Project 4: NYISO Electricity Consumption and Pricing Insights

Introduction

The NYISO dataset provides comprehensive data on New York's electricity market, including pricing, load demand, and grid operations. It includes detailed time series data on electricity consumption, generation, and market prices across different locations. This dataset is widely used for energy forecasting, market analysis, and grid optimization studies.

Key Features of the Dataset

There are 3 main datasets that are provided:

Pricing Data:
Provides locational-based marginal prices (LBMP) reflecting the real-time and day-ahead costs of electricity delivery, including congestion and losses

Power Grid Data:
Contains information on transmission constraints, generation mix, outages, and system stability indicators for efficient grid management

Load Data:
Offers real-time and forecasted electricity demand across NYISO regions, aiding in balancing supply and planning market operations

Some of the main features across all 3 datasets include:

1. **PTID (Pricing Node ID)**: A unique identifier for a specific pricing location in the NYISO grid used for mapping electricity prices and load data.
2. **Integrated Load**: The average total electricity demand in a region over a specific time interval, used to monitor overall consumption.
3. **LBMP (Locational Based Marginal Price)**: The electricity price at a specific node, reflecting generation costs, congestion, and transmission losses.
4. **Marginal Cost of Losses**: The additional cost due to energy losses when transmitting one more unit of electricity across the grid.
5. **Marginal Cost of Congestion**: The extra cost of delivering electricity due to transmission constraints limiting power flow on the grid.

These datasets are considered time series data because the **timestamp** field records the specific time for each data point.

Collection Process

The data is collected through continuous monitoring of the NYISO grid, with sensors and systems recording real-time electricity prices, demand, and generation. This information is then aggregated at regular intervals (e.g., every 5 minutes or hourly).

Structure

The dataset is available in CSV format.

Data Coverage

Start Year: 2000
End Year: Present (the dataset continues to be updated)

Data Access

Details of the dataset are present on the NYISO website.
The dataset is publicly available. We will only consider pricing and load data in this project.
Pricing data can be downloaded from here while corresponding load data can be downloaded from here.

This project will be undertaken in groups of one, i.e. ***individually***.

We will focus on data from 2001 to 2023. Each individual will be assigned a year of non-overlapping data; for example, Individual 1 will work with data from 2001, Individual 2 with 2002, and so on. The dataset allocation will be assigned by the course TA.

Tooling

You are expected to use a combination of `pyspark` and Structured Streaming for processing and MLlib or `sklearn` for analytics in Google Colabs for this project. As always, my recommendation is to upload the dataset to your GWU Google Drive and then process it using pyspark.

You can use any Python library, such as `matplotlib`, etc. of your choosing to create the plots. In addition, all documentation/comments should be in the form of Markdown within the notebook itself.

Analysis

For each piece of analysis, provide an explanation of the process as well as an interpretation of the results. In addition you should aim to improve the performance of your underlying models. The final performance along with the process followed (feature engineering, etc.) should be documented as well.

1. Predict electricity prices (LBMP) based on the historical load demand patterns
2. Detect unusual price spikes in electricity prices (LBMP) due to sudden changes in demand or grid conditions, such as congestion or losses
3. Forecast future electricity demand (Integrated Load) based on historical electricity prices (LBMP) and grid conditions (marginal costs due to congestion or losses)

For each use case data from specific months (March, June, September, and December) via streaming will be used for prediction while the rest of the months will be used for training. You can use this code to stream data for relevant months for prediction.

Due Date

3 December, 2024 - 11:59AM

You can submit the link to your Google Colab via Blackboard in the Assignments section.