DA 204o: Data Science in Practice
Course Project Proposal

Real-Time Dynamic Pricing for Urban Electric Vehicle (EV) Charging Networks

Rahul Kumar, rahulkumar18@iisc.ac.in

Abhishek Malik, abhishek16@iisc.ac.in

Image source: Internet

Problem Definition

- ## Background of the problem
  - Electric-vehicle (EV) charging networks are growing rapidly, but many stations are poorly matched to demand: some remain under-utilized while others suffer long queues. This leads to lost revenue for operators, poor customer experience, and inefficient allocation of capital. We aim to develop data-driven methods to measure utilization, predict demand, and recommend reallocation that maximize station-level utilization and minimize customer wait time subject to business and geographic constraints.
- ## Why is it important?
  - Optimizing utilization enables smarter pricing and better asset allocation, helping operators boost revenue, cut wait times, and ensure chargers are placed where demand is highest.
- ## Objectives of the project
  - To develop an accurate forecasting model to predict charging demand for individual EV charging stations or station clusters.
- ## How can Data Science solve the problem?
  - Leverage historical information on charging stations to produce robust predictions on demand.

Data Science Canvas

Project: EV Charge station Optimization

Team:

Problem Statement

Business Case & Value Added

Forecast hourly EV charging demand to optimize station usage and grid load.

Reduce downtime, improve capacity planning, and support data-driven expansion.

Enable targeted investment by identifying high- and low-performing locations.

Model Selection

Based on data and seasonality of data, we can use:

Tree-Based Time Series Models
- CatBoost
- XGBoost

Classical Time-Series Models
- ARIMA
- SARIMA

Deep Learning Models
- LSTM
- Transformer based models

Model Requirements

Which model requirements must be complied with in order to obtain a valid model?

- **Temporal Awareness:** Must respect chronological order (no random shuffling of training data) to avoid look-ahead bias.
- **Exogenous Handling:** Capability to incorporate external regressors like Temperature, Precipitation, and Day of Week.
- **Robustness:** Must be resilient to sensor noise and outliers (e.g., data spikes from faulty meters).

Skills

What skills are needed to provide the data and model development?

- **Data Engineering:** Ability to build pipelines for resampling event-based logs into hourly time-series tensors.
- **Time Series Analysis:** Understanding of seasonality, lag features, and rolling window statistics.
- **Machine Learning:** Expertise in Gradient Boosting (XGBoost/CatBoost) hyperparameter tuning and regularization.
- **Domain Knowledge:** Understanding of electrical grid constraints (kW vs kWh) and battery chemistry (temperature impact).

Execution & Evaluation

Model Evaluation

Which indicators require quality control and validation and how should they be interpreted? Is real-time monitoring necessary?

- **Primary Metrics:**
- **RMSE (Root Mean Squared Error):** Crucial for penalizing large prediction errors that could lead to grid failures.
- **MAE (Mean Absolute Error):** To understand the "average" error in kWh.
- **Validation Strategy: Time Series Split** (Walk-forward validation). Train on Jan-Mar Test Apr; Train Jan-Apr Test May.

Data Storytelling

What requirements does the target group have for the presentation of the results and how do I effectively communicate this data?

- **Target Audience:** Grid Operators, Station Managers, Investors.
- **Visualization Requirements:**
- **Actual vs. Predicted Plots:** Visual proof of the model's ability to catch peak demand.
- **Feature Importance Charts:** Showing stakeholders that "Temperature" and "Hour of Day" are driving the predictions.
- **Heatmaps:** Visualizing "Efficiency Ratios" (Active Charging vs. Idle Time) to highlight wasted capacity.

Data Collection & Preparation

Data Selection & Cleansing

Which of the available data is relevant? Do the data have to be cleaned up?

- **Relevance:** Focus on `connected_time_start_ts`, `energy_provided_kwh`, `connected_duration_min`, and `station_name`.
- **Cleansing Needs:**
- **Geospatial Filtering:** Remove anomalies like the Utah station (desert climate) from the NYC (temperate) dataset to prevent model confusion.
- **Outlier Removal:** Filter sessions < 0.5 kWh (connection failures) and > 100 kWh (likely data errors).
- Standardised names for locations having more than 1 entries

Data Collection

How and with which methods should additionally required data be collected? What properties has this data to fulfil?

