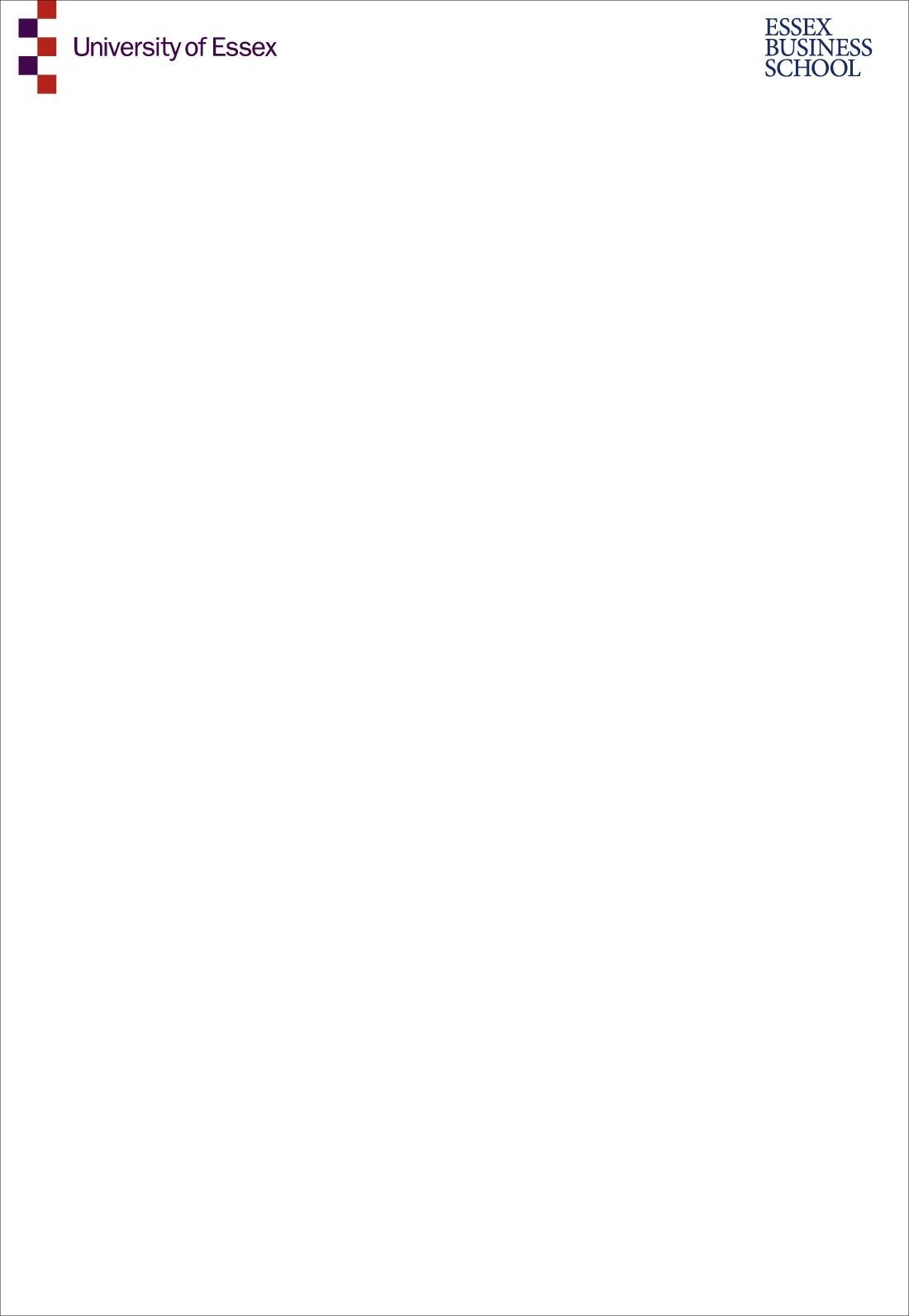
Dissertation (BE982)

**Dissertation (BE982)**

**Unemployment Forecasting using machine learning methods**

**Mages Waran Palani**

Acknowledgement

This thesis would not have been possible without the guidance, help, support and the amount of knowledge provided to me by Dr. Ilias chronopoulos. I would like to thank

Dr. Ilias chronopoulos for his valuable time.

I would also like to thank the University of Essex to provide me with an opportunity to complete my course in the University and allow me to conduct research. The university also provided with the required infrastructure which made it possible to get data and get the results.

Lastly, I would like to thank all my professors, family members and friends who were directly and indirectly involved in the process of my dissertation.

