University of Essex
Department of Mathematical Sciences

MA981: DISSERTATION

Assessing Interdependencies and Forecasting in Financial Time Series: A Granger Causality and Machine Learning Approach"

Abhishek Aher
2322394

Supervisor : Dr. Daniel Felix Ahelegbey

January 5, 2025
Colchester

Contents

List of Figures

List of Tables

CHAPTER 1

Abstract

This report, seeks to interrogate the interdependencies that may exist among some key financial variables in UK markets, particularly banking, real estate, and insurance sectors. Using daily stock prices derived from Yahoo Finance and the London Stock Exchange, we examine the nature of the interaction and spillover effects of these markets. This paper marries traditional statistical models VAR and ARIMA with a sophisticated machine learning approach-LSTM for financial time series forecasting in finding the causal relationship using Granger causality analysis.

My findings indicate that the VAR model is very good at capturing linear interdependencies with high accuracy, while the LSTM model does well in capturing complex, non-linear patterns.We can trust the ARIMA model for one-variable research, but it's not very good at handling multiple variables.In addition, Granger causality study shows the direction relationships, which can teach us a lot about how markets work and what leading indicators really show.

The findings from this study will help financial experts, policy-makers, and investors make informed decisions based on facts that will enable them to minimize their risks and maximize their strategies. This study adds to the ever-growing literature of financial data analysis by fusing old and new approaches in making forecasts and market understanding even better.

CHAPTER 2

Introduction

2.1 Background and Motivation :

The close connections between money markets around the world are still a big problem for the world economy. The case of the UK shows how important these markets are for the growth and stability of a country. This market's relationship with other markets shows how the current financial and social scene controls where money goes and how much risk people are willing to take. Globalization of financial markets has made these connections even stronger.Today, things that happen in one part of the world's economy can affect the whole planet. The financial disaster of 2008, as it were, showed that global financial systems have been so closely interconnected. This, therefore, means that it is important to understand the connectedness of markets to each other and to themselves.

When we look at how these markets depend on each other, we also see more how different parts of each market change one another. This gives us a starting point to make better guesses about money and the economy, and to see what choices we have when making big decisions. British financial markets together with all possible other sectors such as banking, real estate or insurance form some of the most interesting and developing markets in the world. These industries enhance economic development, ensure financial security and provide a means of building stronger infrastructure. The huge financial problem facing the UK during the last years has been not knowing what was to happen with Brexit and the COVID-19 pandemic. Many different industries are vulnerable, such as real estate, banks, and insurance among others, due to these events. They show how changes in one industry can quickly affect others.

However, as these industries become more and more integrated, their behavior collectively increases systemic risks, especially in periods of economic turbulence. The ramifications of, for instance, a downturn in the real estate sector ripple through to banking and insurance sectors rather fast, emphasizing the need for appropriate tools and models that would catch these dynamics. These are complicated relationships to capture from traditional financial models. This, indeed, is a problem that does call for some really advanced tools such as machine learning models, which will not only be able to look at how these dynamic relationships change with time but could also make better predictions about the future behavior of markets. Motivated by these considerations, in this paper, dynamic interdependencies of the UK financial markets are analyzed and predicted. This research tries to give insight into these relationships and provide tools to better foresee market behaviors by taking advantage of state-of-the-art machine learning techniques along with statistical methods. This knowledge is of great value not only for academic purposes but also for practitioners, policymakers, and regulators who seek to mitigate risks and optimize financial strategies.

2.2 Problem Statement:

Financial markets are interrelated in a complex web of time-varying linkages. In fact, the nature of this dependence is susceptible to many sources, like policy changes, global economic conditions, and sudden shocks. This makes the forecast of the behavior of markets or how one event can trigger another a hard task. It is during periods of high market volatility, when the forecasts are the most difficult, that the power of traditional econometric methods often falls short.

Basically, finding and measuring the links to make the exact predictions about market behavior in a prospective perspective is the key problem. Models now very often draw these links in an oversimplified way and, as a consequence, make such a prediction impossible and important details in the actual functioning of markets. The difficulty of this job is further enhanced by the fact that the data concerning finances is not always unambiguous and does not stay put but changes over time. Advanced tools are needed to handle such complexity.

Some of the advanced techniques considered in this study to see the applicability for solving these problems include Long Short-Term Memory (LSTM) networks, Vector Auto regression (VAR), and Granger causality analysis. These methods have been developed for finding hidden links and adapting to changes that take place with financial data over time. Apart from enhancing the accuracy of the prediction, this study also aims at identifying chains of events previously unknown leading to outcomes. This will help in comprehending the complex web of interactions in the market.

This method means that you are able to implement both traditional econometric models and current financial needs of analysis at the same time. It will provide a sound basis for detecting important links and give more accurate forecasts on how the market will change.

2.3 Research Objectives and Questions

2.3.1 Research Objectives

To determine the performance of different models, such as Granger causality, VAR, and LSTM, in predicting future fluctuations within the UK financial markets.

This paper investigates the interconnectedness of the real estate, banking, and insurance sectors, which proves to be supportive during times of economic downturn And to contribute to the knowledge of people regarding how financial markets function so that better tools and plans can be designed to make predictions and decisions. By meeting these goals, the research will fill in the gaps left by earlier studies and provide useful information that financial experts, policy-makers, and regulators can use.

2.3.2 Research Questions

1. How have the UK financial markets developed to become more interdependent with the passage of time?

2. What is the most suitable statistical and machine learning method that can model complicated market linkages?

3. Which of the linkages in the markets are most vital during economic turbulence, and how can these be employed for making businesses resilient?

Advanced tools and methods are required to delve deeply into the facts and model them in order to answer these questions. In this study, the researchers hope to provide a comprehensive picture of how financial markets function and change, with a focus on the roles and impacts of key sectors.

2.4 Scope and Contributions:

2.4.1 Scope

In this study, the daily insurance, banking, and real estate company stock prices are considered in order to study the UK financial markets. Financial action data was utilized for a couple of years obtained from two valid sources: Yahoo Finance and London Stock Exchange. It is due to the fact that these areas, on one hand, constitute the most sensitive sectors for economic growth; on the other hand, due to their combined consequences on market stability.

The paper thus explores all possible causality, employing advanced techniques in time-series analysis, and predicts with it. The study, in a bid to study the functioning of markets and find interdependencies and guess how people will act in the future, uses both machine learning and statistical methods. The models that shall be used to find out how variables affect each other and determine the main factors affecting market success include Vector Autoregression (VAR), LSTM networks, and Granger causality analysis.

Although this study relates principally to UK markets, the methodology and ideas developed should be of value in many other locations. This is a global result and important and useful in that the tools and models developed here can be used in other financial markets. It attempts to help the financial analyst, policy maker, and other interest groups make better choices and create a resilient market by focusing on how different variables are related.

2.4.2 Contributions:

The value of this research lies in the fact that it extends existing knowledge on the interdependencies at large in financial markets and provides indications that can be actionable for stakeholders along methodological, practical, and theoretical dimensions.

Methodological Contributions :

The contribution of this research in terms of methodology has been in identifying the extent to which financial markets are interrelated, developing and using sophisticated techniques such as Granger causality analysis and LSTM networks. By looking at non-linear and changing dependencies, these methods give us more complicated insights than traditional economic models, which tend to make the relationships we're looking at too simple. Using both machine learning and statistics together creates a flexible structure for studying complicated interactions. This opens up a route to deeper insight into the dynamics of the market.

Practical Insights :

These research findings are very useful for analysts and investors, as well as policy makers, in that the findings allow them to understand patterns in the market and make their decisions based on correct information. The study portrays indications of future market behavior by showing patterns and linkages within the data. This thus enables the stakeholders to make strategic choices with reduced risk, which will be of benefit to the market. With such insights, for instance, practitioners can make more accurate predictions and, therefore, decisions on financially optimized strategies.

Data-Driven Decision-Making :

The results of this research enable individuals and organizations to make decisions based on evidence through the light these findings throw on causality relationships and elements that may anticipate outcomes. The ability to identify the most influential variables that affect the financial markets enhances predictability and ascertains the timeliness of actions taken. In a volatile market, this is a key skill for minimizing risk and maximizing returns, so this study is useful for making decisions in real time.

Improved Resilience :

For the purpose of helping to the overall resilience of financial markets, the findings of this study provide insight into areas of weakness and points of potential intervention. The ability of identifying systemic vulnerabilities that are highlighted within the research creates a high degree of applicability in view of current state of affairs from recent global financial crises that have identified the importance of using sophisticated methodologies of analysis.

The study combines recent developments in machine learning with more conventional methods to economics in order to handle the complexity of today's financial systems. It provides both theoretical understanding and practical skills to manage the complexities of interconnected financial systems by providing a complete framework for researching and forecasting the behavior of financial markets to begin with.

CHAPTER 3

Literature Review

3.1 Theoretical Underpinnings of Financial Market Linkages

The linkages among financial markets have been explored by experts in the domain. They put forward concepts to expose how constituents that make up a market reveal interactions with each other .These are a few of the economic theories which can help in understanding these markets, such as Efficient Market Hypothesis and Modern Portfolio Theory. The EMH says that markets work well when prices take into account all the information that is available. When things are really looked at, the idea that these assets are separate from each other often turns out to be false. Instead, investors are often affected by feelings and events in the world economy. Modern Portfolio Theory, introduced by Markowitz[Shojaie and Fox, 2021], stresses the diversification issue and illustrates how the combination of assets with different correlations might help improve the portfolio returns while reducing its risk.(for simplified version [Mangram, 2013]).This theoretical framework brings forth the necessity to understand interdependencies because it is a guidance in creating a strategy regarding both risk management and investment decisions. In fact, identifying the relationship among the assets allows an investor to create a more resistant portfolio to market vagaries. In the area of financial markets, and more so in the UK, such Interconnections are further influenced by the different economic structures and regulatory environments.

For instance, the banking industry clearly depends on the profitability of real estate since a good portion of their assets are mortgages .The insurance business is also closely linked to banking and real estate through claims, investments, and underwriting processes. Three models, VAR, Granger causality analysis, and others build on the above foundation to create useful real-world tools to study how things are connected from [Zaremba and Aste, 2014][Oliveira et al., 2024].

Additionally, behavioral finance adds a new layer to comprehending how money is linked.In behavioral theories, errors, herd behavior, and how people think affect how markets work.

