**Can Machine Learning Techniques Use Economic Variables to Predict Levels Of Public Debt?**

**Conor McKenna**

**A Thesis Submitted in Partial Fulfilment**

**of the requirements for the**

**Degree of**

**Master of Science in Data Analytics**

****

**February 2024**

**Supervisor: David Gonzalez**

**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Titles:** | Capstone Project |
| **Assessment Title:** | Thesis |
| **Lecturer’s Name:** | David McQuaid, Muhammad Iqbal and David Gonzalez |
| **Student’s Full Name:** | Conor McKenna |
| **Student Numbers:** | Sbs23014 |
| **Assessment Due Date:** | 23/02/2024 |
| **Date of Submission:** | 23/02/2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Acknowledgements**

The completion of this project could not have happened without the collective effort and input of numerous individuals. Firstly, I would like to thank my supervisor, David Gonzalez, whose guidance and support helped keep the research moving. Our sessions from the beginning were always fruitful and David helped clear up any questions and doubts I had.

I would also like to thank David McQuaid, Dr. Muhammad Iqbal and Marina Iantorno for all the effort they put into their evening classes throughout the year that preceded this project.

My sincere thanks are also due to the participants who engaged in the interviews, offering valuable perspectives that were crucial in identifying relevant variables and features to look for when gathering the data required to carry out the work. Of course, this would not have been possible without the generous support of the Irish government and their fantastic Springboard programme which helps to further the career of so many individuals every year.

Contents

[1. Introduction 5](#_Toc159436086)

[1.1. Topic Background 5](#_Toc159436087)

[1.2. Research Objectives 5](#_Toc159436088)

[1.3. Background Research and Related Work 6](#_Toc159436089)

[2. Literature Review 7](#_Toc159436090)

[2.1. Approaches to predicting public debt using machine learning 7](#_Toc159436091)

[2.2. Identifying Common Features in Deep Learning Models for Public Debt Prediction 9](#_Toc159436092)

[2.3. Emerging Economies 10](#_Toc159436093)

[2.4. Short Sample Timeseries 11](#_Toc159436094)

[3. Research Framework 11](#_Toc159436095)

[4. Research Methods 13](#_Toc159436096)

[4.1. Primary Data 14](#_Toc159436097)

[4.2. Secondary data 15](#_Toc159436098)

[4.3. Data Preparation 15](#_Toc159436099)

[4.3.1. Summary Statistics 15](#_Toc159436100)

[4.3.2. Handling Null Records 16](#_Toc159436101)

[4.3.3. Correlation Coefficients (Pearson/Spearman) 16](#_Toc159436102)

[4.3.4. Decomposition of the Data 17](#_Toc159436103)

[4.4. Linear Regression 18](#_Toc159436104)

[4.5. Random Forest Regressor 19](#_Toc159436105)

[4.6. Gradient Boosting Machines 21](#_Toc159436106)

[4.7. Deep Learning Methods (LSTM) 23](#_Toc159436107)

[5. Evaluation and Analysis 23](#_Toc159436108)

[6. Ethics Considerations 24](#_Toc159436109)

[7. Conclusion 25](#_Toc159436110)

[8. References 27](#_Toc159436111)

Public debt dataset came from IMF

<https://www.imf.org/external/datamapper/GG_DEBT_GDP@GDD/SWE>

Inflation rate came from

<https://www.imf.org/external/datamapper/PCPIPCH@WEO/WEOWORLD/VEN>

# Introduction

## Topic Background

Public debt, the cumulative result of government borrowing over time, is a critical metric that affects the economic stability of nations and, by extension, the well-being of their citizens. Understanding the factors and features that correlate with (or directly influence) public debt levels is crucial for policymakers, economists, and financial analysts.

The aim of this topic is to address the efficacy of economic variables using data analytics and deep learning methodology. There have been many experiments and projects in the economics domain which have tried to address this question using econometric approaches. However this project will be centered around the capabilities of deep learning, time series, and data analytics to draw a conclusion.

The choice of this area of study is driven by the need to unravel the complex dynamics governing public debt. By employing data analytics techniques, the aim is to investigate the impact/correlation of various economic factors on public debt levels.

As seen in the literature, a lot of work has been carried out against data from developed countries ((Boonman et al., 2015), (Dufrénot and Paret, 2018), (Fioramanti, 2008)), and very little has been researched in terms of data from developed countries. There is a need therefore to examine data from more developed countries to see if there is a correlation between their economic variables and the levels of public debt.

## Research Objectives

Below outlined are the research objectives for this project:

1. Can deep learning methods demonstrate a strong correlation between economic variables such as inflation rate, GDP growth, or interest rates and the level of public debt?
2. Can data analytics techniques be used to develop reliable forecasts of *future* public debt levels using *current* economic variables?
3. How can data analytics techniques and visualisation tools be effectively applied to gain deeper insights into the relationships between economic variables and public debt dynamics, with the aim of informing more robust debt forecasting models and data-driven policy decisions?

## Background Research and Related Work

Public debt, the total amount owed by a government to creditors, is a critical economic indicator that influences a nation's financial stability, investment climate, and its overall economic health. In an era where economic conditions are rapidly changing, the ability to accurately predict public debt levels is invaluable for a wide range of stakeholders, including policymakers, investors, financial analysts, and international organizations.

Deep learning, a subset of machine learning characterized by its ability to learn from vast amounts of data through neural networks, offers significant advantages in predicting public debt levels. Traditional economic forecasting models often rely on linear assumptions and may not fully capture the complex, nonlinear relationships between various economic indicators. Deep learning models, such as Long Short-Term Memory (LSTM) networks, are adept at understanding these complexities, making them particularly suitable for time-series data like economic indicators. They can process large datasets, learn from the sequential nature of the data, and capture temporal dependencies that are crucial for accurate forecasting.

**Importance of Predicting Public Debt Levels**

*Policy Making and Government Planning:* Governments need to manage their debt levels to ensure economic stability and sustainable growth. Accurate predictions of public debt can inform fiscal policy decisions, budget planning, and debt management strategies. It allows policymakers to anticipate future economic conditions and make informed decisions regarding spending, taxation, and borrowing.

*Economic Analysis and Research:* Economists and researchers benefit from accurate debt level forecasts to analyze economic trends, assess the effectiveness of fiscal policies, and study the impact of debt on economic growth and inflation. Deep learning models can enhance the quality of this research by providing more accurate and nuanced forecasts.

*Investment Decisions:* Investors and financial analysts use public debt level forecasts to assess the risk profile of government bonds and other debt instruments. Accurate predictions can inform investment strategies, helping investors to make better decisions about asset allocation, risk management, and portfolio optimization.

*International Organizations:* Entities such as the International Monetary Fund (IMF) and the World Bank monitor public debt levels to assess economic stability and risk across countries. These organizations can use deep learning predictions to guide their lending decisions, economic assistance programs, and policy advice to member countries.

*Market Stability and Economic Confidence:* Predicting public debt levels contributes to market stability by providing all market participants with insights into future economic conditions. This transparency can bolster economic confidence, reduce uncertainty, and facilitate a stable investment environment.

In conclusion, the application of deep learning to predict public debt levels represents a significant advancement in economic forecasting. By leveraging the power of these models, stakeholders can gain a deeper understanding of future economic trends, enabling more informed decision-making and contributing to global economic stability. The accuracy and efficiency of deep learning models in handling complex, nonlinear data make them indispensable tools for anyone involved in economic planning, investment, and analysis.

# Literature Review

## Approaches to predicting public debt using machine learning

The study "Government Debt Forecasting Based on the ARIMA Model" (Zhuravka et al.) utilises the ARIMA model to forecast government debt, demonstrating the efficacy of statistical methods in predicting economic trends. It underscores the utility of deep learning and data analytics in understanding the intricate relationships between economic variables and public debt levels. Through its analysis, the paper illustrates how visualization tools can enhance the interpretation of data, aiding in the development of more accurate forecasting models. This aligns with the broader objective of employing data-driven approaches for economic analysis and policy formulation, making it a valuable reference for exploring the correlation between economic factors and public debt dynamics.

The paper focuses on employing the Autoregressive Integrated Moving Average (ARIMA) model to analyze the dynamics and trends of changing public debt and predict its future values. The ARIMA model is crucial for understanding how historical data can be used to forecast future occurrences of public debt levels based on past trends. Before applying the ARIMA model, the study emphasizes the importance of checking the time series for stationarity. It involves determining the integration order necessary to make the series stationary, which is a fundamental step in ensuring the reliability of the ARIMA model predictions. Regression methods are employed for model evaluation, where estimates of the parameters included in the model are obtained. The study uses the autocorrelation coefficient (ACF) and the partial autocorrelation coefficient (PACF) to calculate the autocorrelation order. This could be one technique this project could use. The paper also mentions the use of ARMAX models, which consider lag variables of the studied indicator and exogenous factors. This approach signifies the potential to enhance forecasting accuracy by incorporating external variables alongside historical time series data.

“Sovereign Debt and Currency Crises Prediction Models Using Machine Learning Techniques” (Alaminos et al.) investigates the prediction of sovereign debt and currency crises using machine learning techniques, expanding upon traditional statistical methods for greater accuracy and broader geographic coverage. Methodologies employed include Multilayer Perceptron, Support Vector Machines, Fuzzy Decision Trees, AdaBoost, Extreme Gradient Boosting, Random Forests, Deep Belief Network, and Deep Neural Decision Trees, aimed at robustly testing through a variety of successful classification techniques. The inclusion of a wide array of methods reflects an ambitious approach to explore the complex, non-linear relationships which are often found in financial crises prediction. The critique lies in the balance between complexity and interpretability. While the breadth of methodologies allows for a comprehensive exploration of predictive capabilities, it also introduces challenges in model comparison and the synthesis of insights across different techniques. On one hand, the use of a diverse methodological framework highlights the project's innovative stance, but on the other hand it runs the risk of complicating the extraction of clear, actionable insights due to the varying strengths, weaknesses, and assumptions underlying each technique.

