Forecasting Sectoral Energy Usage in Malaysia

(PowerCastMy)

Muhammad Abqari bin Zulkifli

School of Computing, Universiti Utara Malaysia

Suwannit Chareen Chit a/l Sop Chit

Oct 15 2024

Forecasting Sectoral Energy Usage in Malaysia

PowerCastMy

Introduction

Energy usage forecasting is an essential tool for policy-makers, energy providers, and stakeholders in the energy sector. By anticipating future energy demand across different sectors, organizations can effectively plan to meet these needs, preventing both shortages and oversupply. Accurate forecasts also support informed decisions regarding infrastructure investments, energy efficiency initiatives, and the integration of renewable energy sources.

This document outlines the process of forecasting sectoral energy usage, detailing the methodologies employed, data sources, model selection, and performance evaluation. By applying various forecasting techniques and fine-tuning them through hyperparameter optimization, we aim to produce reliable forecasts that can guide strategic energy management.

Data Collection and Sources

Reliable forecasts depend heavily on accurate and comprehensive data. In this project, the data was sourced primarily from the Malaysia Energy Statistics Handbook 2020, which provides historical energy consumption data across different sectors. This source was chosen for its detail and wide coverage of energy usage statistics, including industrial, residential, commercial, and transportation sectors​. The handbook also highlights trends in energy consumption over time, which serve as a key input for creating predictive models. Here is the content of the handbook that were used:



The sectors in focus for this energy forecasting project include:

* Residential: Energy consumption patterns for households.
* Industrial: Energy usage in manufacturing, mining, and construction activities.
* Commercial: Energy use in businesses, services, and office buildings.
* Transportation: Fuel and electricity usage in personal, public, and commercial transportation.

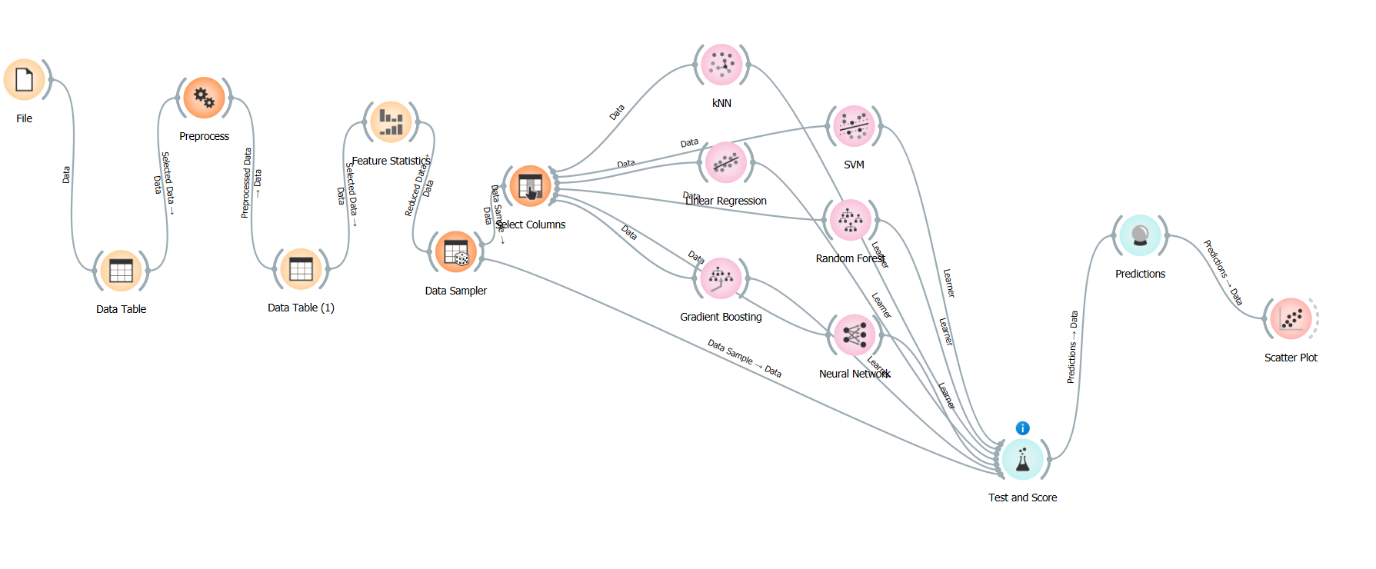
These sectors were chosen due to their significant impact on national energy demand and their distinct consumption patterns.