Abstract

Accurate forecasting of unemployment rates is crucial for effective economic policy and company choices. Traditional econometric models such as Autoregressive Integrated Moving Average (ARIMA) and Vector Autoregression (VAR) are extensively adopted, but they typically fall short of capturing the complex, non-linear connections present in modern economic data. This dissertation studies the application of machine learning algorithms to enhance unemployment forecasting accuracy using the FRED-MD dataset, which comprises 134 monthly economic indicators from the United States.

Several machine learning models were utilised and compared in this paper, including Random Forest, Gradient Boosting, Ridge Regression, Lasso Regression, and Support Vector Machines. To introduce nonlinearity and decrease dimensionality, the dataset underwent substantial preprocessing, including component estimation and feature engineering. The empirical results reveal that machine learning algorithms, especially Random Forest and Gradient Boosting, outperform classic econometric models for projecting unemployment rates. These algorithms effectively capture complex patterns in data, resulting in more accurate estimates.

The findings illustrate machine learning's potential to enhance unemployment estimates and deliver helpful insights for policymakers. The report closes with a discussion of the consequences for economic policy and suggestions for additional research into the application of advanced machine learning algorithms to macroeconomic forecasting.

**Table of Contents**

[1. Introduction 4](#_Toc176880121)

[1.1 Gaps in literature 9](#_Toc176880122)

[1.2 Research Questions 9](#_Toc176880123)

[2. Empirical Methodology 10](#_Toc176880124)

[2.1 Factor estimation 10](#_Toc176880125)

[2.2 Feature Engineering 11](#_Toc176880126)

[2.2.1 Random Forest 11](#_Toc176880127)

[2.2.2 Gradient boosting 12](#_Toc176880128)

[2.3 Regularization 14](#_Toc176880129)

[2.3.1 Ridge Regression 15](#_Toc176880130)

[2.3.2 Lasso Regression 16](#_Toc176880131)

[2.4 SVM 17](#_Toc176880132)

[2.5 Model Training 18](#_Toc176880133)

[2.6 prediction 18](#_Toc176880134)

[2.7 Model Evaluation 18](#_Toc176880135)

[2.8 Cross-Validation for Hyperparameter Tuning 19](#_Toc176880136)

[3. Empirical Results 21](#_Toc176880137)

[3.1 Model Performance Overview 21](#_Toc176880138)

[3.2 Analysis of Result 22](#_Toc176880139)

[3.2.1 Random Forest 22](#_Toc176880140)

[3.2.2 Gradient Boosting 23](#_Toc176880141)

[3.2.3 Ridge Regression 24](#_Toc176880142)

[3.2.4 Lasso Regression 25](#_Toc176880143)

[3.2.5 Support Vector Machines (SVM) 26](#_Toc176880144)

[3.3 Implications of Findings 26](#_Toc176880145)

[3.4 Random Forest Feature Importance 27](#_Toc176880146)

[3.4.2 Gradient Boosting Feature Importance 28](#_Toc176880147)

[3.5 Implications for Economic Forecasting 29](#_Toc176880148)

[4. Conclusion: 30](#_Toc176880149)

[References 31](#_Toc176880150)

**1. Introduction**

Unemployment is an important macroeconomic statistic that determines the overall economic well-being and policy choices. It effects consumer spending, investment levels, and fiscal policies, making it a key element for governments and economic planners. Traditional approaches for projecting unemployment rates have usually depended on econometric models like ARIMA (AutoRegressive Integrated Moving Average) and VAR (Vector Autoregression). These models, although helpful, typically fail to reflect the complicated, non-linear dynamics of labor market disruptions, particularly during times of economic instability and unforeseen shocks, such as the COVID-19 epidemic. Higher levels of worker disengagement result in less tax income and higher government spending on support benefits, retraining programs, and job placement services. This shows the larger social and economic ramifications of erroneous unemployment forecasts.

The limits of standard econometric models have been more evident in recent years, as labor markets have witnessed enormous upheavals caused by fast technology improvements, globalization, and unanticipated shocks. For instance, technological developments have led to both the displacement of jobs and the creation of new positions, changing employment patterns in ways that standard linear models fail to foresee. Globalization has injected additional complications into labor markets, with outsourcing and foreign commerce influencing domestic employment in unforeseen ways. The COVID-19 pandemic further showed the fragility of existing models, since their dependence on past data and assumptions of continuity led to poor predictions in such an unexpected setting.

