Retail

Walmart Sales Forecasting (M5 Competition Dataset)

**Description:** Contains historical sales data from Walmart, including store-level, department-level, and product-level sales with additional metadata (prices, promotions, holidays).

**Forecasting Task:** Predict sales for the next 28 days (multi-step).

**Complexity:** Large-scale, hierarchical time series.

来源：https://www.kaggle.com/c/m5-forecasting-accuracy/data



**可以做：**

**1）商品销量预测（多步回归）** 这是 M5 比赛的官方任务。

·目标变量：sales

·输入特征：滞后销量 (lag\_1, lag\_7, lag\_28)、时间特征（如 wday, month, event\_type）、价格（sell\_price）

·预测目标：接下来的 28 天销量（multi-step prediction）

·模型示例：LSTM / Transformer / XGBoost / ARIMA / Prophet

·评价指标：RMSE / MAE / SMAPE 等

1. **促销影响预测（分类或回归）**

预测 “商品是否因为节日销量会上升”

·目标变量：是否发生销量激增（分类）或激增幅度（回归）

·特征：event\_name\_1, event\_type\_1, snap\_CA，价格变化率，滞后销量

·应用场景：节假日补货、广告投放策略优化

·模型：分类模型（LightGBM, MLP）或回归模型

**3）价格弹性建模（价格变化对销量的影响）**

·目标变量：sales

·特征重点：当前 sell\_price + 滞后销量 + 是否有节假日 + 时间特征

·任务重点：衡量“价格变化”对销量的影响（可加入交互项）

·模型选择：LSTM with sell\_price 或 Transformer 加权机制可解释性强

Transportation

来源：TLC Trip Record Data - TLC

纽约市高频网约车（FHVHV）行程数据，主要来源于 Uber/Lyft 等平台。每一行表示一笔乘客订单（一次乘车）。 以“上车时间 pickup\_datetime”为时间戳，统计单位时间内订单量

Time：2009-2025

我下载了2024年1—6月的，应该够了

**可以做：**

**1）每小时订单量预测**

构造时间序列：pickup\_datetime → 小时级聚合 → trip\_count

·输入特征可用：小时/星期几（hour, weekday）、上一小时订单数（lag\_1, lag\_24, rolling\_mean\_7h）、是否节假日、是否早晚高峰（可用 hour 构造）

·预测目标：未来 1 / 3 / 6 小时的订单数

·适合模型：LSTM, Transformer, TimesNet, XGBoost

**2）区域级订单预测（空间 + 时间）**

对 PUlocationID 做分组预测，每小时每区域的叫车量

Energy

特征丰富 通常包含 load、temperature、humidity、weather\_condition 等多种天气变量

时间粒度适中 多为 每小时/每天数据，适合 LSTM/Transformer 建模

应用真实 来源于巴拿马电力系统，适合展示模型在现实场景下的效果

来源：<https://www.kaggle.com/datasets/saurabhshahane/electricity-load-forecasting?select=continuous+dataset.csv>



**可以做：**

**1）多变量回归预测 - 预测未来电力需求**

输入过去 N 天的用电 + 气象数据，预测未来 1 天或多天电力需求

·输入特征：T2M\_toc, QV2M\_toc, PS\_toc, WS10M\_toc 等气象特征 + 滞后 nat\_demand

·输出目标：下一天 nat\_demand

·模型建议： LSTM, Transformer, TimesNet, XGBoost

·滑动窗口：过去 7/14/28 天输入 → 预测未来 1 天或多天

Climate

About Dataset

THE MISSION

The story behind the dataset is how to apply LSTM architecture to understand and apply multiple variables together to contribute more accuracy towards forecasting.

THE CONTENT

Air Pollution Forecasting

The Air Quality dataset.

This is a dataset that reports on the weather and the level of pollution each hour for five years at the US embassy in Beijing, China.

The data includes the date-time, the pollution called PM2.5 concentration, and the weather information including dew point, temperature, pressure, wind direction, wind speed and the cumulative number of hours of snow and rain. The complete feature list in the raw data is as follows:

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.

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



**可以做：**

**1）多变量时间序列预测**

用过去 N 小时的天气和污染数据 → 预测未来 1~6 小时的 pollution

·输入变量：dew, temp, press, wnd\_dir, wnd\_spd, snow, rain

·时间特征：可构造 hour, dayofweek, is\_weekend

·模型建议：LSTM, Seq2Seq, Transformer

·预测目标：未来 1 / 3 / 6 小时的 PM2.5 浓度

**2）空气质量等级分类（回归 → 分类）**

根据污染浓度将空气质量分为等级（如优/良/轻度/中度/严重）

·方法：对 pollution 做分箱，构造 AQI 等级

·模型：分类模型（MLP/LSTM）

·应用场景：城市空气预警建模

Finance

来源：yfinance



**可以做**

**1）股票收盘价预测**

基于过去 N 天的价格、成交量、技术指标 → 预测未来 1 天或多天的收盘价

·输入变量：Open, High, Low, Close, Volume + 派生特征

·输出目标：Close / Adj Close

·模型建议：LSTM / GRU / Transformer / XGBoost

·滑窗策略：过去 30 天 → 预测未来 1 天收盘价（回归）

**2）方向性预测（涨/跌分类）**

将收益率变为涨跌标签，构建分类任务：1=涨，0=跌

·模型：MLP, RNN, Transformer

·应用：量化交易决策辅助