“Modelling Sovereign Credit Ratings: Evaluating the Accuracy and Driving Factors using Machine Learning Techniques” (Overes and van der Wel) outlines various machine learning techniques used to predict sovereign credit ratings, including Multilayer Perceptron, Classification and Regression Trees, Support Vector Machines, Naïve Bayes, and Ordered Logit models. It details the process of transforming categorical credit ratings into numeric values for model compatibility and discusses the use of SHAP values for interpreting the impact of individual variables within complex models. The section also covers model evaluation methods, emphasizing the use of cross-validation to assess predictive accuracy. This comprehensive approach demonstrates a sophisticated attempt to leverage machine learning for financial analysis, providing a rich framework for understanding the determinants of sovereign credit ratings. Section 2.1.2 of the document discusses Classification and Regression Trees (CART), a machine learning algorithm utilized for both classification and regression tasks. The simplicity of CART, derived from its ability to find optimal splits based on explanatory variable values to classify observations, makes it well-suited for applications in credit rating predictions.

The CART model is structured with a root, multiple nodes, and leaves, where the initial data split at the root is based on one explanatory variable, leading to further splits at the nodes or decision-making at the leaves. This research employs a binary split approach for the data at each node, utilizing the Gini method to calculate the weighted average Gini impurity for variable effectiveness in separating different categories.

CART offers several advantages, including its capability to handle outliers, perform automatic feature selection, and provide straightforward model interpretation. However, a significant drawback of CART is its susceptibility to overfitting (Egelberg et al.), a common issue that necessitates careful model tuning and validation to ensure robust performance.

Despite CART's benefits, its tendency to overfit requires strategies like restricted growth or pruning to mitigate this risk (Patil et al.). However, in this particular study, neither method improved the cross-validated out-of-sample accuracy, leading to the decision to use an unrestricted CART approach. The optimization details for CART and its implementation in Python's scikit-learn package are provided in the document's appendix.

In summary, while CART is acknowledged for its strengths in model simplicity and interpretability, the careful management of its overfitting tendency is crucial for achieving accurate and reliable credit rating predictions.

In the realm of machine learning, Gradient Boosting (GB) stands out as a refined method that sequentially constructs a composite model from simple, individual models—referred to as basis-models—to enhance prediction accuracy. This technique incrementally corrects errors by focusing on the discrepancies between actual outcomes and predictions, known as residuals, with the goal of minimizing these errors through each iteration. The core mechanism of GB involves adjusting predictions based on the negative gradient of a loss function (Alexandros Agapitos et al.), which quantifies the prediction error, thereby guiding the model towards greater accuracy with each step. This approach is particularly effective in regression scenarios, where the squared-error loss function is commonly employed, allowing the model to directly address the residuals.

However, as discussed in the same paper ("Regularised Gradient Boosting for Financial Time-Series Modelling", Alexandros Agapitos et al.)), extensive iteration of GB can lead to overfitting—a condition where the model becomes overly complex, capturing noise rather than the underlying data pattern, thus diminishing its predictive power on new, unseen data. Previous research has highlighted the dilemma of overfitting within GB models, especially when trained on noisy datasets. Although introducing randomness into the training process has been suggested as a remedy, it can compromise model accuracy in its initial stages.

To counteract overfitting, the article advocates for the use of regularization techniques, specifically shrinkage, which scales down the contribution of each new basis-model, compelling the model to prioritize broader data patterns over noise. Despite these efforts, GB models can still exhibit overfitting, particularly in the presence of noisy data. Addressing this, the article introduces an innovative adjustment to the residuals through a sigmoidal transformation, aiming to mitigate the influence of extreme errors without eliminating them from the training set. This modification, built upon the foundation of shrinkage-based regularization, has shown promise in simulations with synthetic, noisy datasets by decelerating the overfitting process and enhancing the model's generalization capability.

This progression in GB techniques, as elucidated in "Regularised Gradient Boosting for Financial Time-Series Modelling," signals a significant advance in the development of more robust and accurate predictive models, especially pertinent to the intricate and often unpredictable domain of financial time-series data.

## Preparing Datasets for Time series analysis

The preparation of datasets for time series analysis has been widely recognized as a critical step in ensuring the accuracy and reliability of forecasting models (Muhammad et al.). In “Environmental Noise Pollution Forecasting Using Fuzzy-Autoregressive Integrated Moving Average Modelling”, the author explores a new way to preprocess the dataset to improve the time series model’s accuracy scores. They highlight the risk of focusing too much on fine-tuning the models without first preprocessing the underlying data that feeds into the model. To tackle the issue of data uncertainty, which can be particularly prevalent in time-series data, the paper introduces a novel approach. This method involves a systematic strategy for preparing data that contains uncertainty —through the use of triangular fuzzy numbers (TFN). TFNs are employed to refine the data preparation process, enabling the transformation of ambiguous data into a structured form that reduces uncertainty and facilitates more precise modeling.

The application of this method is then demonstrated through the development of an ARIMA-based model for predicting ambient noise pollution. By incorporating a systematic data preparation technique that converts uncertain, fuzzy data into clearly defined fuzzy numbers, the authors were able to significantly improve the model's forecasting accuracy.

Given the temporal nature of time series data, which is characterized by observations collected at successive points in time, the process of dataset preparation involves unique challenges and considerations that distinguish it from other types of data preprocessing.

One fundamental aspect of preparing a dataset for time series analysis is the handling of missing values. It has been observed that financial datasets frequently contain gaps due to various reasons such as errors in data collection, transmission losses, or unrecorded observations (Feng and Cong; John et al.). The imputation of missing values is, therefore, a vital process, and several methods have been proposed to address this issue. Techniques such as linear interpolation, where missing values are filled based on linear relationships between available data points (Noor et al.), and more sophisticated approaches like the use of machine learning algorithms to predict missing values, have been extensively explored.

In “Comparison of Linear Interpolation Method and Mean Method to Replace The Missing Values In Environmental Data Set” for example, the authors compared the use of linear interpolation vs the mean method and found that, after analysing the mean absolute error (MAE), root mean square error (RMSE) and coefficient of determination (R2), linear interpolation produced a more accurate value. However “Endometrial Cancer Individualized Scoring System (ECISS): A Machine Learning‐Based Prediction Model of Endometrial Cancer Prognosis” (Shazly et al.) proves that linear interpolation is not the only method, and the use of machine learning algorithms to predict missing can produce superior results to more traditional approaches.

Another significant consideration in dataset preparation is the detection and correction of outliers. Outliers can significantly distort a dataset’s information (Wada), leading to inaccurate forecasts. Various statistical methods can be employed to identify outliers, including z-score analysis and the interquartile range method. Once detected, outliers may be corrected through methods such as trimming, where extreme values are removed (Buckley and Georgianna). Extreme caution needs to be exercised when handling potential outliers, particularly in small datasets because to transform or remove a data point altogether might result in the loss of key information that is, in fact, relevant.

The importance of seasonality and trend decomposition in preparing time series datasets cannot be overstated. Time series data often exhibit patterns that repeat over time (seasonality) and long-term trends. Decomposing the dataset into its constituent components—trend, seasonality, and residual—is crucial for understanding the underlying dynamics of the data. Techniques such as the classical decomposition method as used in “An Evaluation of Alternative Forecasting Methods to Recreation Visitation” (Chen et al.) and the more advanced Seasonal and Trend decomposition using Loess, as seen in “Forecasting the Short-Term Metro Ridership with Seasonal and Trend Decomposition Using Loess and LSTM Neural Networks” (D. Chen et al.) have been widely applied. These methods enable the isolation of the seasonal component and trend from the time series, facilitating more accurate forecasting by allowing models to account for these patterns explicitly.

The literature shows that stationarity is another critical concept in the preparation of time series data. Many time series models, such as the Autoregressive Integrated Moving Average (ARIMA), assume that the dataset is stationary, meaning its statistical properties do not change over time. The presence of trends or seasonality can violate this assumption, necessitating the use of differencing or transformation techniques to stabilize the mean and variance of the series. The Augmented Dickey-Fuller test is commonly used to test for stationarity (Mushtaq; ), guiding the appropriate preprocessing steps to achieve stationarity.

Finally, Normalization and standardization of time series data are also essential preprocessing steps, especially when dealing with variables of different scales. Normalization adjusts the data within a specific range, typically between 0 and 1, while standardization transforms the data to have a mean of zero and a standard deviation of one. These processes are particularly important in the context of machine learning models, where feature scaling can significantly impact the performance of the algorithms. The literature suggests that standardisation is often chosen over Normalisation because it better suits the activation functions used by deep learning models (McNally et al.)

In conclusion, the preparation of a dataset for time series analysis is a multifaceted process that requires careful consideration of several critical factors, including missing value imputation, outlier detection and correction, decomposition of seasonality and trend, ensuring stationarity, and normalization or standardization of the data. These steps are indispensable for the creation of accurate and reliable time series models, underscoring the importance of thorough dataset preparation in the broader context of time series analysis.

## Identifying Common Features in Deep Learning Models for Public Debt Prediction

“State Debt Assessment and Forecasting: Time Series Analysis” (Zhuravka et al.) investigates the critical issue of growing state debt and its implications for the financial systems of countries, with a focus on developing an effective state debt management system through forecasting. It emphasizes the importance of predicting short-term future state debt levels using time series analysis, specifically through an ARIMA (AutoRegressive Integrated Moving Average) model.

