ELECTRICITY CONSUMPTION PREDICTION WITH MACHINE LEARNING

Ahmet Furkan Yorulmaz[1], Arda Erturhan[2], Batuhan Gümüş[3], Melis Can[4], Vehbi Ozan Güzel[5]

[1]*Software Engineering Department, Istanbul Atlas University, Istanbul (220504004)*
[2]*Software Engineering Department, Istanbul Atlas University, Istanbul (220504001)*
[3]*Software Engineering Department, Istanbul Atlas University, Istanbul (220504021)*
[4]*Software Engineering Department, Istanbul Atlas University, Istanbul (220504019)*
[5]*Software Engineering Department, Istanbul Atlas University, Istanbul (220504032)*

Abstract

This project aims to develop a machine learning model to predict household electricity consumption and enabling more efficient energy management. By leveraging data on power usage, voltage, and sub-metering information, the model provides accurate forecasts of electricity consumption patterns. The approach integrates Exploratory Data Analysis (EDA) with RapidMiner and advanced modeling with TensorFlow for prediction. This approach aims to optimize energy usage, lower costs and promote sustainability by delivering actionable insights for smarter and more efficient energy practices.

1. INTRODUCTION

The increasing emphasis on sustainable energy management highlights the importance of accurate household energy consumption predictions. This project aims to utilize machine learning techniques to develop a model that forecasts electricity usage based on features such as power consumption, voltage, and sub-metering data. Such predictions are vital for optimizing energy efficiency, reducing costs, and supporting global sustainability goals. This project desires to put forth actionable insight into consumption by fusing strong data preprocessing, exploratory analysis and neural network modeling together. A successful implementation promises to benefit households and energy providers alike, paving the way for smarter energy utilization strategies.

2. RELATED WORK

The issue of forecasting electricity consumption has roots in traditional energy management challenges. Early approaches relied on basic statistical models to predict consumption based on historical data. As demand for energy grew and consumption patterns became more complex, these methods faced limitations in scalability and accuracy. The emergence of smart grids and smart meters in the 21st century, alongside the development of automatic remote metering systems that collect and analyze consumption data in real time, introduced richer datasets [1]. This paved the way for machine learning techniques to more effectively model dynamic usage patterns and optimize energy management. Predicting electricity consumption is crucial for efficient energy management, cost reduction and environmental sustainability. It helps optimize resource allocation, reduce waste and meet growing energy demands while integrating renewable energy sources and smart grid technologies.

This problem occurs due to fluctuations in energy demand. Demand is affected by factors such as seasonal changes and time of day. Models such as Linear Regression, Lasso Regression, Random Forest, Extra Tree Regressor, ANN and XGBoost are some solutions offered for dynamic consumption prediction [2][3]. Some improvements can be made to existing solutions for better predictions. One approach is to incorporate real-time weather data and socio-economic information. Additionally, developing hybrid models that combine different algorithms can enhance robustness. Another improvement involves leveraging machine learning models such as Long Short-Term Memory (LSTM) networks for time-series data. These models can better capture temporal dependencies, improving prediction accuracy.

3. OVERVIEW, METHODS AND TOOLS

In this project, our primary approach to solving electricity consumption prediction was the combination of robust data preprocessing with state-of-the-art machine learning modeling techniques. The goal was to accurately model consumption patterns and provide actionable insights into energy usage.

The initial step was to clean and preprocess the dataset, which contained minute-level household electricity usage data. The dataset included variables like global active power, voltage, and sub-metering values. [4] Given the raw data's inconsistencies, such as missing values and noise, preprocessing was critical for accurate analysis and future modeling.

We used RapidMiner for the initial data cleaning and visualization due to its user-friendly interface and powerful preprocessing capabilities. The cleaned data was then further analyzed using Python, where we conducted pilot visualizations and feature analysis to uncover trends and patterns.

Using Jupyter Notebook, we tracked data for every minute and reduced our dataset to include only the data from the top of each hour to make it more manageable. We then split our dataset into input features and target variables. A neural network model was created using TensorFlow's Keras API. The model consists of three layers:

- an input layer, where the input shape is defined based on the number of features in the dataset
- two hidden layers with ReLU activation functions
- L2 regularization to prevent overfitting; and an output layer with a single neuron using a linear activation function for regression output.

