Overview of the 2nd project

- Best writing award: HUANG Yuxin, LEI Yunxin, AN Tianyuan, LIN Fengshan, LIU Zongxuan (paper 1)
- Best technique award: LI Aoran, MA Yijia, WENG Langting, ZHOU Tianying (paper 2)
- Best overall award: YANG Tianhao, JIA Yaoyao, JIANG Xiaoyue, HUANG Yuxuan (paper1)
- Best overall award: SUN Peiran, LUO Xinyang (paper 2)

- Congratulations! You can gain some **bonus** for your grades!
- By the way, the grade of the warm-up project does not count a lot for the final score; you still have the opportunity to earn a **bonus** during the following projects!
- Besides, please *submit your project reports / review / rebuttal in time*, otherwise, you may lose some point.

Project 3:
Kaggle ——
G-Research Crypto Forecasting

MAFS6010Z, Fall 2023

Background

Cryptocurrencies, such as Bitcoin and Ethereum, are among the most popular assets for speculation and investment, yet have proven wildly volatile. **Fast-fluctuating** prices have made millionaires of a lucky few, and delivered crushing losses to others. Could some of these price movements have been predicted in advance?

- Task: use your machine learning expertise to forecast **short term returns** in 14 popular cryptocurrencies.
- Difficulty: extreme volatility of the assets, the non-stationary nature of the data, the market and meme manipulation, the correlation between assets and the very fast changing market conditions.

Data

Millions of rows of minute-by-minute cryptocurrency trading data dating back to 2018

Data features:

- **timestamp**: Timestamps in this dataset are multiple of 60, indicating minute-by-minute data.
- **Asset_ID**: The asset ID corresponding to one of the cryptocurrencies.
- **Count**: Total number of trades in the time interval (last minute).
- **Open**: Opening price of the time interval (in USD).
- **High**: Highest price reached during time interval (in USD).
- **Low**: Lowest price reached during time interval (in USD).
- **Close**: Closing price of the time interval (in USD).
- **Volume**: Quantity of asset bought or sold, displayed in base currency USD.
- **VWAP**: The average price of the asset over the time interval, weighted by volume.
- **Target**: Residual log-returns for the asset over a 15 minute horizon.

Your Job

Predict price returns across 14 major cryptocurrencies, in the time scale of minutes to hours.

- Your predictions will be evaluated by how much they **correlate** with real market data collected during the future three-month evaluation period.
- In advance, you are encouraged to perform **additional statistical analyses** to have a stronger grasp on the dataset, including autocorrelation, time-series decomposition and stationarity tests.

Prediction Targets and Evaluation

- **Predicting Targets:** predict returns in the near future for prices P^{α}, for each asset a.
  - Log returns over 15 minutes: $R^{a}(t) = log(P^{a}(t+16) \ / \ P^{a}(t+1))$
  - Weighted average market returns: $M(t) = \frac{\sum_{a} w^{a} R^{a}(t)}{\sum_{a} w^{a}}$

$$\beta^{a} = \frac{\langle M \cdot R^{a} \rangle}{\langle M^{2} \rangle}$$

  - Target: $\text{Target}^{\alpha}(t) = R^{\alpha}(t) - \beta^{\alpha} M(t)$

weights w^{α} given by the 'weight' column in the Asset Details file; bracket $\langle \cdot \rangle$ represent the rolling average over time (3750 minute windows)

Refer: https://www.kaggle.com/code/cstein06/tutorial-to-the-g-research-crypto-competition/notebook

Prediction Targets and Evaluation

- **Evaluation metrics: weighted Pearson Correlation Coefficient (wPCC)** of your prediction and the real data value.

$$wPCC = w^{\alpha} \frac{cov(\mathrm{Target}^{\alpha}, \mathrm{Real}^{\alpha})}{\sigma_{\mathrm{Target}^{\alpha}} \sigma_{\mathrm{Real}^{\alpha}}}$$

weights w^{α} given by the 'weight' column in the Asset Details file; cov is the covariance; σ is the standard deviation.

Refer: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Notes

- Be careful of the missing values. Rows with nulls in the test set ground truth are ignored for scoring purposes.
- **The danger of overfitting** should be considerable.
- The **volatility** and **correlation structure** in the data are likely to be highly **non-stationary**.
- Changes in prices between different cryptocurrencies are highly **interconnected**. For example, Bitcoin has historically been a major driver of price changes across cryptocurrencies but other coins also impact the market.

