Analysis of Stock Market using Machine Learning

June 17, 2024

Abstract

This report presents a comprehensive analysis of stock market prediction using a fusion of technical, fundamental and sentiment analysis through machine learning techniques. The project is divided into three main parts: technical, fundamental and Sentiment analysis. In the technical analysis segment, various deep learning models such as Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) networks were employed to predict stock price movements based on historical stock data. In the fundamental analysis segment, data from Alpha Vantage was utilized to extract financial statements and calculate essential fundamental ratios. These ratios were then used to train machine learning models for predicting stock price movements. The results obtained from both analyses were promising, demonstrating the potential of integrating technical and fundamental analysis in stock market prediction.

Contents

1 Introduction

The stock market is a dynamic and complex system influenced by a multitude of factors, including economic indicators, investor sentiment, geopolitical events, and company-specific fundamentals. In your project, we explored three critical aspects of stock market analysis: technical analysis, fundamental analysis and Sentiment Analysis. Let's delve into each of these components:

Technical Analysis

Technical analysis involves studying historical price and volume data to identify patterns, trends, and potential future price movements.Leveraging deep learning models like Artificial Neural Networks (ANNs) and Long Short-Term Memory (LSTM) networks, we successfully predicted price movements based on historical patterns.

Fundamental Analysis

Fundamental analysis, on the other hand, focuses on evaluating a company's intrinsic value by examining its financial statements and other relevant information. By analyzing income statements, cash flow statements, and balance sheets, we calculated essential fundamental ratios. These ratios provide insights into a company's financial health, profitability, and growth prospects.

Sentiment Analysis

The objective of this project is to utilize Transformers-based models for summarizing financial news articles and analyzing sentiment. These models are powerful tools in natural language processing, capable of summarizing lengthy texts and extracting sentiment from them.

2 Technical Analysis:

In the technical analysis segment of the project, historical stock data including opening price, closing price, high, low, and volume was collected. Various machine learning models were implemented to predict stock price movements based on this data. ANN and LSTM networks were chosen for their ability to capture complex patterns in sequential data.

2.1 Data Collection and Preprocessing:

Historical stock data was obtained from Yahoo finance and preprocessed to remove any null values in the dataset. This historical dataset had data about Open, High, Low, Close and Volume of a particular stock.

Using the python library named pandas technical analysis we calculated various technical indicators belonging to different criteria. Like for Trend indicators I used moving averages and macd, for Momentum Indicator I used Stochastic Oscillator, Commodity Channel Index and RSI, for Volatility Indicator I used Bollinger band and for Volume based indicators I used Chaikin Oscillator and VWAP.

Since indicators like moving averages, VWAP and Bollinger band are highly correlated with the closing price of stock. Instead of using their actually value we decided to use their slope.

2.2 Model Selection and Training:

2.2.1 Traditional Machine Learning Models

In addition to the deep learning models (ANN and LSTM), I explored traditional machine learning algorithms. These models are computationally less intensive and often used for various tasks, but their performance was very poor as compared to deep learning models. Here are the models I tried:

Perceptron: 54%
Logistic Regression: 59%
K-Nearest Neighbour: 53%
MLP (Multi Layer Perceptron): 57%
This results might look good if you just look at the accuracy but the problem here was that model was only making one type of prediction and the other one's precision, recall and F1-Score all were zero. Which is not good for any model. To solve this issue we decided to use more complex model i.e. Deep Learning model wihch can capture more complex relationships.

