Energy Consumption Forecasting

Prasuk Jain, Pratik Dnyaneshwar Kale, Karthik Mahalingam, Vedanth Prasanna Bharadwaj
School of Computing and Augmented Intelligence
Arizona State University
Tempe, Arizona, 85281

***Abstract*- This report presents a comprehensive approach to forecasting energy consumption using advanced data mining techniques. The project aims to enhance decision-making and policy implementation for sustainable energy management by developing accurate predictive models. Our methodology integrates data preprocessing, exploratory analysis, and the application of various machine learning models including KNN, Random Forest, and LSTM. Key challenges such as data quality, integration of diverse variables, and adaptation to non-stationary energy patterns are addressed. Performance metrics like Mean Absolute Error is employed to evaluate each model's effectiveness. The study utilizes a dataset provided by the U.S. Energy Information Administration, consisting of timestamps and energy consumption metrics, enabling the analysis of temporal and spatial energy consumption trends. This project significantly contributes to understanding energy patterns and optimizing resource management, thereby supporting the transition towards more efficient and environmentally friendly energy systems.**

I. Introduction

In a time where technological growth intersects with heightened ecological awareness, precise energy consumption forecasting is key for efficient energy distribution, grid stability, and the incorporation of sustainable energy sources. As the world's energy needs escalate, there is a requirement for advanced models capable of accurately projecting consumption trends. This report investigates the use of sophisticated data mining methods for such forecasting, striving to create predictive models that enrich resource management decisions, infrastructural development, and policy-making. Utilizing a detailed dataset from the U.S. Energy Information Administration, the effectiveness of various forecasting techniques like K-Nearest Neighbors, Random Forest, and LSTM networks is examined. The report will methodically outline the diverse methodologies employed, thoroughly tackle the intricate and multifaceted aspects of sophisticated energy use modelling, and thoughtfully explore the substantial, far-reaching economic and environmental ramifications that precise, reliable, and strategic forecasting can have, ultimately guiding and shaping informed, innovative strategies for effective energy management in the coming years and beyond.

A. Background

As the global demand for energy continues to escalate, the imperative to optimize energy management becomes increasingly critical. The effective forecasting of energy consumption stands at the forefront of this endeavour, serving as a cornerstone for strategic planning and sustainable development. This project explores the application of sophisticated data mining techniques to predict energy usage patterns, thereby facilitating more informed decision-making across various sectors.

B. Problem Statement

Energy consumption forecasting involves predicting future energy needs based on historical data. Accurate predictions are essential for utility companies, policymakers, and stakeholders to ensure efficient energy distribution, maintain grid stability, and incorporate renewable energy sources effectively. However, the complexity of influencing factors and the dynamic nature of energy consumption present substantial challenges in developing reliable models.

C. *Importance*

The significance of this project is multifaceted. Foremost, it supports the optimization of energy resources, reducing wastage and enhancing supply chain efficiencies. It also plays a pivotal role in infrastructure planning, where accurate forecasts inform the development and maintenance of energy systems. Environmentally, precise forecasting aids in the integration of renewable energy, contributing to reduced carbon footprints and promoting sustainable practices.

D. *System Overview*

This study employs a multi-model approach to predict energy consumption. The system architecture integrates data collection, preprocessing, feature extraction, and model training phases into a cohesive framework. Each component is designed to work synergistically, enhancing the overall predictive accuracy of the system.

E. *Data Collection*

The data utilized in this project originates from the U.S. Energy Information Administration, comprising detailed records of electricity usage across different sectors and time periods. This dataset includes variables such as time stamps of energy usage and corresponding consumption metrics, which are crucial for developing our forecasting models.

F. *Components of ML System*

Our machine learning system comprises several key components: data preprocessing modules for cleaning and normalizing data, feature engineering tools to extract and select relevant features, and a suite of predictive models including Random Forest, LSTM, and Gradient Boosting Machines. The choice of models is based on their ability to capture temporal dependencies and handle large datasets effectively.

