UNIVERSITY OF LIVERPOOL

COMP702 MSC PROJECT

STOCK PRICE PREDICTION

SPECIFICATION AND DESIGN PROPOSAL

Name: VENKATASHIVA PASUPULETI

Student ID: 201737217

Statement of Ethical Compliance:

This project falls under ethical compliance category A0. All the data used are publicly available, and there will be no involvement of human participants. I confirm that the project will follow the ethical guidelines outlined by the institution.

Supervised by: RAMYA CHAVALI

Project Description:

Stock price forecasting is the key area in finance. It plays significant role in investment strategies, risk management and economic planning. Accurate stock price forecasts can lead to informed investment decisions, optimised portfolio management, and improved risk mitigation measures. When we look at it clearly, stock price prediction is a time series forecasting problem, with each price point acting as a response to prior market behaviours. But the nature of financial markets is highly dynamic influenced by a wide range of factors such as market sentiment, economic indicators, geopolitical events and firm performance. These complexities make it difficult to drive a stock price prediction and requires sophisticated models capable of detecting linear and nonlinear trends in time series data.

Traditional statistical models such as ARIMA (AutoRegressive Integrated Moving Average) and its seasonal equivalent SARIMA have been widely used in time series forecasting due to their ease of use and effectiveness to model linear combinations. A more modern model called Facebook Prophet expands on these principles and is made to deal more strongly with outliers, seasonal effects, and missing data. However, the complex nonlinear data found in financial data are frequently difficult for these models to represent. On the other hand, deep learning models have shown great promise in solving this problem, especially long-term and short-term memory networks (LSTM), which are part of recurrent neural networks (RNNs). Time series forecasting tasks are a great fit for LSTM networks because of their ability to retain long-term dependencies.

So, for this project I'm using the Sensex dataset to build and compare deep learning approaches (LSTM) and conventional statistical models (ARIMA) to determine which method is more effective for stock price predictions.

Why Sensex Data?

Sensex also known as S&P BSE Index consists of 30 reputable and stable company stocks that are listed in the Bombay Stock Exchange of India. This index represents a wide range of industries which includes technology, finance, healthcare, consumer goods, and energy and it is oldest and widely followed index in India. Historical data of Sensex index is well documented and readily available, and the dataset offers a diverse sample of stock price changes caused by both local and global economic situations, making it a great choice for this project.

Aims and Requirements

Aims:

- Develop and implement a traditional statistical time series forecasting model ARIMA for stock price prediction.
- Develop and implement a deep learning-based LSTM model for the same purpose.
- Compare the performance of the two models using various evaluation metrics.
- Analyze the reproducibility of the models across different stocks.

Requirements:

Essential:

1) **Data Collection and Preprocessing:**
   - Collect historical stock price data for all the stocks in the Sensex index from the Yahoo Finance by using the Yahoo finance library and before training the model with dataset preprocess the data to handle missing values and outliers.

2) **Development of Models:**
   - Develop and implement ARIMA model by using statsmodel library from Python for time series forecasting and implementing hyperparameter tuning to optimize performance for a Sensex index.
   - Create a deep learning-based LSTM model using TensorFlow/Keras to forecast stock prices by using the historical data for training.
   - Optimize the LSTM model using techniques such as dropout regularization, early stopping, and learning rate scheduling to enhance its predictive capabilities.

3) **Model Evaluation and Validation:**
   - Generate stock price forecasts for a predefined future period by using both the models.
   - Evaluate the model's performance by using different metrics like RMSE, MAPE, MAE and compare the accuracy of both models.
   - Conduct cross-validation to ensure models are not overfitting.
4) **Analysis, Documentation and Reporting:**
   - Conduct the through analysis of both the model's strengths, weakness and real-world compatibility in financial forecasting.
   - Create visualizations comparing forecasted vs. actual stock prices.
   - Document every step of the project right from the data collection to comparison and result.
   - Create a detailed report of the findings including model suitability and reproducibility

Desirable:

- Integrate additional features such as sentiment analysis from financial news data to improve forecasting accuracy.
- Create a user-friendly interface to visualize stock price predictions and model performance.

Key literature and background reading:

Stock price forecasting is a popular topic with a large amount of research work and literature present in traditional statistical methods and modern machine learning and deep learning techniques that have proven effective. Understanding the theoretical foundation is important which helps in guide the implementation process.

