Analyzing The Trends In Stock Volatility Through The Application Of Abduction Methods

Yermal Koustubh Rao

November 2023

Contents

1 Who Am I?

Yermal Koustubh Rao (200100176) is a final year undergraduate pursuing his bachelors in the Computer Science and Engineering department at Indian Institute of Technology, Bombay (IITB), graduating with a Honors degree in 2024.

This document is the final project report for the course CS 781: Formal Methods in Machine Learning, offered by Prof. Supratik Chakraborty at the Indian Institute of Technology, Bombay (IITB) in the Autumn Semester (2023 - 2024).

2 Introduction

Machine Learning in the recent times has proven itself to be an indispensable tool in performing various tasks without human intervention. Due to its easy scaling as well as the fast improvements in GPU Architectures, larger models are trained with larger datasets which leads to the model learning and predicting the hardest learnable tasks in very little time. Sometimes these models end up outperforming the best of humanity, as we saw AI models defeating Chess and GO greats in the recent past.

Due to its prowess people have used machine learning in various fields and walks of life from image detection to speech generation as well as learning to play games. To this end, machine learning is also used extensively in the Fintech field to predict instrument performance like stocks, options etc. LSTMS, CNNs and RNNs have proven to perform really well in predicting prices of instruments in the recent past. But there is a major twist. Machine Learning models are often too complicated to the human mind to grasp and understand, hence certain guarantees cannot be provided by the models.

For example, a model performing really well price prediction up until now may end up crashing at the next moment leading to losses in the order of millions. Since hedge funds and high frequency firms deal with liquidity and transactions in the order of millions in a single day, a single outlier prediction might lead to great losses. Hence, machine learning models aren't used to the expected levels in the finance industries as it is very hard to give solid confirmations for large learning models. Hence, to this day statistics is used majorly to price instruments and predict its motion with time.

We observed this phenomenon throughout the course were proving strict properties of learning models is indeed a difficult task. The approaches that we learned were very loose in terms of the approximations made or would require heavy calculations to be made for large models. This is inefficient as the learning models are continuously learning with time, hence the predictions for now might not be very useful later on for the same model.

To this end, in this report we see how we can get strict confirmations on some of the properties of a machine learning model which predicts stock volatility.

2.1 Problem Statement

To find and understand the minimal set of days from the recent past which sufficiently explains the volatility trend of a stock for the day. Volatility trend here refers to the close open difference in the stock price for a given day. The day ends on a positive trend when closing price is greater than the opening price and negative trend otherwise.

We first train a simple FFNN (FFNN was sufficient, CNN were useless anyways) on the stock data of 99 firms from the finance sector, predicting the trend for the current day given the history of the stock performance since the last few days. We then use the tools obtained to get a valid, sufficient, minimal set of days from the history which is sufficient to correctly predict the current day trend.

2.2 Motivation

There are several factors which affect the performance of a stock for a given day. Analysis made before the market opens lead to a difference between today's opening price and yesterday's closing price. This happens because firms usually tend to announce major company policies towards the end of the market or during weekends, mostly Friday's to afford stability during the market hours. Events like elections, budgets, government policies, calamities etc. can greatly affect the performance during market hours.

It is a difficult task to judge which events in the recent past actually affect's a company in a visible way, some events may lead to a bearish performance and some events may lead to bullish performance following the events. Some might have little to no effect in regards to a company's performance. It is exactly this question upon which I wanted to spend time and use the tools and theory taught during the course, to get into it deeper.

This minimal set, gives us a deeper understanding of how certain events are related to a certain firm and how this gets portrayed in the market. To limit the scope of this vast field, we commit this research only towards companies from the finance industries. Although one might expect to find several events which affect the finance sector as a whole in common, certain events/days might be particular for that bank/firm. For example commercial banks might be affected by changes in tariff rules but private banks might not. Hedge funds might be immune to many of these events, at least in theory due to the hedging of their risk, which might be detrimental to the common bank firms.

It is exactly this relation between events and volatility trends that we wish to understand in this project. We will find some common set of events which affects the industry as a whole as well certain particular events which effects only a few set or a smaller fraction of the finance sector.

These results can be leveraged in the future for better understanding and hence prediction of the relation between events and finance companies/banks. In effect preparing oneself better and hence saving oneself from huge downward trends. This would also end up leading to the development of a stronger economy and hence better life style of the human race.

