Using Machine Learning to Analyse News for Movement in Stock Prices
Finsearch 2022

Sahil Gawade	Yash Salunkhe	Shlesh Gholap	Aryan Goyal
21b030013	210020156	210070080	21d180006

Auguest 28 2022

Contents

1 Motivation behind the project

It is well known that the stock market is erratic, dynamic, and nonlinear. Because of several (macro and micro) variables, including politics, international economic conditions, unforeseen occurrences, a company's financial performance, and others, it is very difficult to anticipate stock prices accurately.
However, with the introduction of Machine Learning and its strong algorithms, the most recent market research and Stock Market Prediction advancements have begun to include such approaches in analyzing stock market data.
There is a lot of data to look for patterns in as a result of all of this. In order to identify stock market patterns, financial analysts, academics, and data scientists continue to investigate analytics tools.
Our team is interested in studying and exploring the technological elements of finance. In the future, there will undoubtedly be a lot of interest in the idea that the stock and cryptocurrency markets may be anticipated to some extent. In the present day, financial professionals play a significant part in determining the global economy's future course, thus we feel this undertaking can teach us a lot about the markets of the future.

2 Time Series Forecasting

2.1 What is Time Series Forecasting?

Forecasting or predicting the future value over a period of time is known as time series forecasting. It involves creating models based on historical data and using them to draw conclusions and direct future tactical decisions. It's not always an accurate prediction, and forecast probabilities might vary greatly—especially when dealing with the variables that frequently fluctuate in time series data and with outside influences.

The responsibility is on analysts to understand the boundaries of analysis and what their models can enable because there really isn't a clear-cut set of guidelines for when you should or shouldn't utilise forecasting. Every model won't fit every set of data or provide a response to every query. When data teams have the necessary data and forecasting capabilities to tackle a problem statement, they should use time series forecasting.

On the basis of the past, the future is predicted or estimated. A time order dependence between observations is added by time series. This dependency of functions is a source of more knowledge as well as a limitation.

2.2 Forecasting considerations

The quantity of data available should be taken into account first because having more observational points would improve your analysis. This applies to all forms of analysis without exception, and time series analysis forecasting is no different. The amount of data is crucial to forecasting, perhaps even more so than other techniques. It is directly based on historical and recent data. Your predicting will be less accurate the less data you have to extrapolate from.

Your forecast's time frame is also important. With fewer variables, a shorter time horizon is significantly simpler to forecast than a longer one. The variables become more unpredictable as you travel farther out. Alternately, if you change your time horizons, having fewer data can occasionally still work with predicting. You can produce a short-term projection if you have a lot of short-term data but little long-term recorded data.

Forecast temporal frequency: Forecasts can frequently be performed at lower or higher frequencies, enabling the use of data up- and down-sampling.

2.3 Training different models on any particular index

The ARIMA(Autoregressive integrated moving average) model is a well-liked and widely utilized technique for time series forecasting. It is one of the most widely used models for predicting data from linear time series.

Since this model is well-known to be reliable, effective, and has a significant potential for short-term share market prediction, it has been employed extensively in the fields of finance and economics. The two most popular methods for predicting time series are exponential smoothing and ARIMA models, both of which offer complementary approaches to the issue.

While ARIMA models seek to characterise the auto-correlation in the data, exponential smoothing methods are based on a description of the trend and seasonality in the data. ARIMA model has been used extensively in the field of finance and economics as it is known to be efficient and has a strong potential for short-term share market prediction.

The model we used was the LTSM(Long short-term memory) Model on the Nifty50.

3 Our Model

The following features were implemented in our model(demonstrated in the video submission). We ran the model on the Nifty50 index for this project.

3.1 Sentiment analysis

Sentiment analysis is a natural language processing method for determining the positivity, neutrality, and negativity of data. Using NLP, computers can decipher the meanings hidden in text, pictures, and other types of data.

