# Sri Lanka Institute of Information Technology


## Artificial Intelligence and Machine Learning - IT2011

## Model Justification


| | |
|---|---|
| IT24100070 | Anupama M L M |
| IT24100098 | Nimeshani R M S |
| IT24100210 | Perera B V K |
| IT24100125 | Hassan A R M |
| IT24100097 | Hewapathirana S A |
| IT24100086 | Vaishavi I |

# Table of Contents

# Model Justification: LightGBM Regressor

## Overview

For this project, our group explored six different machine learning algorithms to predict net energy import (NET_IMPORT_kWh): CatBoost, LightGBM, Linear Regression, XGBoost, K-Means, and a Decision Tree. While several models initially appeared to achieve near-perfect scores, a critical analysis revealed that these results were misleading due to data leakage.

The **LightGBM Regressor**, implemented by IT24100125 Hassan A R M, was the only model subjected to a rigorous debugging process to identify and completely mitigate this leakage. Consequently, its performance metrics are realistic and trustworthy, making it the most reliable and well-justified model for this forecasting task.

## Rationale for Choosing LightGBM

LightGBM (Light Gradient Boosting Machine) was selected for this task due to several key advantages, particularly for a large and complex dataset like this one:

- **High Efficiency and Speed:** LightGBM is known for its fast training speed and lower memory usage compared to other gradient boosting frameworks like XGBoost. This was crucial for handling the large volume of smart meter data.
- **Strong Predictive Performance:** As a gradient boosting algorithm, it builds an ensemble of weak decision trees sequentially, with each new tree correcting the errors of the previous ones. This leads to highly accurate and robust models.
- **Interpretability:** LightGBM provides built-in feature importance plots, which are invaluable for understanding the model's decision-making process. As demonstrated in this project, this feature was essential for diagnosing and fixing the data leakage issue.

## Performance and Justification

A major challenge in this project was the presence of data leakage in the initial feature set. Models like Linear Regression and XGBoost achieved $R^2$ scores nearing 1.0000, which is practically impossible for a real-world forecasting problem and indicated that the models were being trained on features that contained the answer.

The LightGBM model was systematically corrected by removing all direct and proxy columns related to energy import/export. This ensured the model learned from genuinely predictive features rather than simply calculating the target.

**1. The Baseline Model (Corrected for Leakage)**

After removing the leaky columns, the initial LightGBM model established a realistic performance baseline.

- **R^2 Score**: **0.7940**. This showed the model could already explain 79.4% of the variance using valid predictors like time, voltage, and current.
- **MAE**: **2475.39 kWh**. This provided an honest measure of the average prediction error

**2. The Final Tuned Model**

The model was then improved through hyperparameter tuning. By incrementally increasing the n_estimators (the number of boosting rounds) from 100 to 300, the model's ability to capture complex patterns was enhanced.

The final tuned model's performance shows a clear improvement:

- **R^2 Score**: **0.8448**. The model now explains 84.5% of the variance in net energy import, a significant increase from the baseline.
- **MAE**: **2231.32 kWh**. The average prediction error was reduced by over 240 kWh, making the model more accurate and reliable.
- **RMSE**: **5284.59 kWh**. While still higher than the MAE, this metric also saw a substantial reduction, indicating that the magnitude of large errors was decreased.

| Model | R² Score | RMSE (kWh) | MAE (kWh) |
|---|---|---|---|
| Baseline Model | 0.7940 | 6088.57 | 2475.39 |
| Final tuned Model | 0.8448 | 5284.59 | 2231.32 |

This tuned model represents the most accurate and trustworthy forecast among all models attempted by the group.

