CSC3094 - PROJECT PROPOSAL

PREDICTIVE MODELLING FOR STOCK TRADING

UTSAV KAILASH KOTHARI (210100637)

3450 Words [excluding Title, Note, Ethics, References and Glossary]

The stock market plays a crucial role in the global financial ecosystem by enabling the raising of capital for a listed company and promoting economic growth. Companies, which are called issuers, put their shares, also called stocks, on the market so that anyone can buy and sell them. These stocks represent partial ownership in the company, and the values of the share change depending on supply and demand. Investors, who can be individuals or institutions, buy these shares with the hope of making a profit from a possible increase in share value in the upcoming time. The advantage lies on both sides, the companies get the money to grow, and the investors can profit from the success of companies they believe in, which could mean getting a big return on their money. This effect of making money creates even more economic action, starting a growth cycle that keeps going. Stock markets are very important for making sure that prices are found quickly and clearly. These are exchanges where stocks are bought and sold that are regulated. They make sure that deals are fair and run smoothly by bringing buyers and sellers together. This is good for both companies and investors.

I became fascinated by the stock market by watching my father's careful research and investment strategies at home. He dedicated many hours to studying financial reports, analysing charts, and talking about market trends with his colleagues. Early exposure to the market sparked my interest in how it works and how it can help drive economic growth.

I got more interested in the world of finance when I was sixteen years old. I convinced my dad to invest in Bharti Airtel Ltd. by telling him about what I thought was deep market knowledge I got from online forums and social media posts. This choice, based only on market buzz and uninformed excitement, cost a lot. The subsequent drop in the stock price caused a large loss of money later. This event was a lesson of how dangerous speculative trading can be and how important it is to do a lot of research before making a decision. It showed how important fundamental analysis is, including studying a company's finances, future plans, and competitors. It also showed how important technical analysis is to understand market trends and find good times to buy and sell. This loss stressed how important it is to plan your investments for the long term instead of letting short-term market changes and your emotions get in the way.

The Indian stock market is experiencing an influx of young investors, aged 15-30, drawn in by the accessibility and ease of online platforms like Zerodha and Groww where you can start investing in under an hour. While the large number of young investors is good for the market, it has also shown a major knowledge gap in financial matters because investment-related learning is not taught in the Indian education system. People participate in the market because stock prices and company news are easy to find, but it doesn't always have the accuracy needed to make the right decisions. Investors often make bad investment decisions because they can't figure out how to use technical analysis, fundamental analysis, and the constantly changing market dynamics.

The problem is made even harder by the fact that existing research on price prediction models isn't very useful. There are a lot of research papers available on the internet with the same concept but are

particularly in regards to the US stock market, which is different from the Indian market in many ways. Some of these differences are the way regulations work, investor behaviour, cultural/festival drive economic performance, trade/foreign policy, political stability and other economic factors. Hence, models need to be made that are specifically made for India and take into account these unique aspects.

This project proposes an Indian stock market-specific Predictive Price Modelling(PPM) system. Machine learning is used to generate insights and predictions to help investors make informed decisions in this solution. The model will use stock prices and volumes, news sentiment analysis, and economic indicators. The model processes this data to find patterns and relationships humans may miss. These insights can be used to predict future stock prices, but they're not guaranteed and should be used cautiously with other investment strategies.

Biases and emotions can affect human decision-making. PPM reduces these biases by providing data-driven insights to supplement human judgement. An investor may invest in a company out of personal preference and overlook important financial indicators. PPM can objectively assess the company's financial health, past performance, and future growth potential for more balanced investment decisions. PPM aids "Risk assessment" by identifying sectors/companies with higher volatility through historical data analysis helping investors diversify portfolios and mitigate risk. For mutual funds, PPM can help optimise equity investments by analysing predicted stock performance in chosen sectors, potentially boosting returns for investors. PPM can provide market insights by predicting prices for several companies within a sector. Analysing these predictions, like those for Infosys, TCS, Wipro, and HCL Tech., provides an overview of the Indian IT sector, helping investors understand its health and direction before investing. Additionally, analysing predicted price movements across sectors can help identify undervalued sectors or emerging trends for strategic portfolio allocation.

