Predicting Stock Market Prices Using Machine Learning and Deep Learning Models

Dr. Sumit Pundir (Associate Professor), Kartikay, Aman Prajapati, Manish Kumar Singh, Tanish Gupta

Department of Computer Science, Graphic Era Deemed To Be University (*UGC Affiliated*), Dehradun, Uttarakhand

*Abstract*—Financial markets are beneficial to individuals in various domains such as financial, corporate, businesses and banking. Today, artificial intelligence is playing a pivotal role in financial markets by using its vast set of capabilities and computing resources. This technology is widely used in financial forecasting, valuation, business analysis, resource planning, investment strategy and other business fields. Traders and investors are using machine learning models to predict trends in financial instruments. Since artificial intelligence is being extensively used today in finance, it becomes imperative to encapsulate the contemporary understandings of machine learning and deep learning. This facilitates thorough examination and comparison of various machine learning models and techniques in financial domain. This article investigates various techniques and algorithms such as Long short-term memory (LSTM), Autoregressive Moving Averages (ARIMA). Prophet (developed by Meta) and NLP based sentiment analysis model to predict the movement of stock market. The key findings and takeaways from this research paper are as follows: (a) provides an overview of financial models in machine learning and deep learning; (b) provides a general framework for cost estimation and allocation; (c) using and testing the performance of various trading strategies and combination models to predict market value and compare the results to analyze which strategy-based model delivers best performance.

# Introduction

# Algorithmic trading is a process that uses mechanical, modified premarketing techniques to make decisions representing factors such as time, price, and volume. Such businesses attempt to use the speed and computing resources of personal computers to compete with human brokers. Only one transaction is possible every five days. Program trading strengthen these opportunities using well planning, testing and execution.

# Black Box trading uses algorithms which follows established patterns and guidelines for trading. This business can generate income at a very low rate and frequently. The process of importing job reports into the program is based on time, price, value or a mathematical model. In addition to providing excellent results to investors, algorithmic trading eliminates the influence of human emotions on trading, making trading more liquid and profitable.

The USP of trading robots is that they simplify trader's job and help the trader make money quickly with minimum efforts. Moreover, the current economy is a "prerequisite" for survival in the future financial market. Market reports indicate that the world algorithmic trading market is expected to extend from US$11.1 billion in 2019 to US$18.8 billion in 2024. Therefore, the future of algorithmic trading was never here. Demands for the project were due to the lack of "simple but productive workers" who could be used by "ordinary people".

# Literature Survey

The literature review of various proposed and implemented machine learning algorithmic trading techniques has been described in this segment. It explains the research of existing systems and software for businesses with machine learning algorithms. Current machine learning algorithms include pure random forests, and probabilistic regression, genetic algorithms such as deep MLP neural networks, support vector machine regression (SVR) and random forests, and gradient boosting decision. Constraint of present systems and software are as follows.

**Linear Regression:**

Linear or Simple Regression is used in stock or financial market forecasting for predicting the succeeding value of stock returns and forecast the upcoming worth of the stock; closing price, opening price, volume etc. A model is used based on one or more features such as .. stock value. Regression modelling aims to model the relationship between dependent and independent variables. Regression models create a line of best fit that express the relationship between independence and achievement.

In this regression model the process involves, a straight line which is represented by equation (1), i.e. drawn to make sure the line intersects as high as possible in terms of the number of features present in the data set. When the values ​​of a data set are plotted on a graph, a straight line is drawn between the points so that the distance or square of the difference between each point and the straight lines is as small as possible. For each given value of x, an imaginary line is utilized to estimate and calculate the value of y. This method of predicting values is called linear regression. Various parameters like RMSE, MAE, MSE and R-squared are used to evaluate the outcomes and further verify how well the model fits the straight line.

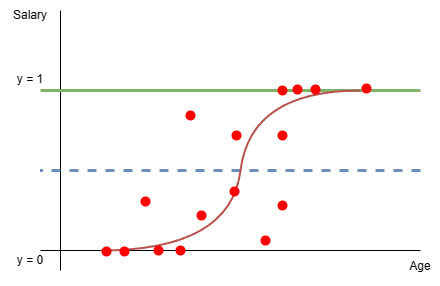
𝑂= 𝑆+𝐾 (1)

where 𝑂 is the resultant, slope is represented by S, and 𝐾 is the constant.

