Appliance Energy Prediction Report

Submitted by: Pragya Gupta
PGDM-Research & Business Analytics

Abstract

This project focuses on predicting the energy consumption of appliances in a household using environmental and temporal variables. Data preprocessing, exploratory data analysis (EDA), and statistical modeling were applied to identify significant predictors of appliance energy usage. A Multiple Linear Regression (MLR) model was trained, but the performance was limited, with an R-squared of 0.19, indicating the need for advanced regression models such as Random Forests or Neural Networks.

1. Introduction

Energy consumption prediction is critical for designing energy-efficient systems and reducing costs. This dataset, collected over 4.5 months, includes household temperature, humidity, and weather data. The aim is to predict appliance energy consumption using statistical and machine learning approaches.

2. Dataset Description

The dataset contains measurements of indoor temperature and humidity across multiple rooms, together with weather data from a nearby weather station. The target variable is 'Appliances' (energy consumed in Wh). Initially, 27 independent variables were available; after removing multicollinearity and insignificant predictors, 18 variables were retained for modeling. The dataset also contained outliers, treated using the IQR method, and scaling was applied via MinMaxScaler.

3. Methodology

The following steps were carried out: • Data Cleaning: Removed the date column, checked for nulls and duplicates, treated outliers via IQR.
• Data Scaling: Applied MinMaxScaler to normalize feature ranges.
• Exploratory Data Analysis: Conducted univariate, bivariate, and multivariate analysis. Correlation heatmaps, scatter plots, violin plots, and pairplots were used to examine relationships.
• Multicollinearity Check: Variance Inflation Factor (VIF) was calculated; features 'rv1' and 'rv2' were dropped.
• Modeling: Built a Multiple Linear Regression (MLR) model using statsmodels. Stepwise elimination was performed to drop insignificant features ($p > 0.05$).
• Diagnostics: Residuals were checked for normality and homoscedasticity.

4. Results & Evaluation

The final MLR model retained 18 statistically significant predictors out of 27. The R-squared value was 0.190, meaning only 19% of the variance in appliance energy consumption could be explained by the model. Residual analysis revealed heteroscedasticity, indicating model assumptions were not fully satisfied. This suggests that linear regression is not the optimal method for this dataset.

5. Conclusion & Insights

The Multiple Linear Regression model demonstrated limited predictive power. Key environmental and weather-related features influenced appliance energy usage, but the low R-squared indicates that more complex models are required. Future work should include advanced regression techniques such as Random Forests, Gradient Boosting, or Neural Networks, which can better capture nonlinear relationships and interactions among features.