In such unpredictable circumstances, the assumption of linearity and historical data continuity, which forms the foundation of ARIMA and VAR, becomes less valid, resulting in projections that may not adequately represent real-world labor market dynamics. These models generally depend on the idea that historical patterns can accurately predict future results, but when faced with large-scale, non-linear disturbances, they fail to adapt. Moreover, these models are typically unable to manage the high-dimensional, large-scale datasets that are currently ubiquitous in macroeconomic research. The sheer number and complexity of current data—from social media trends to real-time economic indicators—require more flexible, sophisticated ways.

To overcome these constraints, there has been an increasing movement toward the implementation of sophisticated machine learning (ML) approaches in macroeconomic forecasting. Unlike conventional models, ML approaches such as Random Forest and Gradient Boosting are especially well-suited to handle massive datasets and reveal nuanced patterns without having strong parametric assumptions. These models may uncover complicated, non-linear correlations between variables, making them more flexible to quickly changing labor market situations. Machine learning algorithms may also include a greater variety of data sources, including real-time data streams, allowing more responsive and accurate projections.

One major achievement in this discipline is the creation of hybrid machine learning models. For instance, Yurtsever (2023) presented a unique hybrid model that combines Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), both of which are deep learning architectures meant to capture time-dependent patterns in data. This hybrid model, evaluated on unemployment data from the United States, United Kingdom, France, and Italy, shows better accuracy above individual LSTM and GRU models. The hybrid technique exhibited greater predictive accuracy, notably for the U.S. and U.K., however in Italy, the GRU model beat the hybrid one. This demonstrates the potential of machine learning to give more accurate and dynamic solutions for unemployment forecasting in diverse scenarios.

Tattikota & Srinivasan (2021) introduced a method that combines econometric models like ARIMA with machine learning techniques such as Random Forest and K-Nearest Neighbors (KNN). Their study showed that this hybrid approach more effectively captures patterns in unemployment and inflation compared to traditional econometric methods. This highlights a broader industry trend where integrating machine learning with conventional econometric models has proven more robust in managing non-linear economic data.

Vinaya et al. (2023) conducted a comprehensive study using supervised machine learning models such as SVM, Random Forest, Gradient Boosting, and Extreme Machine Learning to predict future unemployment rates. Their findings demonstrated that machine learning models offer superior accuracy compared to traditional methods, especially in scenarios with volatile economic conditions and non-linear relationships. Vinaya's research underscores the growing adoption of machine learning techniques in macroeconomic forecasting, highlighting their effectiveness in delivering more precise unemployment estimates from complex datasets.

Celbiş (2023) emphasized the role of machine learning in predicting rural unemployment across Europe. His research employed various machine learning techniques, such as Random Forest and Gradient Boosting, to examine individual-level factors affecting unemployment in rural settings, focusing on education, age, and gender disparities. The findings revealed that access to training programs and parental education significantly impacted employment chances, with machine learning models offering greater flexibility in uncovering these nuanced relationships. The study demonstrated the effectiveness of machine learning approaches in rural labor markets, underscoring the need for targeted policies to address these disparities in areas with distinct economic challenges.

Recent studies have further demonstrated the effectiveness of these models. Yoon (2021) applied Gradient Boosting and Random Forest techniques to forecast Japan's real GDP growth. The results were benchmarked against the traditional forecasts provided by the International Monetary Fund (IMF) and the Bank of Japan (BOJ). The study found that machine learning models outperformed conventional methods, delivering more accurate and reliable predictions. This accomplishment highlights the capability of machine learning models in macroeconomic forecasting and supports their application in predicting other key economic indicators, such as unemployment rates.

Ahmad et al. (2021) also explored the application of hybrid models that integrate traditional econometric methods with machine learning techniques. Their research focused on predicting unemployment rates across various European countries by combining ARIMA, Support Vector Machines (SVM), and Artificial Neural Networks (ANN). The study demonstrated the remarkable effectiveness of these hybrid models in accurately capturing the non-linear and abrupt changes in unemployment rates induced by the pandemic. This finding reinforces the shift towards more advanced, machine learning-based approaches for economic forecasting, especially in response to significant economic disruptions.

Güler et al. (2024) investigated machine learning models, specifically ANN, SVM, and XGBoost, for predicting unemployment rates in Turkey. Their study revealed that these models outperformed traditional methods, particularly in analyzing the effects of various economic factors such as inflation, currency exchange rates, and labor market dynamics. Using TURKSTAT data from 2005 to 2023, they demonstrated that machine learning models provided more accurate forecasts of unemployment rates compared to conventional statistical methods, which struggled with non-linear relationships during periods of economic volatility. The findings underscore the growing adoption of machine learning techniques in macroeconomic forecasting, due to their ability to handle large datasets, accommodate complex dynamics, and deliver improved predictive accuracy.