3.2 Empirical Evidence of Interconnections across Financial Markets:

Empirical studies have focused on interdependencies between financial markets with an eye towards the effects of monetary policy changes and economic shocks. Using Granger causality, for example; a 2014 Zaremba and Aste study attempts to measure links between asset classes and reveals that they can be quite dynamic [Zaremba and Aste, 2014] Such as studies using VAR models, for example, by Lutkepohl (2005)[Lutkepohl, 2005], demonstrated how lagged relationships of the market variables can predict future movements in the same [Siggiridou and Kugiumtzis, 2016][Lutkepohl, 2005]. Such studies depict time-series analysis of predictive power in understanding the dynamics in financial markets.Within the UK, for example, empirical studies most frequently address the banking, real estate, and insurance sectors. In the work of [Shang et al., 2020], the authors go into detail on the question of causality that exists between these three sectors, finding strong interdependencies, especially during times of economic turmoil (Shang et al., 2020)[Shang et al., 2020]. This emphasizes the meaningfulness of time-series analysis when understanding such subtleties of market dynamics. In contrast, the hierarchical Granger causal model of [Cai et al., 2024] stands for the grouped time series analysis that provides much-valued insight into sector-specific linkages [Clements and Galv, 2024]. Another set of research work has examined The drivers from the macroeconomic indicators and their impact on financial markets. For instance, among those, it is revealed that interest rate volatility, inflation rates, and GDP growth rates significantly impact stock price, real estate value, and insurance premium.

Such macroeconomic factors often turn out to be the common threads running across different financial markets. In that respect also, machine learning methods come to the front. Recent studies, to make predictions for the market behavior have been conducted with integrated LSTM networks together with traditional econometric models that enhanced the accuracy of the forecasts.These work [Olaniyan et al., 2024] benefit from the benefits of sequential models because they can show non-linear interactions and temporal dependencies in a way that more traditional methods usually don't.Twenty-four researchers have shown that machine learning models based on causality make financial predictions more reliable, especially when the market is unstable [Oliveira et al., 2024]. Recent studies using real data have shed light on financial contagion, which is when a problem in one area or business spreads to other areas.For example, studying the COVID-19 outbreak and the financial crisis of 2008 taught us a lot about how linked things make systemic risks worse. The significance of having good models that can predict and lessen these kinds of events is emphasized [Clements and Galv, 2024].

3.3 The Role of Machine Learning in Financial Modeling

According to the findings of a study conducted by SiamiNamini and colleagues (2018), LSTM models achieve much better results than ARIMA models when it comes to predicting financial time series. This is particularly true in environments that are characterized by high levels of volatility.This being their results, it conclusively shows that LSTM learns from experiences and responds quite effectively to fluctuating market scenarios, depicting the flexibility this type of model can bring on board. [Namini and Namin, 2018]

This approach would help a lot in research about money, since the model is able to learn from experience, adapting to the market developments.

Bayesian testing of Granger causality in functional time series becomes an important development; it is of great relevance while dealing with complex time-series data, as seen by Sen et al. (2022)[Sen et al., 2022].

Granger causality analysis was first created for linear settings, but it has been further enhanced with machine learning recently in such a way that non-linear interactions are possible too. Indeed, the models of kernel-based Granger causality and those inspired by causality feature selection did wonders in the capture of the complicated dynamics of the market[Oliveira et al., 2024].

These techniques, combined with Bayesian optimization, create a very significant increase in model performance regarding systematic tuning of hyperparameters. For example, Bayesian optimization allows more efficient exploration of the parameter space, increasing the predictive models' accuracy.

Recent developments have also contributed to hybrid models that embody VAR with deep learning. These newer models exploit interpretability from VAR combined with predictive capabilities of LSTM-a very powerful framework for analyses of interdependencies among financial markets. In fact, various studies, including the one by Babii et al. in 2023 [Clements and Galv, 2024], prove that it can outperform stand-alone techniques on many occasions with respect to the aspect of accuracy and robustness.

More recently, applications have included the use of random forests and gradient boosting among other ensemble learning methods on financial forecasting. Using this group of models lowers the risk of overfitting because they mix results from various models.This gives the models more uses. For an instance, a research [Olaniyan et al., 2024]showed that these methods typically yield good results in predicting the stock market during times of volatility.

Another varietal technique that has been a surefire approach in financial modeling is transfer learning. In this new paradigm, it is possible to make much more accurate predictions with much less data by reusing models already trained on similar datasets.This is especially important in the financial markets, where good data is often hard to come by or costs a lot.

3.4 Gap in Literature

This is particularly relevant in the financial markets as dependable data is often expensive or hard to obtain. Considering the distinct characteristics of the UK economy, including its varied regulatory frameworks and market structures, this exclusion holds significant importance.

Additionally, Granger causality coupled with sophisticated machine learning methodologies, especially LSTM, remains sparsely applied within UK markets. Most the literature has either focused on standard econometric models or machine learning techniques in isolation but without exploiting the potential such a hybrid model may carry to uncover complex interdependencies.Another critical gap is in the assessment of the performance of models during times of high volatility. Disregarding the necessity of strong models adapting to the shocks of the markets, many current studies have welled on average performance measures.

This brings out the importance of building methods that could capture time-varying and possibly non-linear interdependencies within changing economic conditions.Lastly, although hybrid models that which combine econometric and machine learning models have great potential their application to the analysis of sector-specific interdependencies in the UK is limited. More sophisticated models, for instance, have rarely let one assess how macroeconomic shocks affect sectoral links [Oliveira et al., 2024, Clements and Galv, 2024]. By filling these gaps, the practitioner policymaker and the industry practitioner will have practical tools as well as contribute to scholarly knowledge.

There is a crying need for more extensive datasets that put together high-frequency data across different sectors. The absence of these essential datasets frequently hinders the opportunity to examine real-time interdependencies, thereby restricting the applicability of findings to ever-changing market conditions. Moving ahead the research should prioritize the development and an utilization of such datasets to bolster the robustness and relevance of analyzes within the financial markets.

CHAPTER 4

Data Sources and Preparation

4.1 Data Sources and Preparation

4.1.1 Sources of Data

For this study, we use financial data from two reliable sources; 1)Yahoo Finance [Finance, 1997]and the 2)London Stock Exchange[Exchange, 1801]. It reflects the trends in stock prices of a wide range of companies operating as banks, insurance, real estate, and so on. Such a combination enables the consideration of broader trends and relative performance, one of the key insights into how these financial markets have changed over time. This dataset will enable viewing of time trends, assessment of risks, and determination of the best investments. Both of these websites have accurate and thorough data on stock prices, which is very important for any research that wants to find a link between different financial markets.

4.1.2 How often and how long data is stored

We got daily stock prices for a long time as part of the info we gathered. Using daily data helps you keep track of small changes in the markets and gives you enough information to do a good analysis.This is the daily financial dataset from Yahoo Finance and the London Stock Exchange. It runs from January 3, 2000, to the present, containing the stock prices of large firms such as Barclays, HSBC, and Lloyds, among others.

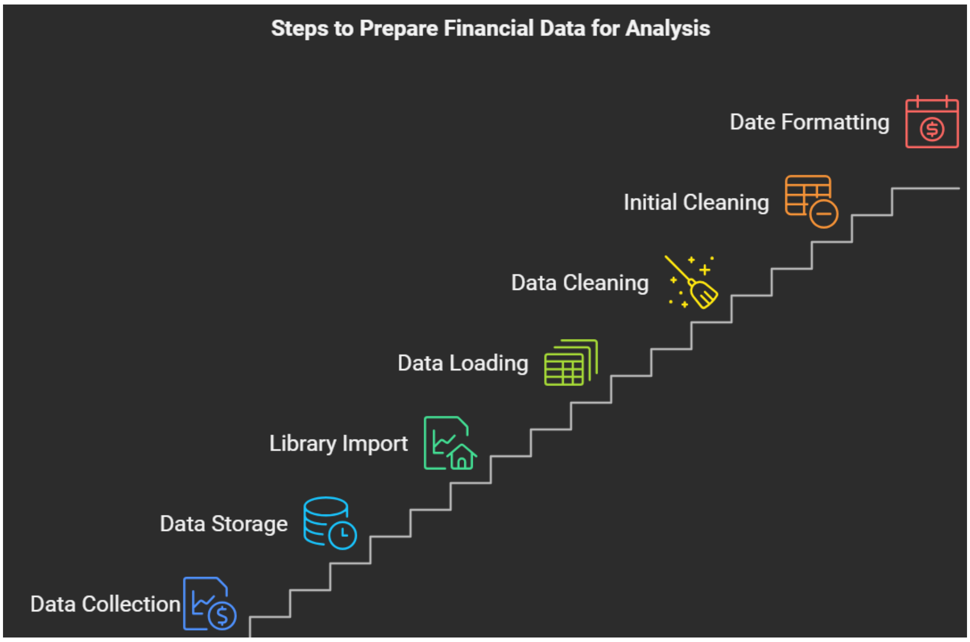

Figure 4.1: Data Resources and Initial Cleaning

4.2 Data Preprocessing

4.2.1 Importing Libraries and Loading the Dataset

CSV files are among the common ways of storing table data, and this is how data in this study were saved. The dataset was placed in a Pandas Data Frame a robust data format in Python that helps in the investigation and manipulation of data with great ease. Dealing with the data in Data Frame format made this activity straightforward. It thus makes the processing and research smooth.

1. Importing Libraries:

To work easily with the data, i have to import important libraries like Pandas and NumPy have been added. for the Visualization purpose , tools like Matplotlib and Seaborn were added for the first study and to make the data easier to see.

2. Loading the dataset:

Data set is in "$Final.CSV$" file is read into the Pandas Data Frame by using the $read_csv()$ method. The original data has been changed by now to make it easier to read and understand. Utilizing the head() method to look at the first few rows helped us check that the dataset was loaded correctly and then the data was organized correctly.

What's the Point of a Data Frame?