3 Tools

This analysis is made along the lines of the following papers, Abduction-Based Explanations for Machine Learning Models by Alexey Ignatiev, Nina Narodytska, Joao Marques-Silva and Using MaxSAT for Efficient Explanations of Tree Ensembles by Alexey Ignatiev, Yacine Izza, Peter J. Stuckey, Joao Marques-Silva. An exact tool for the earlier paper, which was the main basis upon which the following project was undertaken was not available on the internet. Hence, the tool 'XReason' for the latter paper was used. This tool was modified extensively to come up with a working tool for the earlier paper. The authors of the earlier paper were contacted, and upon request a broken prototype of the earlier paper was obtained. A lot of wiring and restructuring was done to come up with a working tool which was the main ingredient of the paper. The source code for XReason can be found at https://github.com/alexeyignatiev/xreason, the code base for 'XPlainer', which is the actual tool developed for the first paper can be requested from the author (they currently only have a prototype version with no documentation).

3.1 Modifications

The tools obtained, both 'XPlainer' and 'XReason' were hard coded for the MNIST dataset, hence augmentations had to be made to accommodate the stock data needed for analysis in this project. To this end, certain configuration files were written which consisted configuration details like, loss functions, target feature, weights, epochs, batch size etc. The models.py was updated to handle deeper layers, the older version had only 2 hidden layers which was not sufficient for the dataset in our case. The Neural Network to SMT encoding was present in the 'XPlainer' repository, but the minimum hitting set algorithm was present in 'XReason', this was stitched.

The loss functions were updated to reshape the model predictions to match the shape of the target predictions. The pipeline was hard coded for MNIST and other datasets which the original authors used for their research, so these had

to be updated to intake 'stock_data' dataset, relevant for this project. These updates were tiny but many in number.

4 Solution

Here is an outline of the solution approach and the analysis pipeline used at arriving at the results.

4.1 Data Processing

The top 99 firms from the financial/banking sector was chosen for the analysis listed on NYSE/NASDAQ in the United States of America (USA). Yahoo Finance's yfinance api, was used to obtain the entire stock history of these 99 firms which were then stored in csv format for each firm.

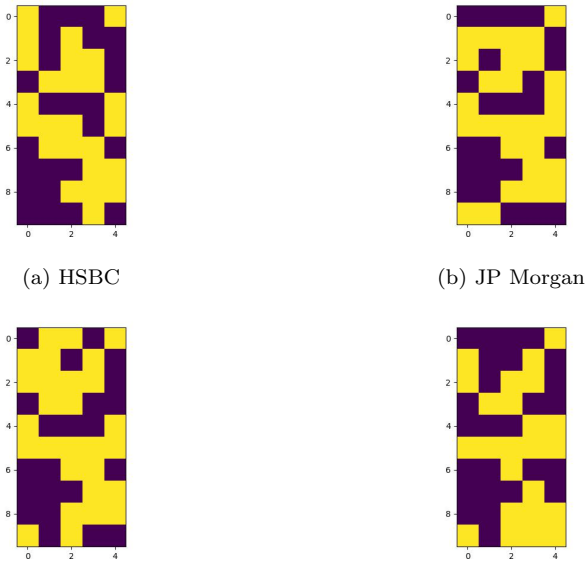
Here are the stock id's of the firms used in the analysis,

'HSBC', 'JPM', 'BAC', 'WFC', 'LYG', 'SPGI', 'MS', 'AXP', 'TD', 'GS', 'HDB', 'BCS', 'MUFG', 'SCHW', 'MMC', 'BLK', 'PGR', 'CB', 'UBS', 'C', 'IBN', 'CME', 'BX', 'AON', 'SMFG', 'ICE', 'MCO', 'SAN', 'ITUB', 'KKR', 'AJG', 'USB', 'BNS', 'PFH', 'BBVA', 'ING', 'APO', 'PNC', 'AFL', 'AIG', 'MET', 'BSBR', 'MFG', 'COF', 'TFC', 'TRV', 'BK', 'MFC', 'ALL', 'AMP', 'BBD', 'PRU', 'ARES', 'CSGP', 'ACGL', 'PUK', 'NDAQ', 'NU', 'BBDO', 'SLF', 'HBANM', 'WTW', 'HBANP', 'DB', 'EFX', 'COIN', 'CBRE', 'HIG', 'IX', 'NWG', 'TW', 'TROW', 'DFS', 'RJF', 'STT', 'BRO', 'FCNCA', 'CBOE', 'MTB', 'INVH', 'OWL', 'BEKE', 'ROL', 'MKL', 'WRB', 'LPLA', 'FITBI', 'FITB', 'RYAN', 'EG', 'PFG', 'FITBP', 'ASBA', 'RKT', 'KB', 'CINF', 'SLMBP', 'HBAN', 'L'

This data contained various information for a trading day like opening price, closing price, high, low, traded volume etc. To limit the scope of the analysis, we were only interested in the closing price opening price difference. Hence a finer dataset was made which had only one point for a trading day, 1 if closing price was greater than the opening price else 0. This indicator would roughly tell about the stock volatility trend for that day.

Here are some images of 50 day (10 weeks) images of stock trends for some well known firms from the finance sector. The representation with five columns were semantically and visually more useful as any week has only five trading days at max, from Monday to Friday. Yellow being a positive trend and purple being a negative trend.