Project 3:
Kaggle ——
M5 Forecasting

MAFS6010Z, Fall 2023

Background

- M5 forecasting: the 5th Makridakis Competition.
- Task: Forecasting (accuracy) and estimating the uncertainty distribution of the realized values of the same series
    - **Accuracy task**: Can you estimate, as precisely as possible, the **point forecasts** of the unit sales of various products sold in the USA by Walmart?
    - **Uncertainty task**: Can you estimate, as precisely as possible, the **uncertainty distribution** of the unit sales of various products sold in the USA by Walmart?
- Aim:
    - Identifying the most appropriate method(s) for different types of situations requiring predictions and making uncertainty estimates
    - Comparing the accuracy/uncertainty of ML and DL methods versus those of standard statistical ones

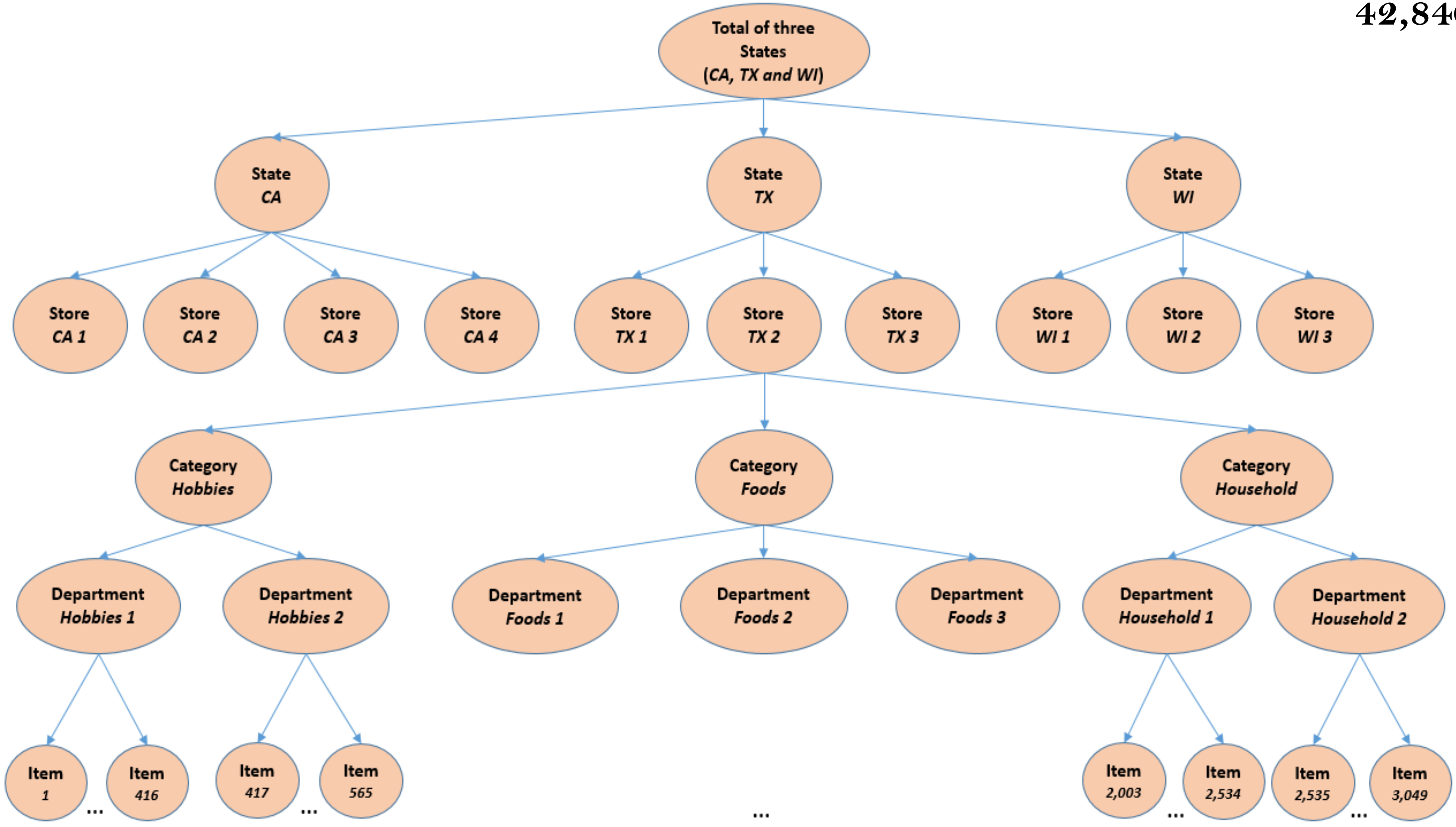
Data

- 42,840 time series data from Walmart (sales data from 2011-01-29 to 2016-06-19).
- **hierarchical sales data:** starting at the **item level** and aggregating to that of **departments**, **product categories** and **stores** in three geographical areas of the US: California, Texas, and Wisconsin.
- **explanatory variables** are also included; such as price, promotions, day of the week, and special events (e.g. Super Bowl, Valentine's Day, and Orthodox Easter) that affect sales which are used to improve forecasting accuracy.
- The majority of the more than 42,840 time series display **intermittency** (sporadic sales including zeros).

Refer: https://mofc.unic.ac.cy/m5-competition/

Data Organization Overview

In total:
42,840 time series



Your Job

- **Accuracy task:** forecasting daily sales of each products for the next 28 days. (m5-forecasting-accuracy)
- **Uncertainty task:** 28 days ahead probabilistic forecasts for the **median** and four **prediction intervals (PIs)** (50%, 67%, 95%, and 99%). (m5-forecasting-uncertainty)

➢ The two task using the same dataset.

Evaluation metrics

- **Accuracy task: Weighted Root Mean Squared Scaled Error (RMSSE)**