1. **Traditional Time Series Forecasting:**
    - **a.** Adhikari and Agrawal (2013) provide an in-depth review of ARIMA models in stock market forecasting, highlighting its ability in capturing linear correlations in financial time series.
    - **b.** The statsmodels Python package, which we'll use to create these models, is well-documented by Seabold and Perktold (2010) and provides a solid framework for statistical computation**.**
2. **Deep Learning Models (LSTM-Based RNNs)**
    a. Recently, LSTM networks have shown promise in financial forecasting: Fischer and Krauss (2018) demonstrated that LSTM networks could predict major index stock prices better than traditional methods, such as ARIMA
    b. Bao et al. (2017) researched how well LSTM networks could model highly volatile and non-linear patterns in financial data, outperforming other deep learning models such as Convolutional Neural Networks (CNNs) and standard RNNs.
3. **Evaluation Metrics and Model Comparison**
    a. RMSE, MAPE, and MAE are common metrics that quantify prediction accuracy, particularly when comparing models with various underlying assumptions. Studies by Hyndman and Athanasopoulos (2018) have

emphasized the importance of these metrics in assessing model performance, especially when comparing models with different underlying assumptions.

b. When A. Durgapal and V. Vimal (2021) tested several time series models using a range of measures, they observed that no single model always performed better under all circumstances. This result supports the goal of the project of using several assessment criteria to evaluate ARIMA and LSTM model performance in a comprehensive manner.

Development and Implementation Summary:

Development Environment and Implementation Language:

- I will implement the project using Python as it is the preferred language for both time series analysis and deep learning tasks due to its extensive libraries and strong community support.

Libraries for Model Development:

- **ARIMA**: I will use the statsmodels library to implement the ARIMA model. This library provides robust tools for statistical modeling, time series analysis, and forecasting, making it ideal for developing the ARIMA model.
- **LSTM**: To build and train the LSTM (Long Short-Term Memory) network, I will utilize the TensorFlow and Keras libraries. These libraries are widely used in deep learning for their flexibility and support for complex neural network architectures.

Data Handling and Visualization:

- **Data Handling**: Libraries like pandas and numpy will be used for data cleaning, manipulation, and transformation—essential steps in preparing the data for modeling.
- **Visualization**: To visualize stock price trends and model predictions, I will use libraries such as matplotlib and seaborn. These libraries provide high-quality plotting capabilities and are effective for displaying time series data.

Development Environment:

- I will use Jupyter Notebooks/Google colab for the development process. This environment allows for interactive coding, data visualization, and easy documentation of the workflow.

Implementation plan:

I will organize the project into a series of phases, with each phase focusing on a specific part of the implementation.

- **Phase 1:** Collect and preprocess historical stock price data, handling missing values, normalizing data, and performing exploratory data analysis.
- **Phase 2:** Develop the ARIMA model by testing stationarity, optimizing parameters, fitting the model, and evaluating its performance.
- **Phase 3:** Build and train an LSTM model by designing the architecture, preparing data, tuning hyperparameters, and evaluating its performance.
- **Phase 4:** Compare and analyse the performance of the ARIMA and LSTM models, focusing on accuracy, efficiency, and robustness, with visualizations of predictions.
- **Phase 5:** Document the development process, prepare a comprehensive report, and present findings and conclusions.