Table 4.1: Dataset Columns Description

Column Name	**Description**
`date`	Date of observation (from **01-01-2001** to **24-10-2024**)
`BARC`	Stock price of Barclays
`HSBA`	Stock price of HSBC
`LLOY`	Stock price of Lloyds
`NWG`	Stock price of NatWest Group
`STAN`	Stock price of Standard Chartered
`Unnamed: 6`	Empty column (no data)
`Unnamed: 7`	Empty column (no data)
`AV`	Stock price of Aviva
`DLG`	Stock price of Direct Line
`HSX`	Stock price of Hiscox
`LGEN`	Stock price of Legal & General
`PRU`	Stock price of Prudential
`BLND`	Stock price of British Land
`DLN`	Stock price of Derwent London
`GPE`	Stock price of Great Portland Estates
`HMSO`	Stock price of Hammerson
`LAND`	Stock price of Landsec
`SGRO`	Stock price of SEGRO
`SHC`	Stock price of Shaftesbury

- Rows and columns are easyto get to.

- Handling lost or inconsistent data is made easier, and it's easy to use the transforms that are needed for more in-depth study.

- This step set the stage next steps, which were to clean the data, line up the time marks, and use machine learning models.

4.2.2 Cleaning up data

To begin, the data had to be checked to make sure it was correct: 1. Any gaps in the data were filled with a method that uses the most recent known number. This made the data complete. 2. Outliers: We got rid of rare and extreme numbers so that the results wouldn't be misleading. 3. Align the time: Put all the data on the same schedule so it's easy to see how they compare.

4.2.3 Initial Cleaning

People who filled out a lot of sections were deleted.In order to get things done quickly, we used the forward-fill method to add the last few numbers. For completeness, this makes sure that all the data is right.

1. Removing Unnecessary Columns: Columns that were considered irrelevant to the analysis were removed. This helped in reducing noise and focusing on relevant data.Certain columns like 'GPE', 'HSX', 'HMSO', 'BLND', 'SGRO', and 'SHC' were deemed irrelevant and dropped.

2. Missing Value Handling: We got rid of columns that had a lot of empty entries,so we drop column who carrying more than 50 percent missing values. To keep things going smoothly, the forward-fill method was used to fill in the rest of the missing numbers . This makes sure that the information is complete without adding any mistakes.

3. Formatting the Date Column: The 'date' column was transformed into a proper datetime format to allow for time-series analysis and to align the time indices. Also, invalid rows, for example, those containing corrupted dates, were cleaned to maintain integrity.

4.2.4 Basic Statistical Analysis

To understand the dataset's structure, we computed key descriptive statistics:

1.Mean : The average value of the dataset provides an understanding of central tendency.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4.2.1}$$

2.Standard Deviation: Indicates the spread of the data around the mean, helping identify variability.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{4.2.2}$$

3.Minimum and Maximum Values: Highlighting the range of data to identify potential outliers.

We made sure we had a basic idea of the dataset by doing this analysis. This makes it ready for more in-depth exploratory analysis.

4.3 Exploratory Data Analysis (EDA)

Figure 4.2: Exploratory Data Analysis

4.3.1 Descriptive Statistics

Together with the initial data, more in-depth metrics, such as skewness and kurtosis, concerning each variable were analyzed. By doing this, one is assured to find non-normal distributions that might affect the functioning of the model.

4.3.2 Correlation Analysis

A correlation matrix was used to identify linear correlations between numerical components. Creating a heatmap that displays these correlations made it easier to identify features that are closely related and may affect model performance or lead to multicollinearity. The formula used for correlation (Pearson's correlation coefficient):where:

- $\mathrm{Cov}(X, Y)$: Covariance between variables X and Y.

- σ_X and σ_Y: Standard deviations of X and Y, respectively.

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{4.3.1}$$

There is a strong and positive linear link between two variables, which is measured by the correlation coefficient. It's worth looking into possible dependencies that have high positive or negative associations.

Key Insight:

The correlation analysis provided a clear picture of how variables relate, helping prioritize features for modeling and further analysis. A correlation matrix was made to identify the way different financial variables in the data set relate to each other. A correlation matrix depicts how strongly variables are related with one another:

- The numbers go from -1 to 1. If close to 1, it would reflect that a very strong positive relationship exists in the variables, hence indicating that as it goes up, it would without fail go up.
- A number close to -1 suggests that the relation is highly negative, in which case if one variable increases, the other variable tends to decrease.
- Where a number gets closer to 0, that would say that there isn't much relationship between the variables involved.

The heatmap for the correlation matrix is given below:

Significant Findings:

High Positive Correlations

According to a correlation coefficient of 0.94, BARC and NWG move heavily together, which is in line with similar market trends in the banking sector.

It's clear that LGEN and PRU are closely connected a ratio of 0.94 suggests this probably because both work in the insurance industry.

Negative Correlations:

Date and NWG (-0.81): displays an inverse trend over time, indicating that the variable has been dropping over time.

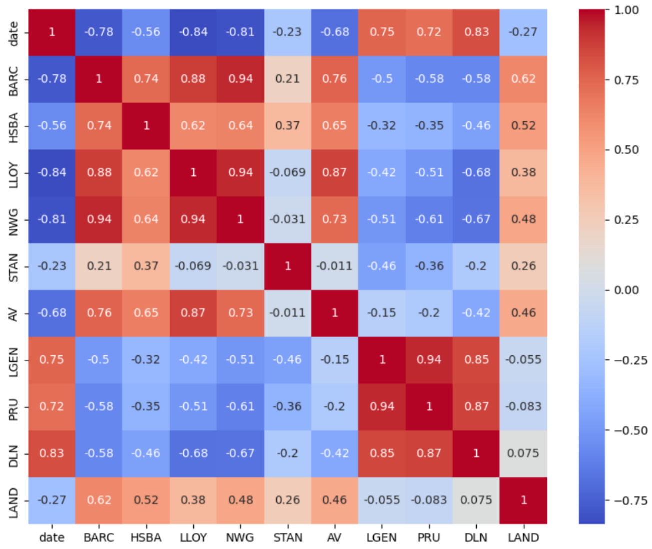

Figure 4.3: Correlation Heatmap

Ill-Fated Connections:

When looking at LAND and most other variables, it shows lower correlations with other variables, which is a reflection of how the real estate market behaves in particular.

The correlation matrix shows how many different areas are connected, like banks, insurance, and real estate. it is helpful For these reasons,

1. Selecting features means finding important factors that affect other variables and lowering multicollinearity.

2. Modeling Insights: Helping to improve machine learning models by looking at how traits depend on each other.

4.4 Stationarity Testing and Log Returns Calculation

4.4.1 Stationarity Testing

All good models, with a special emphasize on the time series, require that data be stable. Hence, statistically significant variables are mean and range, which remain constant over time. Because the normal stationarity tests passed, the conditions were met. They were changed into difference or logarithm tests if they were not stable. One of the prerequisites for modeling time-series, whether the data is stationary or not? for that the Augmented

Dickey-Fuller(ADF) test was done for every column's time series. The missing values from each column were removed, but not for the date; then it calculated the p-value and ADF Statistic it means column was stationary and prepared for analysis if the p-value was below 0.05; otherwise, it should have been stabilised with transformations such as differencing. This will ensure that the dataset meets the criteria for an accurate model.

$$ADF = \frac{\text{Standard Error of } Y_{t-1}}{\text{Coefficient of } Y_{t-1}} \tag{4.4.1}$$

4.4.2 Log Returns Calculation

So for achieving the stationarity we are using Log Returns ; Log returns were calculated to stabilize variance and make the data closer to stationary. The process involved: 1. Getting the right multidimensional columns out of the collection. 2.Log returns refer to the calculation of the difference between two numbers using the natural logarithm and then increasing by 100 to get the percentage of change. 3. Removing or filling in missing values created by the first preprocessing step of change. 4. Adding the log return entries to the dataset such that it can be used for modeling later.

This change improved our data's suitability for time-series models like VAR and ARIMA by meeting the standards for stationarity.

$$\text{Log Return} = 100 \times (\log(P_t) - \log(P_{t-1})) \tag{4.4.2}$$

where:

- P_t: Value (e.g., price) at time t,
- P_{t-1}: Value at time $t-1$,
- $\log$: Natural logarithm.

CHAPTER 5

Methodology

5.1 Introduction

The present chapter goes through the methodology of the study at great length, right from preprocessing to using advanced machine learning models for making predictions and analysis. Much emphasize is laid on clarity of approach so that the same could be comprehensible by several people with different technical backgrounds.

The chapter starts with the explanation of how the raw dataset was created, including the process for handling missing values, styling data for time-series analysis, and transforming non-stationary data into stationary data. Then, various models are implemented to analyze and forecast the movement of the financial markets, such as statistical models and machine learning models. We go into great detail on how to apply VAR, LSTM networks, and Granger causality analysis and show the capability of each in bringing out both linear and non-linear relationships.

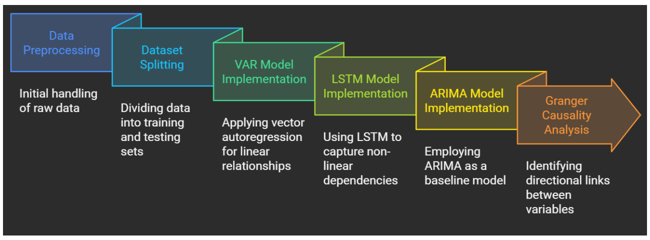

Figure 5.1: Flow of Methodoly for Modeling Algorithm

It bridges this gulf between technical complexity and useful application by giving each step in such a way that readers not necessarily having a technical background can also grasp how to pursue the method. Also, all the evaluation metrics that have been used in the experiment in order to make judgments about performance will be reviewed here to allow a proper understanding of study results.

5.2 Splitting the Data

For a more accurate model review, the dataset is split into training and test subsets:

Training Data

(90 percent): For building models and making them work better.

Testing Data

(10 percent): Set aside for testing on data that hasn't been seen yet, or "out-of-sample evaluation."

This split will ensure the model generalizes well and does not overfit.

5.3 Model Implementation

5.3.1 Vector Autoregression (VAR)

Purpose:

A linear relationship between the different time series was identified using a VAR model. One predicts the value of a variable by looking at its historical values as well as the values of the other variables in the list.

