Project Name: Electricity Price Prediction

Phase 3: Development Part 1 - Loading and Preprocessing the Dataset

Introduction

In this phase, we will begin building the electricity price prediction model by loading and preprocessing the dataset. The dataset contains historical electricity prices and various relevant factors. We'll follow a systematic process to prepare the data for analysis, which includes data loading, data cleaning, and data transformation.

Data Loading

We will start by loading the historical electricity prices dataset. To do this, we need to import the necessary libraries and read the data from the provided CSV file.

Data Exploration

It's important to get an initial understanding of the dataset. We'll inspect the data to understand its structure and the types of information it contains.

Data Cleaning

Data cleaning is crucial to ensure that the dataset is free from errors, missing values, and inconsistencies.

Data Transformation

In this step, we will perform data transformations to make it suitable for analysis. This may include converting categorical variables to numerical representations, scaling numerical features, and creating new features.

python

Saving the Preprocessed Data

Once the data is loaded, cleaned, and transformed, it's a good practice to save it for further analysis and model development.

Code:

```
# Import libraries

import pandas as pd


# Load the dataset

data = pd.read_csv('electricity_price_prediction_dataset.csv')
```

```
# Display the first few rows of the dataset
print(data.head())

# Check the data types of each column
print(data.dtypes)

# Summary statistics
print(data.describe())
# Check for missing values
missing_values = data.isnull().sum()
print(missing_values)

# Remove rows with missing values or replace them as needed
data = data.dropna()

# Check for duplicated rows
duplicates = data.duplicated().sum()
print("Number of duplicated rows:", duplicates)

# Remove duplicates if necessary
data = data.drop_duplicates()

# Save the preprocessed data to a new CSV file
data.to_csv('preprocessed_electricity_data.csv', index=False)
```

Output:

```
       DateTime Holiday  HolidayFlag  DayOfWeek  WeekOfYear  ...  ORKWindspeed  CO2Intensity
ActualWindProduction  SystemLoadEP2 SMPEP2
```

```
0  01/11/2011 00:00     NaN          0         1         44  ...        9.30       600.71              356.00        3159.60
54.32

1  01/11/2011 00:30     NaN          0         1         44  ...       11.10        605.42              317.00        2973.01
54.23

2  01/11/2011 01:00     NaN          0         1         44  ...       11.10        589.97              311.00        2834.00
54.23

3  01/11/2011 01:30     NaN          0         1         44  ...        9.30       585.94              313.00        2725.99
53.47

4  01/11/2011 02:00     NaN          0         1         44  ...       11.10        571.52              346.00        2655.64
39.87


[5 rows x 18 columns]

DateTime                 object

Holiday                  object

HolidayFlag               int64

DayOfWeek                 int64

WeekOfYear                int64

Day                       int64

Month                     int64

Year                      int64

PeriodOfDay               int64

ForecastWindProduction    object

SystemLoadEA             object

SMPEA                    object

ORKTemperature           object

ORKWindspeed             object

CO2Intensity             object

ActualWindProduction     object

SystemLoadEP2            object

SMPEP2                   object
```

```
dtype: object
      HolidayFlag    DayOfWeek    WeekOfYear         Day        Month         Year   PeriodOfDay
count  38014.000000  38014.000000  38014.000000  38014.000000  38014.000000  38014.000000
38014.000000
mean      0.040406     2.997317     28.124586     15.739412     6.904246   2012.383859    23.501105
std       0.196912     1.999959     15.587575     8.804247     3.573696     0.624956    13.853108
min       0.000000     0.000000     1.000000     1.000000     1.000000   2011.000000     0.000000
25%       0.000000     1.000000    15.000000     8.000000     4.000000   2012.000000    12.000000
50%       0.000000     3.000000    29.000000    16.000000     7.000000   2012.000000    24.000000
75%       0.000000     5.000000    43.000000    23.000000    10.000000   2013.000000    35.750000
max       1.000000     6.000000    52.000000    31.000000    12.000000   2013.000000    47.000000
DateTime                  0
Holiday               36478
HolidayFlag               0
DayOfWeek                 0
WeekOfYear                0
Day                       0
Month                     0
Year                      0
PeriodOfDay               0
ForecastWindProduction    0
SystemLoadEA              0
SMPEA                     0
ORKTemperature            0
ORKWindspeed              0
CO2Intensity              0
ActualWindProduction      0
SystemLoadEP2             0
SMPEP2                    0
```

```
dtype: int64

Number of duplicated rows: 0
```

Conclusion

In this phase, we have loaded and preprocessed the historical electricity prices dataset, preparing it for analysis and model development. The preprocessed data can now be used in the next phases for feature engineering, model selection, and model training.