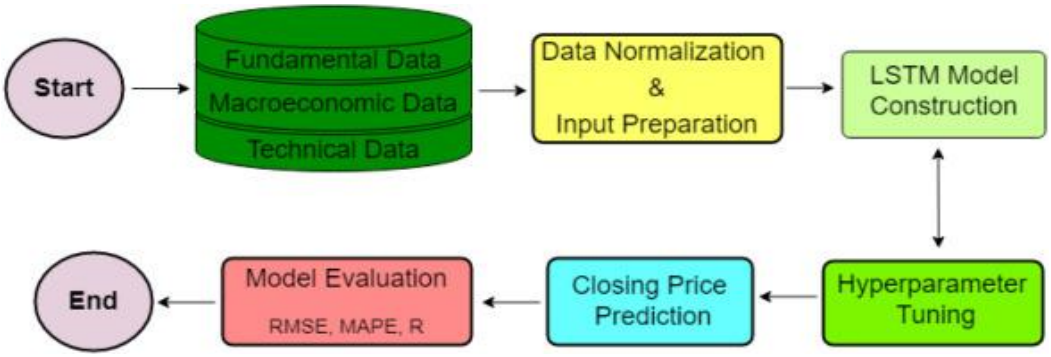

Fig: LSTM Model

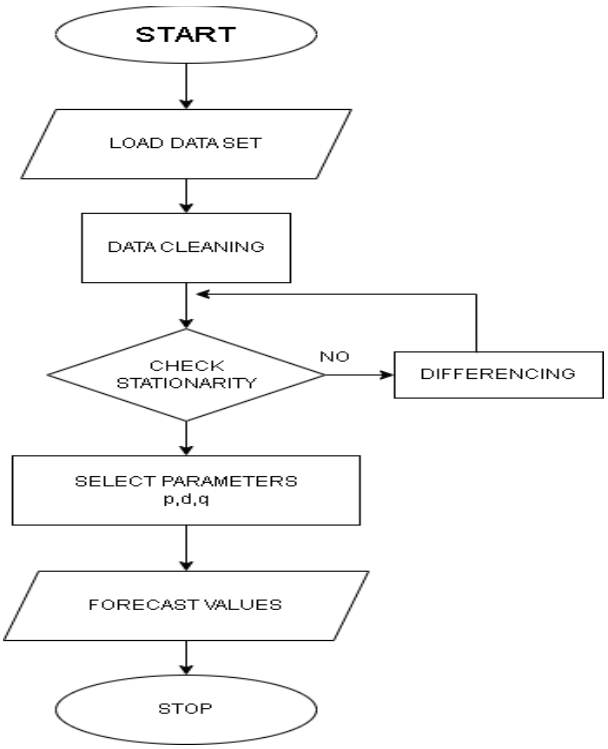

Fig: ARIMA Model

Workflow and Organization

I'll follow the Agile workflow with iterative development and regular review checkpoints. Key activities will be organized into weekly or bi-weekly sprints to ensure continuous progress and timely completion.

- **Version Control:** The project codebase will be managed using Git, with regular commits to a GitHub repository for version control and collaboration.
- **Task Management:** I'll use google sheets to manage tasks, deadlines, and project milestones.
- **Progress Reviews:** I'll connect with the supervisor to review progress, discuss challenges, and plan subsequent tasks whenever necessary.

Data Sources

In this project I will use historical stock price data for Sensex index which I will get by using the Yahoo Finance library in Python. The Yahoo Finance library in python provides access to publicly available financial data through its API which allows users to retrieve stock prices, historical data, and other financial metrics. The use of Yahoo Finance data is legal with its terms and conditions and we can use if for non-commercial, research, and educational purposes.

The dataset which is collected from the Yahoo finance library includes daily stock prices, volumes, and other relevant financial indicators for Sensex index over a specified period. Since the data is publicly available and does not contain any personal information, there are no confidentiality or anonymity concerns. I won't misuse the data and will ensure that the data is used responsibly and only for the intended purpose of model development, analysis, and evaluation in this project.

Testing and Evaluation

1. **Model Testing:**

- **Unit Testing:** Individual components of the data preprocessing pipeline, ARIMA model, and LSTM model will undergo unit testing to ensure they function correctly. This includes verifying data handling processes, checking model parameter optimization, and validating LSTM architecture setup.
- **Integration Testing:** I will perform integration testing to ensure that different components (data preprocessing, model training, prediction, and evaluation) work seamlessly together as a cohesive system. This will verify that the data flows correctly

between modules and that the entire pipeline operates as expected without errors or inconsistencies.

- **Performance Testing:** I will conduct performance testing to evaluate the computational efficiency and scalability of both models. This will involve testing the models on datasets of varying sizes to assess their training time, memory usage, and inference speed, ensuring they meet the project's performance requirements.
- **Cross-Validation:** I will apply cross-validation techniques to test the models' performance on different subsets of the data. This will help in assessing the generalizability of the models and avoiding overfitting.

2. Model Evaluation:

- The ARIMA and LSTM models will be evaluated using metrics such as RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error) to assess their predictive accuracy on the test dataset.
- I will compare the performance of both models against a baseline to determine their effectiveness in stock price prediction.
- The evaluation will also include visual analysis by plotting predicted values against actual values to visually assess the models' forecasting accuracy.

Project ethics and human participants

In this project I'll use the historical stock price data obtained from the Yahoo Finance Library. It is available publicly and it does not include any personal or sensitive information. There is zero human participation in collecting the information. Though there are no data privacy concerns it is my duty to handle data ethically during and after the project.

1. **Current Use of Project Data:**
    a. **Purpose-Driven Data Use:** I'll use the data exclusively for processing, analysing, and visualizing for research, educational and non-commercial purposes as part of this project.
    b. **Data Security and Storage:** All the data will be securely stored in a controlled environment, and it is secured from the unauthorised access. Data access will be restricted to me and supervisors of this project.
    c. **Transparency and Reproducibility:** All the models, findings and results will be shared transparently by following all the open science principles.