Although traditional econometric models such as ARIMA and VAR are frequently utilized, they have significant limitations, particularly in capturing the complex, non-linear patterns often observed in unemployment data. These models generally assume that past trends will continue into the future, making them less effective when faced with unexpected economic shocks, such as the COVID-19 pandemic. This reliance on historical data can result in inaccurate predictions when the economic environment shifts abruptly. Additionally, VAR models, which aim to enhance forecasting accuracy by incorporating multiple macroeconomic variables like GDP, unemployment, and inflation, can be susceptible to overfitting, especially if too many variables or lags are included. Overfitting can degrade prediction accuracy, particularly in volatile economic conditions, as the model's complexity may hinder its ability to generalize beyond the sample data. While VAR models are designed to capture interdependencies among variables, their practical application often falls short in fully addressing the non-linear economic dynamics that influence unemployment rates during periods of significant disruption.

In support of the transition to machine learning, Maehashi and Shintani (2020) performed an in-depth analysis of machine learning techniques applied to Japanese macroeconomic data, highlighting their advantage over traditional autoregressive models, especially for medium to long-term forecasts. Their findings emphasize the crucial need for models that incorporate non-linear dynamics and interactions between variables factors often overlooked by conventional econometric methods.

Similarly, Cicceri et al. (2020) utilized machine learning techniques to forecast economic recessions in Italy, focusing on changes in GDP. Their study revealed that advanced machine learning models, such as Nonlinear Autoregressive models with exogenous variables (NARX) and Support Vector Regression (SVR), significantly improved prediction accuracy compared to traditional statistical methods. These models excelled at forecasting periods of recession, demonstrating the superior capability of machine learning techniques in capturing complex economic patterns that conventional models often miss.

Adding to the evidence of machine learning’s forecasting capabilities, Gabrikova et al. (2023) utilized ensemble learning methods, including CART, CHAID, and discriminant analysis, to predict the duration of unemployment among jobseekers in Slovakia. Their study demonstrates the effectiveness of ensemble techniques in unemployment forecasting and underscores the increasing adoption of machine learning for complex economic predictions. By combining CART, CHAID, and discriminant analysis, the ensemble model achieved a prediction accuracy of 78%, significantly surpassing traditional methods. The study also emphasized the model’s ability to identify long-term unemployment risks, which is valuable for policymakers. This further supports the expanding body of research advocating for advanced machine learning techniques to enhance the precision and reliability of unemployment forecasts, especially in volatile economic conditions.

Expanding on these advancements, this dissertation explores the application of machine learning techniques, including Random Forest, Gradient Boosting, Ridge Regression, Lasso Regression, and Support Vector Machines, to enhance the precision of unemployment rate forecasts. Utilizing the FRED-MD dataset, which encompasses 134 monthly economic indicators from the United States, this research aims to determine whether machine learning methods can deliver more accurate and reliable predictions compared to traditional econometric models.

Historically, ARIMA and GARCH models have been commonly used to forecast unemployment rates over the past two decades. For example, Dobre and Alexandru (2008) applied ARIMA in Romania, while Floros (2005) used GARCH in the UK. Similarly, Mladenovic et al. (2017) employed ARIMA models for unemployment estimation in the EU28, and Kurita (2010) utilized ARIMA for forecasting unemployment in Japan. These studies illustrate the continued reliance on traditional econometric models, despite their known limitations in addressing non-linear dynamics and complex relationships within economic data.

The limitations of traditional econometric models like ARIMA and GARCH have led researchers to explore machine learning techniques for forecasting unemployment. Machine learning algorithms excel in processing large datasets and uncovering complex, non-linear relationships that conventional models may overlook. Dzhunkeev (2022) highlights the growing importance of these models, noting their superior performance in predicting unemployment during economic downturns, a period when traditional models frequently fall short.

As machine learning models continue to advance, they have demonstrated significant advantages over traditional econometric approaches in several key areas. Firstly, these models offer remarkable flexibility, allowing for adjustments through hyperparameter tuning that enhances their ability to generalize and make precise predictions across diverse economic conditions. Additionally, machine learning models excel at identifying non-linear relationships and complex interactions between economic variables, which linear models like ARIMA and GARCH often miss.

Furthermore, integrating machine learning with big data analytics enables the inclusion of a broader range of factors in forecasting models, thereby improving accuracy. For example, modern unemployment predictions can now incorporate real-time data from sources such as social media sentiment, online job postings, and economic policy announcements, providing a more comprehensive view of the variables affecting unemployment trends.

