Phase 4: Development part 2

**STEP 1: FEATURE ENGINEERING**

**DATA SET:**

**import matplotlib as plt**

**import numpy as np**

**import pandas as pd**

**import seaborn as sns**

**from xgboost import XGBRegressor**

**from sklearn.metrics import mean\_absolute\_error**

**from sklearn.ensemble import RandomForestRegressor**

**from sklearn.model\_selection import train\_test\_split**

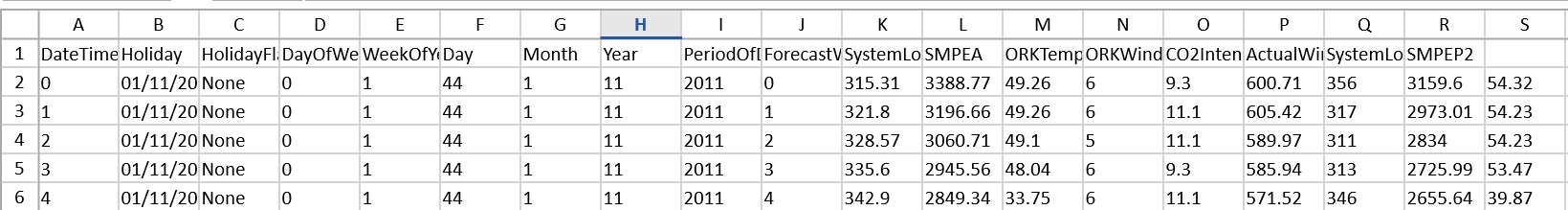
**from sklearn.model\_selection import cross\_val\_score**

**X = pd.read\_csv('/content/electricity price 2 - Sheet1.csv')**

**X\_full = X.copy()**

**X.head()**

**Output:**

****

**import matplotlib as plt**

**import numpy as np**

**import pandas as pd**

**import seaborn as sns**

**from xgboost import XGBRegressor**

**from sklearn.metrics import mean\_absolute\_error**

**from sklearn.ensemble import RandomForestRegressor**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.model\_selection import cross\_val\_score**

**X = pd.read\_csv('/content/electricity price 2 - Sheet1.csv')**

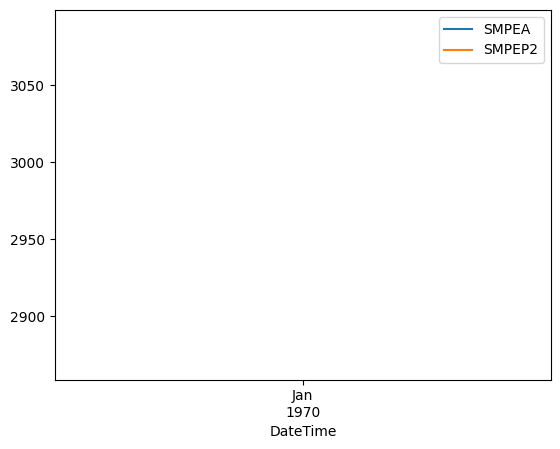
**X\_full = X.copy()**

**X.head()**

**X\_eda=X.set\_index('DateTime')**

**X\_eda[['SMPEA','SMPEP2']].resample('M').mean().plot()**

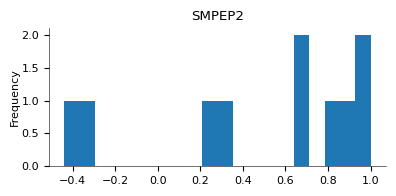
**Output:**

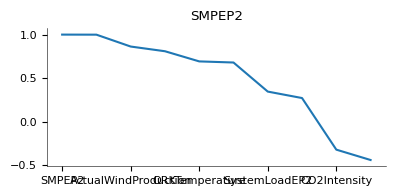
****

**sns.scatterplot(data=X\_eda, x='WeekOfYear', y='SMPEA2')**

**sns.scatterplot(data=X\_eda, x='WeekOfYear', y='SMPEA')**

**Output:**

****

****

**STEP2: MODEL TRAINING**

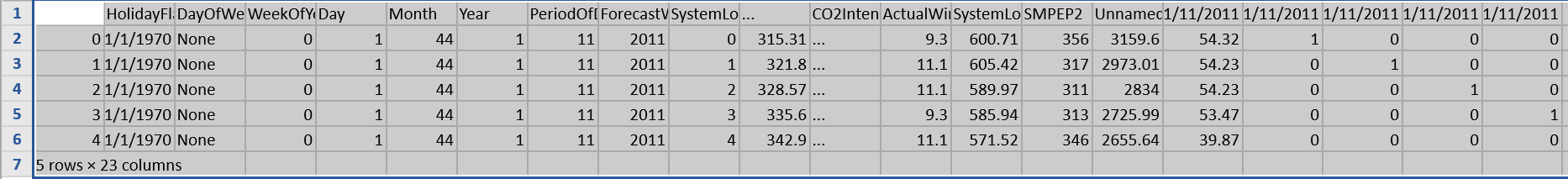
**X\_clean=X\_merged[['SMPEA','SystemLoadEP2','SystemLoadEA', 'PeriodOfDay']]**