PPM should be used with other fundamental and technical analysis methods because it is not foolproof. However, its ability to analyse massive amounts of data and identify complex relationships makes it a valuable tool for investors of all levels looking to bridge the knowledge gap and invest in the Indian stock market with confidence.

Work Citation	Summary of Work	Relevance to the Project
Beyaz, E. (2018). EFFECTIVE STOCK PRICE FORECASTING USING MACHINE LEARNING TECHNIQUES WHILST ACCOUNTING FOR THE STATE OF THE MARKET. International Journal of Engineering and	This study examines the effectiveness of various machine learning techniques in forecasting stock prices, taking into account market conditions.	The project relies on a thorough analysis of machine learning model performance in different market states. My predictive modelling strategies are based on my research on stock price prediction accuracy. Focusing on market conditions fits with my project's goal of creating a strong model for the Indian stock market, which has unique growth and volatility patterns. Knowing how model performance changes across markets can help you choose and improve Indian stock price prediction algorithms. This source enhances my study's theoretical framework and model development methodology. This ensures that my models apply to the complex Indian financial market. Beyaz's market sentiment indicator variables could help me analyse Indian financial news sentiment. Considering how local market sentiment affects stock

Applied Sciences, 10(2), 115-122.		prices could improve the model's prediction accuracy, which is crucial in a volatile market.
Guo, Y. (2022). Stock Price Prediction Using Machine Learning. Master's thesis, KTH Royal Institute of Technology.	This research focuses on predicting stock prices using machine learning algorithms, highlighting the comparative performance of different models.	This paper compares machine learning models for predicting stock prices and can help shape the algorithmic framework of my dissertation. This study helps me choose the right algorithms for the Indian market by comparing their performance to real-world data, which is a key part of my research on predictive price modelling. Guo's work backs up the choice of machine learning techniques since they are good at predicting stock prices. To make the model more accurate and reliable in the fast-paced Indian stock market, it is also important to look at the pros and cons of each algorithm. This study shows how important it is to choose the right model, showing that you need to find a good balance between managing complexity and performance to make sure the model you choose understands the Indian market.
Patel, M. (2020). Stock Price Prediction Using Machine Learning. Birla Vishvakarma Mahavidyalaya (Engineering College), Gujarat Technological University.	This paper explores machine learning models for predicting stock prices, focusing on the Indian stock market and comparing various algorithms' performance.	This paper offers valuable insights for PPM projects in the Indian context. Analysing the performance of various algorithms on individual stocks and indices(NSE) with different liquidity levels and market volatility helps in data selection. For example, concentrating on liquid stocks with readily available historical data may be better than illiquid ones. The research also covers feature engineering methods like scaling numerical features or creating market trend features to help you pre-process and select features. It emphasises the importance of understanding Indian stock market data's unique characteristics for model accuracy. Machine learning models have limitations in absolute price prediction, according to the research. My project should emphasise the uncertainties and the importance of using PPM as a supplement to other investment strategies. A robust and informative model can be created using these insights.
Baker, M., & Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. The Journal of Finance, 61(4).	Investor sentiment and stock returns are examined in this study, suggesting that optimistic sentiment can over-value stocks and pessimism can undervalue them.	The importance of investor sentiment and stock returns is very relevantly mentioned here. News articles, social media, and other data sources can help spot overconfidence or pessimism, allowing to adjust predictions and mitigate market bubbles and crashes. This is extremely important to my project while incorporating sentiment analysis in the growing world of propaganda to influence stock prices. My model can adjust predictions for stocks highly sensitive to sentiment changes by analysing news articles, and financial reports for sentiment indicators. This approach matches Baker and Wurgler's global findings and adapts it to the Indian market's unique market structure and investor behaviours, where investor sentiment may be even more important.
Jayalakshmi, V., Jothi, S., & Shankar, K. (2019). A Robust Predictive Model for Stock Price Prediction Using	This research proposes a deep learning model that combines natural language	This paper explores a deep learning model using NLP techniques, showing that PPM solutions can use different model architectures and sentiment analysis. Analysing news sentiment and historical data could improve my model's input. Their focus on the Nifty 50 index illustrates Indian market data selection. The fact that my project may target specific sectors or indices emphasises the importance of carefully selecting and preparing data based on its