Despite the significant advantages offered by machine learning models, they do come with challenges, such as the need for large datasets and substantial computational resources. Nonetheless, ongoing advancements in processing capabilities and the creation of more efficient algorithms are gradually overcoming these limitations, making machine learning an increasingly viable and valuable tool for economic forecasting.

While traditional econometric models like ARIMA and GARCH have provided essential insights into unemployment forecasting for many years, the integration of machine learning represents a significant advancement. Leveraging machine learning's capabilities enables researchers and policymakers to gain a more nuanced understanding of unemployment dynamics, which can lead to better-informed economic decisions and more effective labor market policies. As a result, machine learning is set to become a pivotal component in the future of macroeconomic forecasting.

## **1.1 Gaps in literature**

In addressing shortcomings found in the current literature, my study provides many crucial additions. First, while previous studies have utilized hybrid models combining econometric and machine learning techniques, I have expanded this by integrating more advanced methods, such as Ridge Regression and Lasso Regression, along with Random Forest and Gradient Boosting, to provide a more robust approach to forecasting. This allows for a more thorough comparison of models, notably in capturing the non-linear dynamics of unemployment. Also, I’ve employed polynomial feature engineering, which is not typically addressed in earlier research, to better account for non-linear interactions between unemployment and macroeconomic factors. Moreover, the paper deals with high-dimensional data more successfully by adding factor models to minimise dimensionality, enhancing model efficiency. Finally, I’ve employed substantial cross-validation and hyperparameter tuning approaches, ensuring the models are optimized for accuracy, which boosts the predictive capabilities compared to past research that didn’t fully utilize these techniques. Through these approaches, I’ve addressed the constraints discovered in the previous research and pushed forward the bounds of unemployment forecasting using machine learning.

## **1.2 Research Questions**

1. How do machine learning models such as Random Forest, Gradient Boosting, Ridge Regression, Lasso Regression, and Support Vector Machines compare in terms of accuracy when predicting unemployment rates?

2. Can machine learning approaches, along with feature engineering and dimensionality reduction (e.g., polynomial features and component models), outperform classic econometric models like ARIMA and VAR in unemployment forecasting?

3. To what degree can hyperparameter tweaking and cross-validation boost the predictive powers of machine learning models in unemployment forecasting compared to earlier studies?

4. What role do non-linear interactions between macroeconomic factors and unemployment play in increasing the accuracy of unemployment projections using machine learning models?

5. How do hybrid models that integrate machine learning techniques with classic econometric approaches perform with solely machine learning-based models in terms of predicting unemployment during times of economic instability?

6. What are the important macroeconomic factors that substantially impact unemployment rates, and how can machine learning models detect and prioritize these variables via feature importance?

# **2. Empirical Methodology**

In this section, we prepare by loading and processing a large high-dimensional macroeconomic dataset, which consists of 134 monthly U.S. economic indicators. The dataset is sourced from the FRED-MD database, which is curated by the Federal Reserve Bank of St. Louis. As described by McCracken and Ng (2016), this dataset is specifically designed to support empirical macroeconomic research by providing a standardized and consistent set of economic indicators.

The data is loaded from a CSV file with date vectors and transformation codes (‘tcode’) extracted. The data undergoes preprocessing where missing values are handled using methods such as interpolation or imputation depending on the transformation code associated with each series. Outliers are identified and removed to ensure robust model estimation.

## **2.1 Factor estimation**

In this section we have utilized the factor models to reduce the dimensionality of the dataset by estimating latent factors that captures the common variation across multiple time series. This approach described by stock and Watson (2002)

Where contains the factors loadings, is the vector of latent factors are time , and is the vector of idiosyncratic error term.

Where PCp​(r) is the penalized objective function used to select the number of factors. N is number of observed variables and T is the number of time period. represent the estimated factors loadings for the i-th variable, while represents the estimated factors at time t. r is a regularization parameter that controls the trade-off between fit and complexity. as a penalty term that increases with the number of parameters, typically a function of N and T.

## **2.2 Feature Engineering**

In this section Xpoly is a variable representing the set of features that have been engineered to include both the original factors (Fhat) and their polynomial transformations. The Xpoly is to introduce non-linearity into the feature set. By adding squared terms, you allow the models to capture quadratic relationships between the predictors and the target variable.

A group of black text

Description automatically generated

This work used and compared many machine learning models, such as Random Forest, Gradient Boosting, Ridge Regression, Lasso Regression, and Support Vector Machines.