$$\boldsymbol{RMSSE} = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}(Y_t-\widehat{Y}_t)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(Y_t-Y_{t-1})^2}}$$

where $\boldsymbol{Y}_t$ is the actual future value of the examined time series at point *t*, $\widehat{\boldsymbol{Y}_t}$ the generated forecast, *n* the length of the training sample, and *h* the forecasting horizon.

$$\boldsymbol{WRMSSE} = \sum_{i=1}^{42,840} w_i * RMSSE$$

where w_i is the weight of the i_{th} series of the competition. A lower WRMSSE score is better.

Evaluation metrics

- **Uncertainty task: Weighted Scaled Pinball Loss (WSPL)**

$$\mathbf{SPL(u)} = \frac{1}{h}\frac{\sum_{t=n+1}^{n+h}(Y_t - Q_t(u))u\mathbf{1}\{Q_t(u) \le Y_t\} + (Q_t(u) - Y_t)(1-u)\mathbf{1}\{Q_t(u) > Y_t\}}{\frac{1}{n-1}\sum_{t=2}^{n}|Y_t - Y_{t-1}|}$$

where $\boldsymbol{Y_t}$ is the actual future value of the examined time series at point *t*, $Q_t(u)$ the generated forecast for quantile u, *n* the length of the training sample, *h* the forecasting horizon, and **1** the indicator function.

➢ Given that forecasters will be asked to provide the **median**, and the 50%, 67%, 95%, and 99% **PIs**, *u* is set to u1=0.005, u2=0.025, u3=0.165, u4=0.25, u5=0.5, u6=0.75, u7=0.835, u8=0.975, and u9=0.995.

$$\boldsymbol{WSPL} = \sum_{i=1}^{42,840} w_i * \frac{1}{9}\sum_{j=1}^{9} SPL(u_j)$$

where w_i is the weight of the i_{th} series of the competition and u_j the j_{th}out of the examined quantiles. A lower WSPL score is better.

Weighting

M5 involves the unit sales of various products of **different selling volumes and prices** that are organized in a **hierarchical** fashion. Therefore, you must provide accurate forecasts across all hierarchical levels, **especially for series of high importance**, i.e. for series that represent significant sales, measured in US dollars.

To that end, the forecasting errors computed for each participating method (both RMSSE and SPL) will be **weighted** across the M5 series based on their **cumulative actual dollar sales**, which is a good and objective proxy of their actual value for the company in monetary terms.

Refer: https://github.com/Mcompetitions