Team Project 4

Step 1: Baseline

We started off by figuring out how to submit our results to the competition using the evaluation server. For our baseline, we went with Logistic Regression (well, technically LinearRegression in the code, but you get the idea). We just threw in all the features without tweaking anything mostly to see where we stood. It turns out we're only allowed five submissions a day, which made testing new ideas a bit tricky. Our baseline got us a score of **1.157**. Not amazing, but it gave us something to improve on.

Step 2: LightGBM + Validation

After that, we moved to the LightGBM V1 model. We picked LightGBM since it's quick, works great with tabular data, and handles non-linear stuff better than our first model. Our first try with LightGBM shot our score up to **2.515**.

Next, we realized the dataset's temporal, so we switched to time series validation. Randomly shuffling the data just didn't match real-world test conditions. With time series validation, we could see how the model would actually perform. We also added a cap so volatility wouldn't go over 120% that way, our predictions stayed under control. Even with these changes, our leaderboard score was a bit better but training took way longer because we had to run a bunch of models.

Step 3: Feature Engineering

Then we jumped into feature engineering. We plotted feature correlations and picked out only the ones that crossed our threshold. This trimmed down the number of features and focused on the strongest signals for training. After this feature selection, our results got better, so clearly, it helped.

Step 4: Metrics and Sharpe Score

To make sense of all our experiments, we wrote functions to calculate and print our Sharpe variant score for both the classic and time series models. This way, we could test out new features or strategies without burning through our precious five daily Kaggle submissions just to see if something worked.

---

github link : https://github.com/Arishkanyam/ML-Project-4