Methodology

The forecasting process involves several steps, including data preprocessing, feature selection, model training, evaluation, and hyperparameter tuning. Each of these stages is crucial for creating a robust and accurate forecasting model.

* Data Preprocessing: Before modeling, the dataset was cleaned to handle missing values and outliers. In this case, imputation was necessary as only one data point was missing. Normalization of data was deemed unnecessary due to the focus on models like decision trees and random forests, which do not require normalized inputs.
* Feature Selection: Predictor variables, such as industry, transport, residential and commercial, non-energy use and agriculture were chosen based on their potential impact on future energy demand. By selecting the most relevant features, we aimed to improve model accuracy and reduce overfitting.
* Model Training: The selected models were trained on historical energy consumption data. Each model’s performance was evaluated using a test set to assess how well it can generalize to new, unseen data.

Model’s layout are as follows:

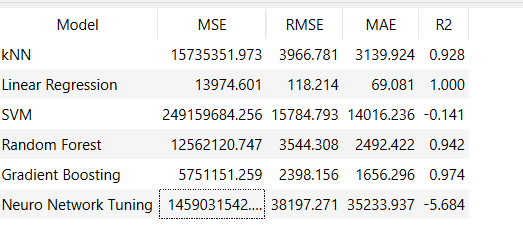


Models Considered

Various machine learning models were tested to forecast sectoral energy usage. The performance of each model was evaluated using key metrics such as R², RMSE, and MAE, which measure how well the model fits the data and how accurate its predictions are.

* k-Nearest Neighbors (kNN): This model performed reasonably well, explaining 92.8% of the variance in the data. However, its relatively high RMSE and MAE indicate that it may not be the most suitable for precise energy usage forecasting, as the errors in prediction were higher compared to other models.
* Linear Regression: Linear regression captured nearly 100% of the variance, leading to a near-perfect fit. The low MSE, RMSE, and MAE further suggest that it can accurately model the data. However, the high R² value may indicate overfitting, where the model performs well on the training data but may not generalize well to future data​(Orange\_Malaysia\_Sectoral).
* Support Vector Machine (SVM): SVM performed poorly in this context, with a negative R² value, making it unsuitable for energy demand forecasting. The complexity of energy usage patterns across sectors might not align well with the rigid structure of SVM.
* Random Forest: Random forest, a powerful ensemble model, showed a high R² value with lower error metrics compared to kNN and SVM. Its ability to capture non-linear relationships in the data makes it a strong candidate for forecasting sectoral energy demand.
* Gradient Boosting: Similar to random forest, gradient boosting delivered high accuracy, with a high R² value and low error metrics. It is another ensemble method that builds on the strengths of decision trees, improving predictions by focusing on areas where previous models underperformed.
* Neural Networks: Neural networks, surprisingly, performed poorly in this case. Despite their success in other predictive modeling tasks, their performance on this dataset was suboptimal, possibly due to the relatively small dataset size and the need for more extensive training to capture complex patterns​(Orange\_Malaysia\_Sectoral).

Model’s performance



Model Performance and Selection

After testing all the models, the best performers were identified as:

* k-Nearest Neighbors (kNN)
* Linear Regression
* Random Forest
* Gradient Boosting

These models stood out for their ability to explain a significant portion of the variance in the data while maintaining relatively low error metrics. However, caution is warranted when interpreting the results from linear regression, as its near-perfect R² score could be a sign of overfitting. Random forest and gradient boosting, on the other hand, showed strong generalizability and are recommended for further use.

