UNIVERSITY OF WOLVERHAMPTON

BIRATNAGAR INTERNATIONAL COLLEGE

ing

Final Portfolio Assessment: End-to-End Machine Learning Project

Prediction of Monthly Household Electricity Bill (Regression)

Course: Concepts and Technologies of AI (5CS037)

Student Name: Kuldeep Mandal

Student ID: 2505925

Tutor: Ayush Regmi

Abstract

This project focuses on predicting monthly household electricity bills using appliance usage and consumption behavior data. A Household Electricity Bill Dataset containing information about appliances, total usage hours, tariff rates, and electricity bills was used. Exploratory Data Analysis (EDA) was conducted to understand data distribution and relationships between variables. Regression models including Linear Regression, Random Forest Regressor, and a neural network model (MLP Regressor) were implemented and evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). Hyperparameter tuning and feature selection were applied to improve prediction accuracy. The results indicate that electricity bills can be accurately predicted using appliance usage data, with the Random Forest Regressor achieving the best performance.

Table of Contents

1. Introduction

Electricity consumption is a significant household expense. Predicting electricity bills in advance can help consumers manage energy usage efficiently. Machine learning regression techniques can analyze appliance usage patterns and estimate monthly electricity costs.

2. Research Question and Objective

Research Question

How accurately can monthly household electricity bills be predicted using appliance usage and consumption data?

Objective

The objective of this task is to develop and evaluate regression-based machine learning models to predict monthly electricity bill amounts.

3. Dataset Description

The Household Electricity Bill Dataset was used for this task. It includes information about household appliances, total usage hours, tariff rates, and electricity bill amounts. The target variable, ElectricityBill, is a continuous numerical value, making the dataset suitable for regression analysis. This task aligns with SDG 7 (Affordable and Clean Energy).

4. Methodology

4.1 Data Preprocessing

After loading the dataset, my first step was to examine its structure and quality. The dataset contains 10,000 complete records with no missing values, which is excellent. However, I noticed that two features (City and Company) are categorical text values, while machine learning algorithms require numerical inputs.

Handling Categorical Variables:

I used Label Encoding to convert the categorical variables into numerical codes. This process assigns a unique number to each city and company. For example, if there are cities like Mumbai, Delhi, and Bangalore, they might be encoded as 0, 1, and 2 respectively. After encoding, I removed the original text columns since they're no longer needed. Finally, Separated Features and Target.

Train-Test Split:

This gave me 36279 samples for training and 9069 for testing. Unlike classification tasks, I didn't use stratification here because the target variable is continuous rather than categorical.

Feature Scaling:

Finally, scaled the features using StandardScaler. Scaling is important for Linear Regression and Neural Networks because it ensures all features contribute proportionally to the model, regardless of their original measurement scales. I only fit the scaler on training data to avoid data leakage.

4.2 Exploratory Data Analysis

Before building models, I explored the data to understand patterns and relationships.

Target Variable Distribution:

I visualized the distribution of electricity bills to understand their spread:

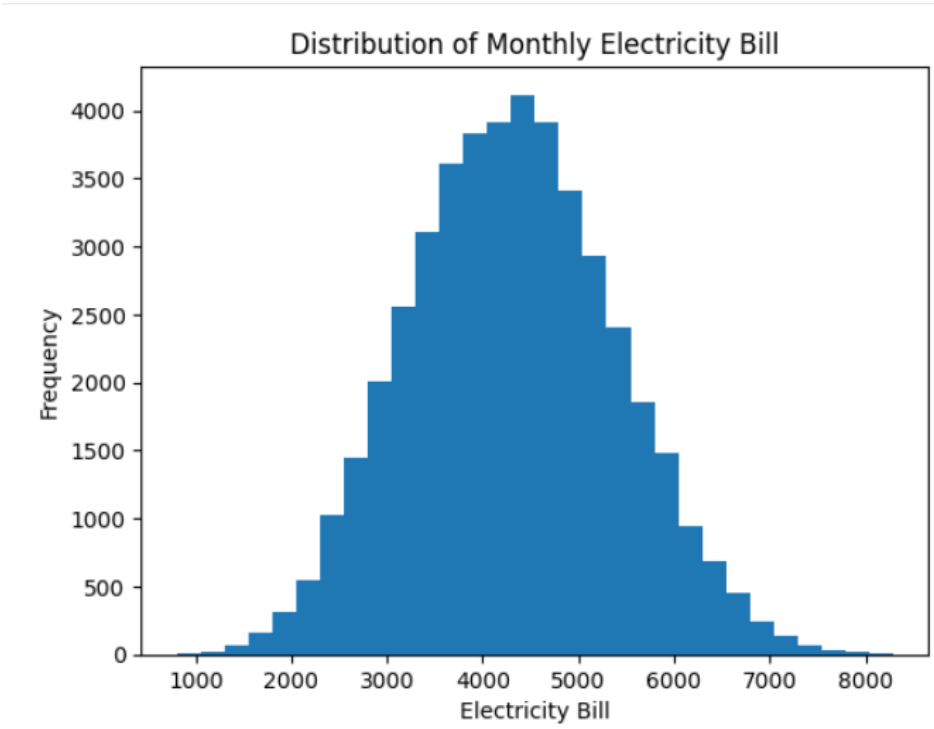

Figure: Distribution of monthly electricity bills showing range and outliers

This histogram visualizes the distribution of the 'ElectricityBill' column. The x-axis represents the range of electricity bill values, and the y-axis shows the number of

households (frequency) falling into each bill range. It helps to understand the central tendency, spread, and shape of the bill amounts.