Implementation:

First, we have split the data into a training set of 90 percent and a test set of 10 percent in applying the VAR. The model was estimated on training data, and performance evaluation was carried out using testing data. We employed statistical techniques to determine the

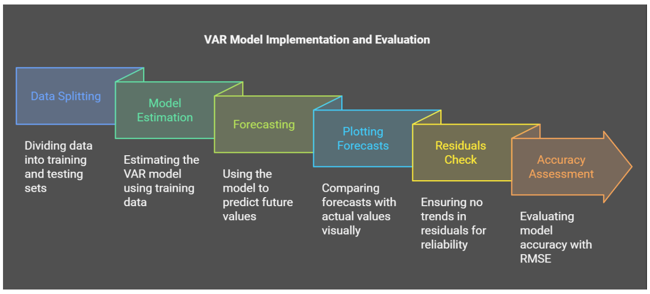

Figure 5.2: VAR model implementation

optimum number of historical observations or lags that would be used by the model in finding the optimum between complexity and good results, and we get our best value at first time so we used value at first lag . We then used the model to forecast future values based on the training data. We plotted these forecasts against the actual values for the testing set using simple time-series plots. We checked that the residuals-or remaining mistakes-did not have any trends to ensure that the model was reliable. Lastly, we have used the Root Mean Squared Error, a widely used metric that indicates how well the predictions matched the actual values, to gauge the model's accuracy.

$$Y_t = c + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \varepsilon_t \tag{5.3.1}$$

where:

- P_t: Value (e.g., price) at time t,
- P_{t-1}: Value at time $t-1$,
- $\log$: Natural logarithm.

5.4 Long Short-Term Memory (LSTM)

Purpose:

LSTM models are particularly effective in capturing non-linear dependencies and long-term memory in sequential data. These capabilities make them ideal for financial time series forecasting.

Implementation:

In order to forecast the log values for every column in the dataset, we used the Long Short-Term Memory (LSTM) model. The data needed transformation so it could be fed into the LSTM, which works with sequential data. A neural network was built with an LSTM layer of 50 neurones, a dropout layer to stop overfitting, and a thick output layer. We used the Adam optimizer to put together the model and trained it for 50 iterations, with a validation split to check how well it was doing during training. For the test data, predictions were made, and the Root Mean Squared Error (RMSE) was found to see how well the model worked. To see how well the model learnt, training and validation loss graphs were drawn. Time-series plots showed how well the model worked by comparing real log returns to the predicted ones for each column. Finally, RMSE results were put together for all columns, which showed clearly how well the LSTM model predicted. An ADF test was performed to check if the time series data in each column was stable, which is needed for time-series modeling. Missing values were removed from each field (except for the date), and the ADF statistic and p-value were calculated. If the p-value was less than 0.05, it was believed that the column was stable and ready for analysis. If it was greater than 0.05, changes like differencing were needed to make it stable. This was done to ensure that the dataset met the requirements for proper modeling.

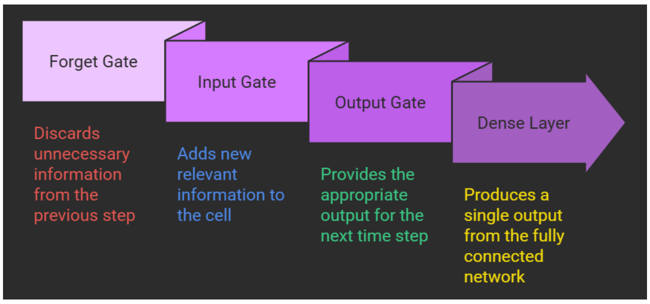

Figure 5.3: working layesrs of LSTM

1. Forget gate : This is used to discard information from the previous step that is no longer needed.

2.Input gate : This helps to add which new information is needed to add to the cell.

3. Output gate : Used to give right output for the next time step.

4. Dense layer: This is a fully connected Neural Network and give only single output at time.

5.5 Autoregressive Integrated Moving Average (ARIMA)

Purpose:

In order to compare more sophisticated techniques like LSTM and VAR, ARIMA was used as a baseline model. It uses moving averages, autoregression, and differencing to predict future values from historical data.

Implementation:

ARIMA model was used to forecast log returns of this dataset. Each column initiated the ARIMA model with fixed parameters of p, d, and q based on the earlier test results(we used (1, 0, 1) for best results). After fitting the model on the training set, the predictions for the test set were made. The differences between the real and predicted values were plotted as a measure of accuracy by using RMSE. A time-series plot was made to show the effectiveness of the model by comparing the real log returns with those predicted for columns where it worked well. Data processing mistakes were recorded, and RMSE results for all columns were put together in a summary table to give a clear picture of how well the ARIMA model worked across the whole dataset.

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{5.5.1}$$

where:Y_t: Current value of the time series. $\phi_1, \phi_2, \ldots, \phi_p$: Coefficients of the Autoregressive (AR) terms, showing dependence on past values $(Y_{t-1}, Y_{t-2}, \ldots)$. $\theta_1, \theta_2, \ldots, \theta_q$: Coefficients of the Moving Average (MA) terms, showing dependence on past errors $(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$. ε_t: Current error (random noise). p: AR order (number of lagged values used). q: MA order (number of past error terms included).

5.6 Granger Causality

Purpose:

To find directional links between variables and ascertain if one may predict another, Granger causality tests were used.

Implementation:

We used the Granger Causality test to determine the variables within the dataset that are related in a particular way. For each pair of variables, the test will check whether one variables past values could be used to make an educated guess at the value of another variable. Provided the p-value for the test across many lags was less than 0.05, it means that one variable "Granger-causes" another. To keep track of these connections, a binary causality matrix was created. A number of 1 meant that there was causality. The result was plotted as a network graph, where nodes are variables and directed lines show which event can cause another. This representation made it easy to identify how variables change over time and showed which ones are significantly directional dependent in the dataset.

$$\text{Var}(Y_t \mid Y_{t-1}, Y_{t-2}, \dots) > \text{Var}(Y_t \mid Y_{t-1}, Y_{t-2}, \dots, X_{t-1}, X_{t-2}, \dots),$$

here, The variance operator is shown by Var.
The variables subscripts are Y_t, Y_{t-1}, and X_{t-1}.
This is what "|" means: "conditioned on."
The "..." show that the sentence can go on. [Granger, 1969]

5.7 Evaluation Metrics

Model performance was assessed using several metrics:

1. Mean Square Error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.7.1}$$

Explanation of Terms:

- n: Total number of observations.
- y_i: Actual value of the i-th observation.
- $\hat{y}_i$: Predicted value of the i-th observation.
- $(y_i - \hat{y}_i)^2$: Squared difference between the actual and predicted values.

This formula calculates the mean of squared errors, making it a common metric to evaluate the accuracy of predictive models while penalizing larger errors more heavily.

2. Root Mean Squared Error (RMSE): Emphasizes larger errors by squaring residuals, making it more sensitive to significant deviations.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.7.2}$$

Explanation of Terms:

- n: Total number of observations.
- y_i: Actual value of the i-th observation.
- $\hat{y}_i$: Predicted value of the i-th observation.
- $(y_i - \hat{y}_i)^2$: Squared difference between the actual and predicted values.

This formula computes the square root of the average of squared differences between actual and predicted values, emphasizing larger errors more heavily than smaller ones.

5.8 Summary

This methodology made it possible to examine and forecast financial market interdependencies in an ordered way. Combining more traditional models like VAR and ARIMA with state-of-the-art machine learning techniques like LSTM provided a comprehensive examination of both linear and non-linear interactions. The investigation was deepened using Granger causality, which showed directional dependencies, and the models' effectiveness was validated by applying strong assessment measures.

CHAPTER 6

Results and Analysis

6.1 Model Performance and Making Comparisons

How Well the Models Worked ? I tested three models to see how well they could predict changes in stock prices:

6.1.1 VAR Model:

This model helped us understand how different factors affect each other over time. It had a few mistakes, which means that it was not pretty accurate. A graph of real versus expected log returns obtained from the VAR model for several financial variables across time shows that, in fact, this model can properly describe linear dependencies: during periods of stability, the predicted values were not very close to the real ones. However, the differences during highly volatile times suggest that VAR is not very good at coping with non-linear dynamics. VAR is a potent tool in investigations into short-run patterns and relationships within multivariate time-series data that are dominated by linear dependencies.