Deep Learning and Natural Language Processing.	processing (NLP) Techniques to predict stock prices. The model is trained and tested on the Nifty 50 index, demonstrating promising results.	goals and target market. This research provides a framework for predicting Nifty 50 stock prices using advanced models and NLP. Their approach supports my project's goal of improving prediction accuracy by showing how deep learning captures market complexities and sentiment analysis understands investor behaviour. Integrating these insights could strengthen my Indian market predictions, giving investors a more complete tool.

AIMS AND OBJECTIVES:

Aim 1: Develop a comprehensive predictive model utilising machine learning techniques and sentiment analysis to forecast stock market trends in the Indian Stock Market, specifically focusing on the NIFTY 50 and Sensex indices.

Aim 2: To enhance investment strategies through the integration of a machine learning-based model capable of identifying and exploiting patterns in financial news sentiment, thereby predicting the impact on stock prices of the NIFTY 50 and Sensex indices.

This aim focuses on leveraging natural language processing (NLP) techniques to quantify investor sentiment and its correlation with market movements, providing investors with a nuanced tool for market analysis.

Aim 3: To implement a dynamic risk assessment framework within the predictive model, utilising advanced volatility prediction algorithms such as LSTM, ARIMA and GARCH for the Indian Stock Market.

The aim includes developing a system that not only forecasts stock prices but also incorporates volatility for NIFTY 50 and Sensex, enabling investors to make informed decisions about portfolio risk management and investment diversification strategies.

The Indian stock market presents an exciting and dynamic landscape for investors, but also one with inherent complexities. This project aims to leverage the power of machine learning and sentiment analysis to create a predictive model that can offer valuable insights and support informed decision-making for investors in the Indian stock market.

To fulfil the aims, the following key objectives will be addressed:

1) Data Collection: The first objective of this project is data collection which will be achieved through a two-stepped approach. First, Angel One(A large broker in India)’s Historical API will be used to obtain historical price data for the NIFTY 50 and Sensex indices, providing a comprehensive dataset that encompasses closing prices for each day, trading volumes, and market indicators. Second, web scraping

techniques using Python will be employed to extract relevant news articles from Moneycontrol's extensive financial news channel, incorporating insights on listed companies, economic indicators, and other potential market-influencing factors. Using both quantitative and qualitative data sources helps to understand market dynamics better, which builds a strong base for further analysis and model development. This approach to gathering data shows that the project is dedicated to being accurate and relevant, which is important for creating a model that accurately depicts the Indian stock market.

2) Data pre-processing: The project acknowledges the reliability of Angel One's API stock prices but also acknowledges the possibility of outliers due to market fluctuations or data entry errors. For data quality and reliable analysis, the project will use multi-step pre-processing:

Using statistical methods like interquartile range (IQR) or z-scores to identify outliers in closing prices and sentiment analysis data points (extracted from news articles) across the time series. This would detect outliers across all model features. Outlier Handling: Assessing and choosing the best strategy from several options based on the outliers' nature and context. Removing, winsorizing (capping extreme values) for price and sentiment data, or robust statistical methods less sensitive to outliers can minimise model learning impact. To address missing data points, mean/median imputation or forward fill will be used. Time series analysis requires data integrity and temporal order, which will be ensured by this. Normalisation: closing prices and sentiment analysis data points will be scaled to a common range if further refinement is needed. This can improve model performance and training efficiency by addressing issues caused by features of different scales.

This comprehensive and data-centric pre-processing approach keeps data integrity in mind while ensuring model accuracy and reliability, creating a robust and informative PPM solution.

3) Feature Selection: Based on Patel (2020), the project will incorporate using different technical indicators, such as moving averages and MACD, to track market trends, volatility, and momentum. Additionally, because Baker and Wurgler (2006) pointed out that investor sentiment affects stock prices, sentiment analysis data from news articles will be added as a feature. Statistical methods like correlation analysis and feature importance scores from trained models will be used to find features that are highly relevant to the target variable (stock price) and get rid of the ones that aren't. This holistic approach makes sure that the right features are chosen, which leads to a strong and useful model that can handle the project well.