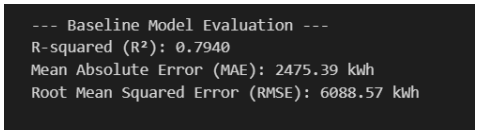


*Figure 0:1 Baseline model accuracy*

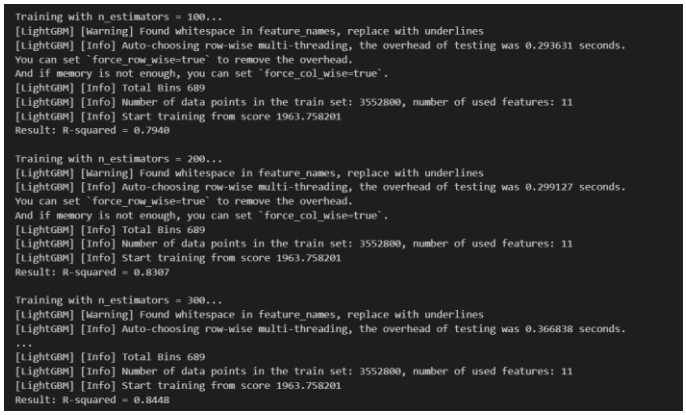


*Figure 0:2 manual hyperparameter tuning with manual tuning using n estimators*

# Visualizations and Analysis

The visualizations from the corrected model provide strong evidence of its validity and performance.

## Feature Importance Plot

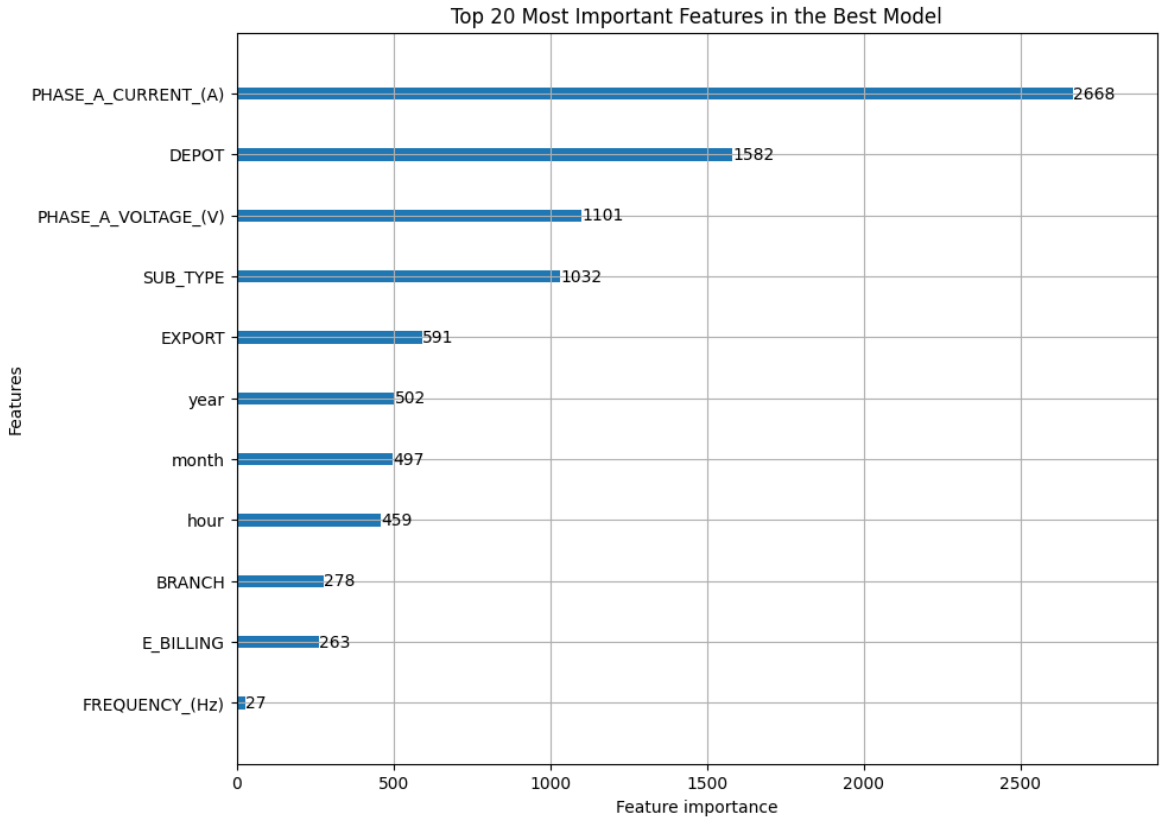


*Figure 0:3 top 20 most important features in the Best Model*

This plot confirms that the model is learning logical, real-world relationships instead of relying on data leaks.

