# Pfizer (PFE) Single Asset Machine Learning Based Trading Strategy

By: Prakhar Mittal (2022B3A70426P)

Email: f20220426@pilani.bits-pilani.ac.in

## Objective

The goal of the project was to develop an out-of-the-box creative trading strategy using a model trained on any data of my choice

## Constraints(and Assumptions)

1. Only Long/Short Positions allowed (not allowed to withhold)
2. Only 1 position of flat 1M $ to be made once a day compulsorily
3. No lookahead Bias
4. Normalized Data
5. 2022-2025 data is reserved for testing only, train-val split is a personal choice

## Folder Structure

1. pfizertrade.ipynb : The submitted working python file w/ jupyter lab locally
2. other_forms.csv: the features extracted from the following documents.
    a. 10K: annual filings
    b. 10Q: quarterly filings
    c. 8K: significant corporate events
3. pfe_data.csv: holds the stock data from our valid range from 2010-2025
4. insider_trading_clean.csv: The features extracted from the SEC filing form 4
5. tda_market_features.csv: The features extracted from extreme events calculations
6. xgb_model.1.3.pkl: the model we trained
7. trade_explanationf20220426p.pdf: this document

# Dependancies

1. yfinance
2. pandas
3. Numpy
4. Matplotlib.pyplot
5. xml.etree.ElementTree
6. Datetime
7. Requests
8. Bs4
9. Transformers
10. Re
11. Os
12. Gtda
13. Scipy
14. Pickle
15. Xgboost
16. Sklearn.metrics
17. sklearn.model_selection

# Data Extraction

The data for PFE was extracted from the yfinance library, whereas the SEC filings were downloaded using sec-edgar-downloader library API which allowed me to automatically download all 10K, 10Q, 8K and Form 4 filings by PFE during the given time period.

We also use sectoral data with the ticker (IBB), again extracted through the yfinance library.

An earlier version of the code called an API to download the JSON responses for each filing instead of the document itself; however that API proved to be far more difficult to handle and was removed. The use of JSON response over full submission files would have saved space on the system but in case of sentiment analysis would have proved far slower and unwieldy.

# Data Cleaning

The values of OHLC components were all shifted by 1 to prevent calculation of any feature on present or future prices. This was done to every instance of price data.

The filing dates of the documents were scraped from within the documents themselves using beautiful soup.

## Non Numeric Data

No NaN rows were deleted (aside from the ones with all values NaN) since XGBoost can handle them automatically.

The binary variables were mapped to 1/0 in case of the value range being true/false and 1/-1 for actionable signal that denoted going long/short

## Feature Engineering

After much deliberation and research the following categories seemed relevant for creating informative features of the data.

1. OHLC based technical indicators
2. Sentiment analysis
3. Extreme Event Detection
4. Other(Options Volatility, Product level indicators, Pharmaceutical pipeline etc)

For the first category we chose:

1. Rolling volatility
2. Momentum
3. RSI
4. Bollinger Z-Score
5. ATR

These are some standard (and in our case normalized) features that provide actionable insights using just the price data. Usually more suitable for short term trading). These were also chosen because during a sanity check with simple strategies, these few indicators did not perform terribly as compared to the rest.

The calculations involved in these metrics were standard to their definition.

For the second category:

1. SEC Filings(10K, 10Q, 8K, 4)
2. News Headlines

3. Social Media

Eventually we ended up choosing only SEC filings due to the unavailability of News headlines for very old data and the unreliability of social media.

From these SEC filings we were unable to process the files themselves due to a lack of time and practicality. The files themselves were nonstandard and required heavy pretrained models that had a working time of up to 100s/file leading to unreasonable timings.

Instead we calculated the standard deviation in the intervals between each filing. The annual 10Ks were almost precise in their timings whereas the 8K were haphazard. Meaning we could create an anticipation flag variable for the days leading up to the file's release. Whereas for 8K we could create a recency variable that would track the recency of the "shock" of the 8k being released.

All kinds of transaction details for form 4 were also cleaned up by decoding transaction categories into buy/Sell actions, position of acting entity etc. This allowed us to receive a value (positive or negative) of the stock transaction. A positive value meant a bullish sentiment by insiders and a negative one meant bearish.

This way we were able to extract actionable sentiment from the forms when we were unable to parse the files themselves.

For the third category:

It is important to realise that we live in a turbulent time, especially due to events like COVID-19 and the environmental degradation, accumulation of pollutants, rising cancer rate. Meaning that intense spikes of extreme events are important actionable intel.

Using the research paper(cited below) we were able to create relevant 'spike' metrics, their interactions and their normalized values that allowed us to calculate the days since the spike of any of our 3 categories. These spikes are calculated using 2 other stocks (JNJ and AZN) and the IBB sector index.

Essentially finding any event like a competitor's discovery of drugs, clinical trial success that reverberates through our stock price.

