Interpretable Insights from Tree-Based Models using LLMs

HAN NGUYEN, SOPHIE ANDERSON, GRACE CAO, PELIN EGRIBOYUN, KEREM FARZALIYEV, HELEN BANG, Break Through Tech AI MIT, USA
ABHISHEK TONDEHAL AND CRISTOBAL MARIN SIEBEL, catalan.ai, USA

ACM Reference Format:
Han Nguyen, Sophie Anderson, Grace Cao, Pelin Egriboyun, Kerem Farzaliyev, Helen Bang and Abhishek Tondehal and Cristobal Marin Siebel. 2024. Interpretable Insights from Tree-Based Models using LLMs. 1, 1 (December 2024), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As energy usage continues to grow, energy conservation has become a pressing concern. Accurate energy demand predictions plays a critical role in balancing supply and demand, optimizing costs, and minimize the costs on energy usage. However, current forecasting models often produce tree-based regression models with inaccurate predictions. To address these challanges, this research with large language models (LLMs) to minimize hallucinations on electricity demand inquiries.

Our primary objective is to predict electricity demand and identify patterns in generation and consumption while ensuring the insights generated are accessible and actionable. The study focuses on electricity data from ERCOT (Electric Reliability Council of Texas), which provides a comprehensive dataset covering factors such as real-time fuel mix, solar and wind power production, and locational marginal prices. This diverse dataset serves as the foundation for building and refining regression models, which in turn inform LLM-based insights.

To achieve this, our approach involves preparing the data by handling missing values, addressing outliers, and constructing meaningful features to capture periodicity and time-based trends. Following data preparation, tree-based regression models such as Random Forests are employed to predict demand accurately. These outputs, along with processed data summaries, are then integrated into LLMs to generate interpretable answers to critical questions, such as projected demand, pricing strategies, and the mix of generation sources.

This project aims to bridge the gap between traditional modeling techniques and the transformative potential of LLMs. By developing automated tools and strategies for combining model outputs with LLMs, this research offers a framework for generating insights that are both reliable and interpretable. Ultimately, this approach has the potential to inform better decision-making in energy management, addressing key challenges in a rapidly evolving energy landscape.

Authors' Contact Information: Han Nguyen, Sophie Anderson, Grace Cao, Pelin Egriboyun, Kerem Farzaliyev, Helen Bang, Break Through Tech AI MIT, Boston, MA, USA; Abhishek Tondehal and Cristobal Marin Siebel, catalan.ai, Boston, MA, USA, hello@catalan.ai.

ACM XXXX-XXXX/2024/12-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 DATA PREPARATION

This milestone was completed on **Saturday, December 14, 2024**. The following steps detail the technologies and tools used, as observed in the uploaded project files, to process and prepare the data for analysis and modeling.

2.1 Understanding the Data

The dataset was analyzed extensively using Python to gain a deep understanding of its structure and content. **Pandas** was employed for data manipulation, enabling efficient exploration of variables and patterns within the dataset. **NumPy** complemented this process by facilitating numerical operations crucial for the initial assessment. Energy generation, pricing, and demand data were reviewed across various categories, including gas, solar, hydro, and coal generation. This examination provided insights into the temporal and spatial distribution of energy sources and their relationship with electricity demand.

2.2 Data Cleaning

To ensure the dataset was consistent and reliable, several data cleaning techniques were applied. **Pandas** was used to address missing values, employing strategies such as forward filling or interpolation where appropriate. Erroneous data entries were identified and corrected to prevent inaccuracies in subsequent analyses. Time-related variables were standardized using the `datetime` module, which allowed for uniform formatting and seamless integration of temporal features in the modeling phase.

2.3 Feature Engineering

Feature engineering was essential to enhance the predictive power of the dataset. Timestamps across multiple datasets were combined and converted into a consistent `datetime` format to ensure uniformity. Average temperatures for each region were calculated to capture weather-related effects on electricity demand. Lag features were created to incorporate past-hour data for predicting current-hour demand. One-hot encoding was applied to categorical variables to make them usable in machine learning models. Correlations between variables were computed, and only those with a correlation greater than 0.5 were retained. A correlation heatmap was generated using **Seaborn** to visualize the relationships among variables and identify significant predictors, as shown in Figure 1.