II. Important definitions and Problem Statement

A. *Data*

The dataset utilized in this project originates from the U.S. Energy Information Administration, which gathers and curates data from over 2,000 U.S. utilities. The primary dataset includes measurements of electricity consumption, specifically recording the date and time (Datetime) and the corresponding energy usage in megawatts (AEP_MW). This data from the year 2017 offers a comprehensive view of energy consumption patterns necessary for forecasting. Prior to analysis, the dataset underwent several preprocessing steps including normalization, handling of missing values through imputation, and outlier detection to ensure data quality and reliability for predictive modeling.

B. *Prediction Target*

The prediction target for this study is the future energy consumption levels, quantified in megawatts. Accurately forecasting this target is crucial for enabling energy providers ad policymakers to make informed decisions regarding energy distribution and resource management.

C. *Variables or Concepts*

This study involves several key variables: Time-Related Features - Hour of the day, day of the week, and seasonality are derived from the Datetime column to capture temporal trends in energy consumption.

Statistical Features - Rolling averages and peak usage statistics are computed to understand and incorporate trends and cycles in energy usage over time.

These variables are anticipated to significantly influence the prediction of future energy consumption levels, providing insights into usage patterns and potential demand spikes.

D. *Problem Statement*

Given: The available historical data on energy consumption, including timestamps and megawatt readings, alongside computational resources for data processing and model development.

Objective: To develop a predictive model capable of accurately forecasting future energy consumption up to 24 hours in advance. This model aims to support strategic energy management and operational planning.

Constraints: The project is constrained by limitations in data quality, including potential inaccuracies and missing data points, computational resource limitations, and the unpredictable nature of external factors that affect energy usage, such as weather conditions and economic shifts.

III. Overview of proposed system/approach

A. *System Architecture*

The architecture of the proposed system is modular, comprising several interconnected components:

1. *Data Preprocessing Module*: Handles initial data cleaning, including missing value imputation and outlier detection. This module also standardizes and normalizes data to ensure consistency and reliability in input data for subsequent analysis.
2. *Feature Engineering Module*: Extracts and selects significant features from the processed data, focusing on both temporal and statistical aspects that are critical for understanding energy consumption patterns.
3. *Model Training and Selection Module*: Employs various machine learning algorithms to develop predictive models. This includes traditional models like Decision Trees and Random Forest, as well as advanced techniques like Long Short-Term Memory (LSTM) networks and Transformers.
4. *Evaluation Module:* Assesses model performance using cross-validation and metrics such as Mean Absolute Error (MAE) and R-squared values to ensure the accuracy and reliability of the forecasts.
5. *Integration and Deployment Module*: Once the most effective model is chosen, it's integrated into the target environment where the system will be used. This involves setting up a pipeline for the flow of real-time or batch data into the model and establishing interfaces or APIs for other systems to access the predictions. The deployment should ensure that the model is scalable, maintainable, and secure. Fig 1 gives a diagram for this.

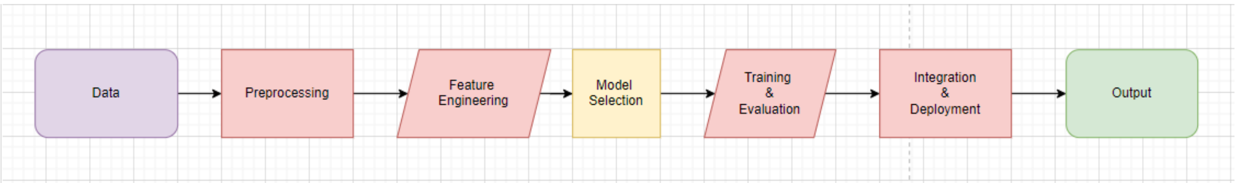

Fig 1: The flow diagram of the system architecture

B. *Data Flow*

Data flows through the system as follows:

1. *Input*: Raw data is inputted into the Data Preprocessing Module where it is cleaned and normalized.
2. *Feature Engineering*: The processed data is then transferred to the Feature Engineering Module where key features are extracted and selected.
3. *Modeling*: The feature-enhanced data is fed into the Model Training and Selection Module for training various models.
4. *Output*: The best-performing model(s) are used in the Evaluation Module to generate predictions and performance metrics are calculated to assess the forecast accuracy.

