explained by the predictive model.

Based on the performance and R-squared values of the three models, we will be using the LGBM model to make predictions for the next year. After training the LGBM on the full dataset, using the best parameters for each meter, we plotted the predicted and actual meter readings (Figure 7).

To make our final predictions, we aggregated the meter type for each building and calculated the total energy consumption. This gives us a holistic prediction of the amount of energy and energy source each building is expected to use in the next year. For a particular building, we can see the predicted energy consumption for each meter compared to the actual meter readings for 2017 (Figure 7).

Conclusion

In discussing our findings and exploring avenues for future research, it's imperative to consider the broader implications of our predictive models' performance and the insights garnered from our analysis. The achieved R-squared values, particularly for the LGBM model, signify a notable level of explanatory power in estimating energy consumption within educational buildings. This level of accuracy holds significant promise for informing strategic decision-making processes aimed at optimizing energy usage and fostering sustainability initiatives within educational institutions. Moreover, the selection of the LGBM model as our primary forecasting tool is underscored by its robust performance across various metrics and its ability to effectively capture

the complex relationships inherent in energy consumption patterns. By leveraging advanced machine learning algorithms, we've not only provided accurate predictions but also gained valuable insights into the underlying factors driving energy usage dynamics within educational facilities. This deeper understanding can inform targeted interventions and policy measures aimed at promoting energy efficiency and reducing environmental impact. However, amidst our successes, it's essential to acknowledge the inherent limitations and challenges encountered in our analysis.

Data inconsistencies, missing values, and potential biases within the dataset may have influenced the performance of our models and the reliability of our predictions. Moreover, the dynamic nature of energy consumption patterns and external factors such as weather variations present ongoing challenges in accurately forecasting future energy usage. Recognizing these limitations underscores the need for continued refinement and validation of our methodologies to ensure the robustness and reliability of our predictions in real-world scenarios. Looking ahead, there are several promising avenues for future research and development in the field of energy consumption prediction for educational buildings. Exploring the integration of real-time data streams, incorporating additional contextual variables, and refining model architectures to capture nonlinear relationships more effectively are areas ripe for exploration. Moreover, leveraging advances in data analytics and artificial intelligence techniques holds the potential to further enhance the accuracy and granularity of energy consumption forecasts, enabling more proactive and responsive energy management strategies. In addition to the machine learning models discussed in our analysis, there is merit in exploring the application of other time series models such as Recurrent Neural Networks (RNN) and Seasonal Autoregressive Integrated Moving Average (SARIMA) in predicting energy consumption within educational buildings as weather seasonality was detected during EDA (Figure 8). RNNs, known for their ability to capture sequential dependencies in data, could offer valuable insights into the temporal patterns and trends inherent in energy usage dynamics. By leveraging historical data and accounting for seasonality, trends, and cyclical patterns, SARIMA models can provide robust forecasts that complement the capabilities of machine learning algorithms. Incorporating these alternative modeling approaches into our analysis could yield a more comprehensive understanding of energy consumption behavior and enhance the accuracy and reliability of our predictions. Thus, exploring the integration of RNNs and SARIMA models alongside machine learning algorithms represents a promising avenue for future research in the field of energy consumption prediction for educational buildings.

Appendix

Figure 1.