**Logistic Regression:**

It is a conservative method of machine learning. Logistic regression divides various independent variables into two or more specific groups using logistic curve variables and predicts the probability of product performance. To analyse product performance using logistic regression according to Equation (2):

𝑍𝑖𝑡=𝛽1+𝛽2𝐸𝑃𝑆𝑖𝑡+𝛽2𝑃𝐵𝑖𝑡+𝛽2𝑅𝑂𝐸𝑖𝑡+𝛽2𝐶𝑅𝑖𝑡+𝛽2𝐷𝐸𝑖𝑡+𝛽2𝑠𝑎𝑙𝑒𝑠𝑖𝑡+𝑉𝑖𝑡 (2) where z = log log (Pr/1-Pr) and Pr = probability of the result being positive.



**Figure 1**. Logistic Regression

**Support Vector Machine (SVM):**

Exercising SVM for stock market forecasting is believably the most important method for stock price prediction because this technique is utilized in both classification and replication algorithm. Comparison of SVM with its alternatives like “Peeling + SVM” and “CC + SVM” describes us that advanced SVM techniques can improve their predictions. Support vector machines involve supervised learning for feature classification using classifiers. Isolated characters are determined when the data at first is mapped into a large-sized feature capacity. He has established that the scattering of data points appearing in N-dimensional space and found the ideal hyperplane. All data points are categorized by their location relative to the plane. Performance of the SVM algorithm can be enhanced by calibrating parameters like regularity, gamma and kernel parameters. Sentiment analysis to analyse marketers' sentiments and directly impact business can also be achieved using SVM. It is quite suitable for high-dimensional datasets and small data.

**Long Short-Term Memory (LSTM):**

LSTM is a deep learning algorithm that use (ANN) artificial neural network. It can be considered as an improved version of recurrent neural network (RNN’s) as they provide feedback connections which are not provided by traditional feed-forward neural networks. They are excellent for tasks such as time series forecasting as they can effectively handle time series data. They can avoid the vanishing gradient problem by using memory cell and gates that can control the flow of information allowing them to selectively retain or discard features. Figure 2 shows diagrammatic representation of structure of LSTM cell. The LSTM cell comprises three gates—the input gate, the forget gate, and the output gate. Mathematically, all gates used in LSTM can be represented by equations (3)-(5).

Input gate (new information arrives in input gate):

𝑖𝑔𝑎=𝜎 (𝑊𝑖𝑝 [ℎ𝑡−1, 𝑋𝑐]+𝑏𝑖) (3)

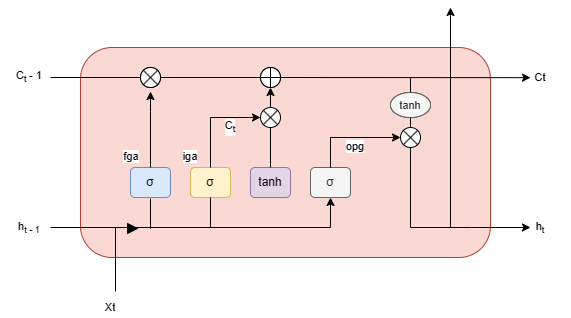
Forget gate (eliminates useless information):

𝑓𝑔𝑎=𝜎 (𝑊𝑓𝑔 [ℎ𝑡−1, 𝑋𝑐]+𝑏𝑓) (4)

Output gate (activation function for final output):

𝑂𝑝𝑔=𝜎 (𝑊𝑜𝑝 [ℎ𝑡−1, 𝑋𝑐]+𝑏𝑜) (5)

where 𝜎 is sigmoid activation function, 𝑊𝑥 here stands for the neuron gate (𝑥) weight, the result of previous LSTM block is denoted by ht-1, 𝑋𝑡 stands for the input, and 𝑏𝑥 stands for the bias.

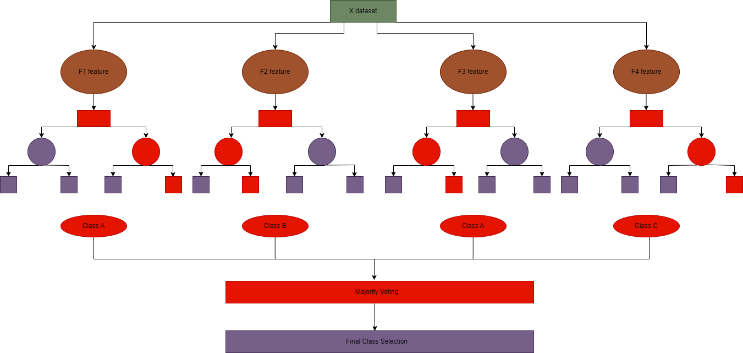


**Figure 2.** Long Short-Term Memory (LSTM)

As depicted in the diagram above the upper part of each memory unit can be connected to transmission line from the model used to process data received by previous cell to current memory cell. The mechanism to store data stream is present at each LSTM node. To provide sufficient time for training connections and allow long-term connections to form, LSTMs store errors at a constant level. Though sometimes neural networks had provided better performance in predicting the time series data for stock market, yet logistic regression models give better results than neural networks in predicting financial problems.