2.2.2 Deep Learning Models

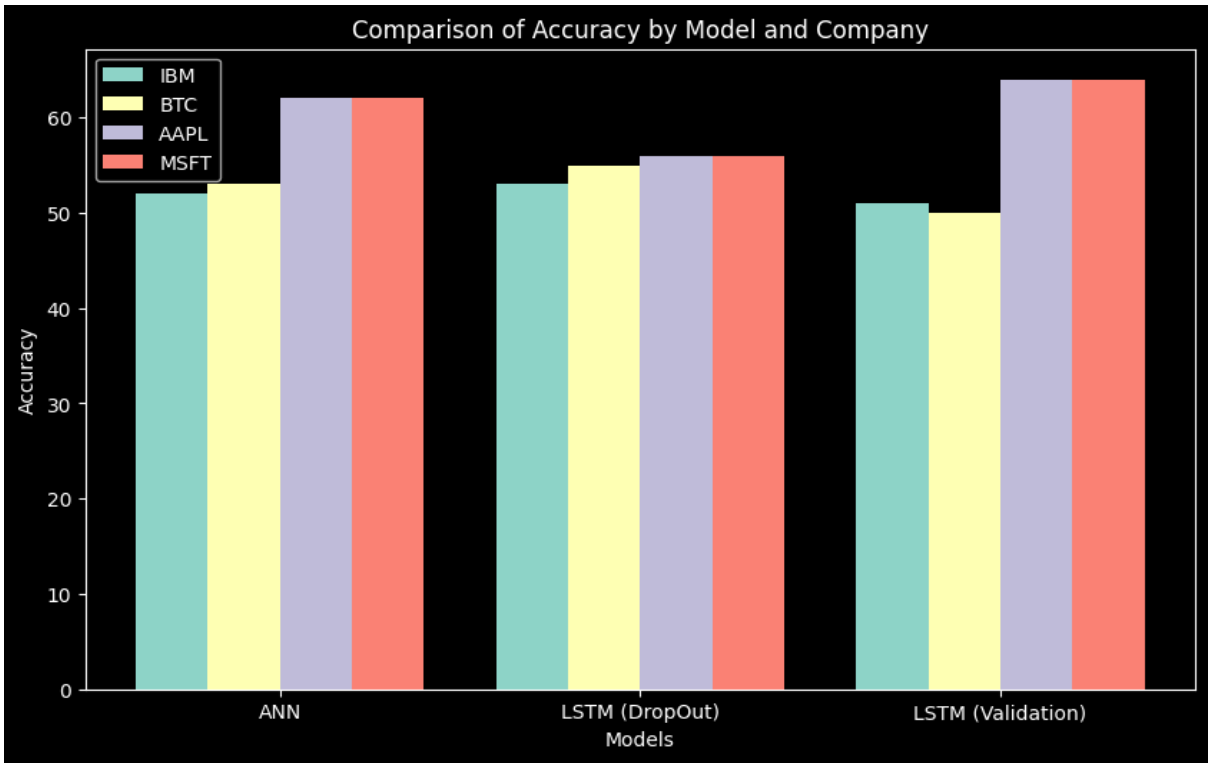
The input architecture for both the model is same, along with passing the current candle information I am also passing information of past candle (in my case I am passing 30 back candles).

- **Artificial Neural Networks (ANNs):** ANNs are powerful tools for learning complex relationships between input features and target variables.
- **Long Short-Term Memory (LSTM) networks:** LSTMs are well-suited for time series data analysis, capturing long-term dependencies within the historical price data.

To address overfitting in my deep learning models (ANN and LSTM), I employed dropout regularization. Here's how I incorporated it:

- **Dropout Regularization:** During training, randomly set a fraction of input units to zero at each update, preventing the network from relying too heavily on specific features. I applied drop out regulation between layers.
  This drastically improved the performance of Model before this model was overfitting and not performing well on training data but after this the accuracy, I get at training time and testing time is almost similar.
- Another way I used to tackle overfitting was saving the best model after each epoch and stopping the training when our training accuracy increases but our validation accuracy decreases or saturates.

2.3 Result and Finding:



Comments: All the three models performed well from accuracy perspective. But ANN and LSTM

model with Validation regulation were more prone to getting overfitted or underfitted because there accuracy may seem to be high for AAPL and MSFT but most of the prediction they make in test dataset were one sided i.e. only predicting one type of signal.
Whereas LSTM model with Drop Out regulation seem to give stable prediction for each stock.
So, from my side I think that best model among all three is LSTM with drop out regulation.
Clearly, if we increase the numbers of deep layers in the model we might also get stable result from ANN and LSTM (Validation) but then training will take a lot of time.

3 Fundamental Analysis

In the fundamental analysis segment of the project, financial data including income statements, balance sheets, and cash flow statements was obtained from Alpha Vantage. Fundamental ratios such as price-to-earnings ratio (P/E), price-to-book ratio (P/B), and debt-to-equity ratio (D/E) were calculated from this data.

3.1 Data Collection and Preprocessing

Financial data was collected for a diverse set of companies across the same industries (in our case we focused on the Tech industry).
After collecting the data, the hardest part came Missing values and inconsistencies. There were a lot of missing values in the data for different companies. Missing values and inconsistencies in the data were addressed through data imputation and cleaning techniques.
By combining all the three income statements, balance sheet and cash flow statements I calculated some important fundamentals ratios like:

- **Current Ratio:**
  Formula: Current Ratio $= \frac{\text{Total Current Assets}}{\text{Total Current Liabilities}}$
  Description: Measures a company's ability to meet its short-term obligations (due within a year) using its current assets. A higher ratio indicates better short-term liquidity.

- **Quick Ratio (Acid-Test Ratio):**
  Formula: Quick Ratio $= \frac{\text{Total Current Assets} - \text{Inventory} - \text{Other Current Assets}}{\text{Total Current Liabilities}}$
  Description: Similar to current ratio, but excludes less liquid assets like inventory to assess a company's most liquid assets' ability to cover short-term debt.