4) Model Selection: The three well-known models are chosen because they are good at capturing different aspects of the Indian stock market:

Long Short-Term Memory (LSTM): Guo (2022) showed that LSTM models are very good at finding long-term dependencies in time series data. This makes them ideal for predicting stock prices where historical patterns and trends play a significant role, particularly for capturing non-linear relationships that exist in the data.

AutoRegressive Integrated Moving Average (ARIMA): ARIMA models are well-known for their ability to look at data that has a straight-line trend. Because of this, they are good at predicting the prices of stocks that

consistently go up or down over time. This project will use what Patel (2020) says about choosing the right parameters for ARIMA models to make sure they work best for the Indian market.

Generalised AutoRegressive Conditional Heteroskedasticity (GARCH): These models are great at showing how volatile the market is(by using a moving average mechanism), unlike other models that assume volatility stays the same. This is very important in the stock market, which is always changing and volatile because volatility affects investment decisions directly.

5) Hyperparameter tuning and Cross-Validation: To finetune the hyperparameters, a mix of Grid Search, Random Search, and potentially Bayesian Optimisation will be used, depending on how complicated the model is and what resources are available. The goal of this process is to find the best configuration for each model's internal settings. After that, cross-validation methods such as k-Fold and Stratified k-Fold will be used to check how general and stable the optimised models are. This method makes sure the evaluation is accurate by separating the data into training and validation sets. It also lowers the chance of overfitting. Cross-validation and hyperparameter tuning will work well together to make the chosen models better and more accurate. This will lead to a model that can be used for other markets and is accurate for the Indian stock market.

6) Back-Testing: Back-testing will be used in the project to see how well the model works with data. This will be done by using stock price data that wasn't used in the training process and running the model on it both before and after tuning the hyperparameters. The model's predictions will then be compared to real prices from the past using well-known performance measures such as R-squared, mean squared error (MSE), and root mean squared error (RMSE). By comparing the results from before and after tuning, an analysis of how well the tuning process improved the model's performance will be done. This thorough method makes sure that the model's ability to predict stock prices in the real world is rigorously evaluated.

With the project’s aims and objectives outlined, a thorough investigation into predictive modelling for the Indian stock market will be conducted. It will use machine learning techniques and thorough data analysis to accurately predict stock prices.

Planning: Click here to view the Gantt Chart:
The Gantt chart briefly outlines the project schedule and is divided into phases in the Gantt chart to ensure smooth flow. It begins with topic selection and research. An in-depth literature review gives the project academic weight. Writing a proposal to ensure a well-researched project, calculated planning, laying out aims and objectives, and checking for ethical considerations. The oral presentation and poster demonstration require much preparation as it clarifies project goals and results. Model building and improvement take up a large time during coding. First, collect and preprocess data, then carefully select financial features and perform statistical analysis. Due to computational demands, model hyperparameters are chosen and fine-tuned over 21 days. The next model evaluation and performance analysis takes 11 days and examines prediction accuracy. Twenty days after Easter are set aside to write the project report, which allows for a full explanation of the literature, methodology, research and findings. The project is finished after adjustments based on supervisor feedback. This organised method supports in-depth academic inquiry, generates interest in the topic, and ensures a successful research project.

Risks: A few general risks are possible in the project and two specific technical risks are mentioned further:

Risk 1: Data Period Limits Generalizability of Back-testing Results - Back-testing models on historical data is essential, but the timeframe may not capture real-time market conditions. The back-testing results may not be generalizable, affecting the model's ability to adapt to market changes.
Mitigation:
- Use rolling window back-testing instead of a single period. Separating the historical data into smaller windows and back-testing the model on each window sequentially can improve the model's generalizability by showcasing its performance across market conditions.
- Real-time model monitoring and adaptation: After deployment, track model performance in real-time. A consideration of retraining or adapting the model to market dynamics, if back-testing results deviate significantly, would be done.