**Random Forest Algorithm:**

Random Forest algorithm is a supervised machine learning algorithm that combines many decision trees. This results in a model that is an aggregate of various models. This greatly reduces the bias and variance of the final model and provides better accuracy in predictions. This process also resolves the problem of overfitting. That models that are combined are generally decision trees. The trees use best-split strategy which provides maximum information gain. The random forest process is shown in Figure 3.



**Figure 3.** Random Forest

Steps for creating a Random Forest Model are given below:

Step 1: Select a random file.

Step 2: Create a base model (decision tree) based on the input N.

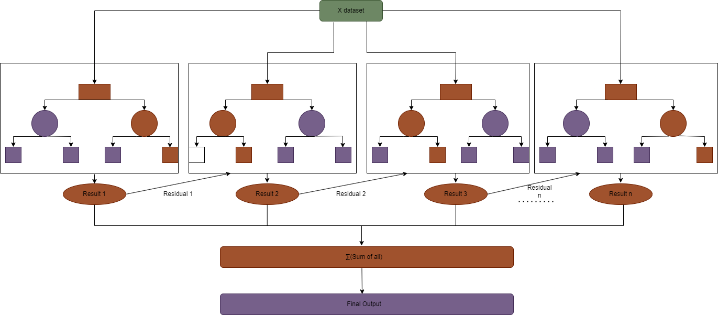
Step 3: Select decision tree to consider.

Step 4: Determine the output of each tree based on the previous step.

Random forest is an efficient algorithm while working on large data sets, but sometimes creating multiple trees reduces the performance of algorithm. Random forest can perform both regression and classification. Random forest methods can be used for a variety of other applications such as estimating the number of instructions.

**XG-Boost Regression Algorithm:**

XG-Boost stands for eXtreme Gradient Boosting It is a ensemble learning algorithm It is generally used for supervised learning tasks such as regression and classifications. It combines predictions of multiple models, often decision tress, to derive its output. This method of combining multiple learners is called ensemble learning. The models that are combined are called base models and can be based on similar or entirely different algorithms. Each model has its strengths and drawbacks. The boosting technique used in XG-Boost aims to reduce errors in previous tree, by process called gradient boosting. The result is the aggregated model with much less bias and variance than individual models. The XG-Boost process is shown in Figure 4.

****

**Figure 4.** XG-Boost Algorithm

# METHODOLOGY

*3.1 Description of Data*

The data for training the models is downloaded from Yahoo Finance. The dataset comprised of past 10 years of data of companies like Google, Apple, Tesla, Microsoft and Tata Motors. The data collected has various attributes such as High, Low, Open, Close, Adjacent close and Volume. For our application only day wise closing price is relevant.

*3.2 ARIMA (Autoregressive Integrated Moving Average)*

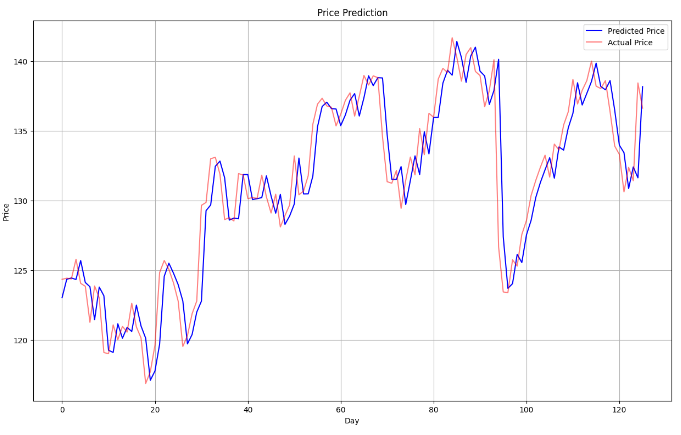
ARIMA stands for Autoregressive Integrated Moving Average. It is a statistical model used in business and statistics to measure events over time. ARIMA is a popular forecasting real-time data such as sales, prices or weather.

ARIMA model is composed of various components. The components are briefly described below.

* **Autoregression (AR):** This is a class of models that operate under premise that current values are affected or influenced by past values.
* **Moving Average (MA):**  incorporates the dependency between an observation and a residual error. It illustrates the connection between a time series and a linear combination of historical error components.
* **Integrated (I)***:*This helps to make the time series data stationary. It replaces the data values by the difference between the current and previous value.