Null values were either removed or filled using the mean of their respective columns. Finally, the data was winsorized to mitigate the influence of outliers. Figures 2 and 3 show the distribution of data before and after winsorization, respectively.

2.4 Data Storage and Output Preparation

Processed datasets were saved in **JSON** format for structured storage and easy accessibility, as seen in the `final_features.json` file. Compatibility between different processing steps was ensured by maintaining consistent data formats and variable names, facilitating seamless integration into modeling workflows.

2.5 Technologies Used

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, os
- **File Formats:** JSON
- **Environment:** Jupyter Notebooks

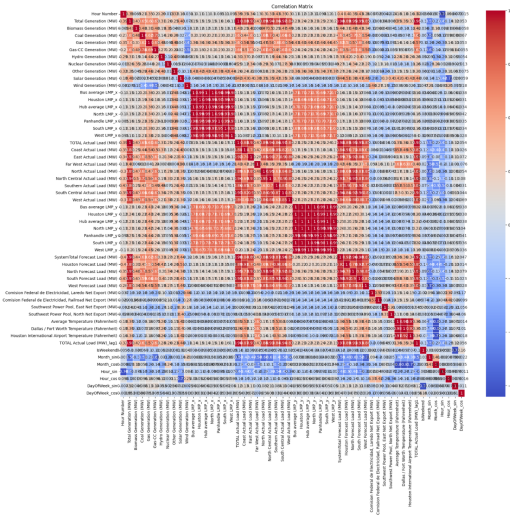

Fig. 1. Correlation Matrix of Features

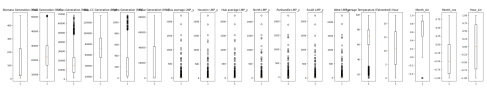

Fig. 2. Features before winsorization

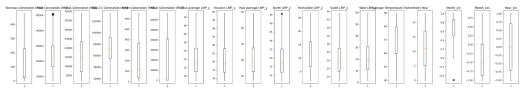

Fig. 3. Features after winsorization

3 MODEL EVALUATION

Key metrics used included **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **R-squared (R$\hat{2}$)**. The evaluation results for different models are summarized below:

3.1 Linear Regression

Linear Regression served as the baseline model for this project. It achieved an **MAE** of 3013.79, an **MSE** of 15,272,932.95, and an **R$\hat{2}$** score of 0.839, indicating reasonable predictive accuracy but room for improvement in capturing non-linear relationships.

3.2 Decision Tree Regressor

The Decision Tree Regressor demonstrated improved performance, achieving an **MAE** of 1928.85, an **MSE** of 7,763,233.63, and an **R$\hat{2}$** score of 0.918. Feature importance analysis revealed that **Average Temperature (Fahrenheit)** was the most significant predictor, as shown in Figure 4.

3.3 Optimized Decision Tree Regressor

Using **GridSearchCV**, the Decision Tree Regressor was optimized with parameters `max_depth=15`, `min_samples_leaf=2`, and `min_samples_split=10`. This resulted in an **MAE** of 1930.27, an **MSE** of 7,320,337.37, and an **R$\hat{2}$** score of 0.923, further enhancing model performance.

3.4 Random Forest Regressor

The Random Forest Regressor outperformed all other models, achieving an **MAE** of 1394.32, an **MSE** of 3,603,917.51, and an **R$\hat{2}$** score of 0.962. This model successfully captured non-linear patterns and interactions among features, providing highly accurate predictions.