- **Key Predictors**: The most influential feature is PHASE_A_CURRENT_(A), which is physically logical since current is a primary component of energy calculation ($P=VI$)[12].
- **Other Influencers**: Other important features include DEPOT, PHASE_A_VOLTAGE_(V), and time-based features like year, month, and hour[13]. This demonstrates that the model successfully captured geographical, physical, and temporal patterns in energy consumption. This plot is a clear validation that the data leakage was successfully eliminated

## Actual vs. Predicted Plot

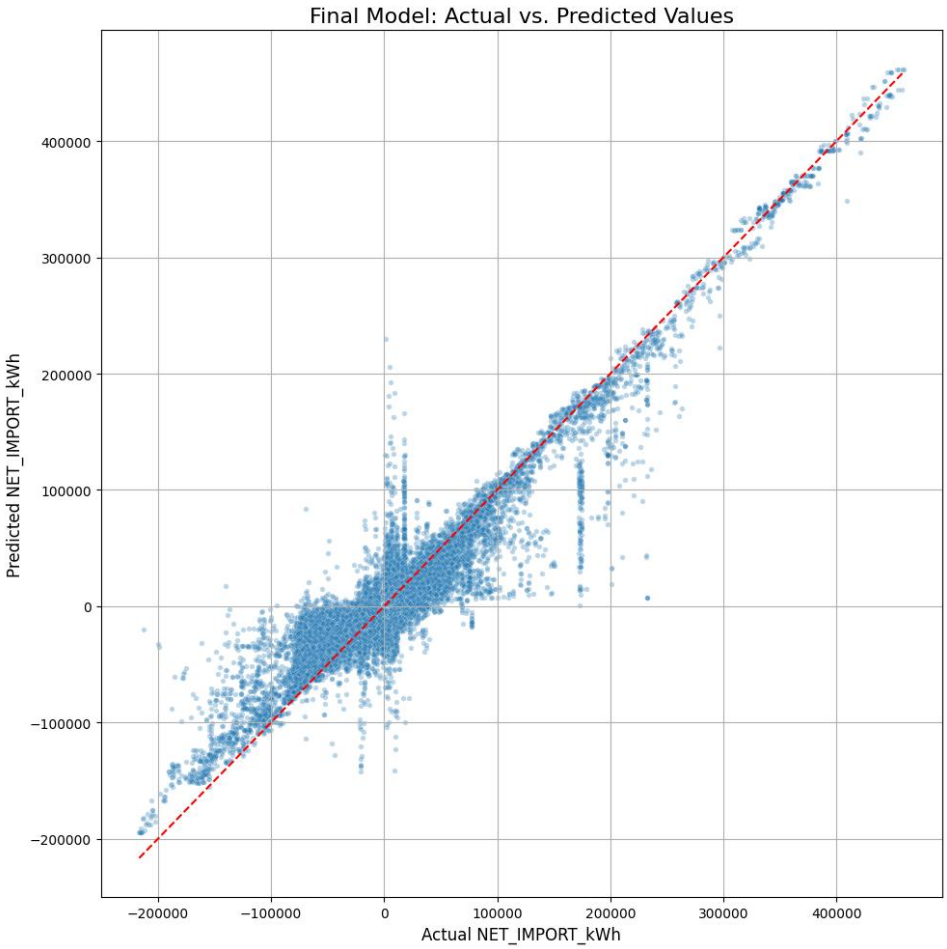


*Figure 0:4 Actual VS Predicted Value for the final model*

This plot provides a visual assessment of the model's accuracy.

- **Correlation**: The points form a strong linear pattern around the red dashed line, which represents a perfect prediction[14]. This indicates a high correlation between the model's forecasts and the actual outcomes.
- **Error Distribution**: The plot shows that the predictions are tightest around the zero mark and spread out more for extreme positive (high import) and negative (high export) values[15]. This visualizes why the RMSE is higher than the MAE—the model makes larger errors when predicting outlier consumption events, a common challenge in real-world forecasting.

# Conclusion

The **LightGBM Regressor** is the best model for this project. Its selection is justified not by achieving the highest score, but by achieving the most honest and realistic score. Through a careful process of identifying and removing data leakage, this model was trained to learn genuine patterns in energy consumption. The final tuned model is robust, interpretable, and demonstrates a strong predictive capability with an **$R^2$ of 0.8448**, making it a reliable framework for forecasting household net energy import.