The study uses data from Ukraine's state debt from December 2004 to November 2020 and, by employing the Hurst exponent, the research calculates the persistence level in the time series data of Ukraine's state debt. The Hurst exponent's value indicated that autoregressive models are suitable for forecasting Ukraine's future debt dynamics.

By incorporating the concept of the Hurst exponent from this paper, the trend or mean-reverting nature of public debt levels can be measured and analysed, aiding in the development of predictive models for effective debt management. This concept is crucial as it lends empirical evidence to the presence or absence of a trend in public debt levels. By providing statistical substantiation, we can objectively determine whether public debt is likely to continue in its current trajectory or revert to a long-term mean.

Out of all the literature reviewed, “Sovereign Debt and Currency Crises Prediction Models Using Machine Learning Techniques” (Alaminos et al.) provided the best insight into what variables and models are more effective than others when analysing public debt. A large range of variables were brought into the analysis and it was carried out across multiple continents. Variables were listed under headers such as “Domestic Macroeconomic Factors” , “Credit Rating Indicators”, “Political Factors”. One critisicm that could be argued is that it incorporates too many variables without analysing the specific ones which make the biggest impact. The author themselves note that there was a “high computational cost” when performing the analyses, as well as a higher degree of complexity when interpreting the outcomes.

The paper concludes that Fuzzy Timeseries and Fuzzy Decision Trees yield the best results of the variables used. Fuzzy decision trees are a type of decision tree which incorporate fuzzy logic. Traditional decision trees classify instances by following a path of decisions based on the features' values until reaching a conclusion at the leaves. These decisions are usually binary (Yes/No) and based on clear thresholds. In contrast, fuzzy decision trees handle ambiguity and vagueness inherent in real-world data by using fuzzy logic to create rules that are not strictly binary (Wang et al.). They allow for degrees of membership to different classes, rather than a strict belonging.

## Emerging Economies

The literature also reveals that the majority of work done so far (with regard to analysing and predicting Public Debt) has focused on data from developing economies (Boonman et al., 2015), (Dufrénot and Paret, 2018), (Fioramanti, 2008)). This may not accurately mirror the patterns observable in developed countries, which is a critical consideration for the scope and objectives of this study. Given that the data being analyzed originates from developed countries, differences in outcomes may emerge due to the distinct economic frameworks, fiscal policies, and market reactions that characterize these nations. The specific features of developed economies necessitate a tailored approach in interpreting findings, as they may diverge significantly from trends and behaviors identified in emerging markets. The importance of recognizing these distinctions lies at the heart of accurately assessing the dynamics of public debt within the context of developed nations.

## Short Timeseries Dataset

Much of the data around economic variables, and in particular, Public Debt itself, comes in yearly snapshots back to the middle of the 20th century for many countries. Time series analysis, on the other hand, often relies on large datasets to predict future values based on past observations. Handling small datasets in time series analysis presents unique challenges, necessitating innovative approaches to ensure the reliability and accuracy of predictive models. This chapter analyses the literature to learn about the limitations of small datasets in a timeseries context, as well as approaches taken by other authors in handling them.

The limited size of small datasets constrains the ability to capture complex patterns and seasonal trends, which are vital for accurate forecasting. This limitation can lead to overfitting, where models perform well on training data but poorly on unseen data. Hence, special attention must be given to model selection, feature engineering, and validation techniques when dealing with small datasets. In “Simulating Time-Series Data for Improved Deep Neural Network Performance” (Yeomans et al.) , for example, the author seeks to avoid over-fitting by simulating time series data.

When working with small datasets, the choice of model becomes paramount. Simple models often outperform more complex ones since they are less likely to overfit the limited data available (Zhang et al.). For instance, Autoregressive Integrated Moving Average (ARIMA) models, which are inherently simpler and require fewer parameters, can be more suitable than more complex machine learning models. Furthermore, simplicity aids in the interpretability of the model, allowing for easier identification of the driving factors behind the observed trends.

Validation is crucial to ensure that the model generalizes well to unseen data. Traditional hold-out methods, where the dataset is split into training and test sets, may not be suitable for small datasets due to the limited number of observations. Instead, time series cross-validation techniques, such as forward chaining, where multiple training-test splits are created sequentially across the time dimension, can provide a more reliable estimate of the model's performance. This method allows for the utilization of more data for training without violating the temporal order of observations. In “Predicting water main failures using Bayesian model averaging and survival modelling approach” (Kabir et al.), the authors use a 5-fold cross validation method to evaluate the performance of the models since the dataset is too small to use the traditional hold out method.

Regularization and Bayesian Approaches

Regularization techniques, such as Lasso and Ridge regression, can prevent overfitting by penalizing large coefficients in the model. These techniques are particularly useful in scenarios where the number of features might be relatively high compared to the number of observations. Bayesian approaches, on the other hand, offer a probabilistic framework that can incorporate prior knowledge about the parameters, helping to guide the model towards more plausible solutions when data is scarce. In “LASSO variable selection in data envelopment analysis with small datasets” (Lee and Cai), the author proposes using the Least Absolute Shrinkage and Selection Operator (LASSO) as a solution. LASSO is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces. By applying LASSO, the study aims to identify and retain only the most significant inputs and outputs.

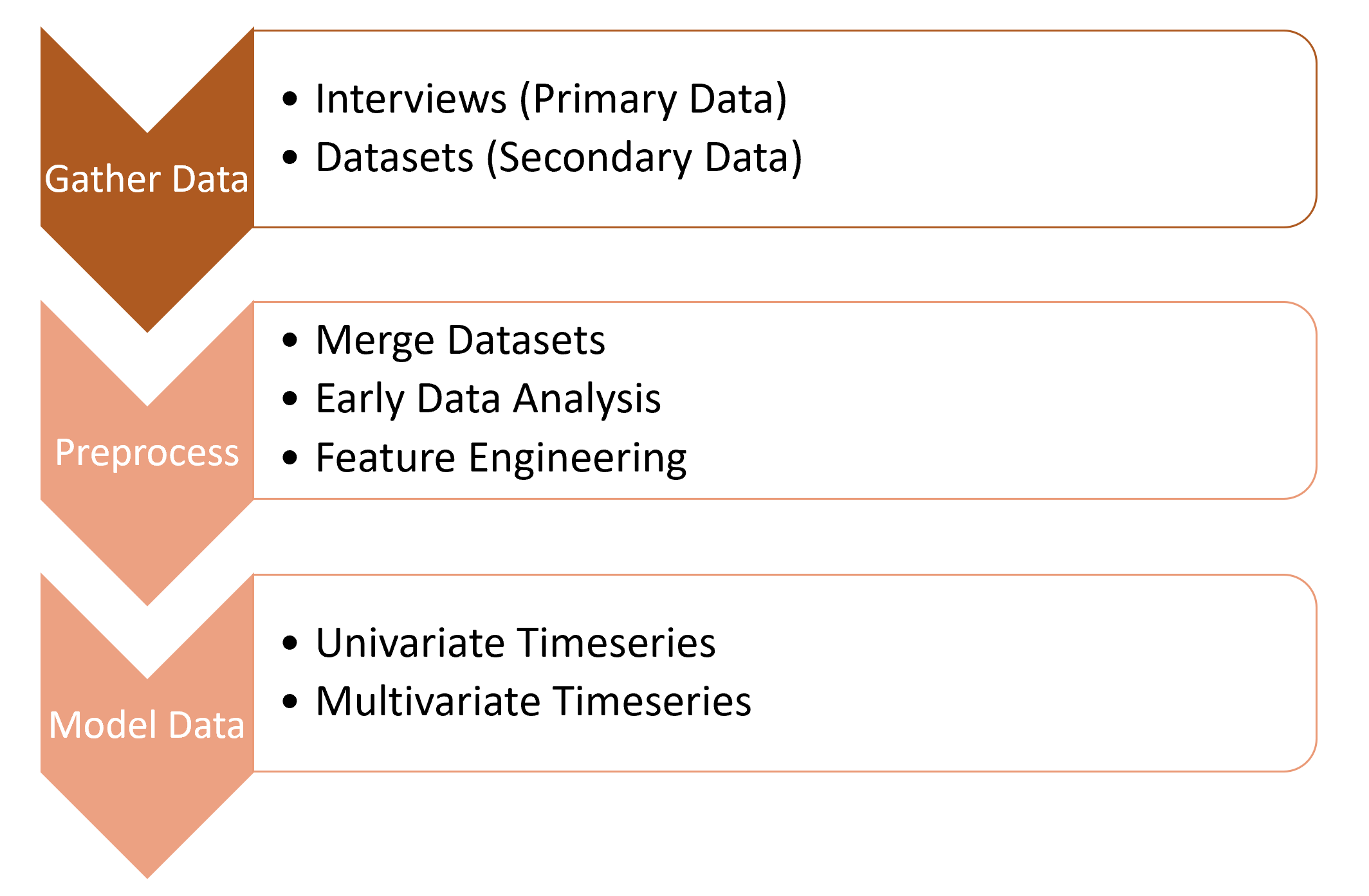
Data Augmentation and Synthetic Data

Data augmentation involves creating additional data points based on the existing data, which can be particularly useful for small datasets. Techniques such as bootstrapping, where random samples are drawn with replacement, can help in assessing the variability of the model's forecasts. Similarly, synthetic data generation methods can be employed to artificially increase the size of the dataset, allowing models to learn a broader range of patterns and interactions.

Conclusion

Handling small datasets in time series analysis requires a careful balance between model complexity, feature engineering, and validation strategies. By opting for simpler models, enhancing the dataset through intelligent feature engineering, employing rigorous cross-validation techniques, and exploring regularization and synthetic data generation, analysts can overcome the challenges posed by limited data. These approaches enable the development of predictive models that are both accurate and robust, ultimately unlocking valuable insights from small datasets in time series analysis.