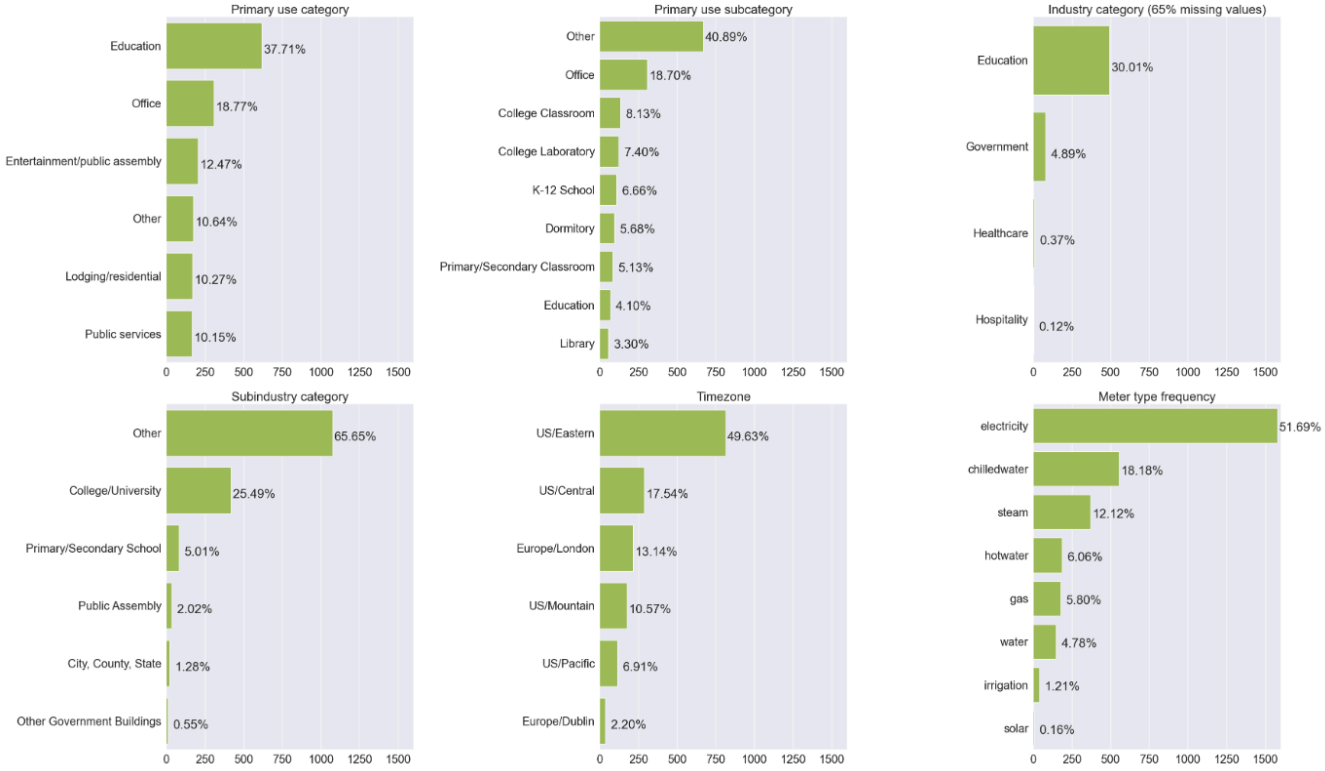

Figure 2.

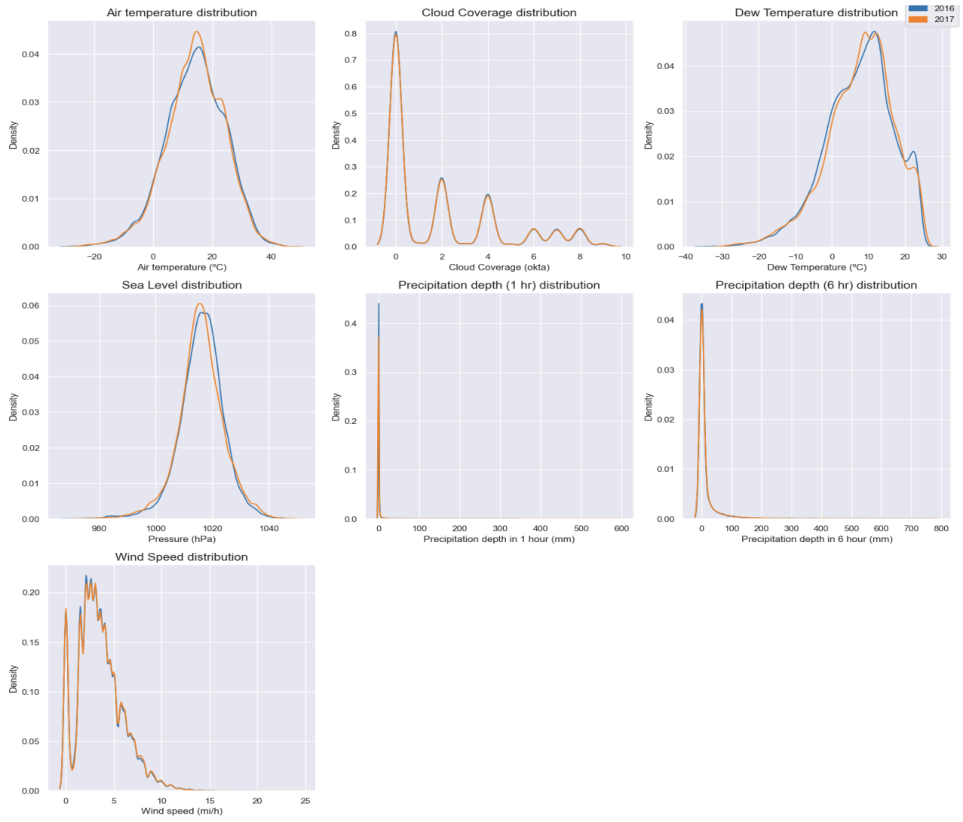

Figure 3.

