AIFI Final Project Abstract

Francesco Tedesco

1 Introduction

This project is focused on evaluating the effectiveness of specific features, known as topological features, in predicting stock market trends over time. We aim to assess how well these features perform in the context of time series forecasting. To ensure an adequate amount of data, we will be selecting daily stock prices for the top-5 market cap companies. So, the tickers selected are: MSFT, AAPL, NVDA, AMZN and GOOGL. The primary objective is to determine if incorporating these topological features into time series models can provide a reliable method for predicting stock prices. As part of our analysis, we have tested two models so far: XGBoost and LSTM (Long Short-Term Memory).

2 Stationarity

In order to assist our model in prediction, we first need to determine whether the data is stationary. If it is not, necessary manipulations will be performed. When plotting the adjusted closing prices for the selected companies, the resulting figure is shown below:

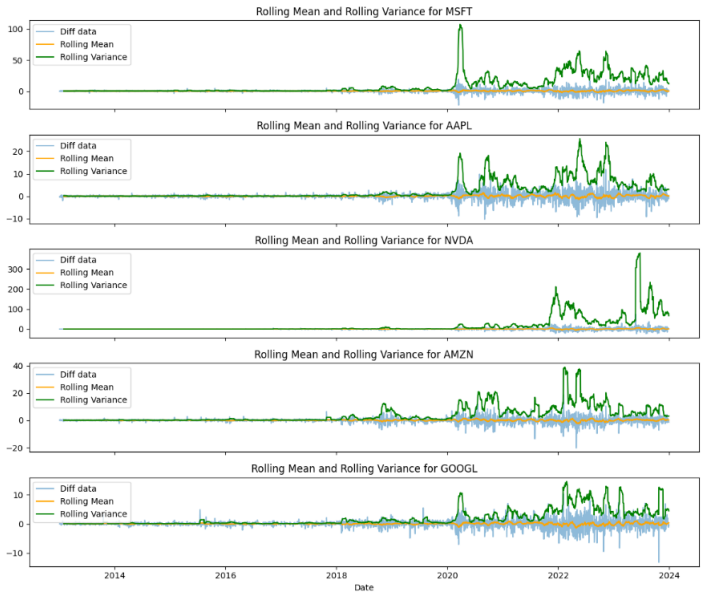


Fig. 1: Figure displaying the closing prices of the selected tickers with rolling mean and rolling variance, using a window size of 20.

As observed, there are common patterns in all the time series:

- The variance increases, indicating that it is not constant. Therefore, the natural logarithm of the closing price will be taken.
- It is evident that the mean is not constant as well. Once the logarithmic transformation is applied, differencing will be performed in an attempt to achieve stationary data.

Now if we perform the Augmented Dickey–Fuller test and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests with a significance level of 0.05, for the correspoinding hypotesis tests we get that the series is stationary. As we can see by the following images both mean and variance tend to be constant.

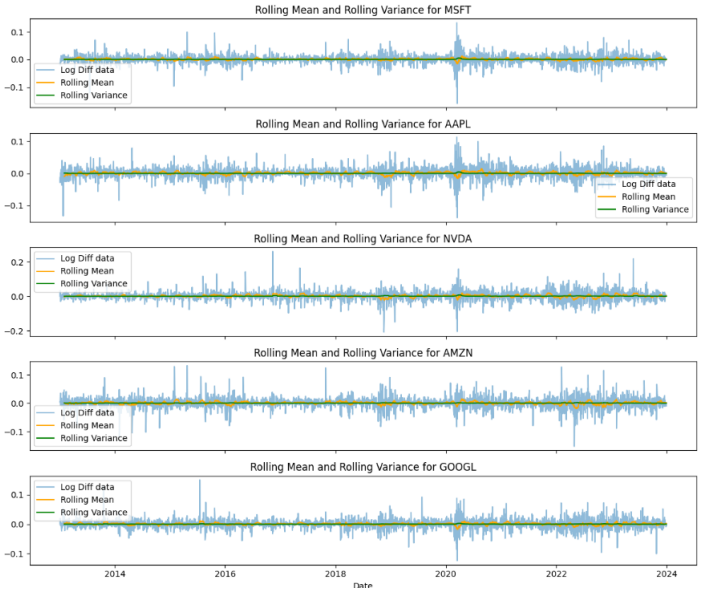


Fig. 2: Figure displaying the closing prices of the selected tickers with rolling mean and rolling variance, using a window size of 20.

3 Feature Engineering

In this section, we will discuss the process followed to build all the features, the topological and non topological.

3.1 Non topological features Features

The non topological features are the following:

1. 5 consecutive lagged features for the target (Logarithm of the Adjusted closing price differenced).
2. Simple moving averages for window sizes 5, 10 and 20.
3. Exponential moving averages for 0.2, 0.6 and 0.8 alphas.
4. The volatility indicator computed as the rolling standard deviation of window sizes of 5, 10, 15 and 20.
5. In order to inform about possible seasonality patterns in the data, the Day, Month and Quarter were encoded using the cyclical encoding technique.

3.2 Topological Features

The topological features are obtained using the giotto-tda library, and the the process to obtain them is the following:

1. **SlidingWindow**: We select a subset of the time series which contains all the features created previously. The subset sizes (window) are 5 and 10.
2. **PearsonDissimilarity**: For each window, which has n features, we get the n x n Dissimilarity matrix which is 1 - Pearson correlation matrix.