- **Additional Sources:**

Grid Load Data: ISO/RTO feeds to correlate station demand with broader grid stress.
Traffic/Events: Local event calendars (e.g., sports games) that might spike usage.

- **Properties:** Data must be timestamped with high precision and consistent timezone formatting (UTC vs Local).

Data Landscape

Software & Libraries

Which software should

Data Integration

In which system should

Explorative Data Analysis

Data Collection and Preparation

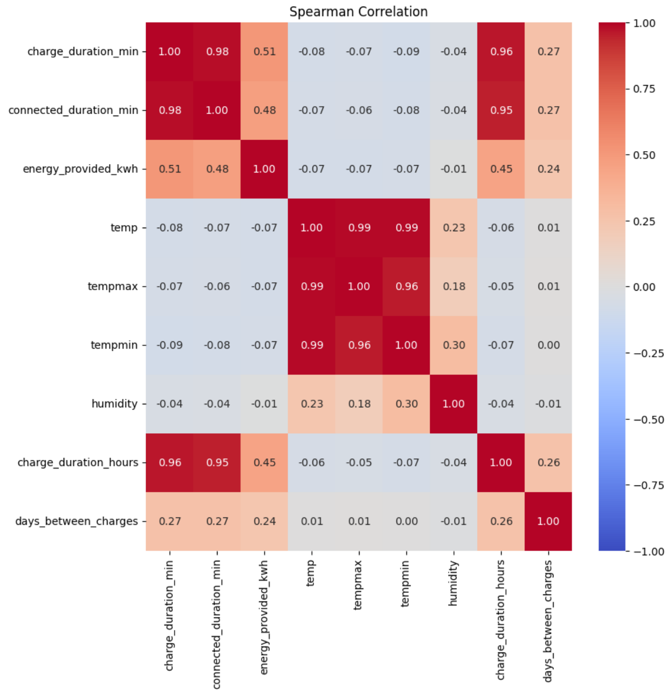
- Data source(s) (where it's from, how it was collected)
  - **EV Charging Session Data** – NYC Municipal Garages (2021–2025)
    NYC Municipal Garage
  - **Garage Location Details** – Google Search (address, latitude, longitude)
  - **Weather Data** – Visual Crossing Weather API (https://www.visualcrossing.com/weather-api/)

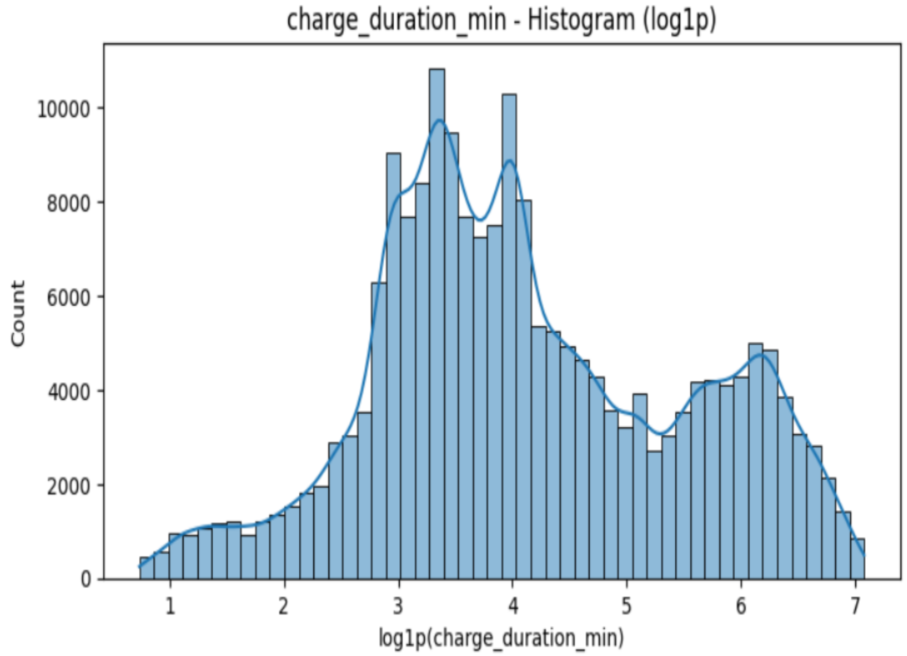
- Description of the data (features, size, format)
  - ~206,000 records with 15 features
  - Data covers EV charging sessions across multiple NYC municipal garages
  - Each record represents a single charging session (time, duration, energy, station info)
  - Location coordinates (lat/long) were mapped using garage names
  - Weather attributes (temperature, humidity, conditions) were linked using date + coordinates

