Electricity Price Prediction Model

The dataset used for this model was downloaded from Kaggle and consists of features such as Holiday, Holiday Flag, Day of the Week, Day, Month, Year, Period of Day, Forecast Wind Production, SystemLoadEA, SMPEA, ORKTemperature, ORKWindspeed, CO2Intensity, Actual Wind Production, SystemLoadEP2, and the target variable, SMPEP2. Using Python in Jupyter Notebook, I employed Pandas, Numpy, Matplotlib, and Seaborn for data import and visualization.

Dataset: Electricity Dataset

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the behavior of the data. Matplotlib and Seaborn libraries were used for this process. Histograms and boxplots were created to study the distribution and detect outliers. Histograms were particularly useful in selecting the most appropriate scaling method. Additionally, pair plots and a heatmap were generated to examine relationships between variables.

Data Cleaning

Handling Missing Values:

All data types were initially numeric, but 10 variables were of object type. These columns were converted to numeric types. Some columns had missing values. The 'Holiday' column had many missing values, so it was removed. Due to a large number of missing values in “ORKTemperature” and “ORKWindspeed,” rows with null values in these columns were removed.

The null values in 'ForecastWindProduction,' 'SystemLoadEA,' 'SMPEA,' 'ActualWindProduction,' 'SystemLoadEP2,' and 'SMPEP2' columns were filled using the median, due to their skewed distributions. The null values in the 'CO2Intensity' column were filled using the mean, considering its normal distribution.

Outlier Detection:

Outliers were identified using boxplots. Data greater than 550 and less than 0 were considered outliers and were removed.

Model Requirements and Preprocessing

The heatmap provided important insights into the relationships between features. Key findings included:

- A high positive correlation between Month and Week of Year.
- A high positive correlation between Forecast Wind Production and Actual Wind Production.
- A high positive correlation between SystemLoadEA and SystemLoadEP2.
- A high positive correlation between ORKWindspeed and both Forecast Wind Production and Actual Wind Production.

These correlations could indicate multicollinearity, increasing model complexity and reducing performance. Therefore, one feature from each highly correlated pair was removed:

- **Month** was retained over Week of Year, as it is more straightforward and useful in most contexts.
- **Forecast Wind Production** was removed to improve model accuracy.
- **SystemLoadEA** (forecasted national load) was removed, keeping **SystemLoadEP2** (actual national system load) for better accuracy.

- **ORKWindspeed** and **Actual Wind Production** were retained, but **Forecast Wind Production** was already removed.
- The **Holiday** column was also removed due to its limited usefulness after considering other features.

Feature Engineering

Features showing cyclic patterns, such as "Month," "Day of the Week," "Day," and "Period of Day," were transformed using sine and cosine functions to preserve their cyclic nature. As there were no categorical features, encoding was not required.

Data Splitting, Scaling, and PCA

The dataset was split into training and test sets, with 20% reserved for testing. Due to skewed distributions, MinMaxScaler from Scikit-Learn was used for scaling. Principal Component Analysis (PCA) was then applied, with a scree plot guiding the selection of components.

Model Application and Accuracy

Since this is a regression problem, models like Linear Regression, Lasso, Decision Tree Regression, Random Forest Regression, and Support Vector Machine Regression were applied. The models were evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score. Random Forest Regression was found to be the most suitable model based on these evaluation metrics, particularly when using scaled data.

Experiments

The models were tested on data at three stages: after data cleaning, after scaling, and after PCA transformation. R^2 scores were calculated for each model and each scenario.

Key Findings

Random Forest Regression consistently emerged as the most suitable model across all scenarios:

- **R^2 score for data after data cleaning:** 0.6393
- **R^2 score for data after scaling:** 0.6502
- **R^2 score for data after PCA transformation:** 0.5839

Therefore, the Random Forest Regression model, using data up to the scaling stage, was selected as the best model for electricity price prediction.