Unfortunately for category 4 it was extremely difficult to obtain data for any of these categories, whether this was due to lack of funds or availability itself.

# Modelling

This problem, with the given constraints, could be modelled as either a classification problem or as chosen, a regression problem.

Ideally a classifier would be able to directly predict the trade signal based on the targets calculated from the actual next day returns:

Next day returns > 0 -> Target = 1; else Target = -1

However in a turbulent quickly changing market, the classifier struggles to detect a pattern and fails quite a lot. After trying both models, I settled on Regression, where I modeled the returns for the next day and predicted them. A positive prediction generated a +1 trading signal and similarly a -1 signal for negative prediction.

The data was also divided into training, validation and testing splits. The final results were calculated on the testing data to ensure fair, out of sample testing.

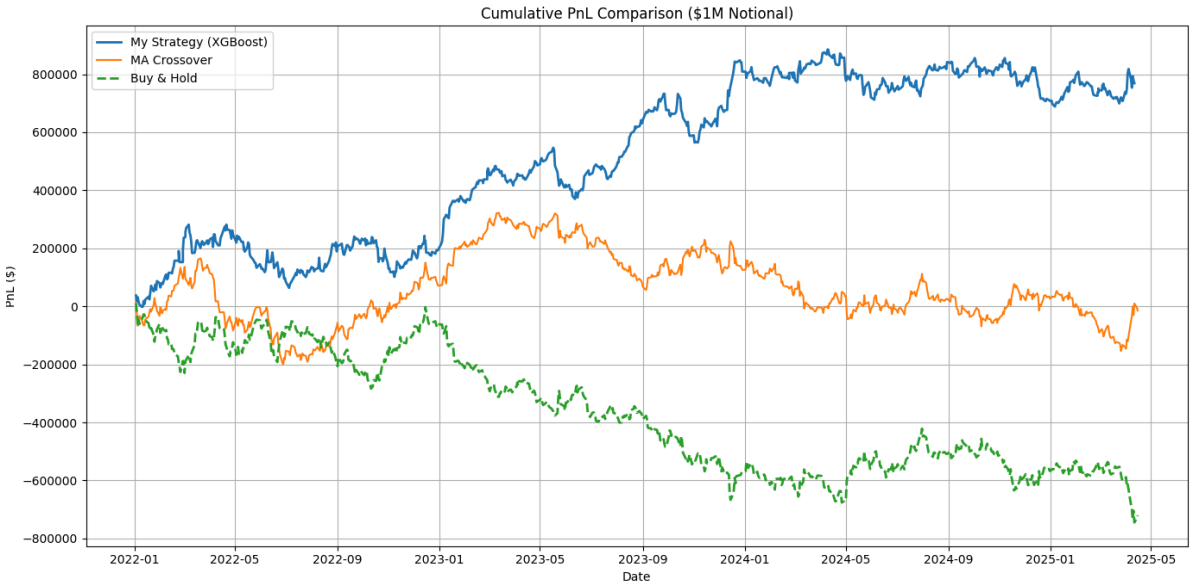
# Results

Although model 1.2 and 1.3 were chosen on the basis of a theoretical hypothesis. More than 25 variations to models and the features were tested rigorously and empirically, model 1.2 was the winner by a huge margin, with model 1.3 close behind.

Model 1.2: only Form 4 data was included
Model 1.3: all SEC filings were included

Model 1.2 showed an impressive 90% profit over 3 years and Profit rate of 1.2 (Sharpe 1.05)
Model 1.3 showed an impressive 76% profit over 3 years and Profit rate of 1.1 (Sharpe 0.95)

Although model 1.2 was better, 1.3 has been submitted due to the theoretical foundation of the inclusion of SEC filings and the sentiments provided by them. No doubt with enough time i could have eventually done better with the model 1.3




*Model 1.3 results (SUBMITTED FILE)*

# Citations/References

**Extreme Events and COVID-19 Paper**
*Chakraborty, I., & Maity, P. (2020).* "COVID-19 Outbreak: An Empirical Analysis of the Impact on the US Financial Markets." *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3554225

**LLM Sentiment Analysis Implementation Paper**
*Huang, A., Dyhdalo, A., & Eder, O. (2023).* "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models." *arXiv preprint arXiv:2104.14841.*

**Pfizer Portfolio Analysis Document**
*Pfizer Inc. (2023).* "Investor Presentation – Pipeline and Portfolio Strategy." Retrieved from Pfizer's investor relations website.

**SEC EDGAR Database**
U.S. Securities and Exchange Commission. (2010–2025). *EDGAR – Company Filings for Pfizer Inc.* https://www.sec.gov/edgar/searchedgar/companysearch.html

**Yahoo Finance**
Yahoo! Finance. (2010–2025). *Pfizer Inc. (PFE) Historical Stock Data.*

**IBB ETF Benchmark**
NASDAQ. (2010–2025). *iShares Nasdaq Biotechnology ETF (IBB).*