Risk 2: Misalignment of GARCH Model Parameters with Market Volatility Patterns - The GARCH model assumes volatility persistence, which may not match sharp volatility shifts in the Indian market. The model's volatility assumptions may not match the data, resulting in inaccurate risk assessments.
Mitigation:
- Implement a dynamic parameter tuning process that periodically reassesses the GARCH model parameters to better align with the current market volatility trends. Adopting a rolling window approach for model calibration can ensure that the model parameters reflect recent market conditions.

General risks:
- Model Overfitting: Machine learning models, such as LSTM, can be at risk of overfitting to the training data, making them less effective on new data. To address this issue, the project will use thorough cross-validation methods and keep a test dataset to evaluate the model's predictive accuracy consistently.

- Technology and tools: Relying on certain technologies, APIs, or platforms may be risky if they face downtime or alter their terms of service. To manage the risk, the project will stay updated on information from these providers and have contingency plans in place, like alternative data sources.

- Accuracy and Reliability: Ensuring precision and dependability Stock price prediction involves inherent uncertainty. It is important for the project to always emphasise the uncertain nature of its outputs and avoid presenting predictions as definite to prevent misleading users.

- Time management: Ensuring efficient time allocation for each project stage is crucial. Utilising project management tools and adhering to the defined timeline can mitigate this risk.

Each of these risks will be monitored throughout the project, with mitigation strategies ready to be deployed as needed to ensure the project remains on track and produces reliable, valid results.

Note: My prior experience as a regular investor in the Indian stock market for over four years has provided me with a solid foundation in many financial concepts and terminology. This background knowledge allowed me to confidently navigate the financial aspects of this project without requiring extensive references in those specific areas. However, I have still incorporated relevant references for technical and methodological aspects to ensure the project's comprehensiveness and credibility. Additionally, GPT was used to form the flow of the document well and improve the clarity of the content(English).

Ethics: The project on predictive price modelling of the Indian Stock Market using machine learning techniques does not raise major ethical issues as outlined in the checklist, specifically:

My project:
1. Will not involve working with **animals** or users/staff/premises of the **NHS**
2. Will be carried out **within the UK or European Economic Area**
3. Will not have any impact on the **environment**
4. Will not work with populations who do not have the **capacity to consent**
5. Will not involve work with **human tissues**
6. Will not involve work with **vulnerable groups** (Children/Learning disabled/Mental health issues, etc.).
7. Will not involve any potentially **sensitive topics** (Examples include but are not exclusive to body image; relationships; protected characteristics; sexual behaviours; substance use; political views; distressing images, etc.)
8. Will not involve the collection of any identifiable personal data

The data used in this project comes from news and financial databases that are open to the public and do not need informed consent. However, the project will make sure that all the data it uses is anonymized, especially the data from Moneycontrol that is used for sentiment analysis, and that any information that could be sensitive about how well a company is doing or how the market is trending is kept secret and honest. The project is dedicated to upholding a high level of ethics, especially when it comes to data privacy, information accuracy, and the proper use of predictive analytics in a business setting.

References:

1. Angel One: https://www.angelone.in/
2. Baker, M., & Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. The Journal of Finance, 61(4), 1645–1680. doi:10.1111/j.1540-6261.2006.00885.x
3. Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short-term memory. PLoS ONE, 12(7), e0180944.
4. Beyaz, E. (2018). Effective stock price forecasting using machine learning techniques whilst accounting for the state of the market. International Journal of Engineering and Applied Sciences, 10(2), 115-122.
5. Bombay Stock Exchange (BSE): https://www.bseindia.com/
6. Dixon, M., Klabjan, D., & Bang, J. H. (2017). Classification-based financial markets prediction using deep neural networks. Algorithmic Finance, 6(3-4), 67-77.
7. Groww: https://www.groww.in/