# Research Framework



The primary research methodology for this investigation involves conducting in-depth interviews with experts in the domain of economics. This approach is chosen deliberately to glean insights directly from seasoned professionals who possess a nuanced understanding of the intricate factors influencing public debt levels. By engaging in dialogues with these experts, the research aims to capture qualitative perspectives, expert opinions, and contextual nuances that may not be fully captured by quantitative data alone. The intention is that they will act as a key input when considering what economic variables and features are particularly relevant to the underlying aim of this project.

*Reasons for Choosing Expert Interviews:*

*Rich Qualitative Insights:*

Expert interviews offer a unique opportunity to delve into the qualitative intricacies of economic variables and their impact on public debt. The depth and richness of information obtained through open-ended conversations can provide an overall understanding of the complex relationships that quantitative data might not fully explain.

*Contextual Understanding:*

Economics is a discipline where contextual understanding is paramount. By engaging in one-on-one interviews with experts, the research can tap into their wealth of contextual knowledge, gaining insights into the specific economic, political, and institutional factors that shape public debt dynamics.

*Validation and Clarification:*

Interviews will provide an opportunity for real-time validation and clarification. Experts will be able to elaborate on complex concepts, challenge assumptions, and provide real-world examples that enhance the clarity and robustness of the research findings.

*Conducting Expert Interviews:*

The selection of interviewees will be purposeful and based on their expertise in economic domains relevant to public debt. The intention is to include economists with a track record of contributing to the understanding of economic variables and public finance. Semi-structured interviews will be conducted, allowing for a balance between a predefined set of questions and the flexibility to explore new avenues as they emerge during the conversation.

Upon the completion of the interviews, the variables decided upon will then guide the selection of secondary datasets. These datasets will need to contain quantifiable measures of the variables identified, thereby anchoring the research in empirical data. The reliance on secondary datasets is a strategic choice, designed to provide a robust quantitative counterpart to the qualitative narratives obtained from the interviews.

This methodical approach ensures that the data collection is both directed and relevant. By allowing the qualitative data to define the scope of the quantitative data, the research is ensured to remain grounded in real-world complexities while benefiting from the precision of empirical analysis.

Statistical analyses are to follow, serving as the empirical backbone of the research. It is here that statistical methods are employed to unearth patterns, correlations, and potential causations within the data. The utility of this step cannot be overstated, as it is through these methods that the validity of the data is corroborated, and its reliability is cemented.

The second last phase involves the deployment of four distinct time series models. Each model is chosen and applied to the dataset, with the intent to predict, understand, and extrapolate the temporal trends inherent within the data. The application of multiple models is crucial, as it allows for a multi-faceted view of the data's temporal progression, thereby enhancing the robustness of the predictions made.

Finally, the outcomes of the time series analyses are to be rigorously evaluated. This evaluation is conducted not in isolation but by juxtaposing the models' results against each other and the original research objectives. The essence of this step lies in its critical appraisal of the models' performance, assessing their predictive prowess, and extracting actionable conclusions.

# Research Methods

A selection of predictive models was carefully curated to explore the intricate relationships between economic variables and public debt levels. Linear Regression, Gradient Boosting Machines (GBM), Random Forest Regressor, and Long Short-Term Memory (LSTM) neural networks were chosen as the primary methodologies for several reasons.

Linear Regression was selected for its simplicity and effectiveness in establishing baseline correlations between independent variables like inflation rate, GDP growth, and interest rates, and the dependent variable, public debt level. This model serves as a foundational analysis tool, enabling straightforward interpretation of economic factors' direct impacts on public debt.

Gradient Boosting Machines (GBM) and Random Forest Regressor were included for their ability to handle non-linear relationships and interactions among variables. GBM, known for its powerful predictive capabilities and flexibility, is adept at capturing complex patterns through sequential model improvement. Similarly, the Random Forest Regressor offers robustness against overfitting and excels in capturing the random nature of economic variables, providing a more nuanced understanding of their effects on public debt.

LSTM neural networks represent the deep learning approach within the project, aimed at answering whether advanced models can identify and leverage long-term dependencies and patterns in economic data for superior forecasting accuracy. LSTMs are particularly suited to model time-series data, making them an ideal choice for analyzing the sequential nature of economic indicators and their influence on the trajectory of public debt.

Together, these models encompass a comprehensive analytical framework designed to develop reliable forecasts of future public debt levels, enhance understanding of the dynamic relationships between key economic variables and public debt, and support the creation of data-driven policy decisions. Through the application of these diverse methodologies, the project seeks to offer insightful visualizations and in-depth analysis, contributing to the broader discourse on economic forecasting and policy formulation.

## Primary Data

In the process of data collection for the project aimed at predicting levels of public debt using deep learning methods, primary data was obtained through interviews to identify useful economic variables for the analysis. Two experts were consulted to leverage their insights on which economic indicators would best predict public debt. The first interviewee emphasised the importance of the inflation rate as a critical indicator, highlighting its direct and indirect impacts on the economy's health and its potential influence on public debt levels. The second expert suggested the Gross Domestic Product (GDP) as a fundamental variable, considering its comprehensive reflection of the economic activity and its correlation with the government's borrowing capacity and fiscal policy. Both interviewees agreed on the value of incorporating data from countries with economic profiles similar to Ireland, specifically mentioning Denmark and the United Kingdom. This comparative approach was proposed to enrich the analysis by understanding the dynamics in economies that share certain fiscal, social, and economic characteristics with Ireland, thus providing a broader perspective on the variables influencing public debt. The consensus on examining the inflation rate and GDP, alongside the comparative analysis with Denmark and the UK, formed the basis for selecting the economic variables to be used in predicting public debt levels through deep learning techniques.

|  |  |  |
| --- | --- | --- |
| **Topics** | **Person 1** | **Person 2** |
| Features | Inflation Rate | GDP and inflation rate |
| Countries | Agreed that Denmark and UK would be appropriate countries | Agreed that Denmark and UK would be appropriate countries |
| Methodology | Agreed | Agreed |

## Secondary data

The decision to source datasets from the International Monetary Fund (IMF) website for the project on predicting levels of public debt was driven by the IMF's reputation for providing comprehensive, reliable, and up-to-date economic data. The IMF's extensive database offers a wealth of standardized financial and economic information across countries, including key indicators like inflation rates and GDP figures. Utilizing data from the IMF ensures the analysis is grounded in high-quality and globally recognized datasets, facilitating accurate and credible forecasts.

## Data Preparation

Two distinct datasets were imported, encompassing vital economic indicators—public debt levels, inflation rates, and GDP figures—for Ireland, Denmark, and the UK. They were transposed to align the data chronologically, and a merge operation was conducted to consolidate the indicators into a single comprehensive dataframe. Plots were then generated, offering a visual representation of how these crucial economic metrics have evolved over time across the three nations. This visualisation step is pivotal, providing initial insights into the interrelations and potential predictive dynamics among the variables.

### Summary Statistics

The summary statistics (.describe()) of the dataset reveal a sample size of 69 observations. The average value of the variable is approximately 55.66, indicating the central tendency of the data. A relatively high standard deviation of about 24.78 suggests there is considerable variability or spread in the data points. The minimum and maximum values of 24.81 and 119.48, respectively, highlight a wide range in the dataset. The interquartile range, from the 25th percentile (36.18) to the 75th percentile (75.05), indicates where the middle 50% of the data lies and further underscores the significant spread of the values. The median of 47.74 is less than the mean, suggesting a right-skewed distribution. Right-skewed data may have outliers or extreme values at the higher end, which can affect the mean. It’s important to investigate these outliers as they can provide insights into the data or indicate data entry errors or other anomalies.

### Handling Null Records

It is important to carefully consider the handling of null records from different angles. The dataset displayed a considerable number of missing values between the years 1936 and 1970, along with a single instance in 1997. In response to this, a strategic decision was made to exclude all records prior to 1970 and apply a backfill method to address the missing value from 1997. This choice is substantiated by the rationale that the economic landscape has undergone significant transformations since the mid-20th century, rendering the data from the pre-1970 era less relevant for contemporary analyses. Modern economic policies, global financial structures, and market dynamics have evolved, diminishing the predictive relevance of historical data from that period for current economic phenomena. Moreover, the integrity and availability of older records often pose challenges due to changes in data collection methods and statistical standards over time. The backfilling of the 1997 record ensures a complete dataset post-1970, facilitating a more robust and relevant analysis while maintaining the continuity and quality of the more recent and applicable data, which is more reflective of the current economic environment and trends.

### Correlation Coefficients (Pearson/Spearman)

The Pearson correlation coefficient measures the linear correlation between two variables, with values ranging from -1 to 1. A value closer to 1 indicates a strong positive linear correlation, a value closer to -1 indicates a strong negative linear correlation, and a value around 0 suggests no linear correlation. The Pearson correlation coefficient for GDP and public debt is approximately -0.095, indicating a very weak negative linear correlation between the two variables. The high p-value (>0.05) suggests that this correlation is not statistically significant, meaning there is little evidence of a linear relationship between GDP growth and public debt as a percentage of GDP in Ireland. The correlation coefficient of approximately -0.350 indicates a moderate negative linear correlation between inflation and public debt. The p-value of about 0.017 is less than the conventional alpha level of 0.05, suggesting that this correlation is statistically significant. This implies that as inflation increases, there might be a tendency for public debt as a percentage of GDP to decrease, or vice versa, with a moderate level of confidence.