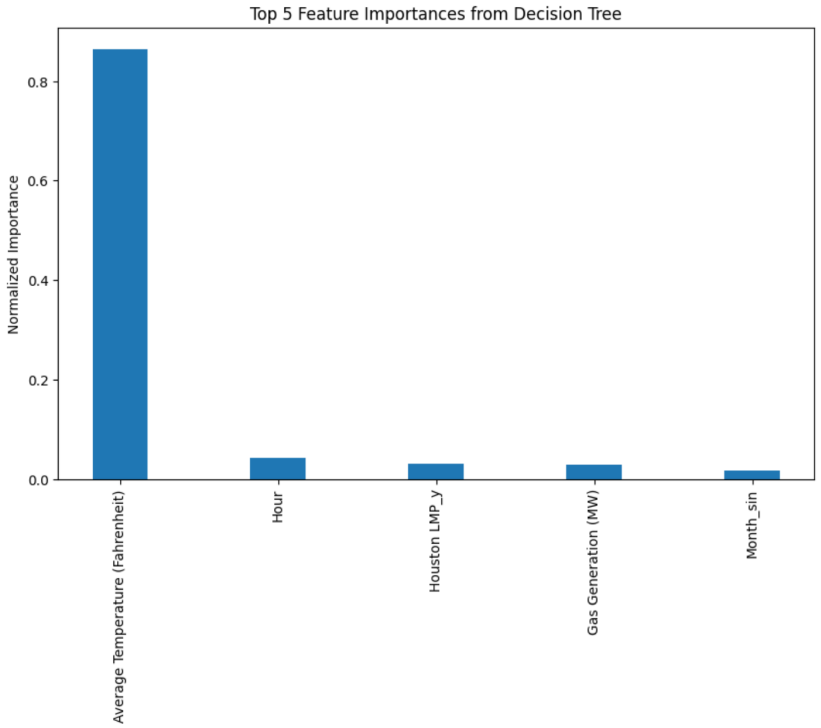

Fig. 4. Top 5 Feature Importances from Decision Tree Regressor

3.5 SHAP Analysis

SHAP (SHapley Additive exPlanations) is a machine learning interpretability framework used to explain individual predictions and overall feature importance by attributing contributions of features to the model's output. It provides a deeper understanding of how each feature influences predictions.

Figure 5 shows the SHAP summary plot, highlighting the average impact of features on the model output. **Average Temperature (Fahrenheit)** emerges as the most significant predictor, followed by **Hour** and **Month_sin**.

Additionally, Figure 6 presents a SHAP force plot for a specific prediction. This visualization shows how individual features contribute to pushing the prediction higher or lower compared to the base value. **Coal Generation (MW)** and **Average Temperature (Fahrenheit)** significantly increase the prediction, while **South LMP_y** has a decreasing effect.

3.6 K-Nearest Neighbors (KNN) Analysis

Historical data (dates before the target date) was separated from current-day data, and Euclidean distances between the current day's scaled feature values and all historical feature values were computed.

The day with the smallest distance to the target date was identified as the most similar day. This approach enabled direct comparisons of electricity demand and feature patterns. It provided an effective mechanism for temporal pattern recognition in the dataset.

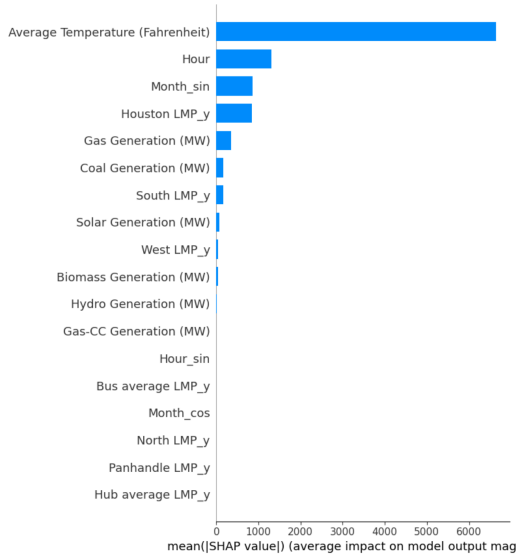

Fig. 5. SHAP Feature Importance Summary

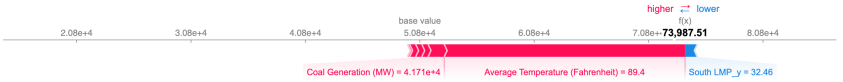

Fig. 6. SHAP Force Plot for a Specific Prediction

3.7 Prompt Engineering using LLMs

Large Language Models (LLMs) were evaluated to determine their capacity for extracting insights from datasets in various formats. Initially, multiple LLMs, including OpenAI's GPT-3, Bard, and Claude, were tested for their ability to process raw data, CSV files, summarized data, and graphs.