3. **VietorisRipsPersistence**: Once we have this n-matrix, which can represent an n dimensional space, we can apply the VietorisRipsPersistence which first create the connections between points at a certain distance, then the persistence is computed to get the information on how the connections created remain stable across different dimensions. (homology dimeniosn)
4. **Amplitude**: Finally, for each dimension, we get how 'strong' are the persitence in the corresponding dimensions.

In this project, we first tried to reach features up to the 3rd homology dimension:

1. Connected Components (0D Homology)
2. Loops or 1D Voids (1D Homology)
3. Voids or Cavities (2D Homology)
4. Enclosed Volumes (3D Homology)

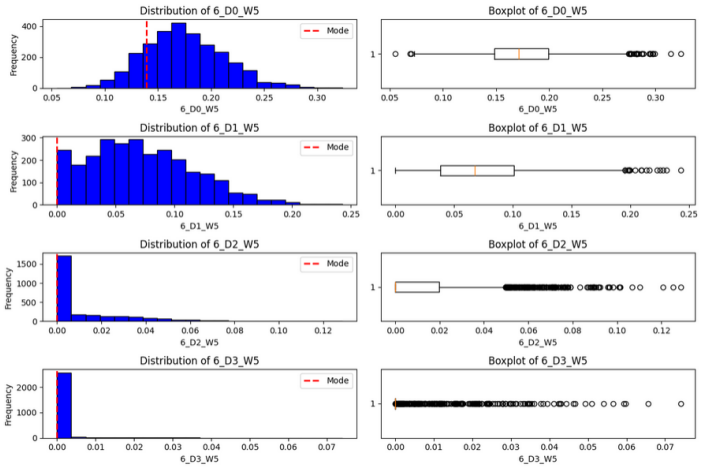


Fig. 3: Figure displaying the amplitude histogram and box plot for window size = 5

In the provided image, it's evident that dimensions 2 and 3 fail to capture meaningful insights about the topological patterns of the data. As this behavior persists in window size 10, only homology dimensions 0 and 1 will be considered.

Having selected the homology dimensions, we can create subsets of features by categorizing the initially created non-topological features into three groups:

1. **Lagged features**
2. **Mean features**
3. **Volatility features**

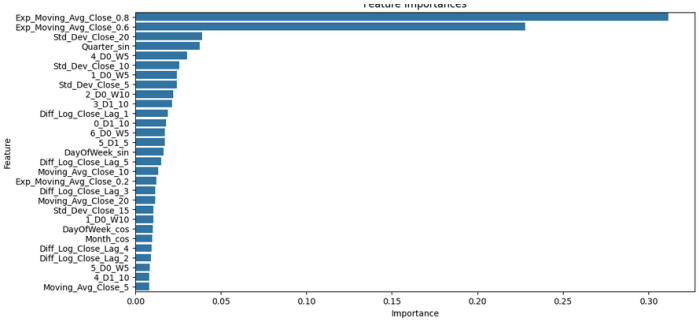
We then consider all possible combinations of these three types of features. For instance, we create a topological feature based only on lagged features, another for lagged-volatility, and so on.

4 Training Process

In this section will be discussed the feature importance of different features created, across XGBoost and LSTM models. Of the 10 years selected initially, 9 will be for training and the remaining one for testing out of sample by recursive prediction. In the 9 first years, during the hyper parameter search / tuning process, will be done over an expanding window in order to try to reduce the variance of the MSE score on train. NOTE: for the moment, only the AAPL stock is analyzed:

4.1 XGBoost

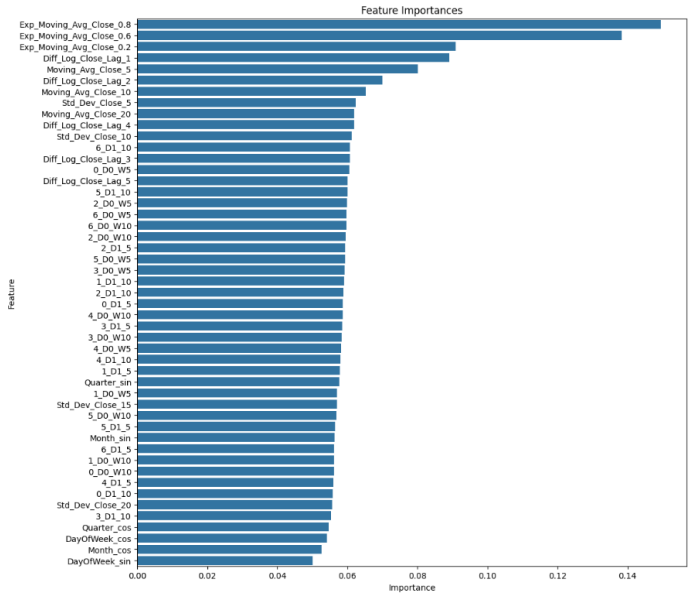
Once a randomized search for the hyper parameters is done, we get the following feature importance's (those ones different than 0):



Looking at the image, we get:

1. The exponential moving averages with alphas 0.6 and 0.8 are by far the most important.
2. The topological features with small window size reach a better feature importance. (4_D0_W5 and 4_D0_W5)
3. We can also see that 4_D0_W5 which combines volatility and lagged features reach a higher importance than most of its initial features.

4.2 LSTM



Here, if we select the first layer of neurons, we can get an idea of the feature importance, though the last image we can achieve pretty similar conclusions as in xgb model.

NOTE: more exhaustive analysis will be done before the presentation, and also more tickers will be analyzed.