The Spearman correlation coefficient measures the monotonic relationship between two variables, which can be more appropriate for non-linear relationships. The Spearman correlation coefficient for GDP and public debt is approximately -0.235, indicating a weak negative monotonic relationship. The p-value of about 0.116 exceeds the conventional alpha level of 0.05, suggesting that this correlation is not statistically significant. This result supports the Pearson analysis, indicating little evidence of a strong relationship between GDP growth and public debt as a percentage of GDP. The correlation coefficient of approximately -0.448 suggests a moderate to strong negative monotonic relationship between inflation and public debt. The p-value is significantly less than 0.05, indicating that this correlation is statistically significant with a high level of confidence. This suggests that as inflation increases, public debt as a percentage of GDP tends to decrease, or vice versa, in a monotonic fashion.

To summarise, there is little evidence of a significant linear relationship between Ireland's GDP growth and its public debt as a percentage of GDP. There is a statistically significant, moderate negative relationship between inflation and public debt as a percentage of GDP, suggesting that higher inflation rates are associated with lower levels of public debt relative to GDP, or vice versa. This relationship is supported by both Pearson and Spearman correlation analyses, with the Spearman correlation indicating the relationship might be non-linear or not strictly linear.

### Decomposition of the Data

The discovery of no residual nor seasonality in the decomposition of the "Ireland\_Public\_Debt\_Pc\_of\_GDP" field through the use of the seasonal\_decompose method from the statsmodels library is a significant finding in the realm of time series analysis. This outcome has several implications for understanding the behavior of Ireland's public debt as a percentage of GDP over time, and it can greatly influence the approach to modeling and forecasting this economic indicator.

Benefits of Lack of Seasonality and Residuals

*Simplification of Models:* The absence of seasonality suggests that the data does not exhibit regular patterns of fluctuation within specific periods, such as monthly or yearly cycles. This simplifies the process of modeling since it indicates that the data can potentially be described by trends without the need to account for repeating seasonal effects. This simplification can lead to more straightforward and computationally efficient models.

*Predictive Accuracy:* A time series without significant residuals implies that the variation in the data can be almost entirely explained by the model used, in this case, an additive model where components are linearly added together. When residuals are minimal, it indicates that the model captures most of the information in the data, potentially leading to more accurate forecasts.

Importance of This Finding

Understanding the characteristics of time series data is crucial before applying forecasting models. The decomposition of the time series into trend, seasonal, and residual components allows analysts to identify the underlying patterns that need to be modeled. In the case of Ireland's public debt to GDP ratio, the absence of seasonality and residuals after decomposition underlines the data's lack of cyclical or irregular patterns, which could otherwise complicate analysis and forecasting.

*Interpretation and Implications*

.

For time series analysis, this finding implies that simpler models, such as linear regression on the trend component, might be sufficient for forecasting purposes, assuming the trend continues without major economic disruptions. However, the absence of seasonality and residuals in historical data does not guarantee that future data will behave similarly. External shocks, policy changes, or significant economic events could introduce new patterns not present in the historical data.

In conclusion, the decomposition of Ireland's public debt as a percentage of GDP revealing no seasonality and minimal residuals provides significant insights for economic analysis and forecasting. It suggests a stable and predictable pattern in the data over the period analyzed, which can aid in simplifying modeling approaches and improving predictive accuracy. This finding is crucial for policymakers and analysts focusing on fiscal stability and economic planning.

## Linear Regression

Linear regression stands as a fundamental starting point for research exploring the relationships between economic variables and public debt levels, chiefly due to its transparency, simplicity, and interpretability.

The methodology adopted, firstly prepares the dataset by selecting relevant features while excluding the target variable (Ireland's public debt percentage of GDP). This preparation is crucial for a clear delineation of independent variables (features) and the dependent variable (target). By employing a train-test split, the model is trained on a subset of the data, thereby mitigating overfitting and enabling an unbiased evaluation of its predictive performance on unseen data.

The choice of a 20% test split in the dataset represents a common practice in machine learning, balancing the need for sufficient training data with enough testing data to evaluate the model's generalizability. This split ensures that the model is trained on a substantial portion of the data (80%), fostering its ability to learn the underlying patterns, while still reserving a meaningful subset (20%) for testing its predictions against unseen data. Choosing a larger test split could provide a more robust evaluation of the model's performance but at the risk of depriving the model of enough training data, potentially impairing its ability to learn effectively. Conversely, a smaller test split might improve the model's learning from a larger training set, but with the trade-off of a less rigorous assessment of its predictive accuracy and application to new data. Increasing the test portion significantly could also reduce the risk of overfitting by limiting the model's exposure to the training data. Thus, the 20% testing split is a strategic choice to prevent overfitting while still providing the model with enough information to uncover and learn the fundamental relationships within the data.

The reported Mean Squared Error (MSE) of 0.35 and R^2 score of 0.67 are very important to understand the model's efficacy. The low MSE indicates a small difference between the observed and predicted values of Ireland's public debt as a percentage of GDP, signifying high accuracy in predictions. The R^2 score, representing the proportion of variance in the dependent variable that can be explained by the independent variables, signifies that about 66% of the variability in Ireland's public debt levels can be accounted for by the model. This substantial figure suggests a strong correlation between the chosen economic indicators and public debt levels, affirming the model's relevance and supporting its use as a foundational tool in economic forecasting and policy-making analysis.

## Random Forest Regressor

The Random Forest algorithm is an ensemble learning method extensively utilized for both classification and regression tasks, which makes it particularly relevant for a wide array of data analytics projects, including the exploration of economic variables' impact on public debt levels. This algorithm operates by constructing multiple decision trees during the training phase and outputting the mean prediction of the individual trees for regression tasks, thereby improving predictive accuracy and controlling overfitting.

The incorporation of the Random Forest algorithm into this research project serves as a sophisticated approach to address the complexity inherent in economic data. Economic variables, such as inflation rates, GDP growth, and interest rates, exhibit nonlinear relationships and interactions that can be challenging to capture with simpler models like linear regression. Random Forest, with its ability to model complex, non-linear relationships between a multitude of predictors and the response variable, provides a robust framework for analyzing how these economic indicators collectively influence Ireland's public debt levels. This capability is crucial for developing reliable forecasts and extracting nuanced insights from the data, which are instrumental in informing policy decisions.

In this specific application, the analysis prepared independent variables by excluding 'Year' and the target variable 'Ireland\_Public\_Debt\_Pc\_of\_GDP' from the dataset, ensuring a focus on the economic indicators of interest. A significant proportion of the dataset, 90%, was allocated for training, with the remaining 10% reserved for testing. This split was chosen to maximize the amount of training data given the complexity and the depth of relationships Random Forest is expected to learn, while still retaining a portion of the data for model evaluation.

A Random Forest Regressor, configured with 100 estimators (trees) and a consistent random state for reproducibility, was trained on this partitioned data. The model's performance was then evaluated on the test set, with Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) serving as the metrics for quantifying prediction accuracy and model efficacy. The relatively high values of MSE and RMSE suggest that the model's predictions deviate significantly from the actual values, which could be attributed to several factors including the complexity of the data, potential underfitting due to overly simplistic model assumptions in the face of complex economic dynamics, or the possibility that important predictors or interactions between variables were not adequately captured. It's also conceivable that the 90% training data split, while intended to maximize learning, might have led to a testing set that is too small or not representative enough of the overall data variability, thereby impacting the model's evaluation metrics.

The adoption of the Random Forest algorithm in this research reinforces the search of a more nuanced understanding and prediction of public debt dynamics in response to changes in key economic indicators. By leveraging its ability to handle complex, nonlinear data relationships, this methodological choice enriches the analysis, providing deeper insights and more reliable forecasts than simpler models might afford. The results obtained, as measured by MSE and RMSE, indicate the model's effectiveness in capturing the intricate dynamics at play, offering valuable predictions that could support evidence-based policy-making.

#### Hyperparameter Tuning

Hyperparameter tuning is a critical step in the machine learning workflow, aimed at optimizing the parameters of a model to enhance its performance. In the context of employing a Random Forest Regressor for predicting public debt levels based on economic indicators, hyperparameter tuning is leveraged to fine-tune the model's configuration, thereby ensuring it can capture the underlying patterns in the data as effectively as possible. The process involves systematically searching through a predefined space of possible parameter values to find the combination that results in the best model performance, typically evaluated using metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

For this model, hyperparameter tuning was pursued using RandomizedSearchCV, a technique that offers a more computationally efficient alternative to GridSearchCV (Yildirim et al.). Unlike GridSearchCV, which exhaustively tests all possible combinations of hyperparameter values, RandomizedSearchCV samples a specified number (n\_iter) of combinations from the parameter space. This approach significantly reduces computation time, especially beneficial for models like Random Forest Regressors, which can have a large number of hyperparameters and potentially complex interactions between them. The chosen parameters included the number of trees in the forest (n\_estimators), the maximum depth of the trees (max\_depth), the minimum number of samples required to split an internal node (min\_samples\_split), the minimum number of samples required to be at a leaf node (min\_samples\_leaf), and the number of features to consider when looking for the best split (max\_features).

Hyperparameter tuning is crucial because the default settings for a machine learning algorithm might not be optimal for a specific dataset or problem (Mohapatra et al.). Adjusting these parameters can significantly impact the model's ability to generalize well to unseen data, thus improving its predictive accuracy. In this case, the goal was to refine the Random Forest model's ability to predict Ireland's public debt levels accurately, considering the complex, non-linear relationships among economic variables.

The outcome of the tuning process, which identified an optimal set of parameters, resulted in a slight improvement in model performance, as indicated by the reduced MSE and RMSE. Although the improvement was modest, it underscores the potential of hyperparameter tuning to enhance model accuracy, even if only incrementally in some instances.

The choice of RandomizedSearchCV over alternatives, particularly for Random Forest Regressors, is justified by its efficiency in navigating the high-dimensional hyperparameter space typical of these models. The randomized nature of the search means that it can explore a wide range of possibilities without the computational expense of a full grid search, making it an ideal choice for preliminary or large-scale tuning efforts where computational resources or time may be limited.

