Trexquant Project

B:

1. **Data Extraction:** Data was extracted in CSV format from https://in.investing.com/equities/pfizer-historical-data

2. **Data Cleaning**: Data was cleaned and sorted based on date

3. **Proper handling of non-numeric features**: Date was converted to *to_datetime* format and sorted. Column 'Vol.' contained string values so 'M' and 'K' were replaced with blanks and multiplied with *10^6* and *10^3* respectively in new columns and the old columns were *dropped*.

4. **Return** and **Target** columns were created. **Return** is a column calculating **percentage change** from one row to the next. As we needed to take only one decision on any given day so **Target** returns a **binary value of 1 and 0**, where **1 represents increase in price the next day** and **0 otherwise.** So it represents next day's return is **positive or negative** (shifted one above).

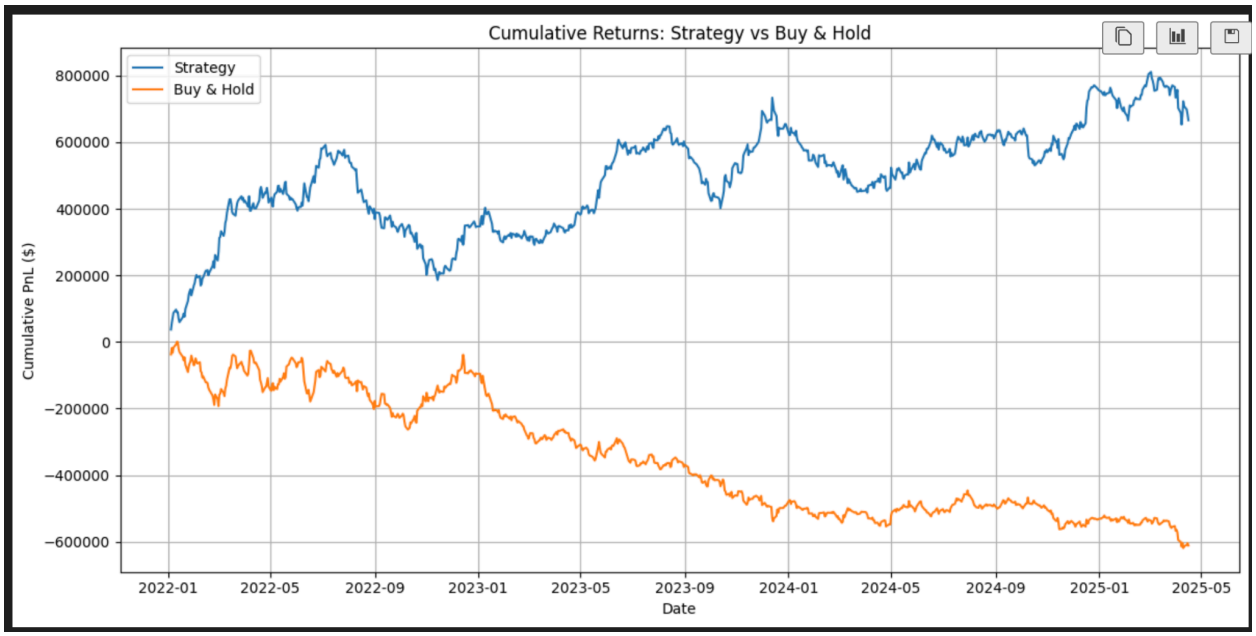
5. **Feature Engineering:** Some standard features were chosen upon testing different ones. Finally the following were used -
   - **Z score:** It tells how far is the price from its mean in terms of standard deviations (Rolling mean and standard deviation were calculated for 10 trading days)
   - **Momentum over 5 Days:** Measures price change over 5 days (short-term momentum).
   - **Momentum over 10 Days:** Similar to above (medium-term momentum)
   - **Bollinger width:** Used as an indicator as it gives a sense of volatility
   - **Volume change:** Measures the change in trading volume from one day to the next.

6. **Hump based Alpha:** It multiplies the **ranked daily negative return** and the **ranked volume spike ratio**, measuring how much volume today exceeds its 20-day average.(It is a metric calculated and used as showed better results and normalization)
   - **avg20** calculates avg daily volume over past 20 days
   - **Volume_ratio** gives a signal telling whether todays volume is above average or not
   - **Rank_Negreturn** ranks daily return with more **negative returns** getting **lower ranks**
   - **Rank_VolumeRatio** ranks days where volume is **most abnormally high** with **lower ranks** (1 = highest volume spike).
   - **Hump_Alpha** combines both ranks by multiplying them emphasizing on days that are **sold off and have unusual volume** to capture rebounds or signals.
   - Inspired by https://platform.worldquantbrain.com/learn/operators

7. Model Training

- **Test/Train split:** Data was split at **2022-01-01** according to problem statement for training and testing respectively avoiding any lookahead bias
- **Scaling:** StandardScaler was applied to normalize features
- **Used XGBoost as it was robust, widely used and produced suitable results apart from others, other models like Ensemble, Light GBM could have been implemented.**
- **positions** logic converts 1 to +1(long) and 0 to -1(short). **prices_test** is a 1D array or list of stock prices, **np.diff(prices_test)** calculates the difference between consecutive prices, **prices_test[:-1]** gives all prices except last one so can divide each price difference by previous.
- **Cumulative pnl** was calculated
- **Buy and hold returns** were also calculated for comparison
- **Visualization** was carried out by plotting cumulative pnl vs buy and hold for comparison

Result



References

https://xgboost.readthedocs.io/en/stable/

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html