8. Guo, Y. (2022). Stock price prediction using machine learning. Master's thesis, KTH Royal Institute of Technology.
9. Moneycontrol: https://www.moneycontrol.com/
10. National Stock Exchange of India (NSE): https://www.nseindia.com/
11. Nelson, D. M., Pereira, A. C., & de Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. International Conference on Neural Information Processing, 450-460.
12. Pate, M. (2020). Stock Price Prediction Using Machine Learning, BIRLA VISHVAKARMA MAHAVIDYALAYA.
13. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42(1), 259-268.
14. Singh, A. K., & Kaur, N. (2023). Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. International Journal of Financial Studies, 11(3), 94.
15. Sonkavde G., Dharrao D., Bongale A., Deokate S., Doreswamy D., Bhat S.(2023). Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications. International Journal of Financial Studies.
16. Upstox: https://www.upstox.com/
17. Zerodha: https://www.zerodha.com/

Glossary:

1) Financial Markets and Analysis:

- Stock Market: A marketplace where investors buy and sell shares representing ownership in companies.
- Technical Analysis: A method of analysing the price and volume history of security to forecast future price movements.
- Fundamental Analysis: An approach to evaluating a security's value by considering its underlying financial health and future prospects.
- Market Capitalization (Market Cap): The total market value of a company's outstanding shares, calculated by multiplying the share price by the number of outstanding shares.
- Indices: Market-wide indicators that track the performance of a specific segment of the stock market. (e.g. NIFTY 50, Sensex)
- NIFTY 50: A benchmark stock market index in India, representing the weighted average of the 50 largest companies listed on the National Stock Exchange of India (NSE).
- Sensex: A stock market index that tracks the performance of the 30 largest and most actively traded companies listed on the Bombay Stock Exchange (BSE) in India.
- Angel One: An Indian online stockbroking platform.
- Zerodha: An Indian online stockbroking platform.
- Groww: An Indian online stockbroking platform.
- Sentimental Analysis: The process of identifying and understanding the emotional tone of the text, often used to gauge market sentiment.
- Risk Assessment: The process of identifying, analysing, and evaluating the potential risks associated with a project, activity, or situation.

2) Machine Learning and Modeling:

- Machine Learning: A subfield of Artificial Intelligence concerned with creating algorithms that learn from data and make predictions without explicit programming.
- Predictive Price Modelling (PPM): The application of machine learning algorithms to analyse historical data (e.g., stock prices, news articles) and forecast future prices. This approach aims to identify patterns and relationships within the data that can help anticipate future market movements.
- LSTM (Long Short-Term Memory): A type of artificial neural network architecture effective in handling sequential data, often used for time series forecasting.
- ARIMA (Autoregressive Integrated Moving Average): dA statistical model commonly used for forecasting time series data.
- GARCH (Generalised Autoregressive Conditional Heteroskedasticity): A statistical model that helps capture and model the volatility of financial time series data.
- Hyperparameter Tuning: The process of optimising the parameters of a machine learning model to achieve the best possible performance.
- Back-testing: The process of evaluating the performance of a trading strategy or model on historical data.
- Cross-validation: A statistical technique used to evaluate the generalizability of a machine learning model by splitting the data into training and testing sets.
- Mean Squared Error (MSE): A common metric used to measure the average squared difference between predicted and actual values.
- Root Mean Squared Error (RMSE): The square root of the mean squared error, another metric for measuring the magnitude of the differences between predicted and actual values.
- R-squared: A statistical measure that indicates the proportion of the variance in the dependent variable that can be explained by the independent variables in a regression model.
- k-Fold: Splits data into folds, trains on k-1 and tests on 1, repeated k times. Ensures all data is used for both training and testing, improving model evaluation.
- Stratified k-Fold: Similar to k-Fold, but ensures each fold reflects the class distribution of the entire dataset, important for imbalanced classification tasks.
- Outliers: Data points that fall significantly outside the overall pattern of the data distribution.

3) Other Relevant Terms:

- Python: A general-purpose programming language widely used in data science and machine learning applications.
- API (Application Programming Interface): A set of instructions and standards that allow applications to communicate with each other.
- Web Scraping: The process of extracting data from websites.
- Data Preprocessing: The process of cleaning and preparing raw data for use in machine learning models.
- Ethical Considerations: Important guidelines to ensure responsible and unbiased development and application of your project.