The selection of these models was based on their capacity to handle high-dimensional and intricate macroeconomic information, as well as to capture non-linear connections among variables. Machine learning techniques such as Random Forest, Gradient Boosting, and SVM are better than classic econometric models like ARIMA and GARCH in identifying complex patterns in data, particularly when dealing with large fluctuation in unemployment rates. Finally, regularised models like as Ridge and Lasso Regression are effective in reducing overfitting, which makes them well-suited for predicting situations involving big, multicollinear datasets.

### **2.2.1 Random Forest**

**Random Forest** is an ensemble learning method, which means it combines the predictions from multiple models to produce a more accurate and robust prediction. Specifically, it builds multiple decision trees during the training process and outputs the average prediction of all the trees.

In this methodology, the model is applied to the engineered feature set Xpoly to produce predictions for the target variable. Random Forest is an ensemble learning method, which aggregates the predictions from multiple decision trees to improve accuracy and robustness.

Where b is the total number of decision trees in the forest. represents the prediction from the b-th decision tree using the feature set which includes the original factors and their polynomial terms.

The random forest leverage two techniques:

1. **Bootstrap Sampling:** Each tree is trained on a separate subset of the data, which is generated by randomly picking samples with replacement. This guarantees that each tree is trained on slightly different data, increasing the variety of the trees.

2. **Random Feature Selection:** Each split in the decision tree selects a random collection of features from Xpoly​. This unpredictability reduces association across trees, resulting in a more effective and diversified ensemble.

These strategies assist to prevent overfitting and improve the model's generalisability. The final prediction is derived by averaging the forecasts of all the trees in the forest, lowering variance and increasing overall prediction accuracy.

The Random Forest algorithm, first published by Breiman in 2001, is very proficient at managing big datasets with varied properties. It is renowned for its ability to prevent overfitting by randomly selecting data points and features. Random Forest has shown its efficacy in capturing non-linear correlations and interactions between important macroeconomic elements that are often overlooked by standard models, particularly in the field of unemployment forecasting.

### **2.2.2 Gradient boosting**

Gradient Boosting is an ensemble learning approach that develops models in a sequential manner, with each new model seeking to repair the flaws of prior models. Unlike Random Forest, which trains trees individually, Gradient Boosting trains trees one at a time, with each tree focussing on the residuals (errors) of the prior trees.

In this methodology, the Gradient Boosting model to forecast the target variable using the designed feature set X\_poly​. The approach improves accuracy and resilience by combining predictions from different decision trees.

M is the total number of boosting iterations. represents the prediction from the m-th decision tree using the feature set Xpoly.​ is the learning rate or weight assigned to the m-th decision tree, controlling the contribution of each tree to the final prediction.

**The Gradient Boosting model uses the following main techniques:**

**1. Sequential Model Building**: Each tree in the series is trained to correct the cumulative residuals (errors) of the prior models. This is accomplished by fitting the new tree to the negative gradient of the loss function in relation to the model's predictions.

**2. Learning Rate**: The learning rate controls the contribution of each tree to the overall model. A smaller learning rate makes the model more robust to overfitting by reducing the impact of each individual tree, but it also requires more trees to achieve good performance.

**3. Residual Fitting:** instead of fitting a tree directly to the target variable ,each tree is fitted to the residuals, which are the differences between the actual values and the predictions provided by the ensemble of trees constructed thus far. This iterative procedure continues until the model's predictions match the actual target values.

Gradient Boosting may use these tactics to create a powerful prediction model that is especially good at dealing with complicated, non-linear correlations in data. Gradient Boosting achieves excellent accuracy and generalisability by focussing on mistake correction at each stage.

Gradient Boosting is a machine learning technique that was first described by Friedman in 2001. It is a robust ensemble approach that constructs models in a sequential manner, where each subsequent model rectifies the mistakes caused by the preceding models. Unlike Random Forest, which produces trees separately, Gradient Boosting trains trees one at a time, where each tree is meant to minimize the residual errors of the previous models.

This technique is very efficient for predicting intricate, nonlinear patterns, such as the fluctuations in unemployment rates during economic recessions. Gradient Boosting is well acknowledged for its capacity to provide very precise predictions, however it requires meticulous adjustment of hyperparameters to prevent overfitting.

In unemployment forecasting, both Random Forest and Gradient Boosting have been used to predict the complicated links between unemployment and major economic factors, such as inflation, GDP growth, and interest rates. These models often surpass classic econometric methods, especially in periods of economic instability, when conventional models may have difficulty capturing these non-linear processes.