C. *Predictive Models and Techniques*

The system utilizes a variety of predictive models to ensure robustness and accuracy:

1. *Decision Trees and Random Forest*: These models provide baseline performances and handle non-linear relationships within the data.
2. *LSTM*: Capable of capturing temporal dependencies in energy consumption data, making it suitable for time-series forecasting.
3. *Transformers*: Recently, these have been adapted for time-series forecasting due to their ability to process sequences of data in parallel and capture long-range dependencies.

D. *Model Integration and Deployment*

The selected model or ensemble of models will be integrated into an operational environment where real-time data can be fed into the system to generate up-to-date forecasts. This integration will involve continuous monitoring and periodic retraining of the models to adapt to new data and changing patterns in energy consumption.

IV. Technical Details of Proposed Approach/Systems

A. *Data Loading and Preprocessing*

The project sets its foundation by importing a suite of essential libraries tailored for comprehensive data manipulation, visualization, and machine learning tasks. This foundational step lays the groundwork for subsequent analyses and model development, ensuring a robust and methodical approach to the task at hand. With the tools in place, attention turns to the electricity consumption data sourced from the "electricity_cleaned.csv" file. Leveraging the versatile capabilities of the pandas library, this pivotal dataset is encapsulated within a DataFrame labeled *df*, serving as the nucleus for all ensuing operations.

Zooming into specificity, the project rigorously filters the electricity consumption dataset to focus on the "Panther_office_Catherine" building, effectively narrowing down the scope to a singular entity for focused analysis. This meticulous curation is augmented by temporal precision, with only data from the year 2017 onwards retained, aligning seamlessly with the project's temporal scope and ensuring relevance to the forecasting task. To address data integrity concerns, robust strategies for handling missing values within the selected electricity consumption dataset are implemented. Employing the interpolation technique—specifically, the nearest method—missing values are meticulously replaced with estimates derived from neighboring data points, a deliberate effort to circumvent potential biases or inaccuracies that could undermine subsequent analyses and model development.

In parallel, the project seamlessly integrates weather data sourced from the "weather.csv" file into its analytical framework. This auxiliary dataset undergoes a similar process of filtration, with a distinct focus on entries pertaining to the "Panther" site and commencing from the year 2017 onwards. This congruence in temporal scope ensures harmonization between the electricity consumption and weather datasets, a pivotal prerequisite for robust analyses. Leveraging the forward-fill method for imputing missing values fortifies the weather data against potential gaps or irregularities, propagating the last valid observation forward in time to achieve completeness and consistency, thereby bolstering its utility in subsequent analyses. With both datasets primed and prepped, meticulous alignment of the weather data to the hourly cadence of the electricity consumption records is achieved through resampling, enabling seamless integration and facilitating nuanced analyses at the hourly level. Moreover, retaining only numerical columns streamlines computations while eschewing superfluous data that may encumber the analytical pipeline.

B. Feature Engineering

Feature engineering serves as a pivotal precursor to effective machine learning and deep learning endeavours, bridging the gap between raw data and actionable insights. In this project, a comprehensive suite of feature engineering steps is meticulously executed to distill nuanced patterns from the underlying datasets.

Primarily, temporal dynamics are captured through the creation of time-based features, a cornerstone of predictive modeling. By generating dummy variables for both hour and day of the week, the project encapsulates the intricate hourly and daily fluctuations in electricity consumption. These features serve as invaluable predictors, offering granular insights into consumption patterns and facilitating more accurate forecasting.