Initial attempts to input raw data, including Jupyter notebooks and CSV files, revealed significant limitations. For instance, when asked to generate a SHAP summary plot directly from a Jupyter notebook, ChatGPT encountered errors and provided generic responses such as "I can't do advanced data analysis right now." Additionally, when processing raw data files exceeding 10 in number, the models consistently failed to deliver coherent outputs.

Summarized data and visual aids emerged as the most effective input formats. By providing data in concise tables or leveraging pre-generated graphs, the LLMs demonstrated improved comprehension and contextual understanding. For example, when tasked with analyzing feature importance from a Random Forest model, ChatGPT successfully identified "Average Temperature (Fahrenheit)" as the most critical predictor and generated insightful visualizations. However, attempts to forecast electricity demand using lag features or to implement ARIMA models yielded suboptimal results due to a lack of direct access to computational frameworks.

The iterative process of prompt refinement highlighted the need for specificity and clarity in queries. For instance, specifying exact features and target variables, such as "TOTAL Actual Load (MW)," enabled the model to deliver more accurate and contextually relevant predictions. In one case, incorporating lag values into prompts allowed the model to simulate temporal dependencies, though the accuracy of these predictions required further validation against external benchmarks.

4 DISCUSSION

The evaluation of models revealed clear insights into their suitability and performance for our project, which involved a dataset characterized by numerous features and non-linear relationships. From the outset, the goal was to work with tree-based models like Decision Tree Regressors, given their ability to handle non-linear data. However, a linear regression model was also tested to serve as a baseline and assess its performance.

The Linear Regression model, as anticipated, produced modest results with an $\mathbf{R\hat{2}}$ score of 0.81. While it is a simple and interpretable model, its inability to capture non-linear relationships limited its effectiveness for this project. Nonetheless, it provided a useful benchmark to gauge the improvements offered by more complex models.

The Decision Tree Regressor significantly improved upon the baseline, achieving an $\mathbf{R\hat{2}}$ score of 0.92. This model demonstrated the ability to capture non-linear relationships in the data and offered valuable interpretability through feature importance analysis. However, it was observed that Decision Trees can sometimes overfit, particularly when used without additional regularization or ensemble methods.

To address this, the Random Forest Regressor was implemented, which builds upon the strengths of Decision Trees by averaging the results of multiple trees. This approach further enhanced predictive performance, yielding an $\mathbf{R\hat{2}}$ score of 0.95, the highest among the tested models. Additionally, the ensemble nature of Random Forest reduced the risk of overfitting and improved the model's robustness.

In summary, while the Decision Tree Regressor provided strong results and interpretability, the Random Forest Regressor emerged as the most suitable model for this project. Its ability to handle non-linear relationships, combined with its superior predictive performance and robustness, makes it the best choice for electricity demand prediction in this context. Future work could explore other ensemble methods or incorporate additional optimization techniques to further enhance performance.

The findings revealed that LLMs struggled to process raw data, particularly when dealing with more than 10 files of raw input, indicating a limitation in handling unprocessed data directly. While LLMs could work with CSV files, their performance declined as the dataset size increased, especially with dimensions as large as `(23935, 66)`. However, the most effective results were achieved when summarized data or visual representations such as graphs were provided. Summarized data enabled LLMs to interpret patterns and relationships more efficiently, while graphs offered an intuitive medium for insights. Ultimately, **ChatGPT** was selected for its performance across multiple contexts and its ability to generate actionable insights effectively.