Yet, despite their greater accuracy, these models have several limits. Gradient Boosting, in particular, demands tremendous processing resources and accurate hyperparameter adjustment to obtain best performance. However, both Gradient Boosting and Random Forest are still crucial methods for economic forecasting, especially in capturing the nuanced relationships that impact unemployment rates.

## **2.3 Regularization**

In high-dimensional datasets, models tend to overfit by collecting noise in the data rather than recognising the genuine underlying patterns. To prevent overfitting, regularization incorporates a penalty term into the model’s loss function. This penalty discourages high coefficients and leads to smaller, more generalizable models.

There are two types of regularization:

* **L2 Regularization (used in Ridge Regression):** This penalty is proportional to the sum of squared coefficients, which decreases the influence of each predictor without fully removing any**.**
* **L1 Regularization (used in Lasso Regression):** This penalty is proportional to the total of absolute values of the coefficients, which decreases certain coefficients to zero, thereby doing feature selection.

Ridge Regression is the inclusion of an L2 regularization term in loss function. This term penalises high coefficients by imposing a penalty proportionate to the square of the coefficients. The regularisation parameter (𝜆) balances good data fit with modest coefficients. Increasing the 𝜆 value reduces overfitting and improves model generalisability by decreasing coefficients.

Ridge Regression gradually reduces the coefficients of less relevant predictors to zero, but it never completely wipes them out unlike Lasso Regression. This shrinkage is particularly effective in high-dimensional situations with multicollinearity since it minimises model variance, making it more stable and dependable for prediction.

### **2.3.1 Ridge Regression**

Ridge Regression, a regularised linear regression model, uses an L2 penalty to prevent overfitting in high-dimensional datasets with multiple predictors or strong correlation. This penalty term reduces the coefficients, lowering the model's variance at the expense of adding some bias, which frequently results in higher predictive accuracy.

Ridge Regression is used in this approach to analyse the engineered feature set X poly, which comprises both the original components and their polynomial modifications. The goal is to provide accurate and robust predictions of the target variable by minimising the loss function while accounting for model complexity via regularisation.

Where is the predicted value of the target variable at time t. is the vector of features at time t. is the vector of coefficients. is the error term.

Ridge Regression, created by Hoerl and Kennard in 1970, is another kind of linear regression that incorporates a regularization element. Unlike Lasso, Ridge imposes an L2 penalty (the square of the coefficients), which helps avoid overfitting by decreasing the coefficients of less relevant variables, but unlike Lasso, it does not push any coefficients to zero.  
  
Ridge is especially beneficial when predictors are strongly correlated, a frequent circumstance in unemployment forecasting where economic factors like GDP growth, inflation, and interest rates commonly display multicollinearity. By decreasing the influence of multicollinearity, Ridge Regression gives more stable and trustworthy coefficient estimates, leading to more accurate unemployment forecasts.  
  
Yet, like Lasso, Ridge needs careful selection of the regularization parameter (lambda). Tuning lambda helps the model balance between fitting the data well and avoiding overfitting. A larger lambda might unduly punish coefficients, leading to a biased model, while a lower lambda may enable overfitting, therefore it's vital to find the value that minimizes prediction error while retaining stability.

### **2.3.2 Lasso Regression**

Lasso Regression (Least Absolute Shrinkage and Selection Operator) is a linear regression method that uses an L1 penalty to reduce certain coefficients to zero. This feature enables Lasso to do variable selection, which is very beneficial when working with high-dimensional data when certain predictors may be useless.

Where, represents the target variable's actual value at time t. The predictor vector at time t is denoted by Xt. 𝛽 represents the vector of coefficients. The regularisation parameter λ determines the penalty based on the absolute magnitude of the coefficients.

Lasso Regression, or Least Absolute Shrinkage and Selection Operator, is a kind of linear regression that adds a regularization component in its cost function. Introduced by Tibshirani in 1996, Lasso is especially successful in circumstances when the number of variables is enormous, and many are likely irrelevant. By adding an L1 penalty (the absolute sum of coefficients), Lasso drives particular coefficients to decrease to zero, thereby accomplishing variable selection.

This property makes Lasso an extremely helpful tool in unemployment forecasting, where several economic variables could be implicated. By picking just the most relevant variables, Lasso simplifies the model, eliminating overfitting while maintaining the most influential predictors. This capacity to concentrate on the core determinants of unemployment guarantees that the model stays interpretable, even when working with complicated, high-dimensional datasets.