The model was trained for 25 epochs with a batch size of 32, and the training process was monitored by setting verbose = 1 to display progress during training. After training, we created plots to evaluate the model's performance during training and its ability to generalize to the validation set. Predictions were then made on the validation set (X_val) using model.predict(), and a scatter plot was created to compare the predicted values against the actual values.

Tools and Techniques

RapidMiner: Employed for data cleaning, including handling missing values and ensuring the dataset's integrity. Key operations included noise reduction, outlier removal and basic statistical analysis.

Python: Utilized for deeper exploratory analysis. We used libraries such as Pandas, Matplotlib and Seaborn to visualize distributions in the data, such as daily consumption patterns and voltage fluctuations.

Data Visualization: Created histogram, box plot, scatter plot and correlation heat-map graphs to identify patterns and anomalies in electricity usage.

4. ARCHITECTURE, ALGORITHMS, MODELS AND DATA

The software architecture for this project follows a modular design that separates data preprocessing, model training, and evaluation tasks. The key technologies used include **Python**, **TensorFlow**, and **Keras** for the development and training of the **Artificial Neural Network (ANN)** model. The code structure is divided into the following modules:

1. **Data Preprocessing Module:** This module handles tasks such as reading the data from the CSV file, cleaning it, and performing necessary transformations like combining **Date** and **Time** columns into a single **datetime** column. Feature extraction from datetime and handling missing data are also part of this module.
2. **Model Training Module:** This module builds the ANN using **TensorFlow** and **Keras**, specifying the architecture, training settings, and optimizer configurations.
3. **Evaluation Module:** Responsible for evaluating the model's performance using metrics such as **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)**, as well as visualizing the results using plots.

This project employs an **Artificial Neural Network (ANN)** for predicting electricity consumption.

Sequential Model: The ANN is built using the Sequential API, which arranges the layers of the network in a linear stack.

- **Activation Functions:**
    - **ReLU (Rectified Linear Unit):** Applied in the hidden layers to capture non-linear relationships.
    - **Linear Activation:** Applied in the output layer since the problem is regression-based and requires continuous predictions.
- **Optimization Algorithm:** The **Adam optimizer** is used for efficient gradient descent optimization, adjusting learning rates automatically.
- **Loss Function: Mean Squared Error (MSE)** is used to measure the difference between the predicted and actual values.
- **Evaluation Metric: Mean Absolute Error (MAE)**, which measures the average absolute difference between predicted and actual values, is tracked during training.

Artificial Neural Network Architecture:

Input Layer: The input layer specifies the number of features in the dataset (after preprocessing and feature extraction).

Hidden Layers:

- **First Hidden Layer:** 64 neurons with **ReLU** activation.
- **Second Hidden Layer:** 32 neurons with **ReLU** activation.

Output Layer: A single neuron with **linear activation**, essential for continuous output prediction.

Data: The dataset for this project contains electricity consumption data, including measurements such as **Global Active Power**, **Voltage**, **Global Intensity**, and **Sub-Metering** for various appliances in the household, along with **Date** and **Time** columns. The dataset is obtained from the **UCI Machine Learning Repository**. The dataset includes multiple years of data, with each row representing a minute of consumption. This makes it a large time-series dataset.

Training and Testing:

The dataset was split into three parts for training, validation, and testing:

1. **Training Set (70%)**: Used to train the ANN model.
2. **Temporary Set (30%)**: This set is further split into two subsets:
    - **Test Set (15%)**: Used to evaluate the final performance of the model.
    - **Validation Set (15%)**: Used to tune hyperparameters and optimize the model during training.

Training Process:

- The model was trained using the training set for **25 epochs**, with a **batch size of 32**. The **Adam optimizer** was used, and **Mean Squared Error (MSE)** was used as the loss function.

Training Duration:

- The training time: 2 minutes.