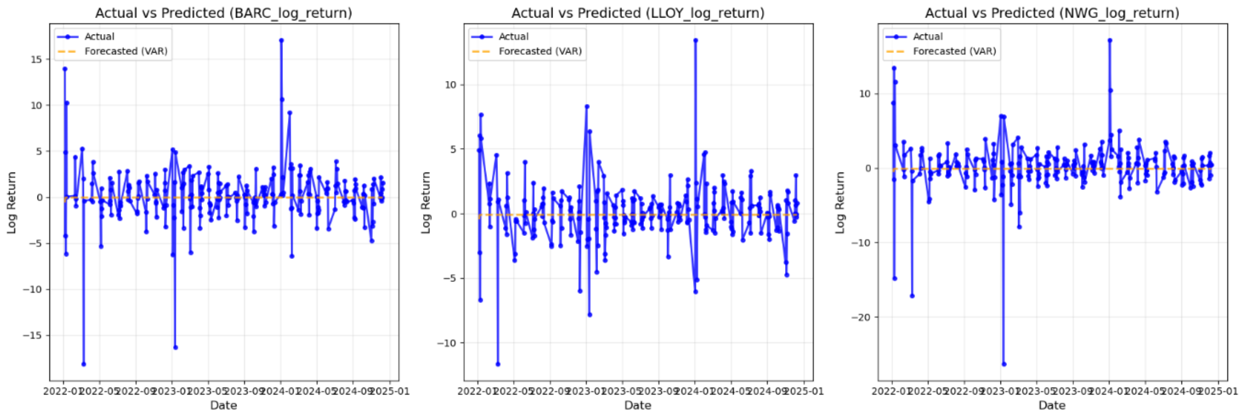

Figure 6.1: Graph for BARC_log_return, LLOY_log_return, and NWG_log_return

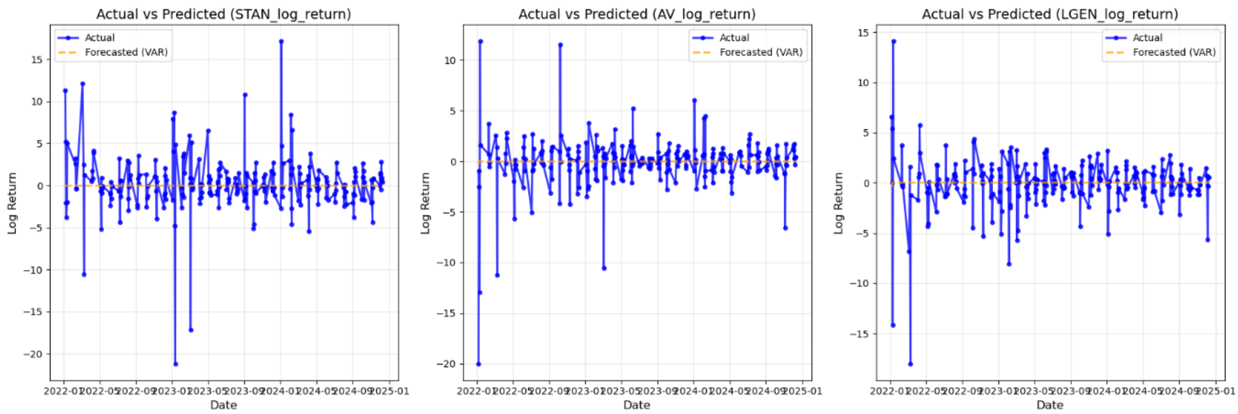

Figure 6.2: Graph for STAN_log_return, AV_log_return, and LGEN_log_return

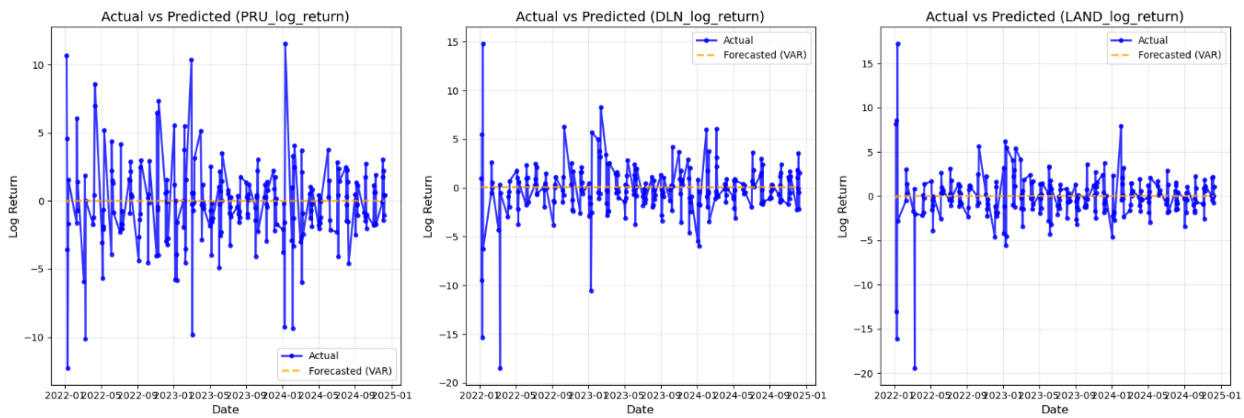

Figure 6.3: Graph for PRU_log_return, DLN_log_return, and LAND_log_return

6.1.2 LSTM Model

This model was very good at dealing with complicated patterns and connections. To work at its best, though, it needed more computer power and careful tweaks.The graph of the LSTM model indicates that it has performed well with non-linear patterns and ongoing connections of financial data. These smooth predicted values and their agreement with real values during times of change show that LSTM can adapt to changing market behavior. This works especially well in the case of long-term predictions where relationships are highly complicated. Unfortunately, this is a very resource-consuming method since it needs to be computed and its hyperparameters need to be tuned.

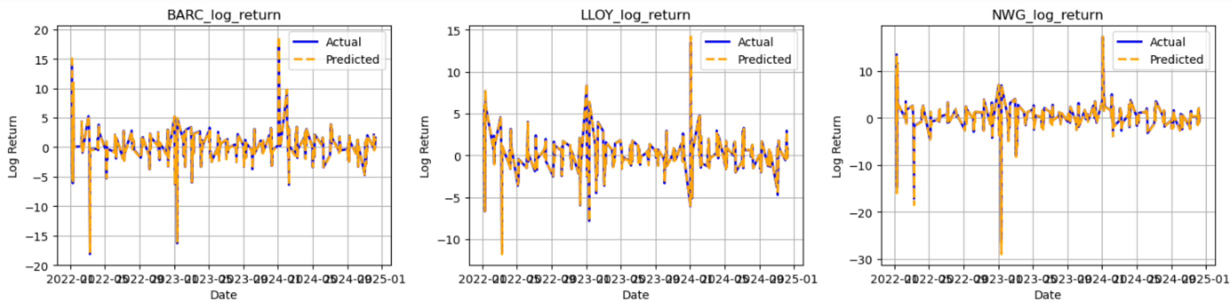

Figure 6.4: Graph for BARC_log_return, LLOY_log_return, and NWG_log_return

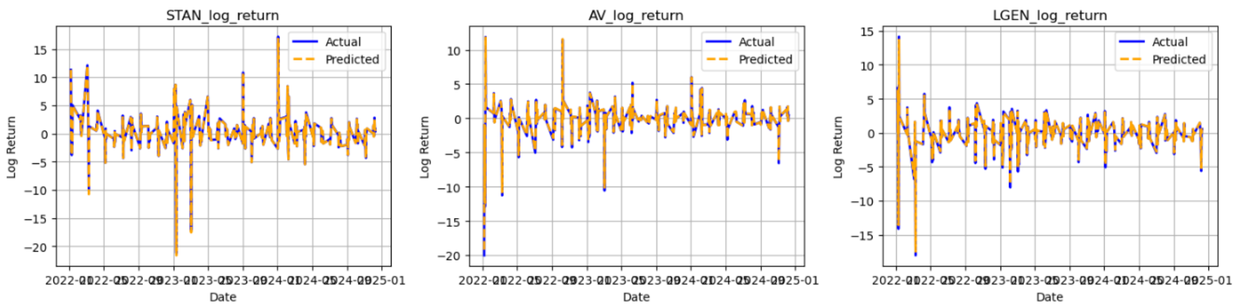

Figure 6.5: Graph for STAN_log_return, AV_log_return, and LGEN_log_return

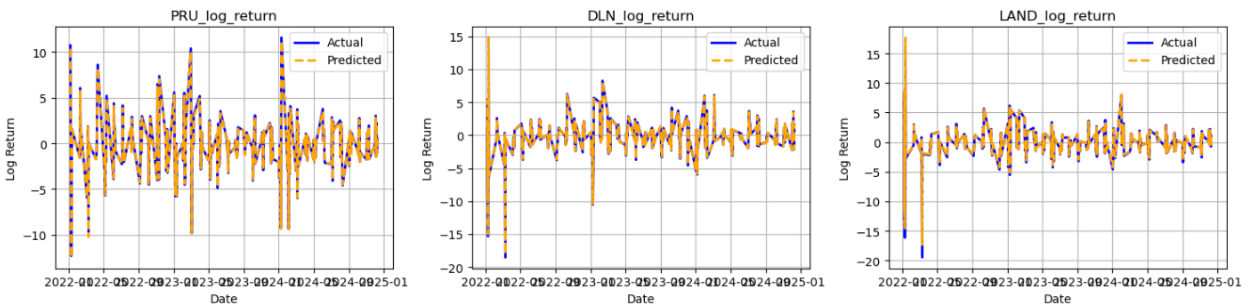

Figure 6.6: Graph for PRU_log_return, DLN_log_return, and LAND_log_return

6.1.3 The ARIMA model

The graph plots a single-variable prediction and real values against the prediction with time. Deficiencies in ARIMA come forward when dealing with multivariate relationships or non-linear trends. It works as an excellent baseline model for univariate time-series data. The high errors of this tool during market volatility prove it is unable to record complex dependencies, which makes it of less use for in-depth market analysis. Nonetheless, ARIMA remains a decent option for simple prediction on a single variable at a time .

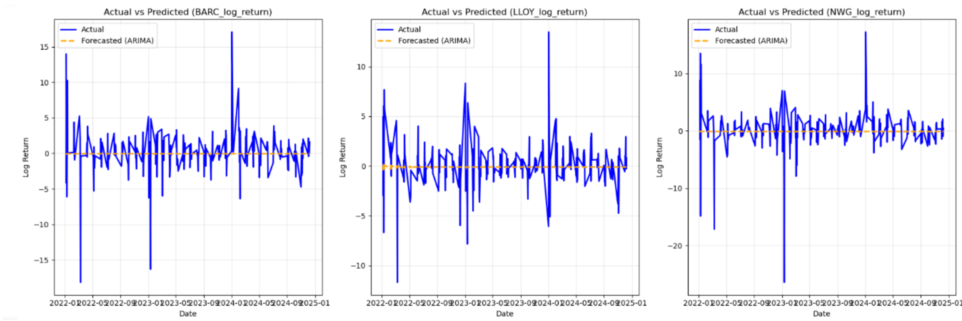

Figure 6.7: Graph for BARC_log_return, LLOY_log_return, and NWG_log_return

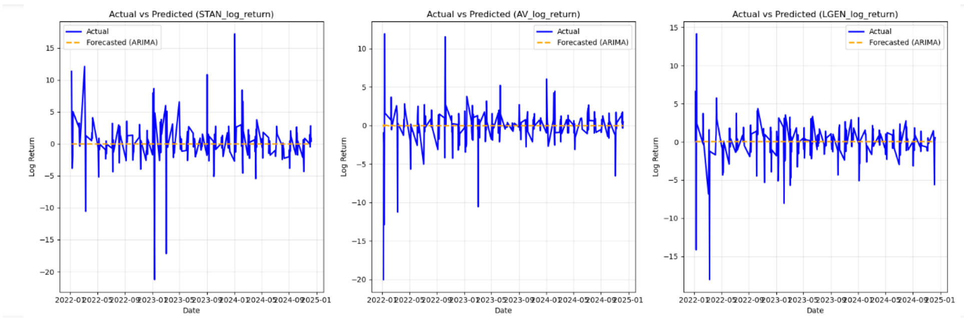

Figure 6.8: Graph for STAN_log_return, AV_log_return, and LGEN_log_return

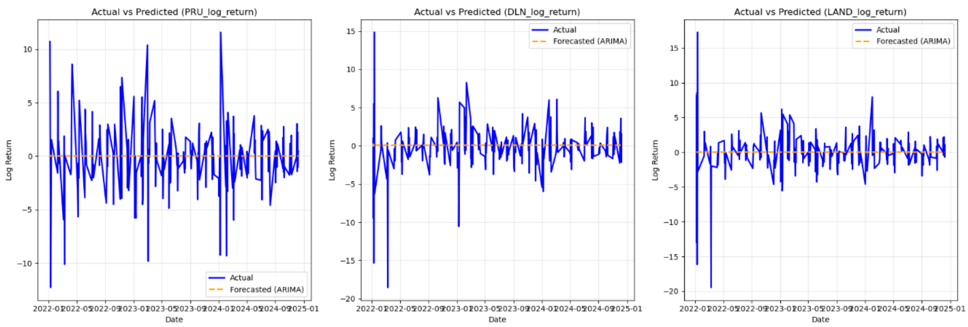

Figure 6.9: Graph for PRU_log_return, DLN_log_return, and LAND_log_return

6.1.4 All Columns Combined Actual vs Predicted Graphs

We create composite graphs for each log return column in this step.