The differenced values are denoted by ‘*d*’ in ARIMA(*p, d, q)* model.

The graph below represents the Actual Prices and Predicted Prices by the ARIMA Model.

x

**Figure 5.** ARIMA Model

Evaluation Metrics of the above model :-

Mean Squared Error (MSE): 5.3531259990

Root Mean Squared Error (RMSE): 2.3136823461

Mean Absolute Error (MAE): 1.6376395425

From predictions of the above model it’s been concluded that ARIMA modelling is generally inadequate for long-term forecasting, such as more than six months ahead, as it relies on parameters that are subject to altercations as they are affected by human thinking.

*3.3 LSTM (Long Short Term Memory)*

In the realm of machine learning, particularly within the domain of time series analysis and sequential data modeling, the Long Short-Term Memory (LSTM) network has emerged as a powerful and widely utilized architecture. LSTMs belong to the family of recurrent neural networks (RNNs), but they are especially made to get bypass the drawbacks of conventional RNNs in terms of identifying and understanding long-term relationships in sequential data.

**Model Overview:** The LSTM (Long Short-Term Memory) model presented in this research comprises multiple layers of LSTM units, each designed to capture and learn sequential dependencies in the time series data.

**Layer Configuration:**

1. **First LSTM Layer:**

**Units:** 50 LSTM units.

**Activation Function:** Rectified Linear Unit (ReLU).

**Return Sequences:** True, indicating the layer returns the full sequence of outputs.

**Input Shape:** Defined by the dimensions of the training data.

1. **Dropout Layer (Regularization):**

The function of dropout layer is to prevent overfitting. In this model a dropout layer having 20% dropout rate is used. This layer is added just after the LSTM layer.

1. **Second LSTM Layer:**

Similar configuration to the first layer with an increased dropout rate and adding more units.

1. **Third LSTM Layer:**

Increased the units and dropout rate even more.

1. **Fourth LSTM Layer:**

The final LSTM layer with highest number of units and dropout rate.

**Output Layer:**

* A Dense (fully connected) layer with one unit is added at the end of the architecture to produce a single continuous output, representing the predicted stock price.

The graph below represents the Actual Prices and the Prices Predicted by the LSTM Model.

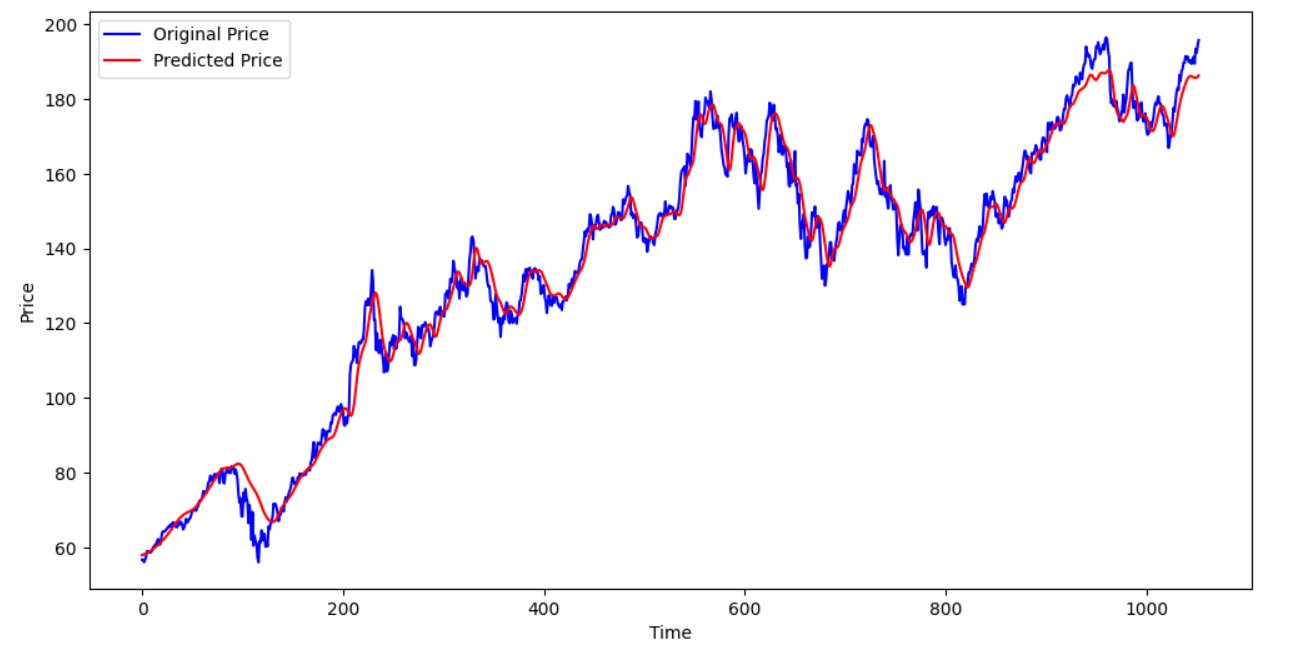
Evaluation Metrics of the LSTM Model :-

Mean Squared Error: 0.0010009753283

Root Mean Squared Error: 0.0316381941382

Mean Absolute Error: 0.0244462605189

From the results, it’s been concluded that LSTM model performs better than the ARIMA Model.