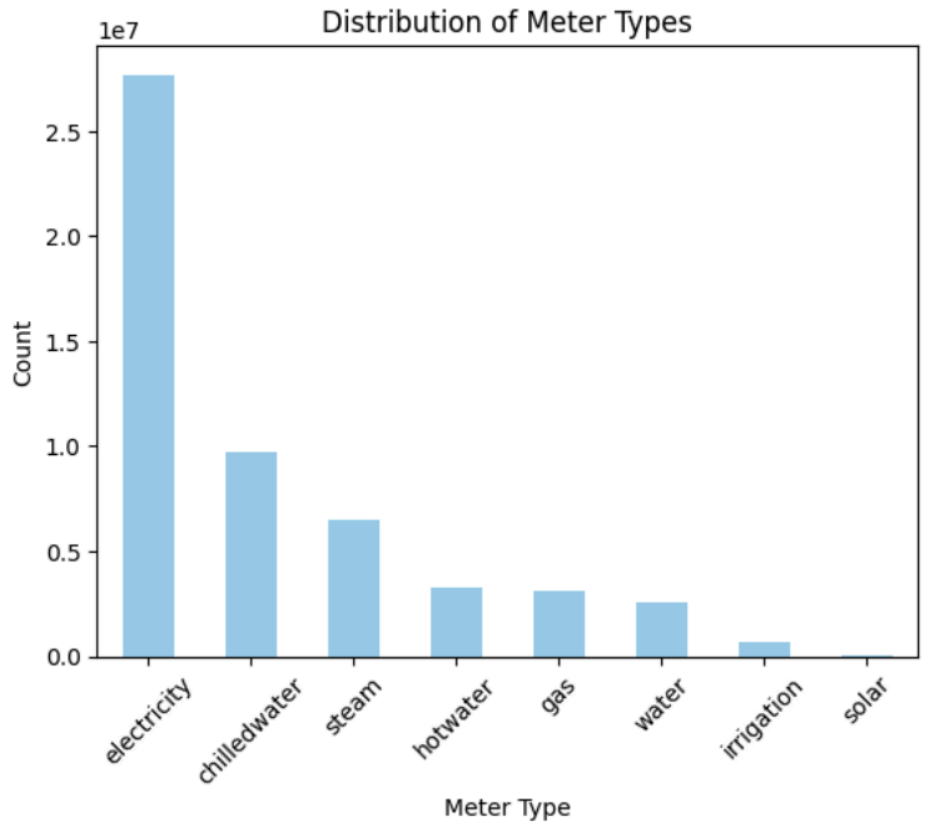

Figure 4.