In this situation, sentiment analysis can be used to examine news headlines concerning a specific stock.
This allows you to infer whether a stock's price is moving in a positive or negative manner.

3.2 LSTM

We used Long Short Term Memory Networks, most typically called as "LSTMs," are a unique class of RNN that can identify long-term dependencies.
RNN stands for Recurrent Neural Networks. Humans do not constantly have to rethink their ideas. Each word in this report is understood in light of the ones that came before it. You don't start again from scratch and discard everything. Your ideas are constant. This appears to be a significant limitation of conventional neural networks that they are unable to perform.
This problem is addressed by recurrent neural networks. They are networks that contain loops, which enable the retention of information.

LSTMs were designed to overcome the limitations of RNNs such as:

- Gradient vanishing and exploding: In RNNs exploding gradients happen when trying to learn long-time dependencies, because retaining information for long time requires oscillator regimes and these are prone to exploding gradients.
- Complex training: RNNs can be very complex to train
- Difficulty to process very long sequences.

Remembering information for long periods of time is intrinsic to LSTM. They don't struggle to learn; rather, remembering information for extended periods of time is basically their default configuration.
All recurrent neural networks have the shape of a series of neural network modules that repeat.

4 Case Study

As an exercise and example, we will try to predict stock prices of RELIANCE by running to slightly different LSTM models on the stock price data available online. The data available gives various information of stock price of RELIANCE of every day when the stock market was open(i.e. all days except Saturdays, Sundays, and National Holidays), from January 2000 to April 2021.

4.1 Method-1

The most crucial part when it comes to apply LSTM models to data is timestamps and features selection. In this method we select timestamps as 240 days (roughly 1 yr) and a multi-feature setting consisting of Opening price, Closing price, High, Low, Last and VWAP(volume-weighted average price). When it came to modify the NNs, we used stacked LSTM.

4.2 Method-2

We define the adjusted closing price and opening price of the stock at time τ by cp_τ and op_τ respectively. Given a prediction day $t := \tau$, we have the following inputs and prediction tasks: Input: We have the historical opening prices, op_τ, $t \in (0, 1, ..., \tau - 1, \tau)$, (including the prediction day's opening price op_τ as well as the historical adjusted closing prices cp_τ, $t \in (0, 1, ..., \tau - 1)$, (excluding the prediction day's closing price, cp_τ). The feature set we provide to the random forest comprises of the following three signals:

1. Intraday returns: $ir_{\tau,m} := \frac{cp_{\tau-m}}{op_{\tau-m}} - 1$
2. Returns with respect to last closing price: $cr_{\tau,m} := \frac{cp_{\tau-1}}{cp_{\tau-m-1}} - 1$
3. Returns with respect to opening price: $or_{\tau,m} := \frac{op_\tau}{cp_{\tau-m}} - 1$

Then we apply the Robust Scaler standardization

$$\tilde{f}_{t,1} := \frac{f_{t,1} - Q_2(f_{.,1})}{Q_3(f_{.,1}) - Q_1(f_{.,1})},$$

where $Q_1(f_{.,1})$, $Q_2(f_{.,1})$ and $Q_3(f_{.,1})$ are the first, second, and third quartile of $f_{.,1}$, for each feature $f_{.,1} \in ir_{.,1}, cr_{,1}, or_{,1}$ in the respective training period. Thus, we obtain three-dimensional standardized features $\tilde{F}_{t-i,1} := (\tilde{ir}_{t-i,1}, \tilde{cr}_{t-i,1}, \tilde{or}_{t-i,1})$

5 References

- https://www.tableau.com/learn/articles/time-series-forecasting
- https://www.influxdata.com/time-series-forecasting-methods/
- https://www.kdnuggets.com/2020/01/stock-market-forecasting-time-series-analysis.html
- https://towardsdatascience.com/exploring-the-lstm-neural-network-model-for-time-series-8b7685aa8cf
- https://randerson112358.medium.com/stock-market-sentiment-analysis-using-python-machine-learning-5b644f151a3e