- **Debt to Equity Ratio:**
  Formula: Debt to Equity Ratio $= \frac{\text{Total Liabilities}}{\text{Total Shareholder Equity}}$
  Description: Indicates the proportion of debt financing compared to equity financing used by a company. A lower ratio suggests a more financially stable company with less reliance on debt.

- **Debt to Asset Ratio:**
  Formula: Debt to Asset Ratio $= \frac{\text{Total Liabilities}}{\text{Total Assets}}$
  Description: Similar to debt-to-equity, but expresses the proportion of debt financing relative to the company's total assets. A lower ratio signifies a more financially sound company with a smaller debt burden.

- **Interest Coverage Ratio:**
  Formula: Interest Coverage Ratio $= \frac{\text{EBIT + Income Tax Expense + Net Income}}{\text{Interest Expense}}$
  Description: Measures a company's ability to service its debt by covering interest payments. A higher ratio indicates a stronger capacity to meet interest obligations.

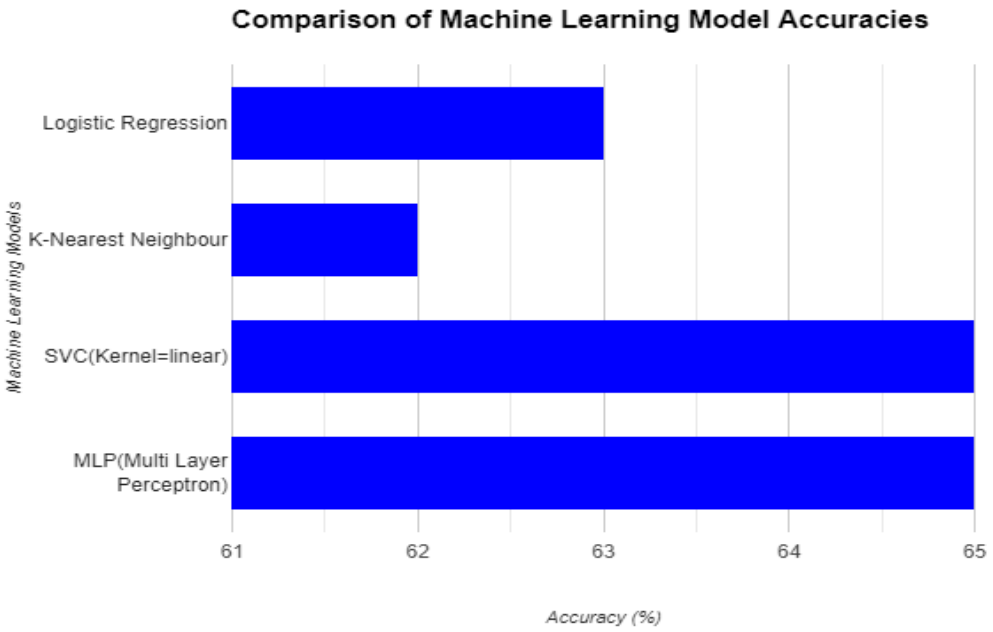
- **Free Cash Flow Conversion Ratio:**
  Formula: Free Cash Flow Conversion Ratio $= \frac{\text{Operating Cash Flow} - \text{Capital Expenditures}}{\text{EBITDA}}$
  Description: Measures the conversion rate of a company's operating cash flow into free cash flow available for debt repayment, dividends, or stock buybacks. A higher ratio indicates better cash flow generation and potential for future investments.

- **Return on Stockholder Equity (ROE):**
  Formula: Return on Stockholder Equity $= \frac{\text{Net Income}}{\text{Total Shareholder Equity}}$

Description: Measures the return generated on the investment of shareholders. A higher ROE signifies a company's efficiency in using shareholder equity to generate profits.

3.2 Model Training and Evaluation:

In fundamental analysis, unlike technical analysis, traditional machine learning model also performed well.
Firstly, I took stock belonging from same industry and used them train to my model, excluding one company which act as our test subject.
The result obtained were very good for most of the models.



4 Sentiment Analysis

4.1 Brief Description of the Code:

The provided Python code snippet imports necessary libraries such as Transformers and BeautifulSoup. It sets up summarization and sentiment analysis models and then proceeds to scrape financial news articles from Yahoo Finance. Specifically, the code accesses the Yahoo Finance website and retrieves news articles related to HDFC Bank (HDB) by using the appropriate URL. Once the articles are fetched, they are processed for summarization using the Pegasus model. Subsequently, sentiment analysis is performed on the summarized content using a sentiment analysis pipeline.