2. **Restrictions of Future Use of Data:**
    a. **Destruction of Data:** After completing the project, the data will be either securely archived for academic purposes or destroyed by following the data retention policies. This ensures that the data is not accessed illegally or repurposed.
    b. **Ethical Reuse of Derived Outputs:** If in case there is any need to reuse the models or findings, it will be done responsibly and ethically and do not promote any high-risk financial activities or misinterpretation of financial forecasts.

BCS Project Criteria

The project main focus is to compare the ARIMA and LSTM models for stock price predictions and it aligns with all the six outcomes mentioned in the BCS, the Chartered Institute for IT.

1. **Application of practical and analytical skills:** The project needs all the practical and analytical skills gained throughout this degree program, such as data collection, preprocessing, model building and evaluation. Building models like ARIMA and LSTM shows the deep understanding application of time series analysis and deep learning knowledge. Data manipulation using Python libraries like NumPy, Pandas, data visualization are the skills applied in this project.
2. **Innovation and Creativity:** The unique feature of this project is its comparative approach which applies two different methods to the same problem of Sensex index price estimation: LSTM, a new deep learning method, and ARIMA, a traditional times series forecasting model. This comparison needs a creative approach to find the model performance, which sheds light on which models perform best under different circumstances when it comes to predicting stock prices.
3. **Synthesis of Information, Ideas, and Practices**: To deliver the best results, the project integrates data sets, concepts and techniques from financial analysis, machine learning, and statistics. It combines advanced neural networks with training techniques and time series analysis concepts, such as stationary testing and parameter optimization. Both use different metrics (RMSE, MAE, and MAPE) and visualizations tools which help the project to evaluate different models to better analyse their advantages and disadvantages in predicting stock prices.
4. **Meeting a Real Need in a Wider Context:** Forecasting the future price of a financial stock is a major issue that affects traders, analysts and investors. To find out whether

the methods provide accurate and reliable forecasts, the study compares ARIMA and LSTM models with stock price forecasting. This information can help in decision making, risk management and strategical planning. The results can be useful for economists as well as academic studies.

5. **Self-Management of a Significant Piece of Work**: This project requires organizational skills, time management, and personal preparation to oversee multiple steps such as data collection, sampling, analysis, and evaluation The project plan outlines how each step takes place builds on the past. The work plan is carefully planned with deliverables and milestones to ensure continuous improvement and timely completion.
6. **Critical Self-Evaluation of the Process:** Throughout the project I'll conduct my own critical assessments to determine how well the selected models, methods and techniques are working. Regular notes will be kept reflecting on decisions made, obstacles encountered and how these influenced the outcome of the project. This ongoing self-assessment will help improve the process and assure an outstanding final product.

UI/UX Mock-up

Building and comparing ARIMA and LSTM models for stock price forecasting using Sensex data in a Jupiter Notebook environment in particular the Google Colab is the current major focus of this project. Right now, no plans in building a user interface (UI) or user experience (UX) part specific to project. But if time permits later in the project, I might consider creating a simple UI/UX to display performance metrics, model predictions in an interesting and dynamic manner. An easy-to-use dashboard that allows people to engage with the data, evaluate model outputs, and obtain insights might be offered by this potential UI/UX.

Project plan

	A	B	C	D
1	**Task Description**	**Start Date**	**End Date**	**Status**
2	Project Planning and Setup	01-08-2024	07-08-2024	Completed
3	Literature Review and Previous Work Study	08-08-2024	21-08-2024	Completed
4	Data Collection and Preprocessing	22-08-2024	04-09-2024	Completed
5	Development of ARIMA Model	05-09-2024	18-09-2024	Ongoing
6	Specification & Design Proposal Submission	06-09-2024	06-09-2024	SUBMISSION
7	Development of LSTM Model	19-09-2024	02-10-2024	Yet to Start
8	Model Evaluation and Comparison	03-10-2024	16-10-2024	Yet to Start
9	Documentation and Video Preparation for Final Presentation	17-10-2024	23-10-2024	Yet to Start
10	Video for Final Presentation	24-10-2024	24-10-2024	SUBMISSION
11	Q&A Sessions on Final Presentation	04-11-2024	08-11-2024	VIVA
12	Dissertation Writing	08-11-2024	14-11-2024	Yet to Start
13	Dissertation Review and Final Touches	15-11-2024	21-11-2024	Yet to Start
14	Dissertation	22-11-2024	22-11-2024	SUBMISSION

Notes:

- **Functional overlap**: Since the ARIMA and LSTM model building phases are independent of each other, they may overlap. That way, both models can be created at the same time.
- **Dependencies**: Basic procedures such as data collection and preprocessing must be completed before the model can be developed. Model construction and tuning must be completed before analysis and documentation can begin.
- **Presentation and topic**: Time allotted for video editing and Q&A sessions will be used for presentation. Once the analysis and presentation are completed, writing of the thesis begins, with additional time allotted for final review and revision.