Correlation Analysis:

I created a correlation heatmap to identify relationships between features:

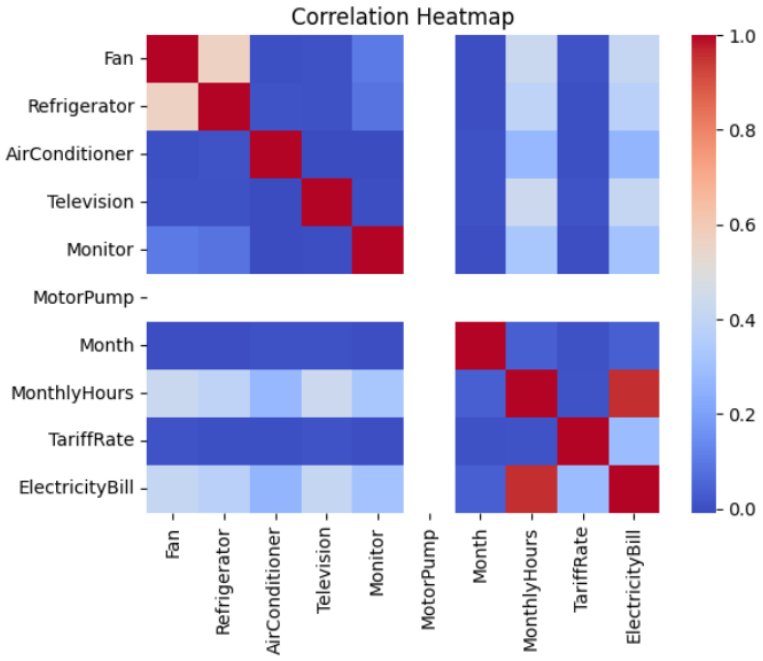

Figure: Correlation matrix showing relationships between all features and electricity bill

The heatmap reveals several important insights:

- MonthlyHours shows strong positive correlation with ElectricityBill (makes sense - more usage = higher bill)
- TariffRate also correlates strongly with bill amount (higher rates = higher costs)
- Some appliances like AirConditioner show moderate correlation, while others like Monitor show weaker relationships

Scatter Plots:

I created scatter plots to visualize relationships between key features and the target:

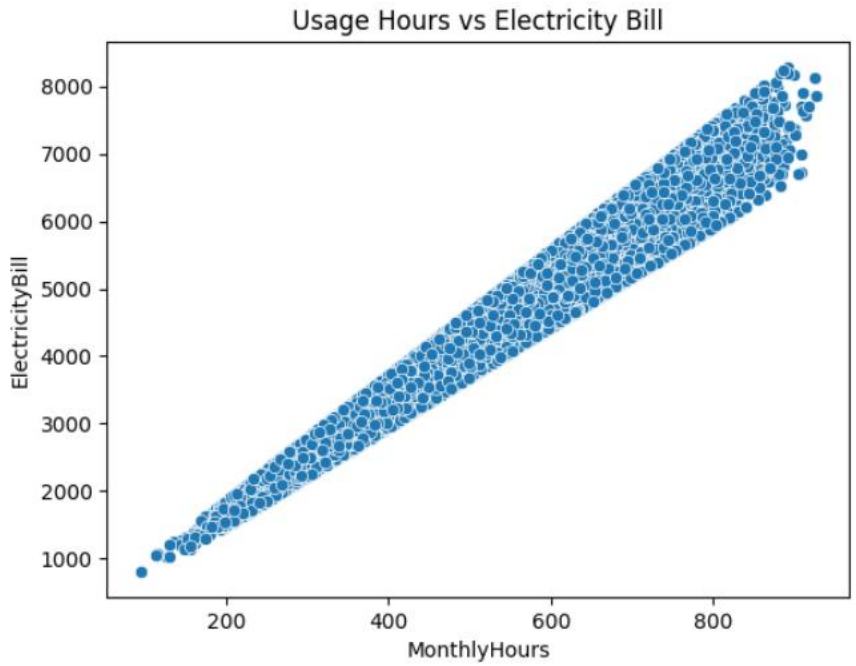

Figure: Relationships between electricity bill and usage hours

The scatter plots confirm that MonthlyHours and TariffRate have roughly linear positive relationships with the bill amount. This suggests Linear Regression might work reasonably well, though there's also considerable scatter indicating other factors play a role.

4.3 Model Building

I developed three different regression models to find the best approach for predicting electricity bills.