This graph helps us visualize how the different models are performing. On each graph, the real log returns are shown as the black line with circles. Also shown are the numbers that three models say they should be:

1. VAR: the squared-off red line shows how two or more time series are linked.

2. Deep learning is used by LSTM (the green dashed line with circles) to predict future values based on past values.

3. ARIMA: This model looks at time series to find patterns and trends (orange straight line with crosses).

It's easy to see from the graph how well each model predicts the real numbers over time.

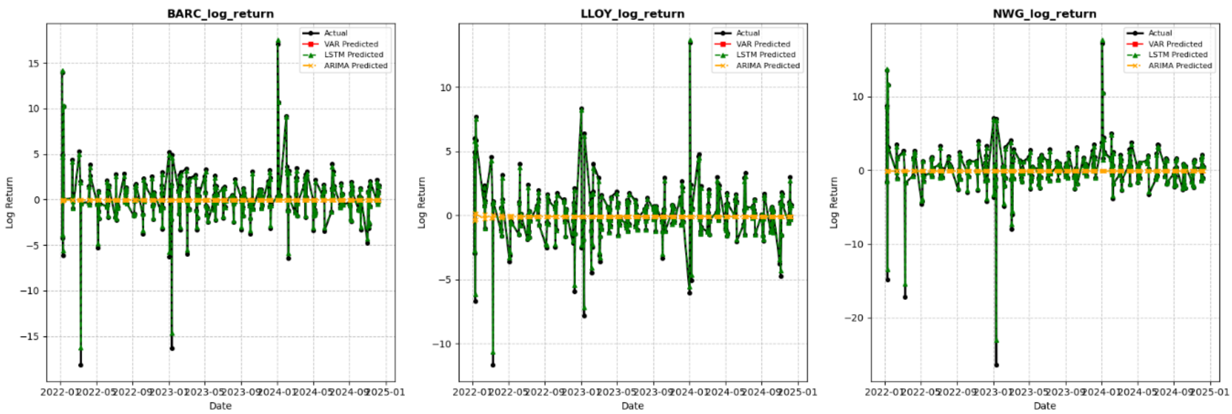

Figure 6.10: Combine VAR,LSTM,ARIMA model Graph For BARC_log_return, LLOY_log_return, and NWG_log_return

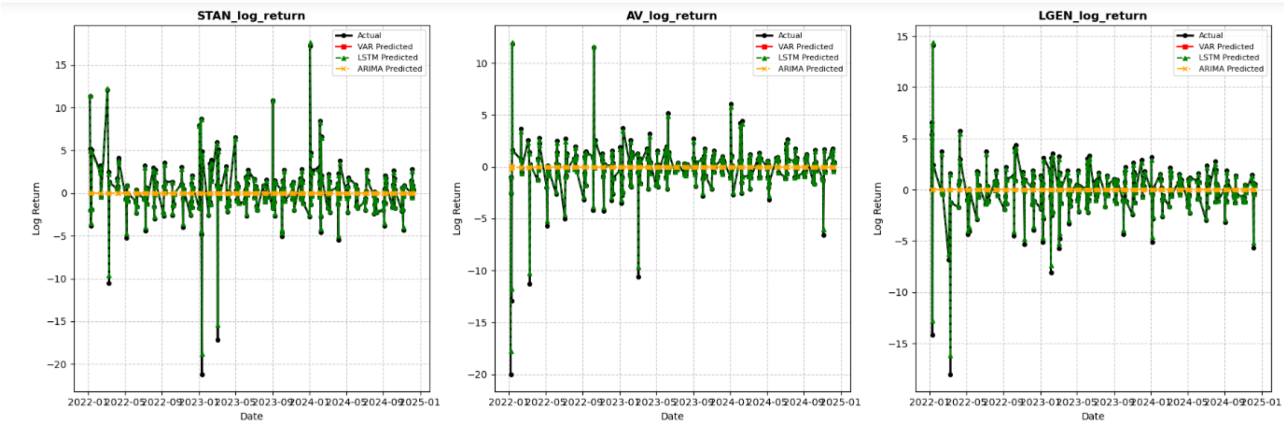

Figure 6.11: Combine VAR,LSTM,ARIMA model Graph For STAN_log_return, AV_log_return, and LGEN_log_return

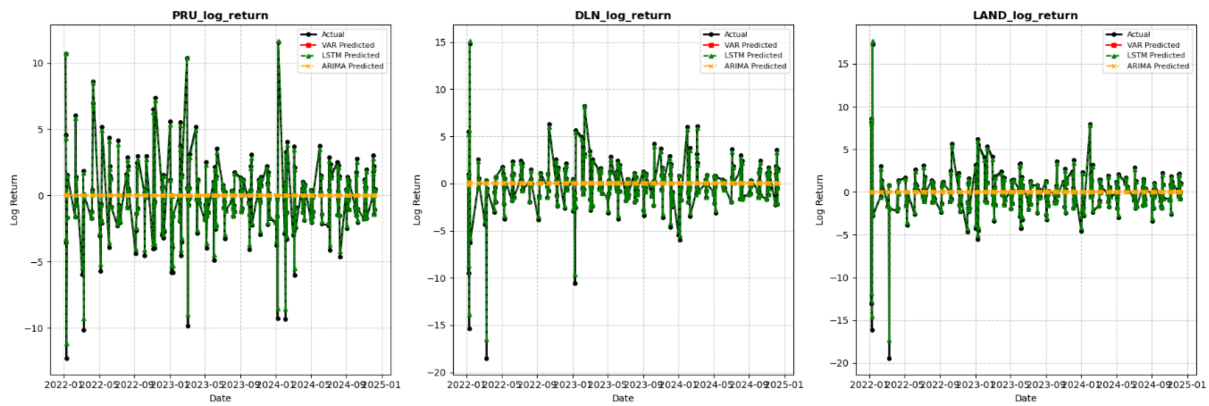

Figure 6.12: Combine VAR,LSTM,ARIMA model Graph For PRU_log_return, DLN_log_return, and LAND_log_return

Below table contains the RMSE values of all models for all columns. The best model for each column is highlighted, as it has the lowest RMSE compared to the other two models.

Column	VAR RMSE	**LSTM RMSE**	ARIMA RMSE
BARC_log_return	3.082023	**0.117089**	3.085677
LLOY_log_return	2.334694	**0.089409**	2.343927
NWG_log_return	3.423654	**0.136008**	3.424583
STAN_log_return	3.338469	**0.110950**	3.338472
AV_log_return	2.616114	**0.056839**	2.615633
LGEN_log_return	2.604680	**0.125941**	2.605387
PRU_log_return	3.041428	**0.142563**	3.039639
DLN_log_return	2.844030	**0.090314**	2.849409
LAND_log_return	2.930616	**0.169963**	2.930959

Table 6.1: Comparison of RMSE values for different models.

- The LSTM model outperformed both VAR and ARIMA on all columns, underlining its capability to exhibit complex patterns in financial time series and describe non-linear dependencies.
- ARIMA and VAR can work well for some variables but show higher RMSE numbers, meaning they have trouble with dynamic relationships and non-linear trends.
- Practical Implications: Drawing a clear difference from the results obtained demonstrates that LSTM would be the best for modeling financial market behavior for time-ahead prediction.

According to the chart, LSTM consistently gets the lowest RMSE, which means it is better at making predictions than the other models.

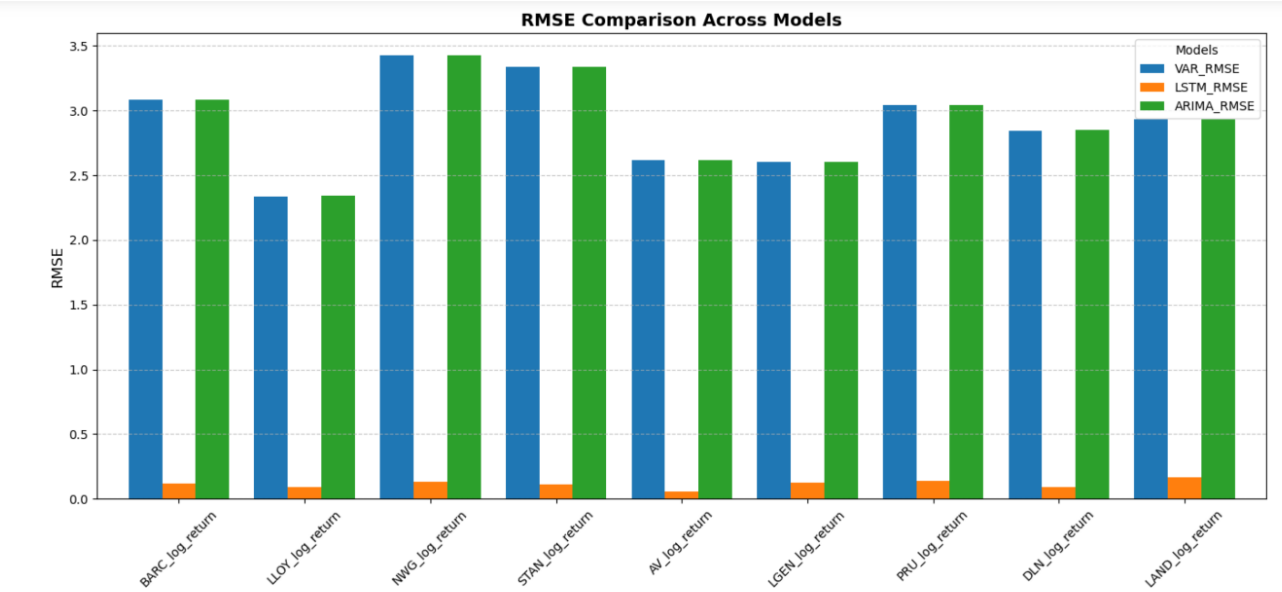

Figure 6.13: RMSE Comparison Across All Models.