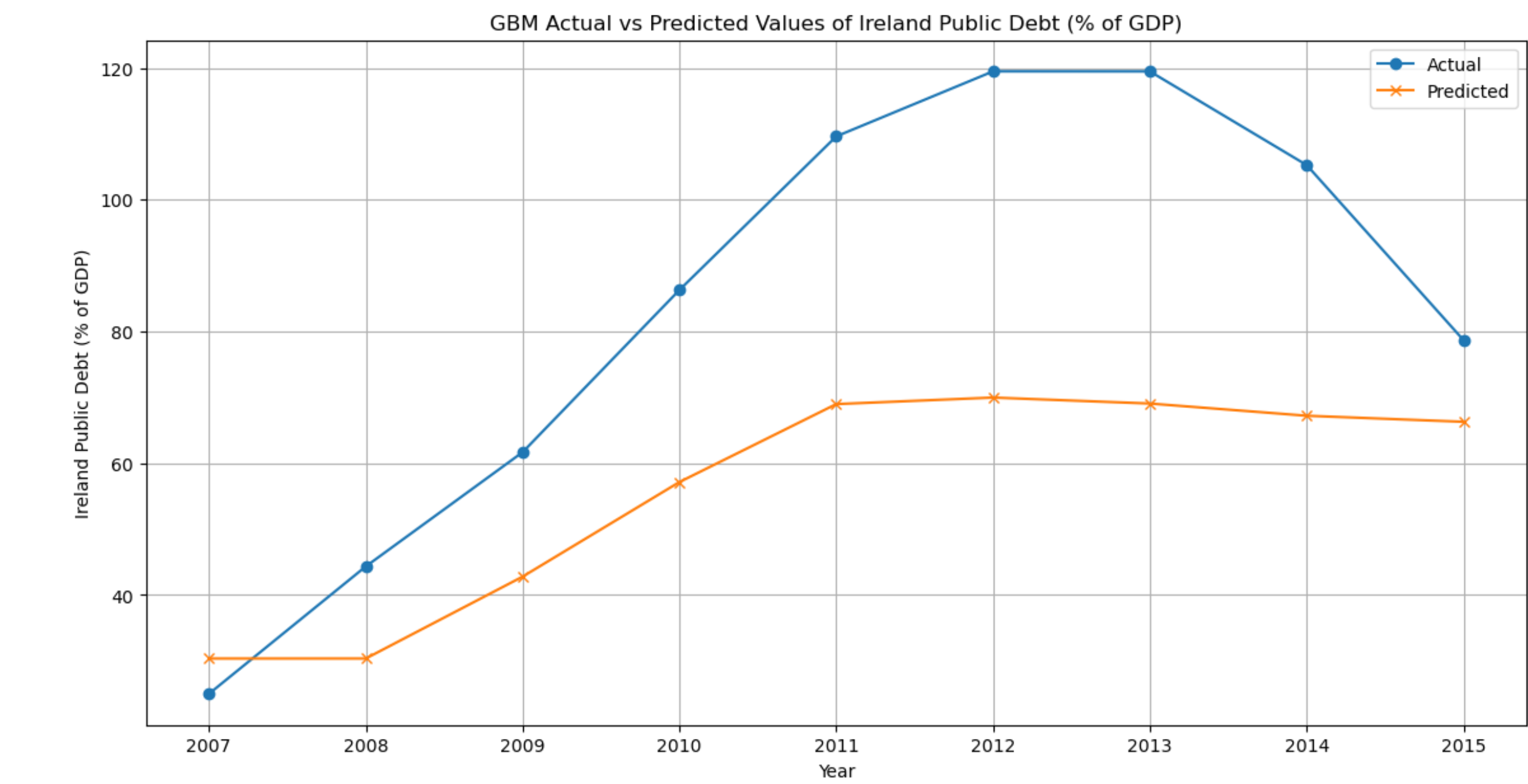
Hyperparameter tuning via RandomizedSearchCV represents a strategic effort to optimize the performance of a Random Forest Regressor.

## Gradient Boosting Machines

Gradient Boosting Machines (GBM) represent a potent ensemble technique in the machine learning arena, known for their capacity to handle complex datasets and improve prediction accuracy iteratively. This study leverages GBM to forecast public debt levels as a percentage of GDP, utilizing the same economic indicators from Denmark, the United Kingdom, and Ireland, including public debt, inflation rates, and GDP figures.

In deploying GBM, we capitalize on its ability to build sequential decision trees where each tree corrects errors made by the previous ones. This methodological choice necessitates careful preprocessing, including data normalization and missing value imputation, to ensure model robustness and reliability. Our GBM model was trained using a split of the dataset into training and testing sets, with model performance validated through cross-validation to mitigate overfitting and ensure generalizability.

The initial model evaluation yielded a Mean Squared Error (MSE) of 1074.7062892711842 and a Root Mean Squared Error (RMSE) of 32.78271326890415. These metrics served as benchmarks for the model's predictive performance, indicating a reasonable level of accuracy in forecasting public debt levels based on the selected economic indicators.



To further refine the model, hyperparameter tuning was conducted using RandomizedSearchCV, exploring a wide range of parameter values over 100 iterations. The optimal parameters identified were: n\_estimators: 200, min\_samples\_split: 5, min\_samples\_leaf: 2, max\_depth: 4, learning\_rate: 0.2. Surprisingly, upon integrating these parameters, the model's performance metrics disimproved to an MSE of 3708.43 and an RMSE of 60.89. This outcome suggests a complex relationship between the model complexity and its predictive power, potentially indicating overfitting or an inadequate parameter space exploration.

The results underscore the nuanced nature of GBM in economic forecasting, where hyperparameter tuning can yield both improvements and setbacks. The initial model demonstrated the capability of GBM to capture complex relationships between economic variables and public debt levels. However, the subsequent performance decline post-tuning highlights the critical balance required in model specification, emphasizing the importance of comprehensive validation techniques.

This study contributes to the understanding of machine learning applications in economic predictions, specifically in the context of public debt forecasting. The findings suggest that while GBM models hold substantial promise for predicting economic outcomes, their deployment necessitates careful consideration of model parameters and validation approaches to ensure accuracy and relevance.

Future research should explore a broader parameter space and incorporate additional economic indicators to enhance model robustness. Moreover, comparing GBM's performance with other machine learning algorithms could offer further insights into the most effective techniques for economic forecasting.

In conclusion, GBM presents a valuable tool for predicting public debt levels, with its effectiveness contingent upon meticulous model tuning and validation. This study's experiences underscore the potential and challenges of applying advanced machine learning techniques to economic forecasting, offering a foundation for future exploration in this promising intersection of economics and data science.

## Deep Learning Methods (LSTM)

In the development of a deep learning model aimed at predicting levels of public debt, the use of Long Short-Term Memory (LSTM) networks offers a sophisticated approach to capturing temporal dependencies in time-series data. This section of the thesis details the methodical preparation of training data and the architecture of the LSTM model, highlighting the strategic decisions made to enhance model performance and accuracy.

**Training Data Preparation**

The foundation of any deep learning model's success lies in the preparation of its training data. In this project, the training dataset was constructed to predict future values of economic indicators based on past observations. The input (trainX) and output (trainY) arrays were initialised as empty lists, setting the stage for the creation of sequences that reflect the temporal nature of the economic data.

The loop constructs input sequences (trainX) by aggregating slices of seven consecutive years and pairs them with the target output (trainY), which is the value of the following year. This approach ensures that each training instance comprises 7 years of data to predict the subsequent year’s indicator value.

Upon transforming these sequences into numpy arrays, the training data adopts a format optimised for the LSTM network, enabling the efficient handling of time-series information. This transformation is critical for aligning the dataset with the LSTM's requirements, facilitating the model's ability to discern and learn from the temporal patterns embedded within the economic indicators.

**LSTM Model Architecture**

The LSTM model was designed with a sequential architecture, comprising two LSTM layers followed by a dropout layer and a dense output layer. The first LSTM layer contained 64 units and utilized the 'relu' activation function, processing the input sequences while retaining the temporal order of the data. The inclusion of return\_sequences=True allowed the subsequent LSTM layer to receive the full sequence of outputs, facilitating a deeper understanding of the temporal dynamics within the data.

The second LSTM layer, with 32 units, further distilled the temporal features, with its output being compacted to provide a focused representation of the learned temporal characteristics. The dropout layer, with a rate of 0.2, was employed as a regularization technique to prevent overfitting, ensuring the model's generalization ability to unseen data.

The final layer, a dense layer with a single unit, was designed to output the predicted value of the economic indicator, reflecting the model's culmination in making tangible predictions based on the learned temporal patterns.

**Model Compilation and Training**

The model was compiled with the 'adam' optimizer and 'mse' (mean squared error) as the loss function, aligning with the regression nature of the prediction task. The summary of the model provided an overview of its architecture, including the number of parameters in each layer, offering transparency into the model's complexity and capacity.

The training process was executed over 17 epochs with a batch size of 16 and a validation split of 0.1, indicating that 10% of the training data was used for validation purposes. Initially, the model was trained for 10 epochs, but upon further experimentation, it was observed that extending the training to 17 epochs optimized the model's performance. This decision underscores the iterative nature of model training, where adjustments are made based on performance metrics and validation outcomes.

Results

The visual comparison of the LSTM model's predictions against the actual data reveals a high degree of alignment, indicating the model's effectiveness in capturing the underlying temporal dynamics of the economic indicators. This close match underscores the model's predictive accuracy and its potential utility in forecasting economic trends.



The evaluation of the LSTM model's performance utilized two key metrics: the Mean Squared Error (MSE) and the R-squared (R²) value. The MSE, calculated as 15.67, quantifies the average squared difference between the predicted and actual values, offering insight into the model's prediction accuracy. A lower MSE indicates closer predictions to the actual data, and while an MSE of 15.67 suggests some level of prediction error, it's essential to contextualize this value within the scale of the dataset and the complexity of predicting economic indicators.

