Individual Household Electric Power Consumption

Members: Madisen LeShoure and Abeeha Fajar

Dataset

Introduction

This dataset contains data and records of electric power consumption within single households, captured at a minute-level over a period spanning nearly four years. The measurements gathered between December 2006 and November 2010 are very specific, allowing this dataset to capture nuanced insights into the history of energy usage of households. The dataset contains 2,075,259 measurements. Key variables are global active power, global reactive power, voltage, global intensity, and sub-metering values corresponding to specific areas within households such as the kitchen, laundry room, and electric water heater/air conditioner. Each area is distinguished by the submetering categories: sub_metering_1: energy sub-metering corresponds to the kitchen, sub_metering_2 corresponds to the laundry room, and sub_metering_3 corresponds to the electric water-heater and an air-conditioner. This dataset allows us to analyze the relationship between global active power and voltage within a household context. By analyzing electric power consumption at a minute-level granularity over nearly four years, we can explore how changes in voltage influences global active power. With variables such as global active power and voltage available in the dataset, alongside additional features we will analyze the relationship between electricity parameters. Through analysis of this dataset, we will uncover insights into how changes in voltage impacts global active power consumption. The dataset does contain missing values, accounting for approximately 1.25% of the rows. Despite this, all date and time timestamps are present. Data cleaning is essential for addressing missing values to ensure data integrity. Overall, this dataset serves as a resource for understanding household energy consumption patterns, especially when trying to understand the implications of household energy efficiency to better sustainability and climate change efforts. This dataset calls for regression. We are using linear regression to find the patterns in the data provided to us with 9 variables. With those variables we are answering the question: How does the global active power change with voltage?

Data Preprocessing: Removal of Missing Values

In the preprocessing of the dataset, our first objective was to make sure the data was clean for the analysis to ensure the quality and reliability of the data.

We began by thoroughly understanding the structure and characteristics of the dataset. There were 9 columns and 1million + rows. Upon analysis using df.insa.sum() we found that the dataset contained missing values in one of the variables. These missing values needed to be removed for inaccuracies in the analysis. We implemented the removal of missing values by filtering out observations containing missing values using appropriate pandas Data Frame methods provided in the code zip. This step ensured that only complete observations were left for further analysis.

It's important to note that while removing missing values simplifies the preprocessing step, it may also lead to a reduction in the sample size and potentially affect the representativeness of the data. But due to our data set having more than a million rows, it was not a huge concern and on top of that we separated the data so we could look more into 2008. Alot of data was missing in other years so we decided to go with this one.

This preprocessing step lays the foundation for further exploration and insights into the underlying patterns and relationships within the dataset.

Model Selection and Implementation

The linear regression model is trained using scikit-learn's LinearRegression() function. The model is then trained using the training data (X_train and y_train) using the fit() method. Where we split the data into 80 and 20% to test and train the data. The model's optimal coefficients are found to minimize the residual sum of squares between observed and predicted values.

The coefficients represent the slopes of the linear relationship between the features and the target variable ('Global reactive power'). Each coefficient corresponds to a feature in the dataset, with seven coefficients corresponding to seven features. A higher coefficient magnitude suggests a stronger influence of the corresponding feature on the target variable.

The intercept term (1.16e-14) represents the predicted 'Global reactive power' when all features are zero, which is not a meaningful interpretation as it's too small but, in the evaluation, we tried improving it.

The linear regression model's performance in predicting the 'Global reactive power' based on the 'Voltage' feature is evaluated using various metrics. R-squared (R2) measures the proportion of variance in the dependent variable (y_test) explained by the independent variable (y_pred). A perfect R2 score of 1.0 indicates that the model perfectly predicts the variance in the dependent variable. A lower MAE (approximately 1.23e-15) indicates better accuracy, while a smaller MSE (approximately 3.06e-30) indicates minimal error between predicted and actual values. These metrics suggest that the model performs exceptionally well on the test data, accurately predicting the 'Global reactive power' based on the 'Voltage' feature. This all shown in the Code provided with the values. After that we used a scatter plot of predicted values against actual values further confirms the model's performance by showing how closely the predicted values align with the actual values.

A tight cluster of points along a diagonal line indicates a strong agreement between predicted and actual values. Overall, the high R2 score, small MAE and MSE values, along with the scatter plot, provide strong evidence of the linear regression model's effectiveness in predicting the 'Global reactive power' based on the 'Voltage' feature.

Evaluation

Cross-validation scores returned from training SDGRegressor model with different learning rates allowed us to gain insights into the model's performance under different hyperparameters. The learning rates tested range from relatively large values (0.01) to extremely small values (1e-5). A large learning rate can lead to faster convergence but risks overfitting the optimal solution, while a small learning rate requires more iterations to converge but offers better stability. The cross-validation scores range from exceptionally high (close to 1) to moderate (around 0.57). Higher scores mean better model performance and more accurate predictions of global active power changes based on voltage. Cross-validation scores close to 1 means that the model is performing exceptionally well, capturing patterns within the data. The SDGRegressor model shows great performance in predicting how global active power changes with voltage, achieving near-perfect scores with larger learning rates. However, the drop in performance with smaller learning rates indicates the need for careful hyperparameter tuning to strike a balance between convergence and model stability. The best learning rate = 0.01 with an accuracy score of 0.9996853252647504.

In conclusion, after experimenting and evaluating different models and hyperparameters for predicting how global active power changes with voltage, the Linear Regression model demonstrated the best performance. Among the tested hyperparameters this model yielded the highest cross-validation and r2 score.

Future steps

For future steps, we would like to do more models where we can use additional evaluation metrics and then plot. We wanted to add matrices like AIC, BIC or model fit and maybe use Logistic regression to compare and predict. Next, would-be in-depth error analysis to understand the data and areas for improvement. Like identifying outliers, residuals and other systematic errors and trying to minimize the outliers while comparing with different variables.