TIME SERIES-BASED ANALYSIS OF ENERGY CONSUMPTION: FORECASTING AND ANOMALY DETECTION USING MODIFIED LSTM AND ISOLATION FOREST

Project Reference No.: *47S_MSC_0138*

College *: Reva University, Bengaluru*

Branch *: Department of Data Science*

Guide(s) *: Dr. Rajeev Ranjan*

Student(S) *: Ms. Madhu Shree M.*

Ms. Dechamma M. P.

Keywords: Long Short-Term Memory (LSTM); Isolation Forest; Household power consumption; Anomaly Detection

1. INTRODUCTION TO THE PROJECT

Global electricity consumption witnessed substantial growth, reaching 24,398 terawatt-hours (TWh) in 2022, nearly three times the consumption recorded in 1981 [2]. This surge in consumption reflects an evolving landscape with increased demands on power infrastructure. As the industries consume most of the power, to forecast and minimize the consumption in the power, the forecasting model will be advisable as it predicts and alerts the user when the consumption is taking a hike. And detecting the irregularities in the consumption pattern such as meter reading discrepancies and power theft using anomaly detection technique. Early detection contributes to the overall integrity and reliability of power management systems. To identify and find a reliable solution for the above problems we are building Forecasting and Anomaly detection models which overcome those issues. The primary objective is to accurately predict and identify anomalies in power consumption, addressing the challenges posed by unpredictable energy usage patterns. Focusing on advanced machine learning techniques, specifically Long Short-Term Memory (LSTM) [3] for forecasting and Anomaly detection with Isolation Forest [4]. The power sector may face challenges including inaccurate power consumption forecasts and difficulties in detecting anomalies such as meter reading

irregularities and power theft. These challenges have far-reaching implications for efficient resource allocation, economic considerations, and the overall reliability of power management systems. By addressing these issues, our project aims to not only fill existing gaps in the literature but also provide practical solutions with direct applications in the power sector. We seek to improve forecasting accuracy to enable better planning and allocation of resources while simultaneously enhancing anomaly detection capabilities to curb power theft and irregularities.

2. OBJECTIVE

- The project aims to advance machine learning techniques to address challenges in forecasting and anomaly detection in household power consumption, ultimately contributing to better energy management and resource allocation.
- By accurately forecasting household power consumption, the project has the potential to enable more efficient energy management practices.
- And by effectively detecting anomalies, we could include optimizing energy usage patterns, reducing peak demand, and identifying potential equipment failures.

3. Methodology

3.1 Data Preprocessing

Cleaning: It is crucial to handle missing values, outliers, and inconsistencies in the dataset to avoid any biases or errors during model training. This involves using techniques like imputation to fill in missing values, removing outliers based on statistical measures, and correcting any inconsistencies in the formatting of the data.

Normalization: Scaling the features to a consistent range is important for improving the convergence and performance of the model. Techniques like Min-Max scaling or Z-score normalization are applied to ensure that all features contribute equally to the model training, regardless of their original scale.

Feature Engineering: Extracting relevant features and transforming the data is essential for enhancing the performance of the model and capturing meaningful patterns. Feature engineering techniques include creating new features from existing ones, converting categorical variables into numerical representations, and selecting the most informative features through methods like Principal Component Analysis (PCA) or feature importance analysis.

3.2 LSTM Model Building

When building an LSTM model, the first step is to design its architecture. This involves deciding on the number of layers, hidden units, input and output layers, and activation functions. It's important to customize the architecture based on the characteristics of

the time series data, such as the length of the sequence and the complexity of the patterns.

Once the architecture is set, the next step is training the LSTM model. This involves feeding historical time series data into the network and adjusting hyperparameters like the learning rate and batch size. To ensure the model's performance is reliable, cross-validation techniques are used to validate its accuracy. During training, a technique called backpropagation through time (BPTT) is used to update the model's parameters and minimize the loss function.

After the LSTM model is trained, it's important to evaluate its performance. This is done to assess how well the model captures temporal dependencies and makes accurate predictions. Common performance metrics used for evaluation include Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), which measure the difference between the predicted values and the actual values.

3.3 Isolation Forest Model Building

Building the Isolation Forest model involves several steps. First, we need to initialize the model by setting hyperparameters such as the number of trees in the forest and the contamination factor. These hyperparameters help determine the proportion of anomalies we expect in the dataset. Proper initialization is important to ensure that the model can effectively identify outliers without being influenced by the normal data distribution.

Next, we move on to the training phase. During training, the Isolation Forest model is fitted to the preprocessed dataset. This involves fitting decision trees to the data and fine-tuning the hyperparameters to optimize performance. The model learns to isolate anomalies by recursively partitioning the feature space until each data point is either isolated or grouped with similar points.

Once the model is trained, we can use it for anomaly detection. By assigning anomaly scores to each data point, the Isolation Forest model can identify anomalies in the dataset. Data points with high anomaly scores are considered outliers, while those with low scores are considered normal. This allows us to effectively detect and analyze anomalies in high-dimensional spaces.

Proposed System

Our proposed work based on the LSTM network with 50 units are used with other training parameters. Our model is trained for 100 epochs using the Adam optimizer and Early Stopping technique, details show that the LSTM architecture is more advanced and optimized, which results in improved performance. Fig 1 shows an LSTM configuration model designed to predict household power consumption.

Input Power Consumption Data: This is the historical data of power consumption from households. It serves as the input for the LSTM unit.

Forget Gate: This gate decides which information from the previous cell state should be forgotten. It looks at the previous hidden state (h(t-1)) and the current input to determine what to keep or discard from the cell state. The output of the forget gate ranges from 0 to 1, where 1 indicates keep completely and 0 indicates discard completely.

