**Stockhack 2025 - Yashas Acharya Submission**

I’m predicting 5-day return and price of the selected equities in the competition. I employed an ensemble of CatBoost, LightGBM and a NN to predict the valued.

My data historical values prices for the stocks. For each of the companies I found dates, last price, open price, high price, low price, volume, bid price, ask price, marker order bid size (which is mostly blank), marker order ask size (which is also mostly blank), average volume 30 days, moving average 20 days, turnover / traded value, RSI 14 day, moving average 20 day, moving average 50 day, moving average 200 day, volatility 30 day, implied volatility mid, current market cap, VWAP. The frequency of the data is daily so it's pretty good. This data goes from 1/1/2015 till last friday closing which was 2/21/2025. I cleaned the data by handling missing values using rolling window averages and applied MinMax Scaling to standardize numerical features for neural network compatibility. To capture temporal patterns, I generated lagged features and calculated the 5-day future return as the target variable.

For training, I used three models: CatBoost, which is a gradient boosting algorithm optimized for categorical data; LightGBM, a highly efficient boosting algorithm for large-scale datasets; and a Neural Network, designed with two hidden layers to capture non-linear relationships. Each model was trained separately on individual stocks using their respective historical data. Optuna was used to tune hyperparameters such as learning rate, tree depth, and regularization for the boosting models, while hidden layer sizes, learning rate, and number of epochs were optimized for the neural network.

The final prediction is derived from a weighted ensemble of the three models, where CatBoost contributes the most with a weight of 0.6742, followed by LightGBM at 0.2760, and the Neural Network at 0.0498. These weights were determined based on cross-validation performance to maximize predictive accuracy.

The final predictions for the 5-day return and price are as follows: ALT US Equity has a predicted return of 9.28% and a price of $7.03, CELH US Equity has a predicted return of -4.93% and a price of $31.01, CVNA US Equity has a predicted return of 7.94% and a price of $241.02, FUBO US Equity has a predicted return of -7.20% and a price of $3.49, and UPST US Equity has a predicted return of -0.59% and a price of $71.35.

The evaluation metrics indicate the model's performance across different stocks. The mean absolute error (MAE) and root mean squared error (RMSE) for each stock highlight the prediction errors. For instance, ALT US Equity has an MAE of 0.0963 and an RMSE of 0.1016, while CELH US Equity has an MAE of 0.1229 and an RMSE of 0.1471. The R-squared values show varying levels of explanatory power, with some stocks having negative values, indicating high volatility and unpredictability in their price movements.

I observed that ensemble models improved accuracy compared to individual models, and CatBoost contributed the most to final predictions. However, the high variance in R-squared scores suggests that further feature engineering is needed. The model can be improved by incorporating additional macroeconomic factors like interest rates and market sentiment. Improving generalization techniques will help mitigate overfitting, and experimenting with Transformer-based models for time-series forecasting could further enhance accuracy.

I also wasn't able to use some intra day data that I found which I could've used in the fine tuning process as well.

**Final Predicted Prices:**

- **ALT US Equity**: **$7.03**
- **CELH US Equity**: **$31.01**
- **CVNA US Equity**: **$241.02**
- **FUBO US Equity**: **$3.49**
- **UPST US Equity**: **$71.35**