When the Forecast Fails: CitiBike Demand Forecasting in the Era of Sudden Disruptions

1. Setting up

https://github.com/zmachine8/citibikenyc

2. Business Understanding

2.1 Identifying Business Goals

2.1.1 Background

CitiBike is a large bike-sharing system whose daily usage varies due to many external factors such as weather, seasonality, weekdays, and longer-term trends. These fluctuations directly affect operational planning: bicycle redistribution, staff allocation, maintenance scheduling, and overall service availability.
To plan resources effectively, the organization needs a clearer understanding of **what influences daily ridership**.

2.1.2 Business Goals

- Understand how different factors (weather, season, weekday, etc.) influence ridership.
- Identify predictable demand patterns throughout the year.
- Use historical data to estimate future ridership.
- Provide interpretable insights that support operational decisions for bike availability and staff planning.

2.1.3 Business Success Criteria

- Ridership patterns and influencing factors are identified clearly and can be communicated to non-technical stakeholders.
- The results help explain why demand changes on different days.
- The insights can support better decision-making for resource allocation (e.g., redistribution, staffing, fleet planning).

- The analysis provides a solid basis for a predictive model with accuracy better than simple baselines.

2.2 Assessing The Situation

2.2.1 Inventory of Resources

Data sources:

- Daily CitiBike ridership counts.
- Daily weather data from Meteostat (temperature, precipitation, wind etc.).
- Date-derived features (weekday, month, season, etc.).

Tools & environment:

- Python ecosystem (Pandas, Scikit-learn, Matplotlib).
- Jupyter/Colab notebooks for computation.
- Statistical and Machine Learning methods for analysis and modelling.

Knowledge:

- Public documentation about bike-sharing systems.
- Prior research showing weather–mobility relationships.

2.2.2 Requirements, Assumptions, and Constraints

Requirements:

- Analyze the influence of external factors on daily ridership.
- Build an interpretable forecast model.
- Produce readable visualizations and explanations.

Assumptions:

- Weather significantly affects daily ridership.
- Historical behavior is informative for predicting future usage.
- Data is sufficiently complete after cleaning.

Constraints:

- Only high-level daily totals — no station-level data.
- No access to event or holiday data unless manually added.
- Limitations on model complexity due to course scope.
- Some missing or noisy values in weather/trip data.

2.2.3 Risks and Contingencies

Risks:

- Missing/incorrect data may distort analysis.
- Outliers caused by events not recorded in the dataset (holidays, storms, lockdowns).
- Overfitting if weather data drives spurious patterns.
- Model performance limited by unavailable variables.

Contingencies:

- Data cleaning, imputation, aggregation.
- Use robust validation (train/test split, rolling windows).
- Add additional calendar features if required.
- Fall back to simpler models if complex ones behave poorly.

2.2.4 Terminology

- **Ridership / Trips:** Number of daily bike trips.
- **tavg, tmin, tmax:** Average/min/max temperatures.
- **prcp:** Precipitation amount.
- **wspd:** Wind speed.
- **Seasonality:** Recurring yearly patterns.
- **EDA:** Exploratory Data Analysis.
- **Regression model:** Predicts numeric values such as daily trip counts.

2.2.5 Costs and Benefits

Costs:

- Time required for data cleaning and processing.
- Computational effort for training and validating models.
- Limited accuracy due to missing external factors.

Benefits:

- Improved understanding of ridership behavior.
- Better operational planning (redistribution, scheduling, fleet size).
- Reduced inefficiencies and operational costs.
- More reliable predictions for different weather and seasonal scenarios.

2.3 Defining Data-Mining Goals

2.3.1 Data-Mining Goals

- Identify which variables have the strongest effect on ridership.
- Detect regular patterns (daily, weekly, seasonal).
- Build a regression model to predict daily ridership from historical data and weather variables.
- Quantify the influence of weather compared to calendar-based patterns.
- Produce visual and statistical summaries useful for decision-making.

2.3.2 Data-Mining Success Criteria

- The predictive model achieves measurable accuracy (e.g., MAE/RMSE better than baseline).
- EDA reveals interpretable and consistent trends.
- The model can generalize to unseen data (test set) without major overfitting.
- The explanation of influencing factors is clear, justified, and backed by data.
- Stakeholders could use the insights for planning purposes.

3. Data Understanding

CRISP-DM divides this phase into four major activities: **gathering, describing, exploring, and verifying data quality**.

3.1 Gathering Data

3.1.1 Outline Data Requirements

The goal of the project is to understand and model **daily CitiBike ridership** and how it is influenced by external factors (weather, seasonality, weekday patterns).
Therefore, the required data must include:

- **Daily trip counts** (ridership)
- **Daily weather measurements** (temperature, precipitation, wind, etc.)
- **Calendar features** (date, weekday, month, season)
- A continuous multi-year time period (to capture yearly cycles)
- No station-level detail is required; system-level totals are sufficient.

3.1.2 Verify Data Availability

Data was collected from two publicly available sources:

1. **CitiBike system data:**
    - Contains daily counts of trips.
    - Covers years **2013–2019, newer format (more limited) 2020-2023.**
    - The dataset contains no personal or confidential information.