Additionally, the project integrates weather features into its analytical framework, recognizing the profound impact of environmental factors on electricity consumption dynamics. Extracting the air temperature data from the weather dataset, a key determinant in heating and cooling demands, enriches the feature space with crucial contextual information. This holistic approach underscores the project's commitment to capturing multifaceted influences on electricity consumption, thereby enhancing the predictive power of subsequent models.

Central to the feature engineering pipeline is the creation of a unified feature matrix *(train_features)*, where temporal and weather features converge to form a cohesive input for predictive models. By concatenating dummy variables for hour and day of the week with air temperature data for the training set, the project constructs a rich feature space that encapsulates the complex interplay between temporal dynamics and environmental conditions. This feature-rich matrix lays the groundwork for robust model development, empowering data-driven insights and informed decision-making in electricity consumption forecasting.

C. Model Development and Evaluation

The project implements and evaluates various machine learning and deep learning models for electricity consumption forecasting:

1. *K-Nearest Neighbours (KNN) Regression*: The K-Nearest Neighbors (KNN) Regression model, a non-parametric approach, predicts electricity consumption based on proximity to the k nearest neighbors in the training data. Trained on this dataset, the KNN model makes predictions on the test data. Evaluation using the Mean Absolute Error (MAE) metric gauges its performance, offering insights into forecasting accuracy. This iterative process ensures a robust evaluation framework, enabling informed decisions on the model's suitability for electricity consumption forecasting tasks.
2. *Random Forest Regression:* The Random Forest model, an ensemble learning technique, aggregates numerous decision trees for prediction. Trained on the provided dataset, the model generates predictions on independent test data. Evaluation of its performance employs the Mean Absolute Error (MAE) metric, providing insights into forecasting accuracy. This comprehensive assessment framework facilitates informed decisions regarding the Random Forest model's suitability for electricity consumption prediction tasks.
3. *Stochastic Gradient Descent (SGD) Regression:* The Stochastic Gradient Descent (SGD) Regression model, a linear regression approach, optimizes coefficients via the stochastic gradient descent algorithm. Trained on the provided dataset, the SGD model predicts outcomes on test data. Model performance assessment utilizes the Mean Absolute Error (MAE) metric, offering insights into forecasting accuracy. This evaluation framework facilitates informed decisions regarding the SGD model's efficacy for electricity consumption prediction tasks.
4. *Decision Tree Regression:* The Decision Tree model, a tree-based approach, partitions the feature space recursively for predictions. Trained on the training dataset, this model forecasts outcomes on test data. Performance evaluation utilizes the Mean Absolute Error (MAE) metric, providing insights into forecasting accuracy. This assessment framework facilitates informed decisions on the Decision Tree model's effectiveness for electricity consumption prediction tasks.
5. *Long Short-Term Memory:* The Long Short-Term Memory (LSTM) Neural Network, implemented with PyTorch, and the Transformer-based model, developed using TensorFlow and Keras, are tailored to capture intricate temporal dependencies within electricity consumption data. The LSTM architecture features an LSTM layer alongside a fully connected layer. Input data is meticulously reshaped to conform to LSTM model specifications, facilitating seamless integration into the training and evaluation processes. Utilizing designated training and testing datasets, both models undergo rigorous training and evaluation, with a particular emphasis on forecasting accuracy. Visualizations depicting actual versus predicted values are instrumental in elucidating model performance, offering valuable insights for further refinement. By leveraging PyTorch's capabilities for LSTM implementation and harnessing the versatility of TensorFlow

and Keras for the Transformer-based model, this research framework underscores the significance of robust model development and evaluation in advancing electricity consumption forecasting methodologies.

6. *Transformer-based Neural Network:* The Transformer-based Neural Network architecture comprises fundamental components, including an Embedding layer, Multi-Head Attention layer, Layer Normalization layer, Global Average Pooling layer, and two Dense layers. Through a specified number of epochs, the model undergoes training on designated training data, with validation conducted using the test dataset. Evaluation of the model's performance on the test set is integral, encompassing the synthesis of predictions on both training and test datasets. To facilitate comprehensive analysis, true labels and model predictions are amalgamated into tables, enabling insightful visualization and interpretation of results. This meticulous approach ensures a thorough understanding of the Transformer-based model's efficacy in capturing intricate patterns and dependencies within electricity consumption data.