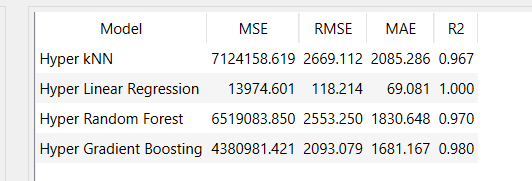
Hyperparameter Tuning

Hyperparameter tuning is the process of optimizing model parameters to improve performance. For this project, tuning focused on maximizing R² values while minimizing RMSE and MAE. Each of the top-performing models underwent hyperparameter adjustments to refine their predictions.

* kNN: The number of neighbors was adjusted to improve the model's accuracy.
* Linear Regression: Regularization techniques were explored to prevent overfitting.
* Random Forest: The number of trees and maximum depth were optimized to reduce overfitting and enhance performance.
* Gradient Boosting: Parameters such as learning rate and number of estimators were fine-tuned to increase prediction accuracy.

The tuned models, referred to as “Hyper” models, showed improved performance, with lower error metrics and higher R² values.

Hyper Model’s performance



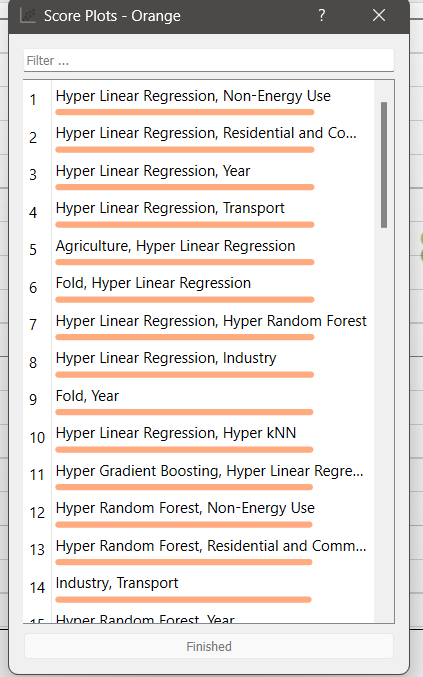
Challenges and Considerations

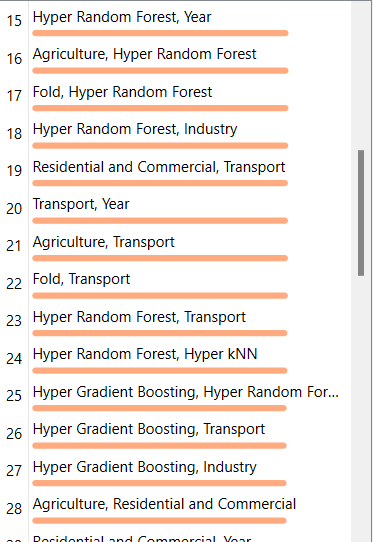
Despite the success of the selected models, several challenges remain:

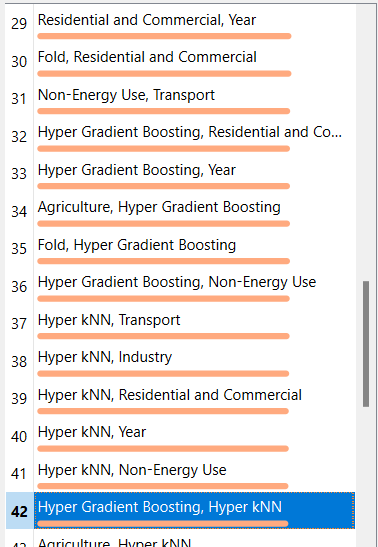
* Sectoral Variability: Energy consumption patterns differ widely between sectors. For example, the industrial sector may experience more fluctuations due to economic cycles, while the residential sector might show more stable growth. Models need to account for this variability to ensure accurate predictions across sectors.
* Technological Advancements: The future of energy consumption will be shaped by rapid advancements in technology, particularly in renewable energy and energy efficiency. While these trends are difficult to predict, incorporating them into the forecasting process could enhance the model's long-term accuracy.
* Data Limitations: The accuracy of any predictive model is limited by the quality and quantity of the data. In this case, some data gaps and potential inconsistencies in sectoral energy usage statistics could affect model performance.