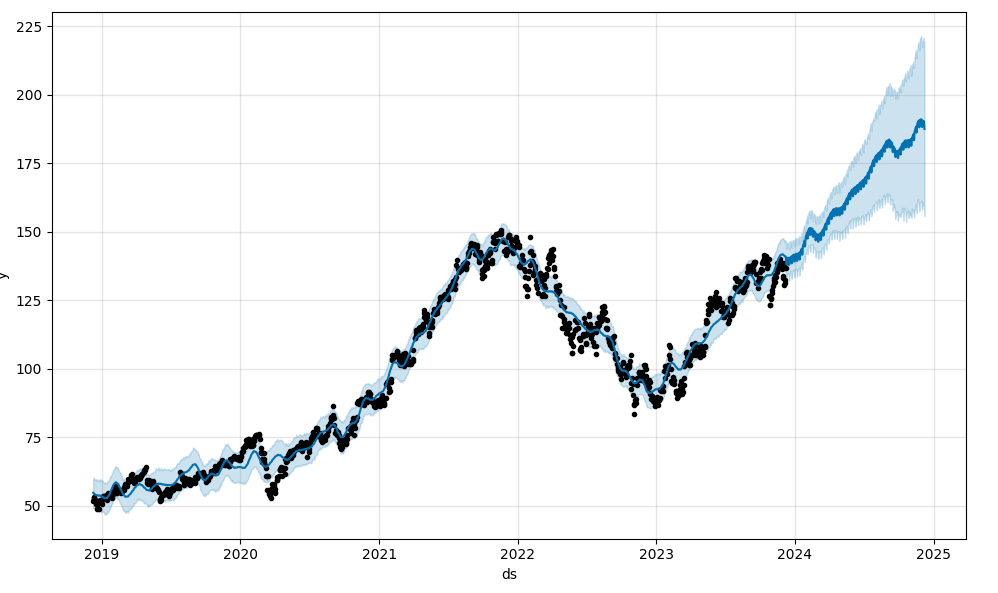
**Figure 6:** LSTM Model

*3.4 Prophet Model*

"Prophet is a method for predicting time series data using an additive model that fits non-linear trends with seasonality on a daily, weekly, and annual basis in addition to holiday effects ('Prophet,' n.d.)".

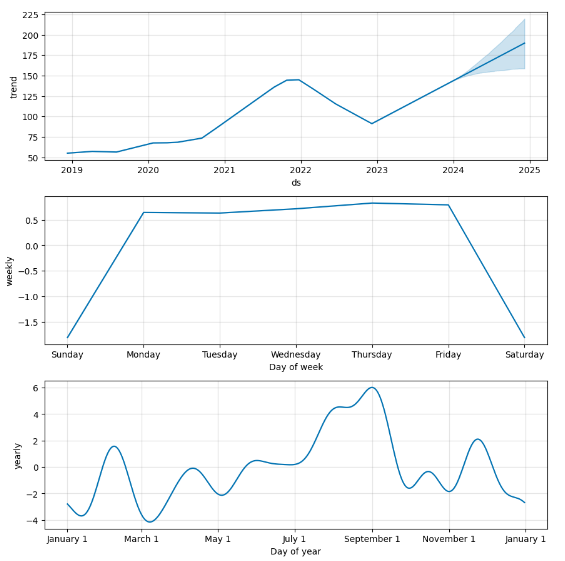
Best for time series that contain strong seasonality and historical data for many seasons. Prophet is resilient to missing data and changes and generally works well.

The chart below shows the forecasts and past prediction by the model.



**Figure 7**: Prophet Model

The figure below shows the trends in stock prices generalised by the model.