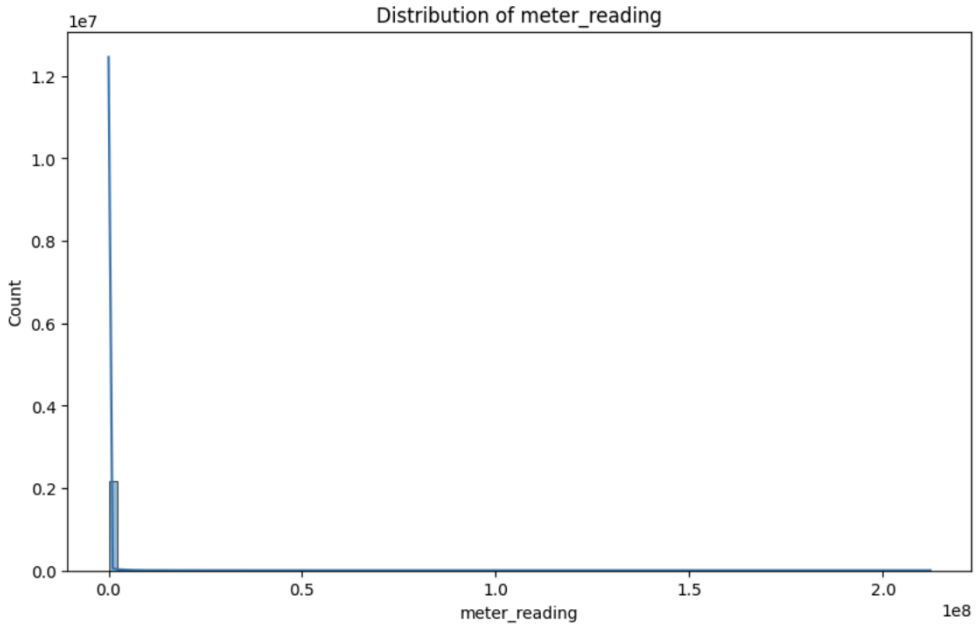

Figure 5.

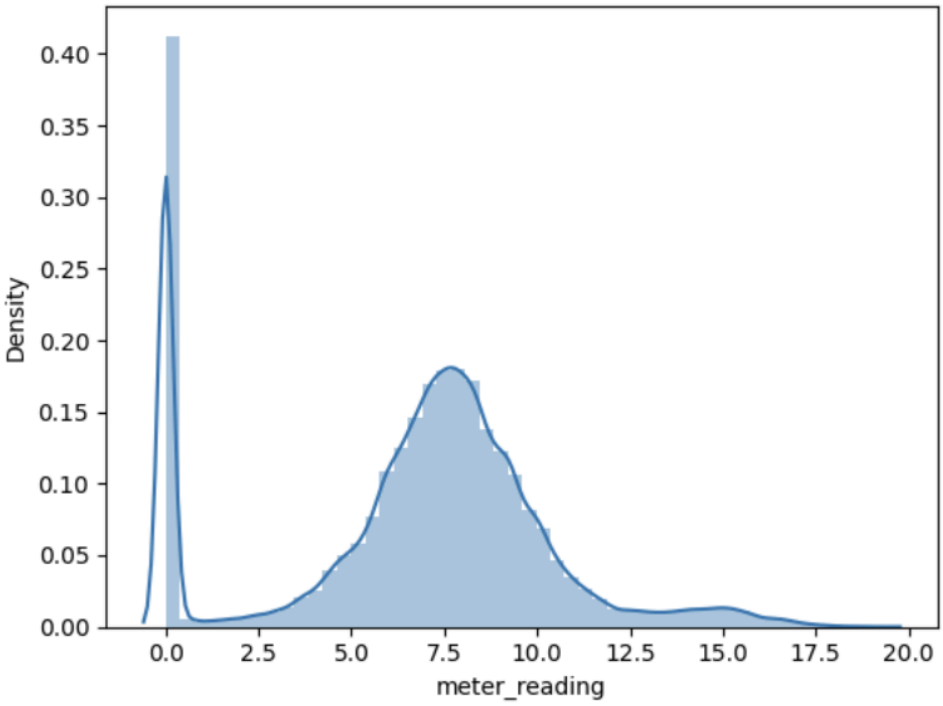

Figure 6.