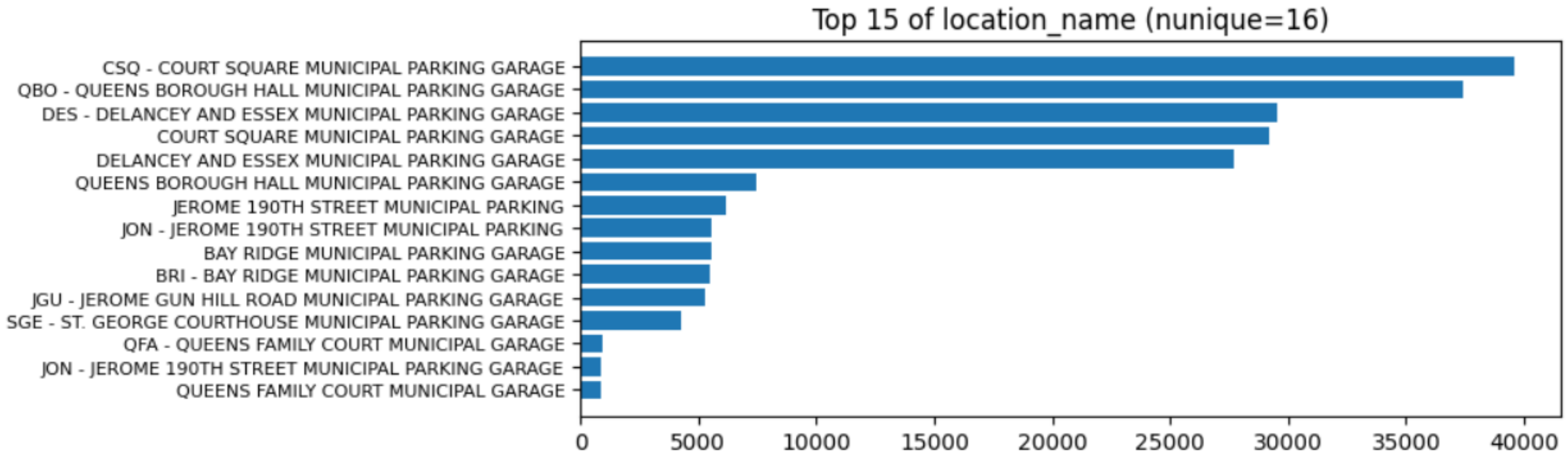
Data Collection and Preparation

- Pre-processing steps done
  - Cleaned and standardized text fields
  - Reindexed data to build a complete hourly time grid
  - Created feature for time since last charging session
  - Added calendar features (day of week, hour of day, month, weekend, etc.)
  - Engineered weather-based features (e.g., is_hot, is_snowing, humidity bands)
  - Aggregated data from session-level to hourly-level
  - Generated time-series lag and rolling window features