4.2 Summarization Process (Pegasus):

4.2.1 Summarization Process (Pegasus):

- The input text is tokenized using the tokenizer associated with the summarization model, specifically "human-centered-summarization/financial-summarization-pegasus".
- Tokenization breaks down the text into subword units and maps each token to its corresponding token ID.

4.2.2 Embedding Layer:

- The tokenized input is passed through an embedding layer, which converts each token into a dense vector representation.

- This dense representation captures the semantic meaning of each token based on its context in the input text.

4.2.3 Encoder Transformer Layers:

- The embedded token representations undergo processing through multiple encoder transformer layers.
- These layers capture contextual information and relationships within the input text, leveraging self-attention mechanisms to attend to relevant parts of the text.

4.2.4 Decoder Transformer Layers:

- The model utilizes a decoder layer to generate the summary, leveraging a variant of transformer architecture optimized for sequence-to-sequence tasks like summarization.
- These decoder transformer layers decode the encoded representations of the input text and generate the summary output by attending to the relevant information.

4.2.5 Output Generation:

- The decoder generates the summary by attending to the encoded representations of the input text, producing a concise summary that captures the key information.

4.2.6 Post-processing:

- The generated summary may undergo post-processing steps such as length restriction, removing redundant information, or refining the summary for coherence and readability.

4.2.7 Output:

- The summarized text, generated using the "human-centered-summarization/financial-summarization-pegasus" model, is returned as the output of the summarization process.

4.3 Sentiment Classification Process (Pipeline Model):

4.3.1 Input Processing:

The input text for sentiment analysis is tokenized using the tokenizer associated with the sentiment analysis pipeline model.

4.3.2 Embedding Layer:

The tokenized input is passed through an embedding layer, converting each token into a dense vector representation capturing its semantic meaning.

4.3.3 Transformer Layers:

The embedded token representations are processed through multiple transformer layers within the sentiment analysis pipeline model.

4.3.4 Classification Head:

The sentiment analysis pipeline model includes a classification head, which takes the output from the transformer layers and computes the probability distribution over predefined sentiment classes (e.g., 'POSITIVE', 'NEGATIVE').

4.3.5 Prediction:

The sentiment analysis pipeline model predicts the sentiment label for the input text by selecting the class with the highest probability from the output layer's probability distribution.

4.3.6 Confidence Score:

Additionally, the pipeline model may provide a confidence score associated with each sentiment prediction, indicating the model's confidence level in its predictions.

4.3.7 Output:

The sentiment label along with the confidence score, if available, is returned as the output of the sentiment classification process using the pipeline model.

4.4 Analysis and Conclusion:

Analysis of Results: The analysis reveals that the Pegasus model excels in generating high-quality and coherent summaries of financial news articles. However, GPT-2 and ProphetNet, while versatile, produced inadequate summaries. The sentiment analysis results provide insights into investor sentiment towards the topics covered in the articles, which can be valuable for decision-making.

Discussion of Limitations and Future Improvements: One notable limitation of the project was the inability to set specific dates for selecting news articles. This limitation could potentially impact the relevance and timeliness of the scraped articles, as financial market dynamics are subject to rapid changes. Future improvements could involve implementing a mechanism to filter articles based on publication dates, ensuring that the selected news articles are up-to-date and relevant to the analysis.

Conclusion: In conclusion, this project demonstrates the effectiveness of using Transformers-based models for summarizing financial news articles and analyzing sentiment. The Pegasus model stands out for its ability to generate concise and informative summaries,whereas t5 also does a decent job while GPT-2 and ProphetNet show potential for improvement. Further research and development in natural language processing for financial applications are essential for enhancing decision-making processes in various domains.

5 Conclusion

In conclusion, this project demonstrated the potential of integrating technical analysis, fundamental analysis, and sentiment analysis using machine learning techniques for stock market prediction. The technical analysis segment employed deep learning models like ANNs and LSTMs to successfully capture patterns in historical stock data and predict price movements. The fundamental analysis component calculated key financial ratios from company statements and used machine learning models to derive insights into a company's financial health and future prospects.
The sentiment analysis segment showcased the power of transformer-based models like Pegasus in summarizing financial news articles and analyzing sentiment, providing valuable information about investor perceptions. While limitations exist, such as the inability to filter news articles by date, the project highlights the potential of integrating these diverse analytical approaches for enhanced decision-making in the financial domain.
Overall, this project underscores the significance of leveraging machine learning and natural language processing techniques to gain a comprehensive understanding of the complex factors influencing stock market dynamics.
In future we will be trying to integrate all three Analysis signal in one to get more precise and robust signal.