**

**Figure 8:** Prophet Model

* 1. *Sentiment Analysis using FinBert*

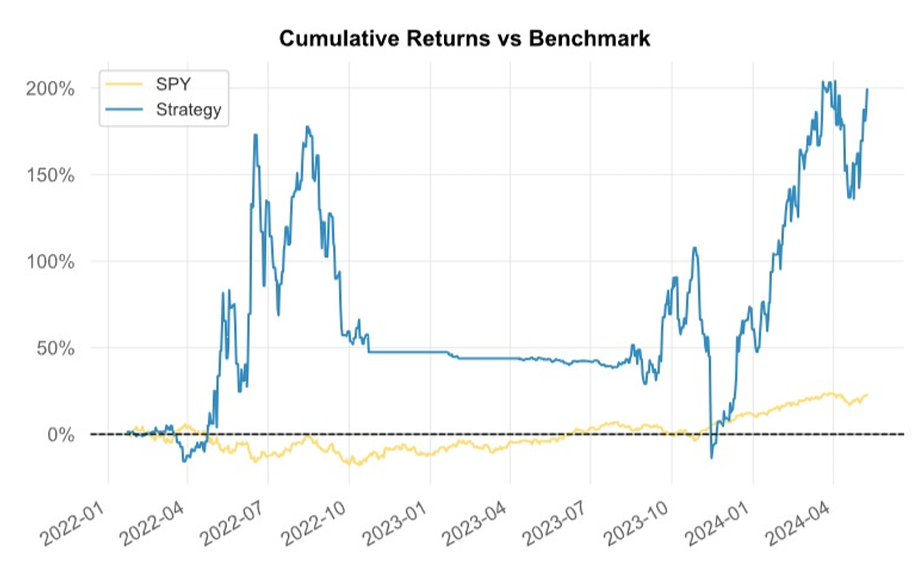
In the rapid development of algorithmic marketing, the integration of emotional analysis with automated systems has become an important innovation. Here we present the best ideas used in the Python trading bot designed to support research thinking to make informed decisions when connecting to trading on the Alpaca platform. The core of the concept is the FinBERT model, a deep learning algorithm designed to analyse financial sentiment. By leveraging the power of FinBERT, our bots can intelligently interpret negative sentiment in financial news, social media and business forums, providing insights for those who decide to trade with high accuracy

Our strategy for delivering emotional analysis follows a carefully designed algorithmic framework. Our bots constantly monitor sentiment data in real time and dynamically measure market sentiment trajectories. When the indicators are based on predefined criteria and conditions in the market, the bot automatically triggers buy or sell signals, optimizing the market in both bullish and bearish scenarios. Our robots are integrated into the Alpaca platform and can trade quickly and efficiently by taking advantage of low-latency analysis opportunities. Through analysis and back testing, we validate the effectiveness of our approach and demonstrate its ability to improve business performance and reduce risk in the business environment.

# Result And Discussion

Work done on the Alpaca API platform demonstrates the success of algorithmic trading using a combination of complex models and techniques. Among them, the LSTM model has demonstrated outstanding performance in volatile markets, becoming a powerful tool for capturing complex moments of business data. Additionally, the ARIMA model ensures stability across a wide range of industries by providing reliable forecasts by experts who analyse different trends and seasonal patterns. To complement this model, Prophet's model shows good performance in processing noisy and irregular data, identifying trends and seasonal products. Additionally, the integration of sentiment analysis techniques (examining articles to obtain sentiment scores) provides immediate results for business sentiment, which improves the entire market and demonstrates that the business is consistent with its business philosophy.

Together, these models and concepts form a unified framework for making informed business decisions on the Alpaca API platform. Opinions provided by opinion analysis about their common adaptation to different sectors and their results in improving decision-making ability and benefiting the business. Combining advanced standards with innovative ideas, the project takes a holistic approach to algorithmic trading, allowing investors to navigate complex financial markets with confidence and precision.



