FINTECH 540 - Machine Learning for Fintech

Fall Semester 2025

First Project Checkpoint

Duke

PRATT SCHOOL *of*
ENGINEERING

Data Understanding and Clarification

- Can we know how the variance and variation values have been calculated? Are these values normalized (using log returns or percentage change) or are they raw values? Are these annualized values? What are the units of these values, are these percentages or dollars/price?
  - This is the description of the dataset: `https://capire.stat.unipd.it/methodology/`
- Can we get an explanation of good/bad variance? How was this calculated? Is good when the stock price goes up and bad is when the stock price goes down?
  - Look at this paper: `https://public.econ.duke.edu/~ap172/Patton_Sheppard_REStat_2015.pdf`.

Feature Engineering and Variable Selection

- One challenge we're facing is feature engineering. Metrics like RV, BPV, and RQ are often correlated and noisy.
    - A good benchmark for these type of vol models is the HAR model `https://statmath.wu.ac.at/~hauser/LVs/FinEtricsQF/References/Corsi2009JFinEtrics_LMmodelRealizedVola.pdf`.
- We are currently unsure whether a volatility forecasting model needs to include all types of data (1 and 5 mins). Should the RQ variable be included in the model?
    - Pick one set of estimated values. RQ is he variation of volatility itself.
- Another challenge is feature selection: what past variables should be used to predict future volatility?
    - Refer to the model above.
- Can we include macroeconomic or sentiment data to predict?
    - Yes, you should think about extending these models that only use past lags of variables.

Feature Engineering and Variable Selection (II)

- Technical Indicators Guidance: Do you have any recommendations for technical indicators that would enrich the DJIA dataset?
  - This is part of the task. Trying different features and see what works. You can start collecting HF price on the same asset and compute Tech Idx using Ta-lib `https://github.com/TA-Lib/ta-lib-python`.
- Feature Engineering for Volatility Forecasting: We are exploring ways to engineer features that capture market regimes or structural shifts. Are there recommended approaches or references for incorporating macroeconomic indicators (e.g., interest rates, CPI, GDP) or sentiment?
  - You should research the best practices and integrate/build upon them accordingly.

Model Selection and Evaluation Strategy

- We have some financial time series data (like returns or realized vol), but we haven't learned time series models like ARIMA or GARCH.
  - Those models can be a benchmark, your goal is to try to go beyond those standard time series approach.
- Should we prioritize model interpretability when scoring the models, or is overall model performance more important?
  - You can prioritize performance in this context, but always relative to a benchmark.

Model Selection and Evaluation Strategy

- Is there any potential answer or metric to see whether the prediction is good enough? (For example, if I choose to use dataset 1 to predict the volatility, do we need to compare our prediction with an official standard?)
  - Having a benchmark is crucial in any type of predcitive task.
- Unsupervised Model Evaluation: For the blockchain alternative, since the algorithm is unsupervised and lacks a target variable, how will this model be evaluated? Should we aim to maximize specific evaluation metrics relevant to our model type?
  - This depends on the task. What is the goal in applied unsupervised learning from the data? For instance, clustering techniques have their own way to evaluate the goodness of a cluster.

Computational Feasibility and Resource Constraints

- Given the resources that we have currently, we are not sure if we have enough computational power to train our models.
  - Have you encountered a bottleneck? Using which dataset?

Volatility Forecasting and Extreme Events

- We did a basic data visualization and observed extreme volatility spikes during the 2008 financial crisis and 2020 COVID-19 period (10–20x higher than normal levels). If we choose to do volatility forecast, how do we balance prediction accuracy in normal and extreme periods (how do we handle the extreme values)?
    - This is expected in financial data. You may want to experiment with different training windows.
- How can we effectively address the Market Regime Classification problem, given the challenge of defining regime boundaries under non-stationary market conditions? Specifically, how can we identify distinct phases of market behavior—such as high-volatility versus low-volatility periods—when volatility levels shift over time and fixed thresholds cannot reliably separate regimes?
    - You can research a bit on how machine learning models have been used to detect regimes. A starting point could be identifying thresholds and labeling your data accordingly. See Isolation Forest `https://scikit-learn.org/stable/`

Blockchain Data and Unsupervised Modeling

- The blockchain dataset is a complicated one that involves different aspects of features. We will probably spend most of our time delving into the data structures and feature engineering.
    - This is expected. Feature engineering and problem design are 80% of the time you spend when solving an ML problem.
- We're working with the Blockchain data. Is there a correlation between these datasets, or would using a single dataset be sufficient for training our model?
    - The PDF describes a bit how these datasets are related. You need to start with an exploration.
- Labeling and Interpretability in Blockchain Modeling: We're unsure how to interpret or validate the results without labeled outcomes.
    - The fact that you don't have a specific label doesn't mean you can't solve a supervised learning problem. The label can be created.

Project Scope, Dataset Choice, and Grading Criteria

- Evaluation Criteria Clarity: We are wondering whether the grading will depend primarily on model performance metrics (e.g., MSE, R2) compared to other teams, or if it will focus more on the creativity and reasoning behind our modeling approach.
  - See the eval template. No comparison across teams in terms of performance.
- Dataset Substitution: Are we allowed to import external datasets to substitute or complement the provided ones? For example, we are considering using specific stock price time-series data to compute the standard deviation of returns and derive our own realized volatility.
  - You can complement, not substitute.
- Should the team decide which dataset to work on now? I answered no. Is that right?
  - Yes, you don't have more time to think about it.

Further Questions? Comments?