V. Experimental Results

A. Data Description

The study integrates three primary data reservoirs for analysis:

1. *Electricity Consumption Data:* Extracted from the "electricity_cleaned.csv" file, this dataset encapsulates temporal sequences of electricity consumption across diverse architectural structures.

2. *Meteorological Data:* Procured from the "weather.csv" file, this dataset encompasses meteorological parameters encompassing temperature, humidity, wind velocity and others, documented at distinct temporal intervals.

3. *Building Metadata*: The "metadata.csv" file furnishes detailed architectural attributes pertaining to multiple edifices, including unique identifiers, geographical coordinates, spatial dimensions, energy utilization patterns, and certification statuses in accordance with Leadership in Energy and Environmental Design (LEED) standards.

B. Feature Engineering

The procedure involves the generation of the subsequent features:

1. Indicator variables representing the hour of the day

2. Indicator variables representing the day of the week

3. Incorporation of air temperature data extracted from the weather dataset, specifically for the designated training and testing timeframes.

These features are amalgamated into a unified feature matrix denoted as *train_features* for the training set and *test_features* for the testing set, respectively.

C. Data Splitting

The partitioning of electricity consumption data for the "Panther_office_Catherine" building into distinct training and testing subsets is delineated as follows:

1. The training set encompasses data recorded during months 1 to 9 (corresponding to January through September).
2. The testing set comprises data recorded specifically within months 10 and 11 (namely, October and November).

D. Evaluation Metrics

The Mean Absolute Error (MAE) serves as the evaluation metric employed to gauge the efficacy of the predictive models. It furnishes an intuitive assessment of the models' predictive accuracy by quantifying the average absolute disparity between the predicted and observed values, normalized as a percentage of the actual values.

E. Baseline Methods

The project undertakes the implementation and evaluation of several baseline methods to forecast electricity consumption:

1. *K-Nearest Neighbors (KNN) Regression*: KNN is a non-parametric technique that predicts electricity consumption based on the k closest neighbors in the training dataset. It's trained on *train_features* and *train.values* and evaluated on `test_features` using the Mean Absolute Error (MAE). From our experiment, we have obtained an accuracy of 77.818% for KNN.
2. *Random Forest Regression:* Random Forest is an ensemble learning method that aggregates multiple decision trees to make predictions. Similar to KNN, it's trained on *train_features* and *train.values* and evaluated on *test_features* using MAE. While effective, classic machine learning models like KNN and Random Forest may not capture intricate temporal dependencies. Hence, the project delves into more advanced methods. We have obtained an accuracy of 81.307% for the Random Forest model in our experiment.
3. *Stochastic Gradient Descent (SGD) Regression:* SGD Regression employs a linear regression model optimized by stochastic gradient descent. Trained on *train_features* and *Y_train*, it's assessed on *test_features* via MAE. By implementing the SGD model, we have obtained an accuracy of 78.674%.
4. *Decision Tree Regression:* Decision Tree Regression recursively partitions the feature space for predictions. Like SGD, it's trained on *train_features* and *Y_train* and evaluated on *test_features* using MAE. In our experiment, we have observed an accuracy of 79.478% for the Decision Tree model.

   Acknowledging the limitations of traditional ML, the project ventures into deep learning models:
5. *Long Short-Term Memory (LSTM) Neural Network:* Implemented using PyTorch, LSTM captures temporal dependencies adeptly. Data is preprocessed, reshaped, and converted to tensors before training for a specified number of epochs. The model's performance is evaluated using MAE on both training and test sets. LSTM model using PyTorch gives us an accuracy of 97.8% for the electricity consumption dataset.
6. *Transformer based Neural Network:* Leveraging TensorFlow and Keras, the Transformer-based model incorporates attention mechanisms for long-range dependency capture. Trained on *X_train_numeric* and *Y_train*, it's validated on *X_test_numeric* and *Y_test*. The evaluation encompasses MAE calculation on both training and test sets. We have observed a loss of 2.3134 in the Transformer Model after running 10 epochs.