**Figure 9: Cumulative Returns vs Benchmark**

A graph of a distribution of monthly returns

Description automatically generated

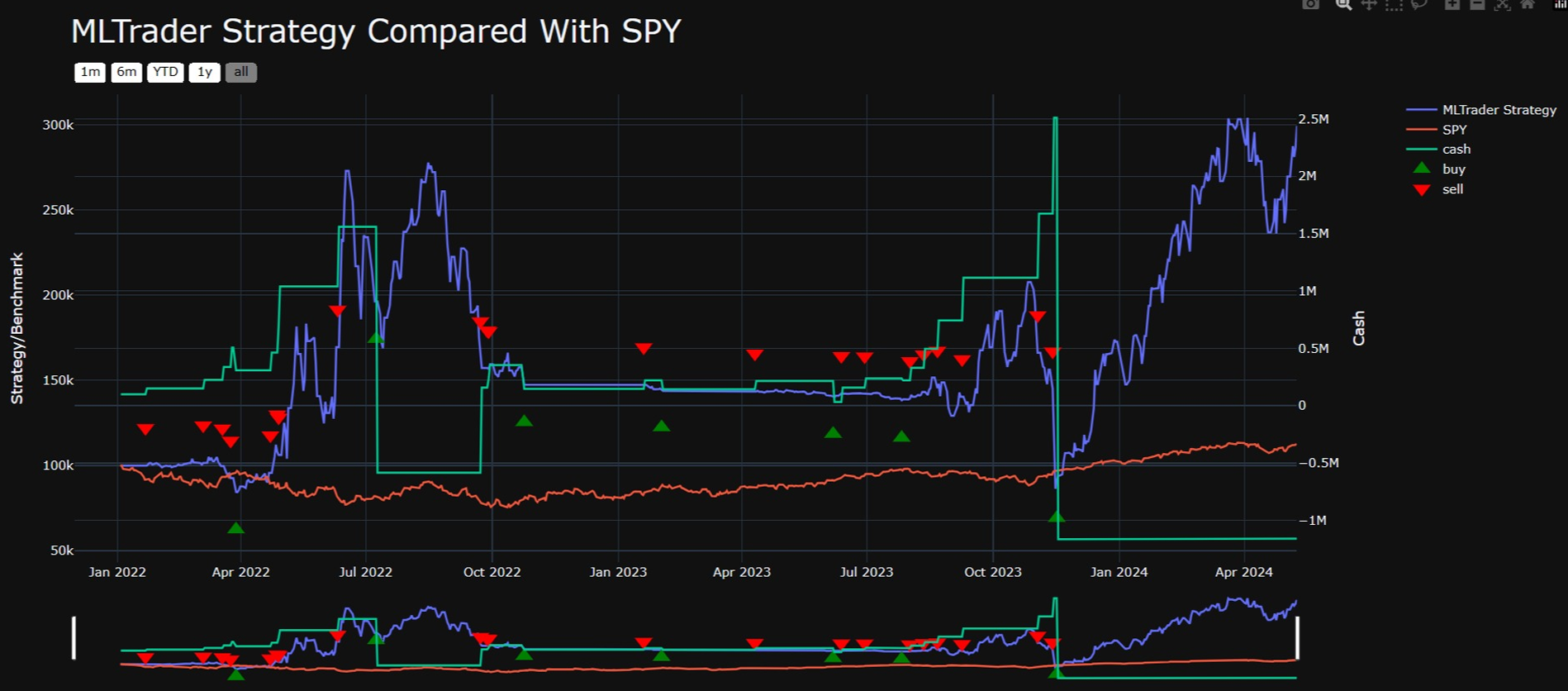
**Figure 10: Distribution of Monthly Returns**

A chart with numbers and a green box

Description automatically generated with medium confidence

**Figure 11: Heat Map Distribution of Monthly Returns**

The chart below demonstrates the performance of the strategy in 5 years of back-testing:



**Figure 12: Back-Testing Results**

This project focuses on the development and testing of algorithmic trading robots aimed at optimizing trading performance using various strategies. These techniques include a variety of techniques, including traditional measurement tools and more advanced machine learning techniques. One of the main ideas used in the project is the crossover movement, which uses the intersection of short-term and long-term concepts to detect future problems. This technique can quickly capture changes in the market and is widely used by investors to manipulate the asset's price difference. Additionally, mean reversion strategies are used to make temporary deviations from the historical average price of the asset to use the time required for the price to return to the mean. This strategy is especially useful in many markets where prices tend to fluctuate around the average price.

Additionally, trading robots use the following algorithms to detect and exploit continuous price movements in the market. The bot is designed to harness the power of price action by identifying and following patterns and maximize profits during the trading period. This strategy uses the principle that changes will occur over time, allowing the bot to benefit from long-term value while minimizing the risk of trading noise and short-term changes. Additionally, sentiment analysis combined with news releases provides additional insight into market dynamics. By analyzing text and providing a sentiment score, marketing bots can gauge market sentiment and incorporate other factors into their decision-making processes, improving their ability to respond effectively to market changes.

After rigorous testing and analysis, this robot business has shown great results, demonstrating its ability to produce results in various industries. The project demonstrates the effectiveness of algorithmic business strategies in streamlining decision-making and maximizing profitability. By combining training, machine learning techniques, and sentiment analysis, trading bots can adapt to different market trends and benefit from the market instantly.

The table below represents the comparison of various strategies:

A screenshot of a graph

Description automatically generated

The graph below represents the results obtained by various trading strategies. Clearly, the results obtained by NLP sentiment analysis stand out.