The tool needed the dataset to be present in one large csv file with the feature names as headers, and it also required the list of quantized/class type features to be mentioned in a separate file with .catcol extension. The history size or the number of days of stock history can be varied according to the user's choice, but the author found that 50 days was apt. Of these 50 trends that we shortlist, the 50th day is the target prediction and the remaining 49 days are



(a) HSBC

(b) JP Morgan

(c) Barclays

(d) Wells Fargo

Figure 1: Binarized stock trends

the past history used for prediction. This brings to an end the data processing stage.

4.2 Model Architecture

It was found that a Feed Forward Neural Network (FFNN) was sufficient for the analysis, and there was no need for more complex architectures like CNN. Intuitively this should make sense as local contexts and patterns observed don't affect the final prediction which lies at the bottom right corner of the image. Of course this could have identified structures in the data but nevertheless unimportant for the task at hand.

The neural networks were designed so that they could adjust with the size of the input, two variants were mainly studied in this project.

- Deep FFNN: with five fully connected layers of sizes n, 2n, 2n, n and 1 where n is the size of the input. ReLU activation was used for each layer with a sigmoid activation at the final layer which predicts the class, positive or negative trend.
- Wide FFNN: with only three layers of sizes n, 4n and 1 where n is the size of the input. ReLU activation was used for each layer with a sigmoid

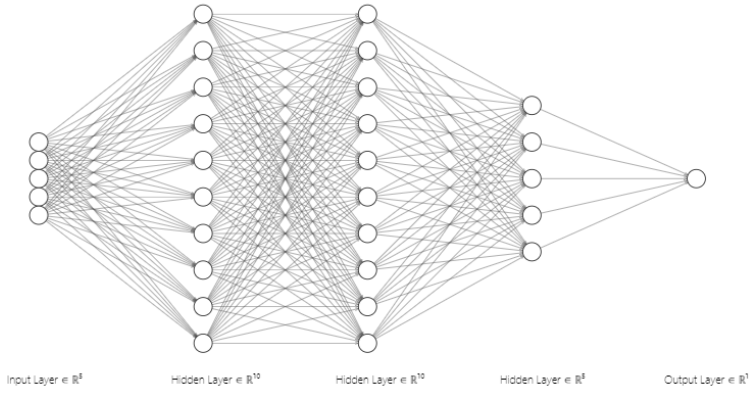


Figure 2: Example network with input size 5

activation at the final layer which predicts the class, positive or negative trend.

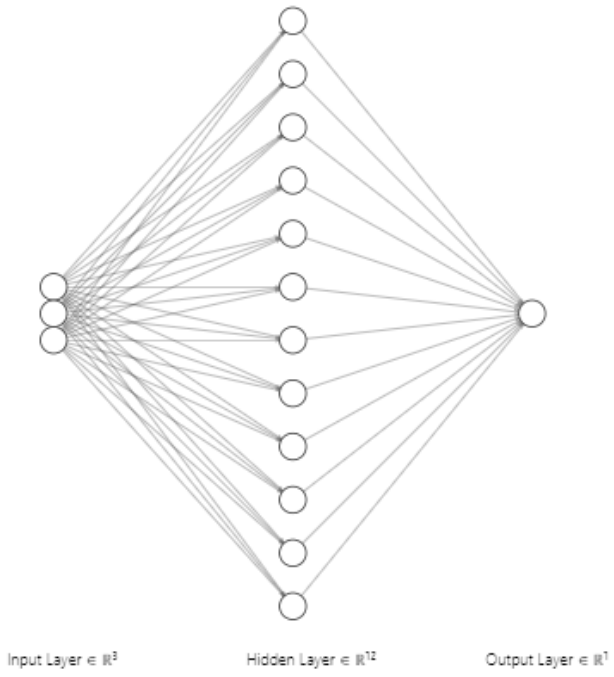


Figure 3: Example network with input size 3

4.3 Training

A training module was implemented to train the above architectures. Configuration files dictated the location for the dataset, location to store the model, number of epochs, model type, learning rate, loss function, optimizer etc. Since we needed the analysis to be particular for some day, the dataset contained 99 points for each of the firm. Experiments were done on the various choices for the length of the stock history used to predict the trend for a particular day. Due to the small size of the data, training was pretty fast. Further the hyper-parameters and the training duration was chosen so as to achieve a 100% accuracy on the training data so that the prediction is accurate for all data points. The trained model was stored as a pickle file.

4.4 MILP Encoding

The MILP encoding of the trained neural network was obtained using the configuration file and the stored pickle file of the model parameters. The encoding obtained was in the form of a text file. And this text file was taken as input to the module which performed the main minimum/minimal hitting set algorithms to find the hitting sets for the input data points. Z3/PySMT Solvers were used.