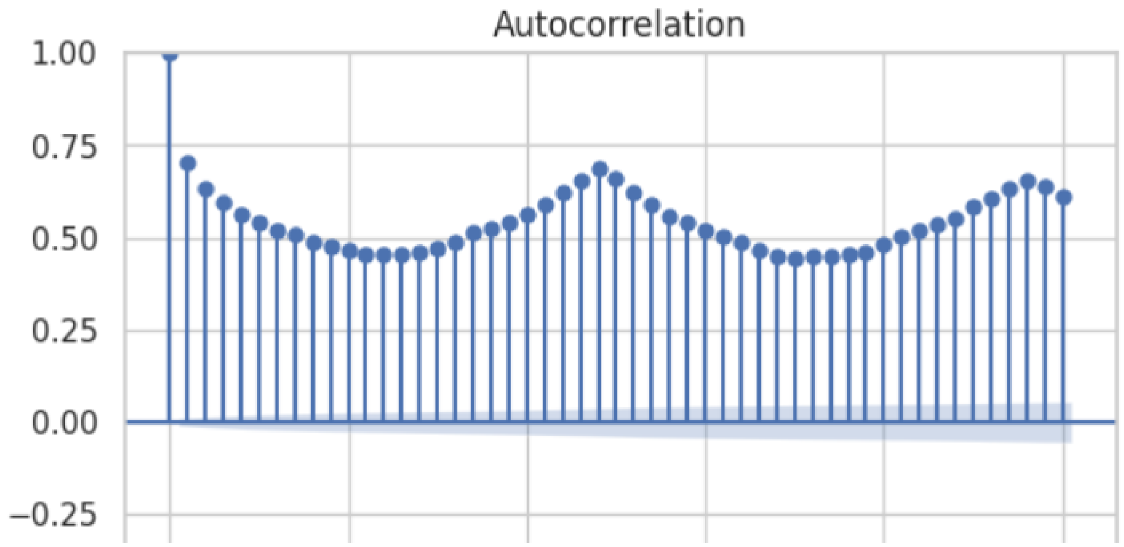
EDA Insights



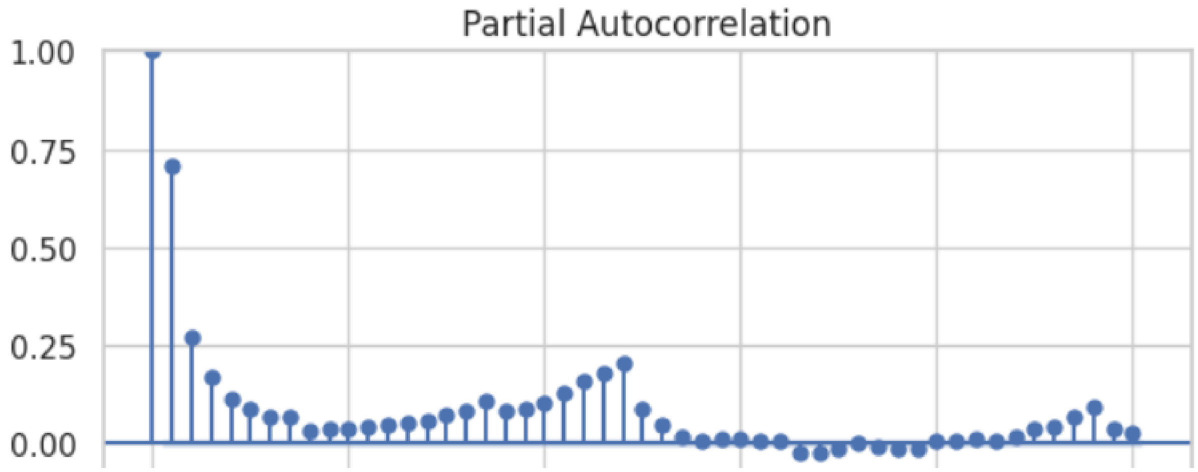




EDA Insights



<Figure size 1000x300 with 0 Axes>



EDA Insights

- All duration-related variables are heavily right-skewed. This indicates many short/average sessions and very few extremely long sessions.
- Charge duration and connection duration have multimodal patterns signifying different users have different charging preference.
- Like duration, enery is also heavily right skewed, incidcating small number of charging events provide unusually large amounts of energy
- Humidity distribution is more centralized indicating moderate weather most of the time.
- Charge duration (hours) still remains skewed even after log-transform. Suggests a small population with very long charging sessions (overnight / long-stay parking).
- Days between charges is extremely skewed. Most customers return within a very short interval (1–3 days), while some return after long gap.
- QUEENS has the highest number of charging sessions (~70k), indicating it is the busiest EV charging locality.
- JAMAICA, ST. GEORGE, and BROOKLYN have comparatively low counts, suggesting uneven charger demand across regions.
- NYC has cloudy weather for most of the time.
- Afternoon is the most common connected time slot, while morning being the second highest.