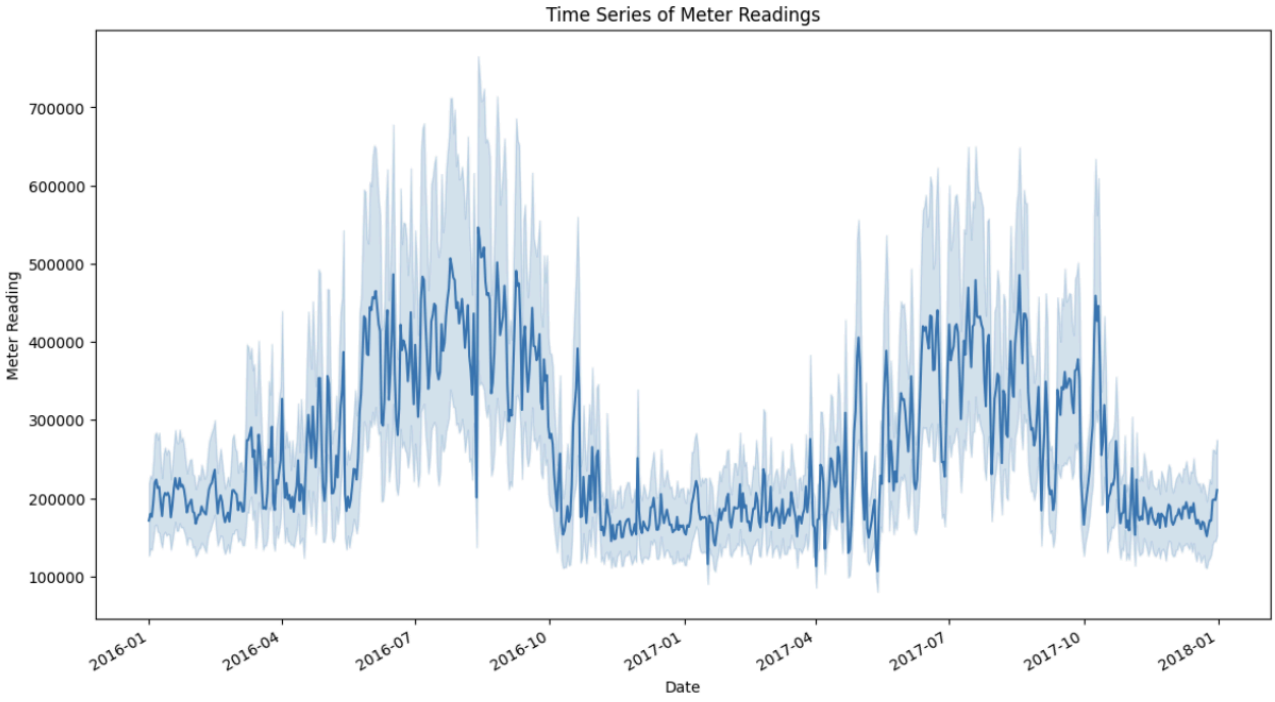

Figure 7.

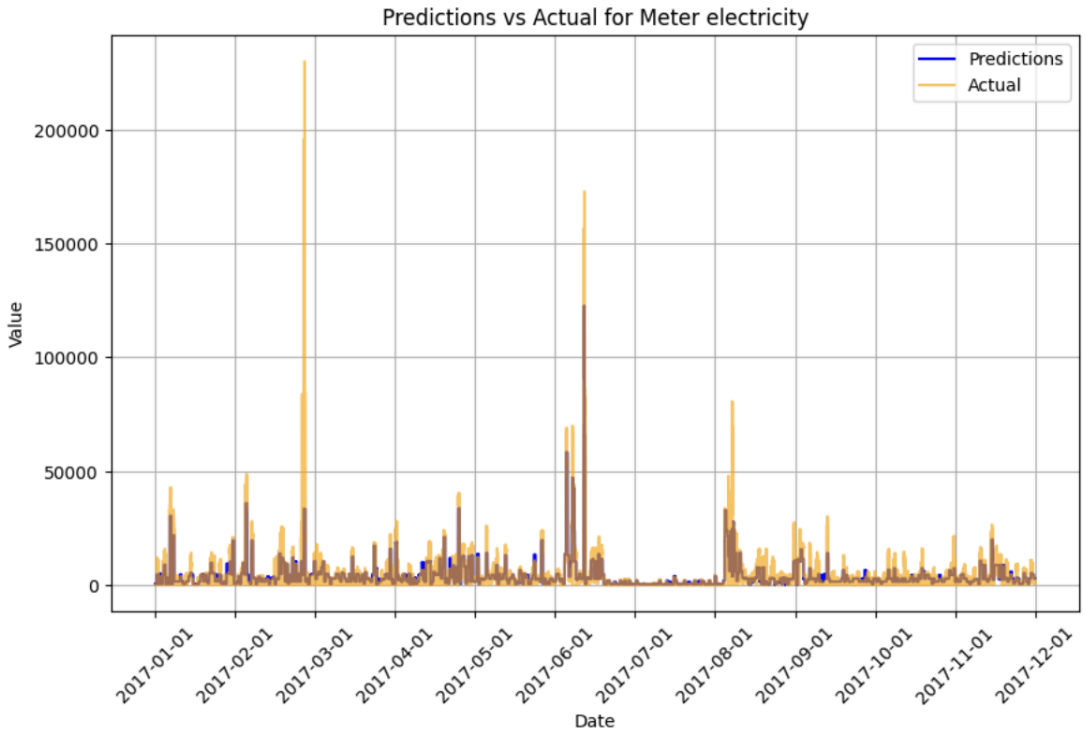

Figure 8.