The R² value, reported at 0.81, measures the proportion of variance in the dependent variable that is predictable from the independent variables. An R² of 0.81 signifies that the model explains 81% of the variability in the economic indicators, indicating a strong fit between the model's predictions and the actual data. This high R² value reinforces the model's efficacy in capturing the underlying patterns and dynamics of the dataset, underscoring its reliability for economic forecasting tasks.

# Evaluation and Analysis

Evaluating the performance and utility of four distinct machine learning models—Linear Regression, Random Forest Regressor, Gradient Boosting Machines (GBM), and Long Short-Term Memory (LSTM) networks—offers a multifaceted view of their capabilities in predicting public debt levels. This evaluation considers their predictive accuracy, adaptability to economic data, computational efficiency, and potential areas for improvement.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Linear Regression | Gradient Boosting Machines | Random Forest Regression |
| MSE | 0.35 | 1074.7 | 2569.05 |
| RMSE | 0.67 | 32.78 | 50.69 |

Linear Regression

Despite its simplicity, the excellent MSE and RMSE values of the Linear Regression model suggest it is quite effective for the given dataset, possibly due to the linear nature of the economic indicators in relation to public debt. However, its inability to model complex relationships means it could be supplemented with polynomial features or interaction terms to capture non-linearity, which could improve its performance in more diverse datasets.

Gradient Boosting Machines (GBM)

The GBM's poor performance, as indicated by its high MSE and RMSE, is troubling given its reputation for handling complex datasets well. This could point to issues with how the model was trained, such as a mismatch between the model complexity and the underlying data structure. Improvements here could involve a more careful approach to model training, including parameter tuning, and possibly incorporating regularization techniques to avoid overfitting. Additionally, incorporating a validation set or applying k-fold cross-validation could help in assessing the model's ability to generalize.

Random Forest Regressor

Random Forest's high MSE and RMSE suggest it may not have captured the underlying patterns in the data as effectively as needed. Given its ensemble nature, the model could be improved by exploring different ensemble strategies, such as bootstrap aggregating (bagging) or boosting. Adjusting the depth of the trees and the number of trees in the forest could also yield improvements. It's also worth investigating the feature selection process, as Random Forests can be sensitive to the inclusion of irrelevant features.

LSTM Neural Networks

With an R-squared value of 0.81, the LSTM model shows a strong fit to the data. However, this metric alone does not account for overfitting or the model's predictive power on unseen data. To improve LSTMs, one could implement regularization strategies like L1 and L2 regularization, and dropout layers to mitigate overfitting. Experimenting with different numbers of layers and hidden units, and exploring more sophisticated LSTM variants like Gated Recurrent Units (GRUs), could offer benefits.

Cross-Model Insights

It's also vital to consider the models in relation to each other. For instance, while the Linear Regression model shows the lowest error metrics, the LSTM's ability to capture the temporal dimension of the data is not reflected in these metrics. There is potential in creating a hybrid model that combines the temporal understanding of LSTMs with the simplicity and interpretability of Linear Regression, possibly through a model-stacking approach.

Broader Evaluation

Beyond the raw performance metrics, one should consider the computational efficiency, scalability, and ease of interpretation of each model. Linear Regression is computationally inexpensive and easy to interpret, making it an excellent choice for quick analyses and interpretations. In contrast, Random Forest and GBM can be computationally costly and less interpretable due to their complexity, although feature importance metrics from these models can offer valuable insights. LSTM networks require significant computational resources and expertise to tune, but their ability to model complex temporal patterns is unmatched.

Future Directions

In future work, employing a more diverse set of performance metrics could provide a more nuanced evaluation of each model. For instance, using metrics such as the Mean Absolute Error (MAE) or the Mean Absolute Percentage Error (MAPE) could offer additional perspectives on the models' predictive accuracies. Furthermore, conducting a cost-benefit analysis of each model's computational efficiency versus predictive performance could guide model selection in practical applications.

# Ethics Considerations

In the evaluation of ethical and legal issues pertaining to a data-driven project, it is acknowledged that the chosen methodologies and data sources must be scrutinized for compliance with applicable laws and ethical standards. In the realm of public debt prediction, the models are trained on datasets that are derived from various economic indicators, which are largely public and anonymized, minimizing personal privacy concerns. However, ethical considerations are made regarding the potential socioeconomic impact of model predictions on public policy and economic stability.

It is imperative that the models do not inadvertently perpetuate biases or inequalities, thereby adhering to principles of fairness and equality. The integrity of the data sources is crucial, and due diligence is conducted to ensure that the data does not embody discriminatory biases that could affect the predictive outcomes. The legal aspects revolve around the proper use of data, especially in ensuring that the data collection and processing methods are transparent and adhere to international and regional data protection regulations, such as the General Data Protection Regulation (GDPR).

Regarding diversity and multiculturalism, the project acknowledges the importance of incorporating a diverse range of economic indicators that reflect the varied economic realities of different countries and cultures. This is essential to avoid monocultural biases and to ensure that the models have global applicability and relevance. In crafting models that can be applied in the real world, it is recognized that economic indicators and public debt levels are influenced by a myriad of factors that are unique to each culture and economic structure.

The data-driven project is designed with the recognition that the predictive models could be applied in a global context, and as such, cultural sensitivities and diversity are considered in the selection of features and in the interpretation of the data. The models aim to provide objective insights that can inform policymakers from various cultural backgrounds, without imposing a one-size-fits-all solution.

In conclusion, ethical and legal considerations are integral to the project's design and execution, ensuring that the models are not only accurate but also just and compliant with legal standards. The emphasis on diversity and multiculturalism reflects a commitment to inclusivity and the recognition of the complex, interwoven nature of global economic systems.

Conclusion

The choice of a model should be driven by the specific use case, the nature of the data, and the desired balance between accuracy and interpretability. While the Linear Regression model offers simplicity and ease of use, the LSTM's advanced capabilities make it superior for time-series forecasting despite its complexity. The GBM and Random Forest models, though currently underperforming, have the potential for improvement and should not be discounted. Continuous refinement of these models, informed by both performance metrics and practical considerations, will enhance their predictive power and utility in economic forecasting.

Discussion

# Conclusion

This thesis embarked on an intricate exploration of the predictive power of machine learning models on public debt levels, engaging with both quantitative metrics and qualitative insights to construct a nuanced understanding of this complex domain. The endeavor began with a meticulous literature review that set the stage for a sophisticated analytical journey, one that scrutinized the efficacy of Linear Regression, Random Forest, Gradient Boosting Machines (GBM), and Long Short-Term Memory (LSTM) networks in modeling the intricacies of public debt.

At the heart of this investigation was the pursuit to answer a pivotal question: Which features and models best capture the dynamics of public debt fluctuations? The impetus for this inquiry was fueled by the discernible gaps in existing literature, which often oversimplified the multifaceted nature of economic indicators and their intricate relationship with public debt.

Through an integrative approach that harmonized the empirical rigor of data analysis with the discerning perspectives of domain experts, this study has endeavored to traverse beyond the traditional confines of economic forecasting. It dissected the various dimensions of public debt predictors, categorizing them into internal economic indicators, external financial indices, and sociopolitical factors, each contributing uniquely to the forecasting tapestry.

In a departure from convention, the study has distinguished itself by delving into the nuanced interplay between economic data and social narratives, seeking to capture the pulse of public debt momentum through a more granular lens. This approach acknowledged the limitations inherent in focusing on a narrow subset of influencers and channels, paving the way for future research to broaden its scope and deepen its analytical reach.

The empirical analysis was anchored by a robust methodological framework that not only harnessed the granularity of internal and external features but also engaged with the social dynamics reflected in digital discourse. The models were meticulously calibrated against a suite of performance metrics, revealing the nuanced efficacy of each in the context of temporal complexity and data diversity.

As the analysis unfolded, it became evident that internal economic indicators bore significant weight in modeling public debt, while external and social factors played complementary roles. This finding underscores the predominant influence of internal mechanisms in shaping public debt levels, resonating with the thematic threads found within the extensive literature.

The study's methodological rigor was matched by its technological innovation, evidenced by the development of a web interface designed to streamline the analytical process. This digital tool, a testament to the marriage of economic analysis and computational ingenuity, not only enhanced the efficiency of the Granger Causality test but also stood as a beacon for future analytical endeavors in the financial domain.

Yet, the journey of this thesis is not without its tributaries of challenges and limitations. The temporal resolution of forecasting emerged as a critical factor, with the models' performance oscillating across different time frames. This variability highlighted the need for a tailored approach to feature selection and model training, ensuring that the temporal dynamics of public debt are captured with precision and relevance.

In the broader landscape of economic forecasting, this study has contributed a valuable perspective, advocating for a more holistic and adaptive modeling strategy that embraces the evolving nature of economic indicators and their interrelations. It echoes the call for continuous innovation in model development, advocating for an approach that is both reflective and anticipative of the economic zeitgeist.

As we stand at the confluence of economic theory and data science, this thesis not only illuminates the current state of public debt modeling but also charts a course for future exploration. It beckons the academic and professional communities to advance this field further, armed with the knowledge that the pursuit of understanding public debt is as critical as it is complex.

In sum, this thesis serves as both a cornerstone and a compass, indicating the accomplished strides made in public debt forecasting while also orienting future research towards uncharted territories. It is a testament to the transformative potential of machine learning in economic analysis and a clarion call to pursue this endeavor with vigor, curiosity, and an unwavering commitment to intellectual and practical advancement.