- Disconnected slot distribution matches connected pattern.
- Night-time and late-night disconnections are lower.
- Most users disconnect after charging is complete, however few of them leave them for longer duration.
- Most users charge for short sessions (<200 mins)
- Most sessions deliver low energy (< 20 kWh)
- Days_between_charges has slight positive correlation with duration.

Model Architecture

- **Time Series Forecasting Using Tree-Based ML**
- **Data Preparation Layer**
  - Combine session data, station metadata, weather data
  - Convert timestamps → hourly time index
  - Reindex to complete hourly grid (per location)
  - Clean missing values (0-fill for sessions/energy, ffill/bfill for contextual variables)
- **Feature Engineering Layer**
  - Temporal Features
  - Lag Features
  - Rolling Statistics
- **Modeling Layer (Supervised ML for Time Series)**
  - Approach 1:Two separate models
    - CatBoost Regressor – Sessions per Hour
    - CatBoost Regressor – Energy (kWh) per Hour
  - Approach 2:
    - XGBoost Regressor – Load/ Energy KWh

Model Architecture

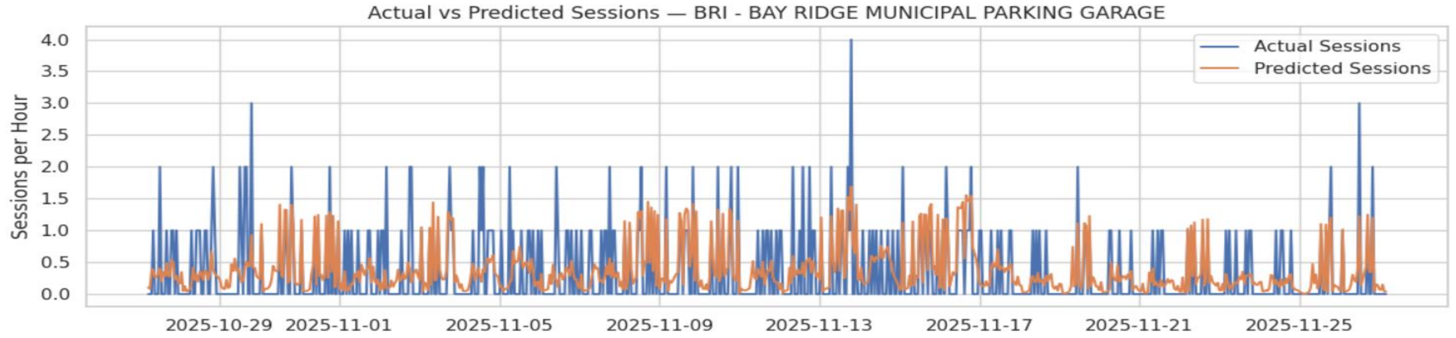
- Forecasting Layer
  - Recursive multi-step forecasting for **n months (n=1)**
  - Generate forecasts for each hour × each location