Contingency Plans:

Risk	**Contingency Plan**	**Likelihood**	**Impact**
Data Source Unavailability	Have alternative data sources (e.g., Google Finance, Quandl) prepared in case Yahoo Finance data is inaccessible. Regularly check for any data access issues and ensure you have backup data sources.	Medium	High
Inadequate Data Quality	Perform data cleaning and validation before model training. Have procedures in place to handle missing values or anomalies. Use multiple sources to cross-check data integrity.	Medium	High
Model Overfitting or Underfitting	Use cross-validation techniques and parameter tuning to mitigate overfitting or underfitting.	Medium	High

	Regularly evaluate model performance using validation datasets and adjust complexity as needed.		
Technical Issues with Software/Tools	Have backup software/tools and ensure compatibility. Keep software updated and access technical support or forums for troubleshooting. Maintain a list of alternative tools.	Low	Medium
Performance Metrics Not Meeting Expectations	Revisit and adjust model parameters, consider feature engineering or additional data preprocessing. Implement different metrics if necessary to better assess model performance.	Medium	High
Limited Computational Resources	Optimize code for performance and utilize cloud computing resources if needed. Have a plan for scaling up computational resources if the need arises. Consider local vs. cloud-based options.	Medium	Medium
Difficulty in Comparing Models	Standardize the evaluation metrics and ensure consistent testing conditions for both models. Document and explain any discrepancies in model performance clearly.	Low	Medium
LSTM Model Training Issues	Ensure sufficient training data and perform regular checkpoints during training. If issues arise, review the architecture, adjust hyperparameters, or consult literature for best practices.	Medium	High
Scope Creep or Changing Requirements	Keep the project scope well-defined and document any changes through a formal change management process. Regularly review and adjust the project plan as needed.	Medium	Medium
Inconsistent Results with Different Stocks	Analyse results thoroughly and consider model adjustments or retraining. Ensure robust validation to assess generalizability. Document findings and limitations transparently.	Medium	Medium

Explanation of Columns:

Risk: Potential issues that could impact your project.

Contingency Plan: Actions to mitigate or handle the risk.

Likelihood: Probability of the risk occurring (low, medium, high).

Impact: Potential effect on the project if the risk occurs (low, medium, high).

References:

1. Adhikari, Ratnadip & Agrawal, R.. (2013). An Introductory Study on Time series Modeling and Forecasting. 10.13140/2.1.2771.8084.
2. 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. https://doi.org/10.1371/journal.pone.0180944
3. G. Rekha, D. B. Sravanthi, S. Ramasubbareddy and K. Govinda, "Prediction of stock market using neural network strategies", *J. Comput. Theor. Nanoscience*, vol. 16, no. 5, pp. 2333-2336, May 2019.
4. J. Brownlee, "How to Create an ARIMA Model for Time Series Forecasting in Python", *Time Series*, January 2017.
5. Krauss, X. A. Do and N. Huck, "Deep neural networks gradient-boosted trees random forests: Statistical arbitrage on the S&P 500", *European Journal of Operational Research*, vol. 2, pp. 689-702, 2017.
6. Mohamed El Mahjouby, Mohamed Taj Bennani, Mohamed Lamrini, Badre Bossoufi, Thamer A. H. Alghamdi, Mohamed El Far, "Predicting Market Performance Using Machine and Deep Learning Techniques", *IEEE Access*, vol.12, pp.82033-82040, 2024.
7. M. Pirani, P. Thakkar, P. Jivrani, M. H. Bohara and D. Garg, "A comparative analysis of ARIMA GRU LSTM and BiLSTM on financial time series forecasting", *Proc. IEEE Int. Conf. Distrib. Comput. Electr. Circuits Electron. (ICDCECE)*, pp. 1-6, Apr. 2022.
8. Seabold, S. and Perktold, J. (2010) Statsmodels: Econometric and Modeling with Python. 9th Python in Science Conference, Austin, 28 June-3 July, 2010, 57-61.
9. S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Comput.*, vol. 9, no. 8, pp. 1735-1780, Nov. 1997.
10. S. Siami-Namini, N. Tavakoli and A. S. Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series", *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018.
11. Thomas Fischer, Christopher Krauss, Deep learning with long short-term memory networks for financial market predictions, European Journal of Operational Research, Volume 270, Issue 2, 2018, Pages 654-669, ISSN 0377-2217