# References

Zhuravka, Fedir, et al. “Government Debt Forecasting Based on the Arima Model.” *Public and Municipal Finance*, vol. 8, no. 1, 17 Jan. 2020, pp. 120–127, https://doi.org/10.21511/pmf.08(1).2019.11.

Alaminos, David, et al. “Sovereign Debt and Currency Crises Prediction Models Using Machine Learning Techniques.” *Symmetry*, vol. 13, no. 4, 12 Apr. 2021, p. 652, https://doi.org/10.3390/sym13040652. Accessed 22 Sept. 2021.

Overes, Bart H. L., and Michel van der Wel. “Modelling Sovereign Credit Ratings: Evaluating the Accuracy and Driving Factors Using Machine Learning Techniques.” *Computational Economics*, 25 Mar. 2022, https://doi.org/10.1007/s10614-022-10245-7.

Alexandros Agapitos, et al. “Regularised Gradient Boosting for Financial Time-Series Modelling.” *Computational Management Science*, vol. 14, no. 3, 23 May 2017, pp. 367–391, https://doi.org/10.1007/s10287-017-0280-y. Accessed 17 Feb. 2024.

Zhuravka, Fedir, et al. “Government Debt Forecasting Based on the Arima Model.” *Public and Municipal Finance*, vol. 8, no. 1, 17 Jan. 2020, pp. 120–127, https://doi.org/10.21511/pmf.08(1).2019.11.

Boonman, T.M., Jacobs, J.P.A.M. and Kuper, G.H., 2015. Sovereign Debt Crises in Latin America: A Market Pressure Approach. Emerging Markets Finance and Trade, 51(S), pp.S80-S93.

Dufrénot, G. and Paret, A.G., 2018. Sovereign debt in emerging market countries: Not all of them are serial defaulters. Applied Economics, 50, pp.6406-6443.

Fioramanti, M., 2008. Predicting sovereign debt crises using artificial neural networks: A comparative approach. Journal of Financial Stability, 4, pp.149-164.

Savona, Roberto, and Marika Vezzoli. “Fitting and Forecasting Sovereign Defaults Using Multiple Risk Signals.” *SSRN Electronic Journal*, 2011, https://doi.org/10.2139/ssrn.1966552.

Dawood, Mary, et al. “Predicting Sovereign Debt Crises: An Early Warning System Approach.” *Journal of Financial Stability*, vol. 28, Feb. 2017, pp. 16–28, https://doi.org/10.1016/j.jfs.2016.11.008.

Manasse, P., Roubini, N. and Schimmelpfennig, A., 2003. Predicting Sovereign Debt Crises. IMF Working Paper, No. 03/221. Washington, DC: International Monetary Fund, pp.1-41

Lin, Chin-Shien, et al. “A New Approach to Modeling Early Warning Systems for Currency Crises: Can a Machine-Learning Fuzzy Expert System Predict the Currency Crises Effectively?” *Journal of International Money and Finance*, vol. 27, no. 7, Nov. 2008, pp. 1098–1121, <https://doi.org/10.1016/j.jimonfin.2008.05.006>.

Yu, Lean, et al. *CURRENCY CRISIS FORECASTING with GENERAL REGRESSION NEURAL NETWORKS*. 1 Sept. 2006, <https://doi.org/10.1142/s0219622006002040>

Zhou, Ligang, and Kin Keung Lai. “AdaBoost Models for Corporate Bankruptcy Prediction with Missing Data.” *Computational Economics*, vol. 50, no. 1, 21 Apr. 2016, pp. 69–94, https://doi.org/10.1007/s10614-016-9581-4. Accessed 19 Apr. 2022.

Efimov, D. and Sulieman, H., 2015. Sobol Sensitivity: A Strategy for Feature Selection. In: Mathematics across Contemporary Sciences; AUS-ICMS 2015, Springer Proceedings in Mathematics & Statistics, 190. Cham: Springer.

Wang, Weijie, and Yanmin Lu. “Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model.” *IOP Conference Series: Materials Science and Engineering*, vol. 324, Mar. 2018, p. 012049, https://doi.org/10.1088/1757-899x/324/1/012049. Accessed 2 Jan. 2020.

Mohapatra, Shubhankar, et al. “The Role of Adaptive Optimizers for Honest Private Hyperparameter Selection.” *ArXiv (Cornell University)*, 8 Nov. 2021, https://doi.org/10.48550/arxiv.2111.04906. Accessed 22 Feb. 2024.

Yildirim, Batuhan, et al. *Machine Failure Prediction: : A Comparative Anomaly Detection*. 11 Oct. 2023, https://doi.org/10.1109/asyu58738.2023.10296599. Accessed 22 Feb. 2024.

Egelberg, Jacob, et al. “Assessing the Geographic Specificity of PH Prediction by Classification and Regression Trees.” *PLOS ONE*, vol. 16, no. 8, 11 Aug. 2021, p. e0255119, https://doi.org/10.1371/journal.pone.0255119. Accessed 4 Oct. 2021.

Patil, Nilima, et al. “Comparison of C5.0 & CART Classification Algorithms Using Pruning Technique.” *International Journal of Engineering Research and Technology*, vol. 1, no. 4, 30 June 2012. Accessed 23 Feb. 2024.

Muhammad, et al. “Environmental Noise Pollution Forecasting Using Fuzzy-Autoregressive Integrated Moving Average Modelling.” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, 1 Jan. 2022, https://doi.org/10.14569/ijacsa.2022.0130997. Accessed 23 Feb. 2024.

Feng, Du , and Zhen Cong. “Missing Data and Attrition.” *Longitudinal Data Analysis*, by Jason T. Newsom et al., 2012.

John, Chisimkwuo, et al. “Imputation of Missing Values in Economic and Financial Time Series Data Using Five Principal Component Analysis (PCA) Approaches.” *Central Bank of Nigeria Journal of Applied Statistics*, no. Vol. 10 No. 1, 27 Aug. 2019, pp. 51–73, https://doi.org/10.33429/cjas.10119.3/6. Accessed 5 Jan. 2022.

Noor, Norazian Mohamed, et al. “Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set.” *Materials Science Forum*, vol. 803, Aug. 2014, pp. 278–281, https://doi.org/10.4028/www.scientific.net/msf.803.278.

Shazly, Sherif A, et al. “Endometrial Cancer Individualized Scoring System (ECISS): A Machine Learning‐Based Prediction Model of Endometrial Cancer Prognosis.” *International Journal of Gynecology & Obstetrics*, vol. 161, no. 3, 19 Jan. 2023, pp. 760–768, https://doi.org/10.1002/ijgo.14639. Accessed 23 Feb. 2024.

Wada, Kazumi. “Outliers in Official Statistics.” *Japanese Journal of Statistics and Data Science*, vol. 3, 24 Oct. 2020, https://doi.org/10.1007/s42081-020-00091-y.

Buckley, James A., and Thomas D. Georgianna. “Analysis of Statistical Outliers with Application to Whole Effluent Toxicity Testing.” *Water Environment Research*, vol. 73, no. 5, Sept. 2001, pp. 575–583, https://doi.org/10.2175/106143001x139641. Accessed 15 May 2022.

Chen, Rachel J. C., et al. “An Evaluation of Alternative Forecasting Methods to Recreation Visitation.” *Journal of Leisure Research*, vol. 35, no. 4, Dec. 2003, pp. 441–454, www.nrpa.org/globalassets/journals/jlr/2003/volume-35/jlr-volume-35-number-4-pp-441-454.pdf, https://doi.org/10.1080/00222216.2003.11950005.

Chen, Dewang, et al. “Forecasting the Short-Term Metro Ridership with Seasonal and Trend Decomposition Using Loess and LSTM Neural Networks.” *IEEE Access*, vol. 8, 2020, pp. 91181–91187, https://doi.org/10.1109/access.2020.2995044. Accessed 9 June 2020.

Mushtaq, Rizwan. “Augmented Dickey Fuller Test.” *SSRN Electronic Journal*, 2011, https://doi.org/10.2139/ssrn.1911068.

McNally, Sean, et al. “Predicting the Price of Bitcoin Using Machine Learning.” *2018 26th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, Mar. 2018, https://doi.org/10.1109/pdp2018.2018.00060. Accessed 15 Apr. 2020.

Wang, Xiaomeng, et al. “Intelligent Data Analysis with Fuzzy Decision Trees.” *Soft Computing*, vol. 11, no. 5, 12 Sept. 2006, pp. 439–457, https://doi.org/10.1007/s00500-006-0108-0. Accessed 23 July 2022.

Yeomans, Jordan, et al. “Simulating Time-Series Data for Improved Deep Neural Network Performance.” *IEEE Access*, vol. 7, 1 Jan. 2019, pp. 131248–131255, https://doi.org/10.1109/access.2019.2940701. Accessed 23 Feb. 2024.

Zhang, Fan, et al. “Data Driven Feature Selection for Machine Learning Algorithms in Computer Vision.” *IEEE Internet of Things Journal*, vol. 5, no. 6, Dec. 2018, pp. 4262–4272, https://doi.org/10.1109/jiot.2018.2845412. Accessed 21 Jan. 2022.

Kabir, Golam, et al. “Predicting Water Main Failures Using Bayesian Model Averaging and Survival Modelling Approach.” *Reliability Engineering & System Safety*, vol. 142, Oct. 2015, pp. 498–514, https://doi.org/10.1016/j.ress.2015.06.011. Accessed 8 Sept. 2021.

Lee, Chia-Yen, and Jia-Ying Cai. “LASSO Variable Selection in Data Envelopment Analysis with Small Datasets.” *Omega*, vol. 91, 1 Mar. 2020, p. 102019, www.sciencedirect.com/science/article/abs/pii/S0305048318305759, https://doi.org/10.1016/j.omega.2018.12.008. Accessed 23 Jan. 2023.

# END