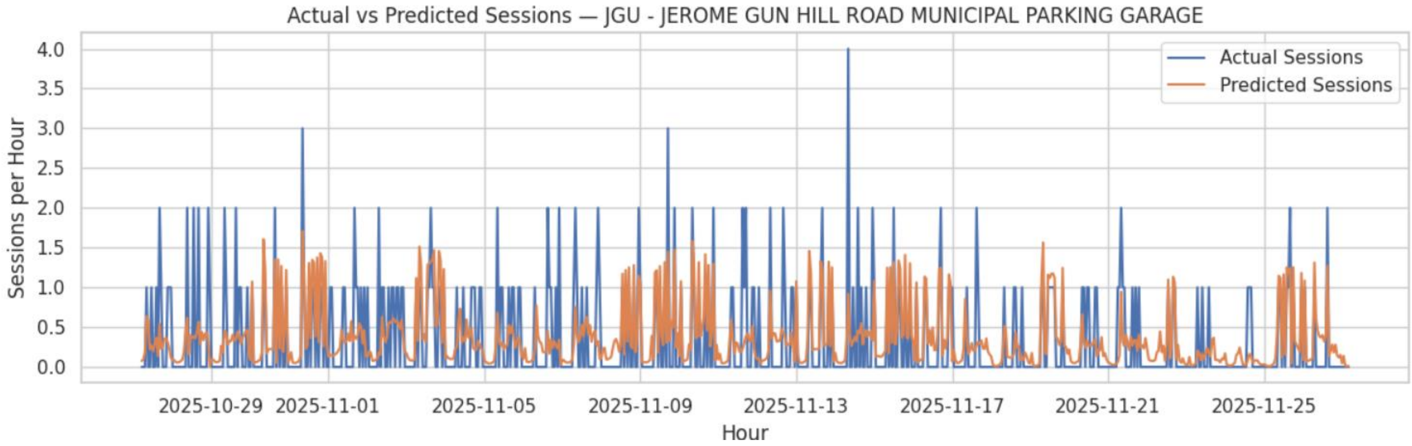
- Output Layer
  - Per-location hourly forecasts
  - Load curves for sessions & energy
  - Actual vs predicted plots
  - Station utilization & peak demand insights

Results

- Approach 1
  - Model predicted both session and energy consumption with low degree of error.

	Mean Absolute Error	Root Mean Square Error
Session	0.294	0.684
Energy Consumption	8.050	18.708