5. EXPERIMENTS, ANALYSIS AND PERFORMANCE

In this project, we performed a series of experiments to assess the performance and efficiency of our machine learning model for predicting electricity consumption. We focused on using an artificial neural network model to predict the global active power based on various time-related features such as hour, weekday, month and period of the day. Below are the key experiments conducted during the project:

1. **Data Preprocessing and Feature Engineering**:
   The first step involved preprocessing the raw electricity consumption data by converting the datetime column and extracting meaningful time-based features. We also performed data cleaning by removing missing values to ensure high-quality inputs for the model.
2. **Model Design and Hyperparameter Tuning**:
   We experimented with different hyperparameters and model architectures. The neural network architecture included two dense layers with ReLU activations and L2 regularization. We tested various regularization strengths and activation functions to observe their impact on model performance. The final model used Adam optimizer and Mean Squared Error (MSE) as the loss function.
3. **Training and Validation**:
   We split the dataset into training, validation, and testing sets to ensure the model's generalization. The model was trained for 25 epochs with a batch size of 32, and the performance was monitored using Mean Absolute Error (MAE) and MSE. We used validation data to prevent overfitting and adjust the model during training.
4. **Performance Metrics and Visualization**:
   During training, we plotted the loss and MAE curves to visualize the model’s learning process and evaluate how well it generalizes to unseen data. These curves showed the reduction in loss and MAE over epochs, indicating the model's improving accuracy over time. The final model predictions were compared with actual values using a scatter plot and residuals were analyzed to understand the errors better.
5. **Period and Monthly Analysis**:
   Additional experiments were conducted by analyzing electricity consumption during different periods (day and night) and across different months. We identified high consumption hours and months based on 75th percentile thresholds for each period. These experiments helped understand the seasonal and time-of-day variations in electricity consumption and gave insights into how to optimize usage.
6. **Common Failure Modes**:
   Common failure modes included overfitting during training, especially with a more complex model. This was mitigated by using regularization techniques and monitoring validation loss. The model also struggled with predicting consumption during unusual events or anomalies that were not present in the training data.

The following graphs illustrate the key results of our experiments:

- **Training and Validation Losses**:
  A plot showing the loss over epochs, illustrating the model's ability to minimize error.
- **Training and Validation MAE**:
  A graph demonstrating the decrease in mean absolute error during training and validation.
- **Residuals Distribution**:
  A histogram illustrating the distribution of residuals, showing how well the model's predictions align with the true values.
- **Daytime vs Nighttime Consumption**:
  Graphs showing electricity consumption during day and night, including threshold-based alerts for high usage.

In summary, our experiments demonstrated the effectiveness of machine learning for electricity consumption prediction. The model was able to handle complex temporal patterns and offered meaningful insights into usage patterns, helping users optimize their electricity consumption.

6. CONCLUSIONS

This project addresses the challenge of accurately predicting household electricity consumption, motivated by the need for better energy management. By utilizing machine learning techniques, specifically neural networks, the project aims to model dynamic consumption patterns and improve prediction accuracy. The key contributions include the development of a tailored prediction model, data preprocessing, and visualization of results, offering actionable insights for optimizing energy usage. The project will also focus on overcoming challenges such as handling high-frequency data, tuning the model and ensuring computational efficiency.

7. REFERENCES

[1] N. Uribe-Pérez, L. Hernández, D. De la Vega, & I. Angulo, "State of the Art and Trends Review of Smart Metering in Electricity Grids," *Applied Sciences*, vol. 6, no. 3, p. 68, 2016.
Available: https://doi.org/10.3390/app6030068

[2] G. V. Reddy, L. J. Aitha, C. Poojitha, et al. "Electricity Consumption Prediction Using Machine Learning," *E3S Web of Conferences* vol. 391, p. 01048, 2023.
Available: https://doi.org/10.1051/e3sconf/202339101048

[3] M. Rambabu, N.S.S. Ramakrishna and P.K. Polamarasetty "Prediction and Analysis of Household Energy Consumption by Machine Learning Algorithms in Energy Management," *E3S Web of Conferences* vol. 350, p. 02002, 2022.
Available: https://doi.org/10.1051/e3sconf/202235002002

[4] G. Hebrail and A. Berard. "Individual Household Electric Power Consumption," *UCI Machine Learning Repository*, 2006. [Online].
Available: https://doi.org/10.24432/C58K54