Algorithm 2: Computing a smallest size explanation

Input: $\mathcal{F}$ under $\mathcal{M}$, initial cube C, prediction $\mathcal{E}$
Output: Cardinality-minimal explanation $C_{\mathcal{M}}$

```
1  begin
2      Γ ← ∅
3      while true do
4          h ← MinimumHS(Γ)
5          if Entails(h, F → E, M) then
6              return h
7          else
8              μ ← GetAssignment()
9              C' ← PickFalseLits(C \ h, μ)
10             Γ ← Γ ∪ C'
11 end
```

Figure 4: Day Frequencies in minimum sets across firms

5 Results and Analysis

The above pipeline was run across 99 of the top financial firms/banks listed in NYSE/NASDAQ to predict the trend for 8th November 2023. A history of 49 days was used for the prediction and subsequently the minimum hitting set was calculated across all the firms.

The days which were the most important were 6th November (which was part of the minimum hitting set for 70/99 firms, 1st November (68/99) firms and 30th October (54/99) firms.

One can notice that the days 6th November and 1st November were a major deciding factor for the stock trend observed on 8th November for the financial firms.

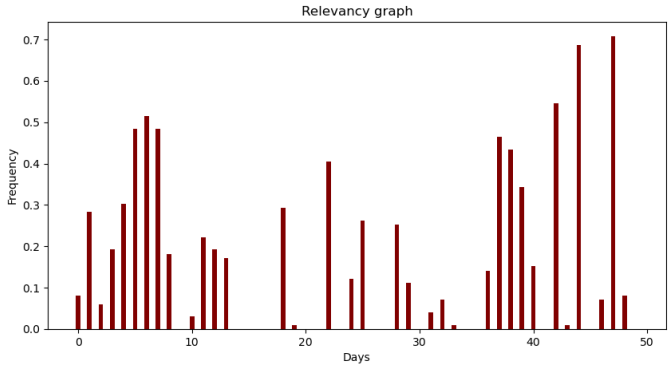


Figure 5: Day Frequencies in minimum sets across firms

One could have assumed that only the very recent past (say the past week) were sufficient to predict today's trend but the graph says otherwise. Points as far as a month ago were relevant for predicting the stock trend.

Surprisingly 7th November was not so relevant as expected. In fact, 7th November was part of the minimum hitting set of only 8 out of the 99 firms studied.

One can hypothesise that an important event might have occurred on 6th November relevant to the finance sector. Or rather an important event occurred on 3rd November (as 6th was a Monday) whose decision was portrayed by the stock market on the 6th.

Indeed, a spike can be observed on 3rd November and an observable change on the graph slope can be seen on 1st November. A talented eye can deduce

Figure 6: NYSE Finance Sector Index

those events particular to these days responsible for having such an impact on the finance sector.

The minimum set for Arch Capital is [1, 7, 22, 42, 44, 47] The minimum set for AFLAC Inc. is [7, 37, 39, 42, 44, 47] The minimum set for American International is [3, 6, 11, 13, 18, 22, 28, 37, 39, 42, 44] The length of the minimal set was 8 on average (8.4) ranging from 5 to 16, out of a total of 49 input features.

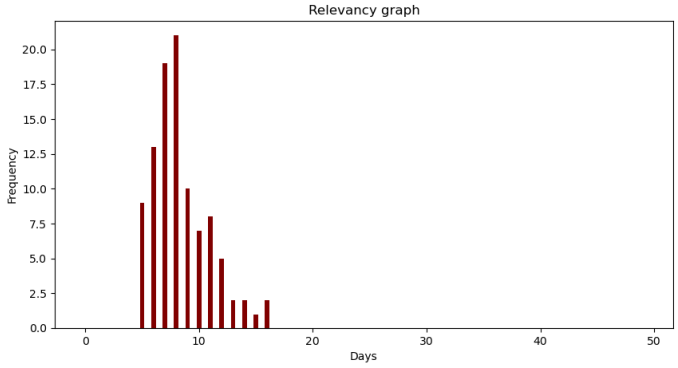


Figure 7: Minimum Hitting Set sizes

Firms like American Express, ICICI Bank, Mizuho Bank and other local banks had the lowest size minimum hitting set. One can hypothesize that local banks are affected by much lower events in general than International banks. Conglomerate like BlackRock had a small size minimum hitting set as well.

Barclays has min hit size 8, Deutsche Bank has 14.

6 Bibliography

- https://github.com/alexeyignatiev/xreason
- Emails with Prof. Alexey Ignatiev and Nina Narodytska
- https://ojs.aaai.org/index.php/AAAI/article/view/3964/3842
- https://pytorch.org/
- https://www.nyse.com/index
- https://finance.yahoo.com/
- https://www.geeksforgeeks.org/