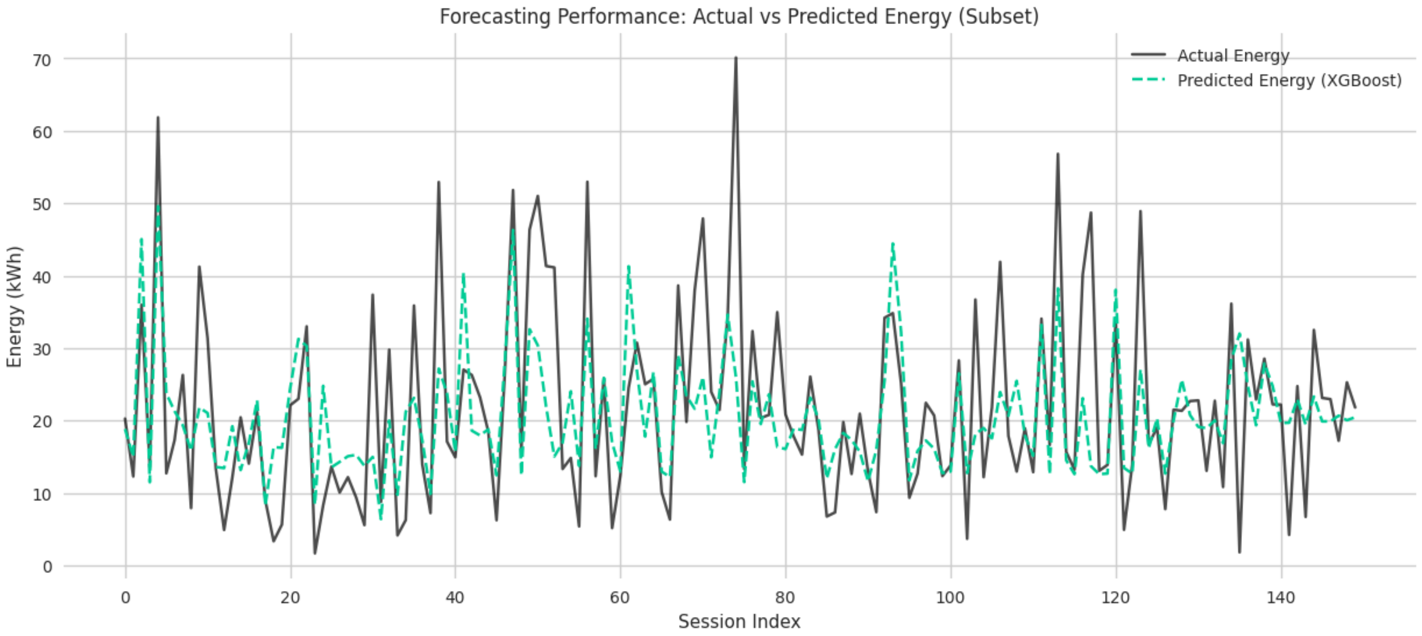




Results

- Approach 2
  - Model predicted load/ energy consumption with low degree of error.

	Mean Absolute Error	Root Mean Square Error
Energy Consumption	8.80	12.10



Insights

- **EV charging demand is highly predictable**
  Daily and weekly seasonality patterns are strong and consistent across stations, enabling reliable forecasting.
- **Forecast accuracy for sessions is extremely high**
  With **MAE < 0.35 sessions/hr**, the model predicts hourly demand almost perfectly at most stations.
- **Energy consumption forecasts are robust**
  Despite natural variation in vehicle behavior, the model achieves **MAE ≈ 8 kWh**, suitable for grid and capacity planning.
- **Some stations show zero or very low activity**
  Multiple charging sites have negligible usage—highlighting opportunities for relocation, optimization, or infrastructure rationalization.
- **High-demand stations exhibit stable peak cycles**
  Consistent peak hours (morning, evening, late-night) help guide staff allocation, maintenance windows, and resource planning.
- **Weather has moderate influence, mainly on energy**
  Temperature and rainfall slightly shift energy consumption, but session counts remain primarily driven by seasonality.
- **The 6-month future forecast shows stable operational trends**
  No drift or unpredictable spikes are observed, making the model suitable for long-term forecasting and decision-making.

Dynamic Pricing

The project implements a **Demand-Based Pricing** where price is a function of predicted utilization relative to capacity.

The Pricing Formula

- Pdynamic=Pbase×(1+α×Upredicted)
  - Where:
  - Pbase: Standard cost of electricity (e.g., $0.25/kWh).
  - α: Sensitivity factor (e.g., 0.5) controlling price aggressiveness.
- Upredicted: The ratio of model-predicted energy to maximum station capacity.
- **Implementation Example**
  - **Scenario:** A station with 70 kWh max hourly capacity.
  - **Forecast:** LSTM/XGBoost model predicts **55 kWh** demand for 6:00 PM.
  - **Utilization (Upredicted):** 55/70=0.78 (78).
  - **Dynamic Price:** $0.25 \times (1 + 0.5 \times 0.78) = 0.3475 per kWh
- **Strategic Outcome:** By publishing this price ahead of time, we encourage price-sensitive drivers to shift usage away from the 6:00 PM peak, smoothing the Duck Curve while capturing higher margins from inelastic demand.

Role and Responsibilities

- Rahul Kumar
  - Consolidated data from multiple raw sources into a unified, clean, and standardized dataset.
  - Ensured consistency across locations, weather inputs, and session-level attributes
  - Worked collaboratively on EDA for gathering key insights
  - Engineered temporal features (lags, rolling averages, seasonality indicators)
  - Built and validated forecasting models for hourly sessions and energy consumption
  - Performed time-based train/validation splits and model evaluation (MAE, RMSE)
- Abhishek Malik
  - Worked collaboratively on EDA, data cleanup and normalization.
  - Feature engineering to generate predictive features from temporal, spatial, and user history data.
  - Trained XGBoost model using Time Series Split and evaluated performance.
  - Model evaluation (MAE, RMSE)
  - Formulated deterministic Dynamic pricing logic using the model’s load prediction.