It clearly shows how well the VAR, LSTM, and ARIMA models can predict log returns for different financial variables. Overall, the LSTM model constantly gets the lowest RMSE across all variables, showing that it is better at dealing with complicated, non-linear relationships in financial data. Although VAR and ARIMA models are better at predicting complex dependencies and changing patterns, their RMSE numbers are higher. LSTM was selected for my paper as the best model, since it is an effective and reliable tool in financial market prediction.

Now , for finding our best column we have to see which column is giving us lowest average for all model ,for that we calculated Average of RMSE for each column which will give us best column which perform best on all model as compare to others.

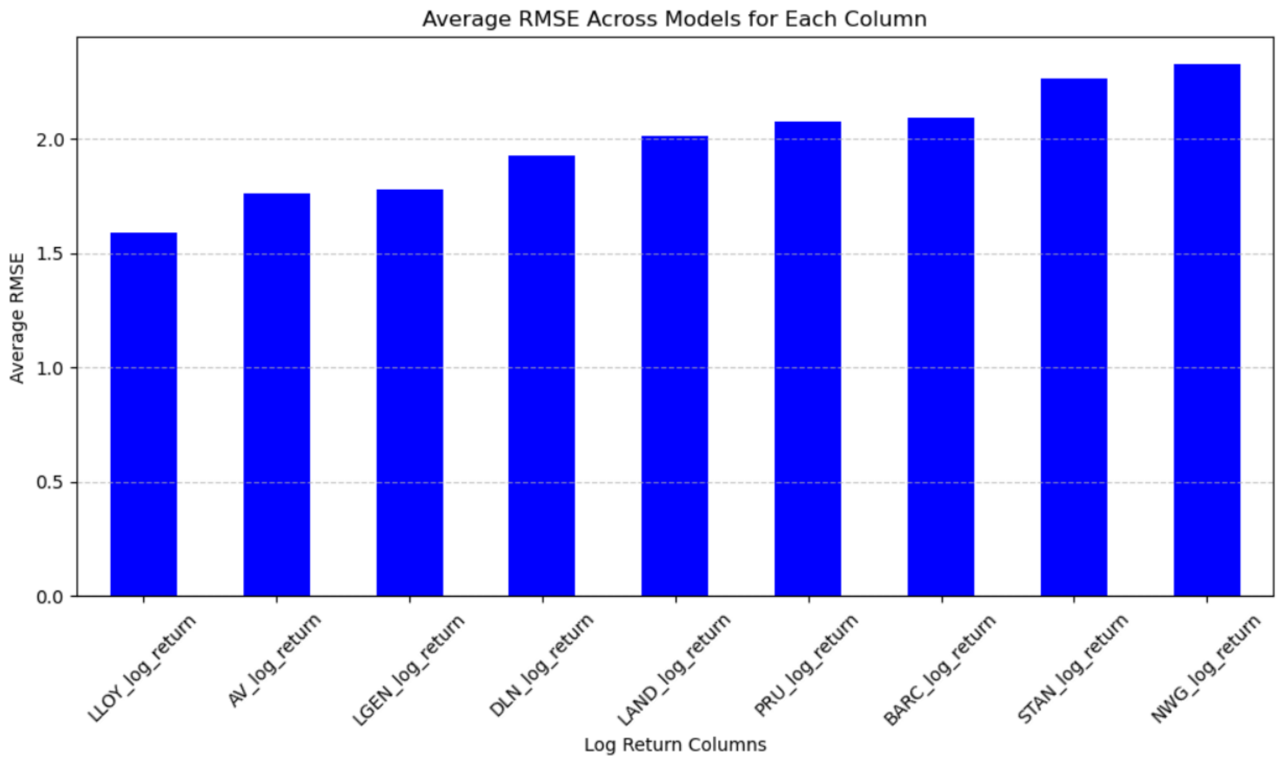

Figure 6.14: Average RMSE Across for each column

Therefore from above result we have successfully showed the best column of our dataset which is from banking ; LLOY.

6.2 What We Can Learn from the Results

Several factors were found to have a bigger effect on the market than others. For instance, some terms in the dataset had a strong link to changes in stock prices, which made them useful for figuring out market trends.

6.2.1 Granger causality and Effect and Connections

To determine which variables Granger cause each other, we employed Granger causality tests: You might learn something about one element from the first one. Each of the other elements influenced the others in turn. They show how different parts of the market are linked more clearly.

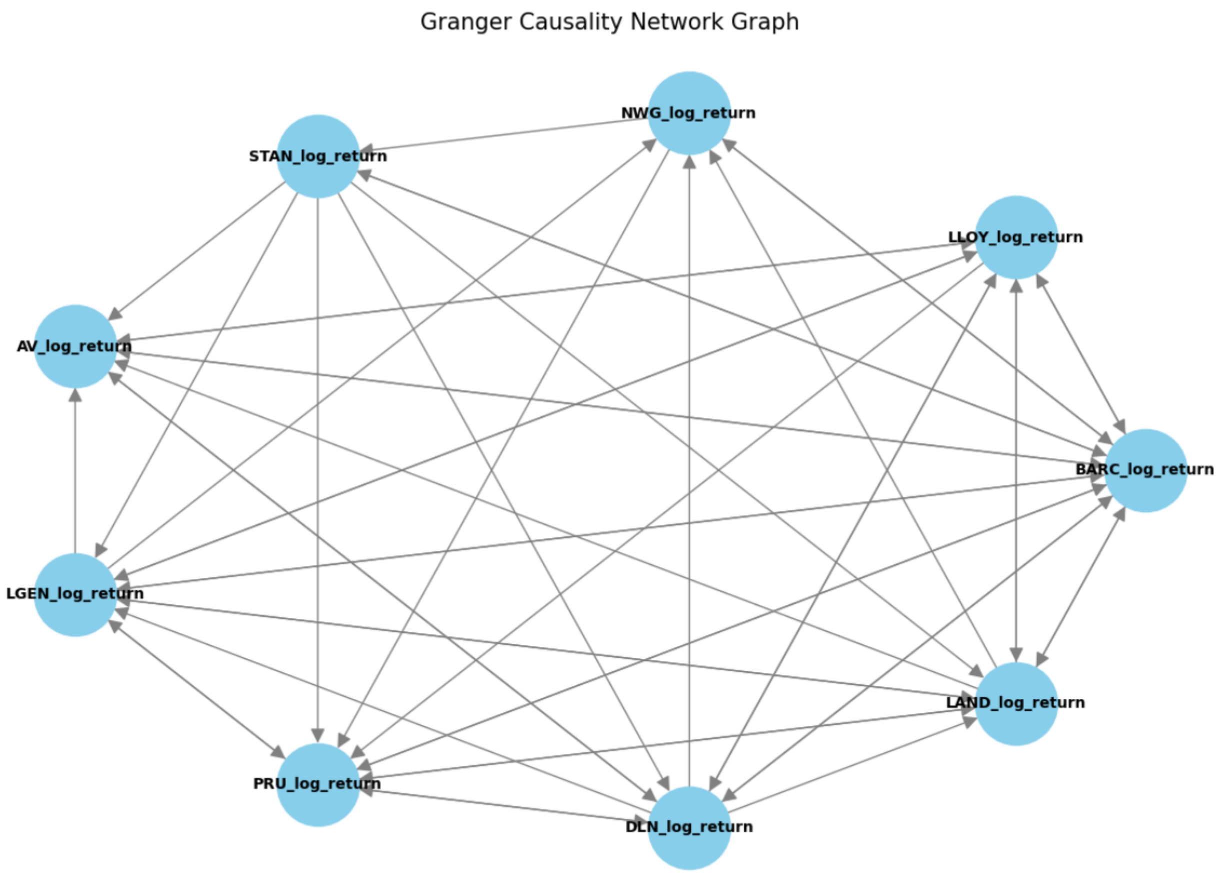

Figure 6.15: Granger Casuality Network Graph

This graph shows the Granger Causality Network, which shows how log-return factors are connected in a directional way. Every arrow shows a Granger causality link, which means that the source node helps determine what the target node will be. Like this: 1.The outgoing lines of 'BARC' show that it affects more than one variable. 2.Strong connections between nodes like 'LGEN' and 'DLN' show how they can predict what will happen to other nodes. The graph shows clearly how the variables are linked in terms of how they can be used to make predictions.