# Conclusion

In summary, the use and testing of various algorithmic trading strategies, including moving averages, mean reversion, pattern following, and sentiment analysis, have proven their results to be good in the market. These strategies provide different tools for different journeys in the market, from identifying patterns to short-term exploitation of price differences and participating in sideways trading. While all ideas have been proven to be good in each business, it is worth noting that ideas that use imagination are the best ideas in creating business results.

The integration of sentiment analysis provides a better understanding of the business environment, allowing marketing bots to quickly react to changes, exchange ideas and exploit emerging markets. By analyzing information and providing sentiment scores, the bot enables the investor to make better decisions by gaining a deep understanding of their emotions. This strategy allows the bot to adapt to changes in market sentiment and make informed trading decisions, ultimately achieving better trading results than other strategies.

Overall, the project emphasizes the importance of using a variety of business strategies and using external business data such as opinion analysis to ensure that the business is good and profitable. Going forward, further refinement and optimization of analytical strategies, as well as ongoing testing and evaluation of additional strategies, will be critical for management to respect and improve business performance in a dynamic economy. Finally, the project demonstrates the great potential of algorithmic trading strategies and the value of integrated thinking to achieve better trading.

**References**

[1] Ash Booth, Enrico Gerding, and Frank Mcgroarty. 2014. Automated trading with performance weighted random forests and seasonality. Expert Syst. Appl. 41, 8 (June ,2014), 3651± 3661. DOI: <https://doi.org/10.1016/j.eswa.2013.12.009>.

[2] Suryoday Basak, Saibal Kar, Snehanshu Saha, Luckyson Khaidem, Sudeepa Roy Dey, Predicting the direction of stock market prices using tree-based classifiers, The North American Journal of Economics and Finance, Volume 47, 2019, Pages 552- 567, ISSN 1062-9408,

<https://doi.org/10.1016/j.najef.2018.06.013>.

[3] Maragoudakis, M and Serpanos, D. (2010), towards stock market data mining using enriched random forests from textual resources and technical indicators. AIAI 2010, IFIP AICT 339, pp. 278-286. <https://link.springer.com/chapter/10.1007/978-3-642-16239-8_37>.

[4]Manoj Thakur and Deepak Kumar. 2018. A hybrid financial trading support system using multi-category classifiers and random forest. Appl. Soft Compute. 67, C (June 2018), 337±349. DOI: <https://doi.org/10.1016/j.asoc.2018.03.006>.

[5] Cain Evans, Konstantinos Pappas, Fatos Xhafa, Utilizing artificial neural networks and genetic algorithms to build an algo-trading model for intra-day foreign exchange speculation, Mathematical and Computer Modelling, Volume 58, Issues 5±6, 2013, Pages 1249-1266, ISSN 0895-7177, <https://doi.org/10.1016/j.mcm.2013.02.002>.

[6] Omer Berat Sezer, Murat Ozbayoglu, Erdogan Dogdu, A Deep Neural-Network Based Stock Trading System Based on Evolutionary Optimized Technical Analysis Parameters, Procedia Computer Science, Volume 114, 2017, Pages 473-480, ISSN 1877-0509,

<https://doi.org/10.1016/j.procs.2017.09.031>.

[7] Maragoudakis, M and Serpanos, D. (2010), towards stock market data mining using enriched random forests from textual resources and technical indicators. AIAI 2010, IFIP AICT 339, pp. 278-286. <https://link.springer.com/chapter/10.1007/978-3-642-16239-8_37>

[8] Manoj Thakur and Deepak Kumar. 2018. A hybrid financial trading support system using multi-category classifiers and random forest. Appl. Soft Comput. 67, C (June 2018), 337±349. DOI: <https://doi.org/10.1016/j.asoc.2018.03.006>

[9] M. Nabipour, P. Nayyeri, H. Jabani, S. S. and A. Mosavi, "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis," in IEEE Access, vol. 8, pp. 150199- 150212, 2020, doi: 10.1109/ACCESS.2020.3015966.

<https://doi.org/10.1051/itmconf/20214003041>

[10] M. H. L. B. Abdullah and V. Ganapathy, "Neural network ensemble for financial trend prediction," 2000 TENCON Proceedings. Intelligent Systems and Technologies for the New Millennium (Cat. No.00CH37119), 2000, pp. 157-161 vol.3, doi: 10.1109/TENCON.2000.892242.

[11] Wen Long, Zhichen Lu, Lingxiao Cui, Deep learning-based feature engineering for stock price movement prediction, Knowledge-Based Systems, Volume 164, 2019, Pages 163-173, ISSN 0950-7051,

<https://doi.org/10.1016/j.knosys.2018.10.034>.