**Predicting Household Energy Consumption Using Random Forest Regression**

**Author**: **Pushkar Yadav**

**Abstract**

The paper discusses the application of machine learning in energy consumption forecasting using Random Forest regression. We will apply the historical household power consumption data in order to forecast total energy consumption with a good level of accuracy. Extensive preprocessing, feature engineering, and hyperparameter tuning have been performed in this work with the intent of optimizing performance. The results depict that the ensemble methods are efficient for the capturing of complex relationships between power-related features and energy usage. It falls into the wider domain of energy management and provides actionable insights for smart energy systems.

**1. Introduction**

In recent times, efficient energy management has become a major concern around the world with the increased awareness of sustainability and resource conservation. Effective prediction, therefore, will be needed in household energy consumption to contribute towards optimizing energy use and limiting wastage. The availability of smart meters and high resolutions of energy data involves a great opportunity for the application of machine learning algorithms for the forecast of future energy needs, hence enabling informed decisions by consumers and providers alike.

The following project deals with the forecast of households' energy consumption by using a Random Forest regression model. Some power-related features in this dataset include global active power, voltage, and intensity, which could be used to predict total energy consumption. Random Forest was chosen because it is a robust method that can be applied to high-dimensional data analysis and makes use of ensemble learning in order to reduce overfitting. A prediction like this will lead to smarter energy distribution systems, more consumption-efficient habits, and altogether help achieve sustainability goals.

**2. Methodology**

**2.1 Dataset**

The dataset utilized for this study is derived from a household power consumption dataset. It includes the following key features:

* **Global\_active\_power**: Active power consumed (in kW).
* **Global\_reactive\_power**: Reactive power (in kW).
* **Voltage**: Electrical voltage (in volts).
* **Global\_intensity**: The amount of current being drawn (in amperes).
* **Total\_energy**: The target variable, representing total energy consumption, which is computed as:

\text{total\_energy} = \frac{\text{Global\_active\_power} \times 1000}{60}

**2.2 Preprocessing**

Data preprocessing involved several key steps:

* **Handling Missing Data**: Missing values in critical features like Global\_active\_power and Voltage were imputed using the **mean** imputation strategy.
* **Feature Engineering**: We created the total\_energy feature to capture the overall energy consumed per unit of time.
* **Data Splitting**: The dataset was split into training and testing sets (80/20 split), ensuring that model evaluation was conducted on unseen data.

**2.3 Machine Learning Model**

A Random Forest Regressor has chosen because the following are nonlinear in nature: relationship among the features with respect to the target variable. Random Forest creates multiple decision trees and then aggregates their prediction. This results in a better generalization compared to a single decision tree. The risk of overfitting decreases.

**2.4 Model Tuning and Evaluation**

To optimize the model's performance, we employed **GridSearchCV**, a systematic method for tuning hyperparameters. We performed a **5-fold cross-validation**, testing the following hyperparameters:

* **n\_estimators**: [100, 200, 300] – Number of trees in the forest.
* **max\_depth**: [None, 10, 20] – Maximum depth of each tree.
* **min\_samples\_split**: [2, 5] – Minimum samples required to split an internal node.
* **min\_samples\_leaf**: [1, 2] – Minimum number of samples required at a leaf node.

The evaluation metric was the **negative mean squared error (MSE)**, with the goal of minimizing prediction error.

**3. Results**

The Random Forest model’s hyperparameter tuning yielded the following optimal parameters:

* **n\_estimators**: 200
* **max\_depth**: None
* **min\_samples\_split**: 5
* **min\_samples\_leaf**: 2

After training, the model achieved a **Test MSE** of 6.451010102085096e-05, demonstrating strong predictive performance on unseen data. Additionally, feature importance analysis revealed that **Global\_active\_power** was the most significant predictor of total energy consumption.

**Visualizations**

**A screen shot of a graph

Description automatically generated**

* **Figure 1**: Feature Importance plot, highlighting the contribution of each feature to the model’s predictions.

A graph of orange lines

Description automatically generated

* **Figure 2**: Actual vs Predicted energy consumption, illustrating the model’s accuracy across different instances.

These visualizations confirm the model's ability to generalize well, providing consistent predictions across a wide range of test cases.

**4. Discussion**

These results hereby prove that Random Forest regression has worked effectively in the prediction of household energy consumption. This model is better suited to handle nonlinear relationships between features inherent in complex data like this. Of course, Global\_active\_power had the highest feature importance, since it's a direct measure of energy usage. The robust performance after tuning that this model demonstrated also suggests that the model may be used fairly effectively in real-world situations where energy forecast accuracy is critical.

This approach, however, has its limit. While the Random Forest model is doing a great job in this case, it does not embed the temporal dependencies in the data. Normally, energy consumption depends on a lot of time-based factors, such as seasons and time of day, which are not clearly modelled here. Future work shall use some sort of time series approach like ARIMA or LSTM to be able to account for these factors and probably increase accuracy. Second, while the use of grid search for hyperparameter tuning has been quite effective, it could have been further augmented with Bayesian optimization-based methods that would have explored the parameter space more effectively.

**5. Conclusion**

This work has demonstrated the potential of machine learning with Random Forest regression in the forecasting of energy consumption for a household. History of power data combined with systematic model tuning resulted in accurate forecasts of total energy consumption. These results can take part in the development of smarter energy management systems, contributing toward more productive and efficient usage and sustainability. Such work should be continued by the incorporation of temporal factors and the implementation of more advanced techniques to tune models.

**6. References**

1. Kaggle. (n.d.). Household Power Consumption Dataset.
2. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
3. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
4. Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

**7. Conference Alignment**

This project aligns with the **International Conference on Sustainable Energy Information Technology-SEIT** because it focuses on the application of machine learning in energy management. The sustainability perspective contributes to providing a solution in optimizing energy consumption, one of the basic issues of sustainable development. Building wiser energy systems is facilitated through this work since it provides solid predictions for energy usage.

<https://cs-conferences.acadiau.ca/seit-24/>