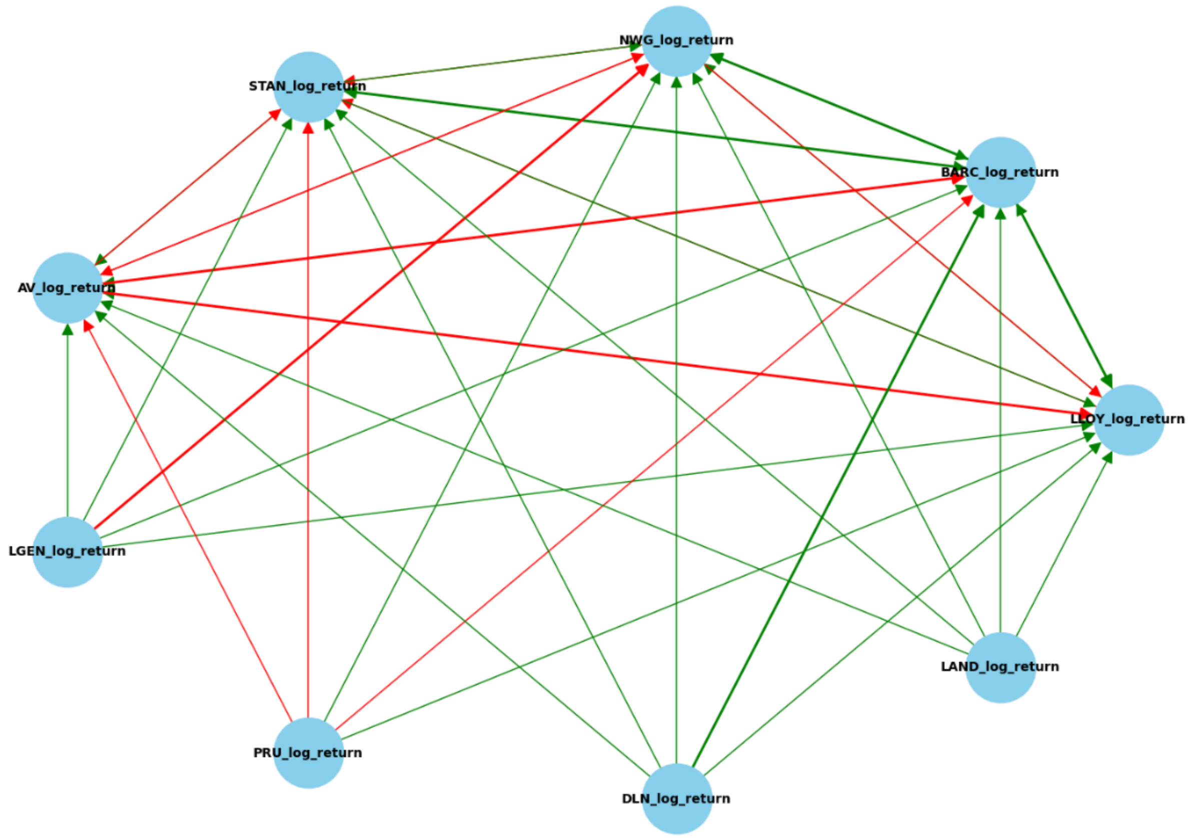

Figure 6.16: Relationship Network Between All Columns

This graph shows the Relationship Network Including All Columns and depicts how different log return factors are related to one another. Notice:

Green arrows depict the relationships in which one variable has a positive impact on another and vice versa. Relationships with red arrows have a negative impact. For instance, BARC , LLOY , and LGEN all have a large number of links, which indicates how strongly they impact other variables. Graphs can show interactions between the factors, hence helping to identify main causes and the way changes are affecting the network. This allows a better understanding of the different dependencies and changing relations across financial factors using this representation.

6.3 Summary

This chapter explained how we got the data ready, tried three models, and figured out what the results meant. We came to the following conclusions, among others ; The VAR model told us the most about how the factors were related.The models acted in different ways. LSTM did well with more complicated patterns, while ARIMA is simpler . By Granger Tests of Causality we will know more about interdependency between our columns and more important in our last section we also calculated positive and negative interdependency between each column .

CHAPTER 7

Conclusions

By carefully analyzing multiple models, this study sought to identify the most effective model for forecasting market movements and to comprehend the interdependencies across the financial markets in the United Kingdom. Through an analysis of the models' benefits and drawbacks, we learned how to successfully use financial time-series data for forecasting and decision-making. Additionally, the study highlights the benefits and drawbacks of different tactics and provides helpful recommendations for practical applications.

7.1 Key Findings

The present study contributes to this rapidly growing area of financial data analysis by providing the professional with useful tools in furthering our understanding of how financial markets work. This provides a sound base for future improvements in the area of financial prediction, based on methods that have proved their worth, and by introducing new techniques. Ensuring that the data was clean, consistent, and well-formatted for time-series analysis was one of the most critical aspects of this work. Handling missing values, aligning time patterns, and transformations such as log returns and differencing were some of the methods used in cleaning the data to prepare it for accurate modeling. This preliminary work assured that all models would perform well and yield useful results.

Markets that depend on each other: 1) Granger causality tests and correlation matrices show that the UK's financial markets are very dependent on each other. It became clear that assets like $BARC_log_return$ and $LLOY_log_return$ were very linked to each other, which means they could change larger market trends. 2) When interdependencies are positive, they show chances for portfolio growth, and when they are negative, they show chances for balancing.

How well the models work: **Vector AutoRegression** (VAR) Did a great job of capturing the linear relationships between several time series. **Long Short-Term Memory** (LSTM) was Good at capturing connections that don't follow a straight line. Even though it looked good, the performance varied across columns, and it took a lot of computing power and however, it was developed with hyperparameters that need to be tuned carefully from overfitting. The **AutoRegressive Integrated Moving Average** (ARIMA) method worked well for one-variable series but not so well for multidimensional ones, with higher RMSE values and less ability to handle complex interdependencies.

7.2 Comparisons of models

The first model is called vector autoregression (VAR): VAR was good at showing how factors were connected in a straight line. It made very accurate predictions with almost no mistakes, which made it very useful for data that had stable relationships.but for non-linear it made a lot mistakes and thats why it did not perform well.

LSTM was great at finding trends that weren't simple or linear. This actually made it great for working with statistics that are always changing and are hard to guess. The LSTM model outperformed both VAR and ARIMA on all Data, underlining its capability to exhibit complex patterns in financial time series and describe nonlinear dependencies. ARIMA(AutoRegressive Integrated Moving Average): In many respects, at least for a time series with only one variable. Clearly, Dealt poorly with the dataset's complex multivariate links.

A Granger Analysis of Causality Granger causality helped us understand how different variables are connected. This method showed how some factors influenced others over time, which helped in the identification of leading indicators and further helped in understanding market movements. It added more depth to research by showing trends that are not easily seen with traditional modeling methods.

7.3 Applications

Results from this study can be very instrumental to financial analysts, lawmakers, and investors in real life. This study portrays how different factors influence and interact with each other, hence enabling one to make improved forecasts and decisions. For example: Granger Causality; It would be feasible to get an insight into the early detection of market movements so that the investor may enable informed choices; **1)Portfolio Maximization :** Utilize growth using positively correlated assets such as $BARC_log_return$.**2)For risk control:** diversification techniques including negatively correlated assets like AV_log_return, **3)Forecasting in the market:** Predicting multi-asset movements benefits much from VAR's capacity to record interdependencies, **4)The non-linear features** of LSTM can be applied in situations including dynamic and irregular patterns, **5)Strategies for Trade:** By using predicted correlations among assets, insights from Granger causality tests and network visualisations can direct active trading decisions, **6)Model Selection:** From simple models, ARIMA to most complex, LSTM, the experts can select any of them based on the complexity of data and type of research.

These tools and techniques now may be used by stakeholders for understanding how markets work, controlling risks, and seizing opportunities.

7.4 Limitations and Future Scope

The constraints of the assessed models point out important areas for development and intellectual stimulation. Since the VAR model generates linear links, it can only partially reflect the non-linear dynamics sometimes observed in financial markets. LSTM is computationally complicated and highly sensitive to hyperparameter modification; so, even if it is powerful for non-linear patterns, demanding significant work and knowledge for optimization is necessary. Although ARIMA is useful for univariate time series with trends or seasonality, it is not very helpful for multivariate applications, therefore lowering its value for datasets marked by complicated interdependencies. These constraints highlight the need of choosing or aggregating models depending on the particular traits of the data and analytical objectives. **Looking to the future;** Combining models like VAR and LSTM to use both linear and non-linear ideas at the same time is called model enhancement. by adding small economic or financial factors ,events all over the globe to the data of the model it will become more correct. New Approaches; To improve the accuracy and dependability of our estimates, we are looking into mixed models or group methods.

7.5 Conclusion

This research demonstrates how crucial it is to pick the appropriate model based on the details of the data and the demand of the application. This information can be best found using the VAR model, which can show how various factors are connected.LST can help with a lot of non-linear trends in a collection. It needs to have enough processing power and not be too tight. People can make better financial market decisions by studying the data more to get better predictions.

Bibliography

[Cai et al., 2024] Cai, R., Wu, Y., Huang, X., Chen, W., Fu, T. Z. J., and Hao, Z. (2024). Granger causal representation learning for groups of time series. *Science China Information Sciences*, 67(5).

[Clements and Galv, 2024] Clements, M. P. and Galv, A. B. (2024). *Handbook of Research Methods and Applications in Macroeconomic Forecasting*. Edward Elgar Publishing.

[Exchange, 1801] Exchange, L. S. (1801). London stock exchange | london stock exchange. Accessed December 21, 2024.

[Finance, 1997] Finance, Y. (1997). Yahoo is part of the yahoo family of brands. Accessed December 21, 2024.

[Granger, 1969] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

[Lutkepohl, 2005] Lutkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Nature.

[Mangram, 2013] Mangram, M. E. (2013). A simplified perspective of the markowitz portfolio theory. Accessed December 21, 2024.

[Namini and Namin, 2018] Namini, S. and Namin, A. S. (2018). Forecasting economics and financial time series: Arima vs. lstm. *ArXiv*, abs/1803.06386.

[Olaniyan et al., 2024] Olaniyan, J., Olaniyan, D., Obagbuwa, I. C., Esiefarienrhe, B. M., Adebiyi, A. A., and Bernard, O. P. (2024). Intelligent financial forecasting with granger causality and correlation analysis using bayesian optimization and long short-term memory. *Electronics*, 13(22):4408.

[Oliveira et al., 2024] Oliveira, D. C., Lu, Y., Lin, X., Cucuringu, M., and Fujita, A. (2024). Causality-inspired models for financial time series forecasting. Accessed December 21, 2024.

[Sen et al., 2022] Sen, R., Majumdar, A., and Sikaria, S. (2022). Bayesian testing of granger causality in functional time series. *Journal of Quantitative Economics*, 20(S1):191–210.

[Shang et al., 2020] Shang, H. L., Ji, K., and Beyaztas, U. (2020). Granger causality of bivariate stationary curve time series. *Journal of Forecasting*, 40(4):626–635.

[Shojaie and Fox, 2021] Shojaie, A. and Fox, E. B. (2021). Granger causality: a review and recent advances. *Annual Review of Statistics and Its Application*, 9(1):289–319.

[Siggiridou and Kugiumtzis, 2016] Siggiridou, E. and Kugiumtzis, D. (2016). Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model. Accessed December 21, 2024.

[Zaremba and Aste, 2014] Zaremba, A. and Aste, T. (2014). Measures of causality in complex datasets with application to financial data. *Entropy*, 16(4):2309–2349.