2. **Meteostat historical weather data:**
    - Contains daily weather measurements for the operational area (New York City).
    - Provides variables such as `tavg`, `tmin`, `tmax`, `prcp`, and `wspd`.
    - Fully available for the same date range.

Both data sources provide complete, accessible, and well-documented datasets.

3.1.3 Define Selection Criteria

Only records meeting the following criteria are included:

- Date range **2013-01-01 to 2019-12-31 for presentational and training/test data, 2020-01-01 to 2023-12-31 for training/test data** (consistent coverage across both datasets).
- Observations must include:
    - A valid date
    - A numeric trip count
    - A full set of weather variables

- Rows with missing weather data or corrupted values are excluded or imputed if minimal.
- The unit of analysis is **one row per day for regression models**.

No station-level, hourly, or user demographic data is used for models, statistical presentations contain daily system-wide trip totals, including aggregated user categories (male, female, unknown), age groups etc.

3.2 Describing Data

3.2.1 Dataset Structure

The combined dataset contains (example column list):

Column	Description
`date`	Calendar date (daily frequency)
`trips`	Total number of CitiBike trips that day
`tavg`	Average daily temperature (°C)
`tmin`	Minimum daily temperature
`tmax`	Maximum daily temperature
`prcp`	Daily precipitation (mm)
`wspd`	Average wind speed (m/s)
`month`	Extracted month (1–12)
`weekday`	Extracted weekday (0–6)
`season`	Derived season label (Winter/Spring/Summer/Fall)

3.2.2 Basic Statistics

Examples:

- Date count: ~2500 rows
- Trips range: Low in winter / peaks in summer
- Temperature distribution: typical NYC seasonal variation
- Precipitation: many zero-rain days; few heavy-rain days
- Wind speed: mostly stable range

3.3 Exploring Data

EDA section

3.3.1 Temporal Patterns

- Strong **seasonality**: ridership peaks in summer and drops in winter.
- Weekly pattern: **Lower on weekends/higher on weekdays**.
- Long-term growth or decline depending on year.

3.3.2 Weather Effects

- Ridership increases on warm, dry days.
- Significant drop on rainy days (`prcp > 5 mm`).
- Extremely cold (<0°C) or hot (>32°C) days reduce ridership.
- Higher wind speed correlates with a small decrease in trips.

3.3.3 Correlation Analysis

- Temperature positively correlated with ridership.
- Precipitation is negatively correlated.
- Season strongly associated with daily trip counts.

3.3.4 Outliers / Special Cases

- Extreme weather events cause visible drops.
- Holiday periods show irregular patterns.

3.4 Verifying Data Quality

CRISP-DM requires checking data validity, consistency, completeness, and accuracy.

3.4.1 Completeness

- CitiBike daily trip counts are complete for all dates.
- Meteostat weather data is mostly complete; occasional missing values were found in:

    - `tavg`, `prcp`, or `wspd` on isolated days.
- Missing weather data was handled by either:
    - dropping affected rows (if few), or
    - imputing based on nearby days.

3.4.2 Consistency

- All datasets use the same date format.
- Weather measurements correspond to the correct location and time of year.
- No duplicated dates after aggregation.

3.4.3 Correctness

- Checking extremes:
    - No negative trip counts.
    - Weather ranges match realistic NYC values.
- Trip values can vary widely but outliers correspond to real conditions (storms, holidays).

3.4.4 Reasonableness

- Weather-driven patterns match expectations.
- Seasonal ridership patterns are logical and align with known CitiBike usage trends.

4. Planning The Project

The project will be organized according to the CRISP-DM framework. The work is divided into five main tasks:

1. **Data Gathering (5 hours)**
   Collect CitiBike daily ridership data and Meteostat daily weather data for the full time period. Validate source formats, ensure correct time zones, and document data lineage.

2. **Data Understanding & Initial Exploration (15 hours)**
   Inspect all variables, compute descriptive statistics, visualize temporal patterns, identify anomalies, and assess the influence of weather and calendar variables.

3. **Data Preparation (20 hours)**
   Clean missing values, correct inconsistent records, engineer calendar features, aggregate data where necessary, and merge both datasets into a unified daily-level table. Perform additional checks on data quality and distributions.

4. **Modelling (20 hours)**
   Develop baseline and advanced models (Linear Regression, Random Forest, Gradient Boosting). Perform hyperparameter tuning, evaluate performance with

MAE/RMSE, analyze feature importance, and compare multiple modelling strategies.

5. **Evaluation, Visualization & Reporting (30 hours)**
   Interpret model outputs, summarize insights, create clear visualizations, discuss limitations, and compile the final written report and presentation.

Total Time: 90 hours.

Methods and Tools

Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and Jupyter/Colab notebooks. Analytical techniques include EDA, correlation analysis, regression modelling, feature engineering, validation, and interpretability analysis.

Comments

Additional time is allocated to ensure careful cleaning, robust model tuning, and more thorough reporting and visualization.