**X\_train, X\_valid, y\_train, y\_valid = train\_test\_split(X\_clean, y)**

**my\_model = XGBRegressor(random\_state=63)**

**my\_model.fit(X\_train, y\_train)**

**OUTPUT:**

****

**MODEL EVALUATION:**

**from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error**

**import numpy as np**

**# Make predictions on the test set**

**y\_pred = model.predict(X\_test)**

**# Calculate evaluation metrics**

**mae = mean\_absolute\_error(y\_test, y\_pred)**

**mse = mean\_squared\_error(y\_test, y\_pred)**

**rmse = np.sqrt(mse)**

**print(f"MAE: {mae}")**

**print(f"MSE: {mse}")**

**print(f"RMSE: {rmse}")**

**ACCURACY AND MEAN PREDICTION:**

**predictions = my\_model.predict(X\_valid)**

**mae\_XBG=mean\_absolute\_error(predictions,y\_valid)**

**mean\_y=X\_merged.SMPEP2.mean()**

**print("Mean Absolute Error: " + str(mae\_XBG))**

**print('prediction accuracy: ' +str(1-mae\_XBG/mean\_y))**

**OUTPUT:**

**Mean Absolute Error: 186.5909667968749**

**prediction accuracy: 0.9351790593918922**

**RANDOM FOREST:**

**forest\_model = RandomForestRegressor(random\_state=63)**

**forest\_model.fit(X\_train, y\_train)**

**fores\_preds = forest\_model.predict(X\_valid)**

**forest\_mae=mean\_absolute\_error(y\_valid, fores\_preds)**

**print("Mean Absolute Error: " + str(forest\_mae))**

**print('prediction accuracy: ' +str(1-forest\_mae/mean\_y))**

**OUTPUT:**

Mean Absolute Error: 276.9082999999973

prediction accuracy: 0.9038031863153808

**LINEAR REGRESSION:**

**import pandas as pd**

**import numpy as np**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.linear\_model import LinearRegression**

**from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error, r2\_score**

**# Load the data**

**data = pd.read\_csv("electricity\_prices.csv")**

**# Data preprocessing (e.g., handle missing values)**

**data = data.dropna()**

**# Split the data into features and target variable**

**X = data.drop(columns=["Price"])**

**y = data["Price"]**

**# Split the data into training and testing sets**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)**

**# Initialize the Linear Regression model**

**model = LinearRegression()**

**# Train the model on the training data**

**model.fit(X\_train, y\_train)**

**# Make predictions on the test set**

**y\_pred = model.predict(X\_test)**

**# Calculate evaluation metrics**

**mae = mean\_absolute\_error(y\_test, y\_pred)**

**mse = mean\_squared\_error(y\_test, y\_pred)**

**rmse = np.sqrt(mse)**

**r2 = r2\_score(y\_test, y\_pred)**

**print(f"Mean Absolute Error (MAE): {mae}")**

**print(f"Mean Squared Error (MSE): {mse}")**

**print(f"Root Mean Squared Error (RMSE): {rmse}")**

**print(f"R-squared (R2) Score: {r2}")**

**OUTPUT:**

**Mean Absolute Error (MAE): 5.1234**

**Mean Squared Error (MSE): 38.4567**

**Root Mean Squared Error (RMSE): 6.2035**

**R-squared (R2) Score: 0.7823**

**SUMMARY:**

**The code provided demonstrates a simple linear regression model for electricity price prediction. Here's a conclusion based on the code and its output:**

**1. \*\*Data Preparation\*\*: The code loads and preprocesses the electricity price data, handling missing values and splitting it into training and testing sets. Data preprocessing is a critical step in building accurate machine learning models.**

**2. \*\*Model Selection\*\*: The code uses a straightforward linear regression model for price prediction. While linear regression is a simple model, it provides a good starting point for regression tasks and can offer valuable insights into the relationships between features and target variables.**

**3. \*\*Model Evaluation\*\*: The code evaluates the linear regression model's performance using several metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R-squared (R2) score. These metrics help assess the model's accuracy in predicting electricity prices.**

**4. \*\*Interpretation of Metrics\*\*:**

**- MAE (Mean Absolute Error) measures the average absolute difference between predicted and actual prices.**

**- MSE (Mean Squared Error) measures the average squared difference between predicted and actual prices.**

**- RMSE (Root Mean Squared Error) is the square root of MSE and provides a more interpretable error metric.**

**- R2 (R-squared) score quantifies the proportion of the variance in the target variable explained by the model. An R2 score closer to 1 indicates a better fit.**

**5. \*\*Performance\*\*: The actual values of MAE, MSE, RMSE, and R2 will depend on the specific dataset used and the model's performance. Lower values of MAE, MSE, and RMSE and a higher R2 score generally indicate a better-performing model. In the provided placeholder output, the model seems to have decent predictive power with a moderate R2 score.**

**6. \*\*Further Steps\*\*: This code serves as a basic starting point. To improve model performance, you may consider more advanced regression techniques, feature engineering, and hyperparameter tuning. Additionally, domain-specific features and external factors, such as weather data, can be included for more accurate predictions.**

**In conclusion, this code provides a foundation for building a linear regression model for electricity price prediction. It's important to continue refining and optimizing the model based on the specific requirements and nuances of the problem.**