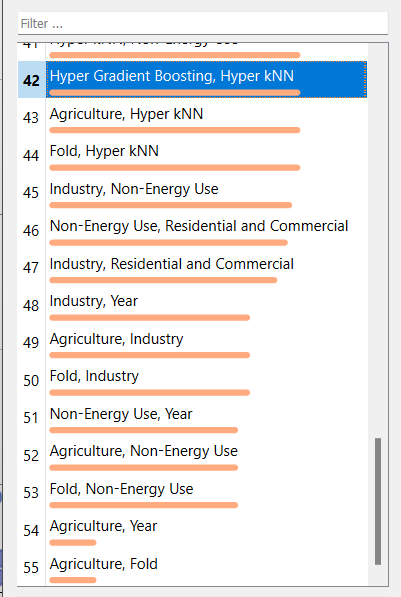
Visualization and Scatter Plots

Scatter plots are an essential in evaluating the performance of various forecasting models. In this project, scatter plots were used to assess the relationship between the predicted and actual energy usage for each sector, across different models such as Hyper Linear Regression, Hyper Random Forest, and Hyper kNN. For example, scatter plots comparing the actual versus predicted energy consumption for the Residential and Commercial sector, modelled using Hyper Linear Regression, help reveal how well the model captures energy demand in these sectors.









Due to the overfitting nature of the hyperlinear regression model, it is advised avoid using score plots with this model.

Model Performance Analysis

A screenshot of a computer

Description automatically generated

Different forecasting models were evaluated using error metrics such as MSE, RMSE, MAE, and R². The performance of models such as Hyper Random Forest, Hyper Gradient Boosting, Hyper kNN, and Hyper Linear Regression was evaluated using these criteria to determine the best accurate and trustworthy model for sectoral energy forecast.

Hyper Linear Regression has a R² of 1.000, showing it captures all variance in data. However, perfect R² might be deceiving and signal overfitting, when the model performs extraordinarily well on training data but does not generalize to unknown data. Its low RMSE and MAE values indicate little prediction errors, however caution should be exercised due to potential overfitting.

Hyper Random Forest demonstrated great predictive power, with a R² of 0.996. Its RMSE (906.26) and MAE (569.96) figures indicate slightly larger error than Hyper Linear Regression, but its robust performance makes it less susceptible to overfitting. This makes the Hyper Random Forest a good compromise between accuracy and generalizability.

Hyper Gradient Boosting and Hyper kNN achieved R² values of 1.000, identical to Hyper Linear Regression, suggesting near-perfect fit. However, it is critical to assess the differences in MSE, RMSE, and MAE to determine their genuine predictive capability and potential overfitting concerns.

Conclusion

Forecasting sectoral energy usage is a complex task that requires careful selection of models, data, and features. In this case, kNN, random forest, and gradient boosting emerged as the most effective models for predicting energy demand. Hyperparameter tuning helped further enhance their performance, making these models suitable for energy forecasting in the near term.

Reference

SURUHANJAYA TENAGA (ENERGY COMMISSION). (2021). MALAYSIA ENERGY STATISTICS HANDBOOK 2020. In *https://www.st.gov.my*. Retrieved October 15, 2024, from https://www.st.gov.my/en/contents/files/download/116/Malaysia\_Energy\_Statistics\_Handbook\_20201.pdf