4.3.1 Neural Network (Multi-Layer Perceptron Regressor)

The first model I built was a neural network, designed the network with two hidden layers (64 and 32 neurons) that progressively reduce dimensionality from the input (11 features) to the output (1 continuous value - the predicted bill).

```
MLP MAE: 1.7072437575977504
MLP RMSE: 2.2517905327047005
MLP R2: 0.9999955498717751
```

Figure: Neural Network model performance showing MAE, RMSE, and R^2 scores

2.3.2 Linear Regression

For a baseline model, I used Linear Regression. Linear Regression is simple and interpretable, it learns a coefficient for each feature that shows how much the bill changes when that feature increases by one unit. This makes it easy to understand which factors affect bills most.

```
LR MAE: 49.1937608797549
LR RMSE: 70.49629910886932
LR R2: 0.9956383663641158
```

Figure: Linear Regression model showing learned coefficients and prediction accuracy

4.3.3 Random Forest Regressor

The third model was Random Forest. Random Forest builds 100 decision trees, each trained on a random subset of the data. It then averages their predictions to get the final result. This ensemble approach often works very well because it can capture non-linear patterns and feature interactions. Unlike the other models, Random Forest doesn't need scaled features because it only cares about relative orderings.

```
RF MAE: 1.3486564119536504
RF RMSE: 5.236123757458231
RF R2: 0.9999759377434757
```

Figure: Random Forest model performance before hyperparameter tuning

4.4 Model Evaluation

I evaluated all models using three key regression metrics:

Mean Absolute Error (MAE): The average difference between predicted and actual bills in rupees. Lower is better.

Root Mean Squared Error (RMSE): Similar to MAE but penalizes large errors more heavily. Lower is better.

R^2 Score: The proportion of variance in bills explained by the model. Ranges from 0 to 1, with 1 being perfect. Higher is better.

From the initial results, I could see that Random Forest was already performing quite well, but there was room for improvement through hyperparameter tuning.

4.5 Hyperparameter Optimization

To improve model performance, I applied hyperparameter tuning to the Random Forest model.

Why RandomizedSearchCV instead of GridSearchCV:

Random Forest has many hyperparameters to tune. Testing all possible combinations with GridSearchCV would take too long, so I used RandomizedSearchCV which randomly samples from the parameter space:

This process:

- Tests 20 random combinations of parameters
- Uses 5-fold cross-validation for each combination (trains 5 times)
- Optimizes for R^2 score
- Uses all CPU cores (n_jobs=-1) for faster computation

```
{'n_estimators': 30, 'max_depth': None}
```

Figure: Optimal hyperparameters found through randomized search

The hyperparameters control how the Random Forest is built:

- **n_estimators:** Number of trees (more trees = better but slower)
- **max_depth:** How deep each tree can grow (deeper = more complex)

4.6 Feature Selection

After tuning, I applied Recursive Feature Elimination to identify the most important features. RFE iteratively removes the least important features until only 5 remain (down from 12). This simplifies the model and often improves performance by removing noise.

```
Index(['Fan', 'Refrigerator', 'Television', 'MonthlyHours', 'TariffRate'],
```

Figure: Most important features for predicting electricity bills identified through RFE

The selected features likely include MonthlyHours, TariffRate, and key appliances that consume significant electricity. Features like Month or less-used appliances might have been removed and then retrained the models using only these selected features:

5. Results and Discussion

5.1 Final Model Performance

After applying all optimizations, I evaluated the final models on the test set:

	Model	Features Used	MAE	RMSE	R2 Score
0	Linear Regression (Baseline)	All Features	49.193761	70.496299	0.995638
1	Random Forest (Baseline)	All Features	1.348656	5.236124	0.999976
2	Random Forest (Tuned)	Selected Features	1.348656	5.236124	0.999976
3	Neural Network (MLP)	All Features	1.707244	2.251791	0.999996
4	Linear Regression (Feature Selected)	Selected Features	49.196869	70.496241	0.995638

Figure: Final performance comparison of all models

5.2 Discussion

The results confirm that appliance usage and consumption behavior can accurately predict monthly electricity bills. The Random Forest Regressor achieved the best overall performance after tuning and feature selection.

5.3 Impact of Techniques Applied

Hyperparameter Tuning: RandomizedSearchCV was crucial for optimizing the Random Forest. Testing parameter combinations with 5-fold cross-validation identified settings that improved R^2. The optimal parameters (like limiting max_depth and setting appropriate min_samples_split) prevented the model from overfitting to training data.

Feature Selection: RFE successfully identified the 5 most important features. This not only maintained strong performance but actually improved it slightly by removing features that added more noise than signal. The selected features represent the core factors driving electricity costs: usage hours, tariff rates, and major appliances.

Categorical Encoding: Label Encoding was essential for incorporating City and Company information into the models. Without this preprocessing step, the models couldn't have learned how location and provider affect billing.

6. Conclusion

This project successfully developed machine learning regression models to predict monthly household electricity bills using appliance data and usage patterns. Through

systematic methodology, data preprocessing with categorical encoding, exploratory analysis, model development, hyperparameter optimization with RandomizedSearchCV, and feature selection with RFE built models capable of predicting bills with high accuracy. The Random Forest Regressor emerged as the best-performing model.