Retail

**Description:** Historical sales data from Rossmann drug stores in Germany, including store attributes, promotions, and state holidays.

**Forecasting Task:** Predict daily sales for each store.

Source：https://www.kaggle.com/c/rossmann-store-sales/data?select=store.csv



**Task：**

**1）sales prediction**

·Target Variable: Sales

·Input Features: lag\_1, lag\_7, day\_of\_week, month, is\_weekend, day, Promo, Open, SchoolHoliday, StateHoliday, StoreType, Assortment, CompetitionDistance, rolling\_mean\_7, rolling\_std\_7

·Prediction Goal: Daily sales forecast for the next 7 or 14 days (multi-step prediction)

·Model Examples: LSTM / GRU / Transformer / XGBoost / SARIMAX

·Evaluation Metrics: RMSE / MAE / SMAPE

**2）Promotion Impact Modeling (Classification or Regression)**

Analyze whether a particular day saw an abnormal sales surge due to promotions or holidays (useful for pre-holiday marketing strategies)

·Target Variable:

- Classification: Whether there was a sales surge (custom threshold, e.g., > mean + 2 std)

- Regression: Magnitude of sales surge (relative to the average over the past 7 days)

·Input Features: Promo, Promo2, PromoInterval, Promo2SinceWeek, Promo2SinceYear, StateHoliday, SchoolHoliday, is\_weekend, lag\_1, lag\_7, StoreType, Assortment

·Model Examples:

- Classification: LightGBM / RandomForest / MLP

- Regression: XGBoost / MLP / LSTM with attention

Transportation

source：TLC Trip Record Data - TLC

https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

**Description:** This dataset contains high-frequency For-Hire Vehicle (FHVHV) trip records in New York City, primarily from platforms such as Uber and Lyft. Each row represents an individual passenger trip. The key timestamp field is pickup\_datetime, which indicates the start time of each ride.

**Time Range:** 2009–2025

**Data Used:** January to June 2024 (sufficient for this analysis)

**Task：**

**1）Hourly Trip Count Forecasting**

·Data Preparation: Aggregate trips by hour using pickup\_datetime, resulting in an hourly time series: pickup\_datetime → hourly → trip\_count

·Input Features:

- Temporal features: hour, weekday

- Lag features: lag\_1, lag\_24, rolling\_mean\_7h

- Contextual features: is\_holiday, is\_peak\_hour (can be derived from hour)

- Prediction Target: Number of trips in the next 1, 3, or 6 hours

·Recommended Models: LSTM, Transformer, TimesNet, XGBoost

**2）Spatio-Temporal Trip Forecasting by Pickup Zone**

·Objective: Forecast the hourly number of ride requests grouped by PUlocationID (pickup location ID)

·Approach: Build a model that captures both spatial (location) and temporal (hourly trends) dynamics in trip demand across NYC

Energy

Features:

- **Rich in variables:** Typically includes multiple weather-related features such as load, temperature, humidity, and weather\_condition.

- **Moderate temporal granularity:** Most data is recorded at hourly or daily intervals, making it suitable for time series models like LSTM and Transformer.

- **Real-world applicability:** Data originates from the Panama power system, providing a realistic setting to demonstrate model performance.Source：<https://www.kaggle.com/datasets/saurabhshahane/electricity-load-forecasting?select=continuous+dataset.csv>



**Task：**

**1）Multivariate Regression for Load Forecasting**

·Objective: Predict future electricity demand using past power usage and weather conditions

·Input Features:

- Meteorological data: T2M\_toc (temperature), QV2M\_toc (humidity), PS\_toc (pressure), WS10M\_toc (wind speed), etc.

- Lagged values of nat\_demand (national electricity demand)

·Prediction Target: nat\_demand for the next day (or multiple future days)

·Model Suggestions: LSTM, Transformer, TimesNet, XGBoost

·Sliding Window Strategy: Use the past 7, 14, or 28 days of data as input to predict electricity demand for the next 1 day or multiple days

Climate

**Mission:**

This dataset aims to support the application of LSTM and other deep learning models for multivariate time series forecasting, enhancing prediction accuracy by learning from multiple correlated variables.

**Content:**

The dataset provides hourly air pollution and weather data collected over five years at the U.S. Embassy in Beijing, China. It includes PM2.5 concentrations and several meteorological features such as temperature, dew point, pressure, wind direction, wind speed, and cumulative hours of snow and rain.

**Features:**

No: row number

year: year of data in this row

month: month of data in this row

day: day of data in this row

hour: hour of data in this row

pm2.5: PM2.5 concentration

DEWP: Dew Point

TEMP: Temperature

PRES: Pressure

cbwd: Combined wind direction

Iws: Cumulated wind speed

Is: Cumulated hours of snow

Ir: Cumulated hours of rain

We can use this data and frame a forecasting problem where, given the weather conditions and pollution for prior hours, we forecast the pollution at the next hour.

Source：<https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate?select=LSTM-Multivariate_pollution.csv>



**Task：**

**1）Multivariate Time Series Forecasting**

·Objective: Predict future air pollution levels based on past pollution and weather data

·Input Variables: DEWP, TEMP, PRES, cbwd, Iws, Is, Ir

·Time Features (to be engineered): hour, dayofweek, is\_weekend

·Prediction Target: PM2.5 concentration for the next 1 / 3 / 6 hours

·Recommended Models: LSTM, Seq2Seq, Transformer

**2）Air Quality Classification (Regression-to-Classification)**

·Objective: Convert pollution values into air quality levels (e.g., Good / Moderate / Unhealthy / Very Unhealthy / Hazardous)

·Method: Apply binning on PM2.5 values to create Air Quality Index (AQI) categories

·Model Suggestions: Classification models such as MLP or LSTM

·Application Scenario: Building urban air quality alert systems

Finance

Source：yfinance API



**Task:**

**1）Stock Closing Price Prediction**

·Objective: Predict the future closing price of a stock based on historical price, volume, and technical indicators

·Input Variables:

- Raw features: Open, High, Low, Close, Volume

- Derived features: Technical indicators such as moving averages, RSI, MACD, etc.

·Prediction Target: Close or Adjusted Close price for the next day or multiple future days

·Model Suggestions: LSTM, GRU, Transformer, XGBoost

·Sliding Window Strategy: Use data from the past 30 days to predict the next day's closing price (regression task)