These approaches present a comprehensive framework for electricity consumption forecasting, ranging from classical techniques to sophisticated deep learning architectures.

VI. Related Work

The project advances time-series forecasting and energy consumption prediction by integrating insights from various studies. It assesses conventional machine learning models like KNN, Random Forest, and Decision Trees, highlighting their limitations. Additionally, it explores the potential of deep learning techniques such as LSTMs, but introduces Transformer-based models for regression to address their shortcomings. Feature engineering methods, including weather data and temporal information, are examined to enhance prediction accuracy. Comparative analyses across machine learning and deep learning models are conducted, employing diverse evaluation metrics and data characteristics. Ensemble methods and hybrid approaches are also explored to improve prediction performance. This research contributes to enhancing energy consumption prediction models, particularly through the novel application of Transformer-based architectures, providing valuable insights for energy management applications.

VII. Conclusion

The technical report effectively demonstrates the performance of various predictive models in forecasting energy consumption, each exhibiting distinct advantages. We employed techniques like K-Nearest Neighbors (KNN), Random Forest, Stochastic Gradient Descent (SGD), Decision Tree, LSTM, and Transformer models. KNN showed a reasonable accuracy of 77.818%, though it suggests the need for improvements to handle complex data patterns. Random Forest improved accuracy to 81.307%, benefiting from its ensemble approach that reduces variance and error. SGD, with an accuracy of 78.674%, faced challenges in modeling non-linear relationships, whereas the Decision Tree model provided a slightly better accuracy of 79.478%. Notably, the LSTM model, implemented using PyTorch, outshone other models with an impressive 97.8% accuracy, demonstrating its superior ability to capture temporal dependencies crucial for accurate forecasting. The Transformer model, though not directly comparable in traditional accuracy terms, indicated potential for progress with a loss of 2.3134 after 10 epochs. These insights highlight the critical need for selecting appropriate models based on data characteristics and forecasting requirements. The exceptional performance of the LSTM model suggests its significant potential in real-time energy management and planning. Future efforts should aim at further refining these models, incorporating ensemble methods, and integrating more detailed real-time data to enhance the accuracy of forecasts. This project not only pushes forward the capabilities in energy forecasting but also aligns with the broader objectives of promoting sustainable and efficient energy management.

REFERENCES

[1] Xu, Zhaoyi & Saleh, Joseph. (2021). Machine Learning for Reliability Engineering and Safety Applications: Review of Current Status and Future Opportunities. Reliability Engineering & System Safety. 211. 10.1016/j.ress.2021.107530.

[2] G. Vijendar Reddy, Lakshmi Jaswitha Aitha, Ch. Poojitha, A. Naga Shreya, D. Krithika Reddy, G. Sai. Meghana, “Electricity Consumption Prediction Using Machine Learning”, E3S Web Conf. 391 01048 (2023).

[3] Elsworth, Steve and Stefan Güttel. “Time Series Forecasting Using LSTM Networks: A Symbolic Approach.” *ArXiv* abs/2003.05672 (2020): n. pag.

[4] Miller, C., Kathirgamanathan, A., Picchetti, B. et al. The Building Data Genome Project 2, energy meter data from the ASHRAE Great Energy Predictor III competition. Sci Data 7, 368 (2020).

[5] https://www.kaggle.com/code/chuanfutan/energy-consumption-forecasting-project/input

[6] Google Drive Link of Code and Datasets: https://drive.google.com/drive/folders/1vzkI7uXeb0yyNoojbAyX41bFlg-QsgDl?usp=sharing