Input Gate: This gate updates the cell state with new information. It first determines which values to update using a sigmoid function and then generates a new candidate vector, which could potentially be added to the state.

Activation Function (Sigmoid): The sigmoid activation function is used in both the forget and input gates. It controls the flow of information by assigning values between 0 and 1, influencing what information is allowed through.

Memory Cell: This component acts as the storage unit and keeps track of information over time. It processes the input data by discarding unnecessary information and integrating new candidate values based on their relevance.

Tanh: The tanh function generates a vector of new candidate values that can be added to the memory cell's state. It ensures that the values are scaled appropriately between -1 and 1.

Output Gate: This gate determines the next hidden state by incorporating information from the current input and the previous hidden state. The resulting hidden state is then passed through the tanh function for normalization and multiplied by the output of the sigmoid gate to select relevant information for output.

Hidden State (h(t-1)): This represents the current state of the LSTM model. It is updated based on the input and previous hidden state, and it influences the output of the model.

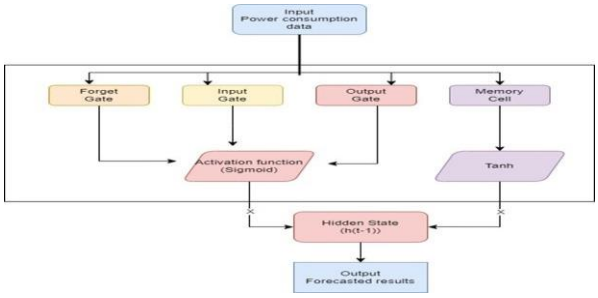

Fig. 1: LSTM configuration model

In isolation forest method we aimed in improving the efficiency and reliability of identifying unusual patterns in power usage data. This will greatly contribute to the progress of the area of study as there are few research exist in this domain.

4. Result and Conclusion

The LSTM model has shown impressive results, as seen in its performance during training and testing. After 100 epochs, the model successfully reduces its loss with each iteration, eventually reaching a training loss of 0.0086. This indicates that the model is able to accurately capture and predict intricate patterns in the data. Furthermore, the model's effectiveness is evident in its low Root Mean Squared Error (RMSE) of 0.0899 for the training set and 0.0799 for the testing set. These results highlight the LSTM's ability to forecast time series data, making it a valuable tool for various applications such as financial forecasting, weather prediction, and energy consumption analysis.

The Isolation Forest for anomaly detection model shows impressive accuracy when it comes to distinguishing between normal and anomalous instances in the dataset. With precision, recall, and F1-score all at a perfect 1.0, the model proves its ability to identify anomalies with complete accuracy, demonstrating its strength and dependability in anomaly detection tasks. This indicates that the model accurately detected all 258 anomalies (True Positives) and correctly classified all normal data points (True Negatives). Moreover, the model did not generate any false outcomes (False Positives) or has not missed any anomalies (False Negatives) and significantly enhanced the performance of anomaly detection models. Moreover, the analysis of anomaly distribution

reveals that anomalies make up approximately 4.99% of the dataset, highlighting the importance of accurate detection methods in uncovering irregularities and reducing risks. Overall, the outstanding performance of the anomaly detection model emphasizes its significance as a powerful tool for proactive risk management and decision-making in various domains, such as cybersecurity, fraud detection, and predictive maintenance.

In this work, we explored the task of forecasting household power consumption using LSTM (Long Short-Term Memory) models and detecting anomalies in the power consumption using Isolation Forest. At first, we pre-processed the data by handling missing values, converting data types, and resampling the data into daily intervals. We then visualized the data using various techniques such as line plots, scatterplots, and histograms to gain insights into the power consumption patterns and correlations between different features. For model selection, we chose LSTM due to its ability to capture temporal dependencies in sequential data effectively. We trained the LSTM model and evaluated its performance using metrics such as RMSE (Root Mean

Squared Error) on both training and testing data. Additionally, we implemented an anomaly detection system to identify unusual patterns in the power consumption using Isolation Forest. By computing precision, recall, and F1-score, we evaluated the effectiveness of the anomaly detection system. Overall, our results indicate that the LSTM model shows promising performance in forecasting power consumption, and the anomaly detection system effectively identifies anomalies in the data, demonstrating the potential for practical applications in energy management and anomaly detection in power systems. With this, one can easily enter the relevant features and quickly get the precise energy usage prediction. This can improve how easily the built model can be used and making it a tool for people looking to manage efficient energy management.

5. Innovation in the Project

This project is innovative because it combines two cutting-edge machine learning techniques to solve important problems in managing household power consumption: isolation forests for anomaly detection and long short-term memory (LSTM) networks for forecasting. Power consumption projections are more accurate when the LSTM model can identify complex patterns and temporal connections in sequential data. This is especially helpful for planning and resource management in the electricity infrastructure. Meanwhile, abnormalities that traditional approaches would overlook, including power theft and discrepancies in meter readings, are expertly detected by the Isolation Forest model. By integrating these two methods, the project offers a complete solution for guaranteeing the accuracy of power usage data and forecasting future demand, helping to create more dependable and effective energy management systems.

6. Scope for future work

To increase this project's effect and applicability, future work can concentrate on several viable avenues. To increase predicting accuracy, enhanced feature engineering could consider outside variables like the weather, economic data, and social events. Better results could be obtained by

using hybrid models that incorporate LSTM with other time series forecasting methods like ARIMA or Prophet. Proactive energy management can be facilitated by putting anomaly detection and real-time forecasting technologies into place, which can offer quick insights. To assess scalability, the models should be applied to bigger datasets—possibly on a regional or national scale—and put into cloud environments for general accessible. Making dashboards easier to use will increase non-technical stakeholders' access to the system. Smart grid integration may provide load balancing and dynamic pricing based on real-time data.