Yet, the performance of Lasso relies strongly on the selection of the regularization strength (lambda). A greater lambda increases the penalty, leading to simpler models but possibly neglecting significant predictors. A smaller lambda might lead to overfitting, therefore careful calibration is important to find a compromise between model simplicity and predictive power

## **2.4 SVM**

The Support Vector Machine (SVM) is a strong supervised learning technique that can do both classification and regression tasks. In the context of regression, SVM seeks to identify a function that deviates from the actual observed targets by no more than a specified margin while remaining as flat as feasible. When combined with proper kernel functions, SVM is extremely successful at managing nonlinear connections in data.

SVM regression also known as Support Vector Regression aims to identify a function 𝑓 ( 𝑋 ) that approximates the target variable 𝑦 within a margin of tolerance 𝜖 . The goal function for SVM regression with a kernel function is written as follow:

Where w is the weight vector. The regularisation parameter C balances the flatness of f(X) with the tolerance for deviations greater than ϵ. ϵ represents the margin of tolerance (insensitivity zone). f(Xi) represents the expected value for the i-th observation.

The inclusion of kernel functions enables SVM to operate in a high-dimensional feature space without explicitly computing the coordinates in that space, making it an appropriate choice for capturing nonlinear connections.

Support Vector Machines (SVM) are another machine learning tool that has shown great potential in forecasting economic indicators, including unemployment rates. SVMs are particularly powerful in high-dimensional spaces and are known for their robustness in avoiding overfitting.

In the context of regression, SVM works by finding a function that deviates from the observed target values by no more than a specified margin, while remaining as flat as possible. This margin is controlled by a parameter known as epsilon, and the regularization parameter (C) balances the model’s ability to tolerate deviations and maintain simplicity.

SVM’s ability to model non-linear interactions is enhanced by kernel functions, which map the input data into a higher-dimensional space without needing to compute the coordinates explicitly. This makes SVM especially effective in capturing non-linear relationships between unemployment rates and various economic variables.

SVM has proven to be a reliable tool for unemployment forecasting, particularly when dealing with large, complex datasets. Its flexibility to handle non-linearities and its robustness against overfitting make it an excellent alternative to traditional econometric models.

## **2.5 Model Training**

A mathematical equation with a number of letters

Description automatically generated with medium confidence

(T train) represents the number of observations in the training set. The optimisation aims to identify the coefficient vector ( ) that minimises the loss function while balancing fit to training data and penalty for big coefficients.

## 2.6 prediction

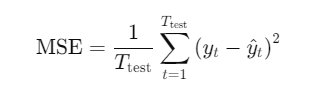
After training, the Ridge Regression model is used to predict the target variable from new data. The optimised coefficients (​) are used to calculate the forecast for a particular observation.

These predictions are then compared to the actual target values to determine the model's performance.

## **2.7 Model Evaluation**

The Ridge Regression model's performance is assessed using common metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), and . These metrics shed light on the model's capacity to generalise to new data as well as the accuracy of its predictions.

**1. Mean Squared Error (MSE):**



**2. Mean Absolute Error (MAE):**

A mathematical equation with numbers and symbols

Description automatically generated

**3. R-squared (R²):**

A mathematical equation with numbers and symbols

Description automatically generated with medium confidence

## **2.8 Cross-Validation for Hyperparameter Tuning**

To ensure the accuracy and generalisability of the machine learning models produced in this work, a meticulous hyperparameter tuning method was used, including cross-validation. Cross-validation is a statistical technique used to assess the proficiency of machine learning models. It is especially valuable for evaluating the model's ability to apply to a separate dataset. By partitioning the data into many groups and training the model on each of these subsets, we can assess its performance with more resilience.

The optimisation of important hyperparameters was performed using a 5-fold cross-validation strategy for all the machine learning models. This approach entails dividing the training data into five distinct subsets, sometimes known as "folds." The model is trained using four of these folds and verified using the remaining fold. This method is iterated 5 times, where each fold is used as the validation set once. Subsequently, the outcomes are averaged to yield a more precise assessment of the model's performance. This methodology aids in mitigating overfitting and guarantees that the model's success is not dependent on a particular subset of the data.

The specific application of cross-validation to each model is detailed below:

**Random Forest:** The number of trees in the ensemble, a crucial hyperparameter, was picked from a set of possible values (50, 100, 150, and 200). For each value, the model was trained and verified using 5-fold cross-validation. The best number of trees was obtained by picking the value that resulted in the lowest cross-validated Mean Squared Error (MSE).

**Gradient Boosting:** The number of boosting iterations, generally referred to as the number of trees, was improved via cross-validation. The potential values for this hyperparameter were also 50, 100, 150, and 200. The model with the lowest average MSE over the five folds was picked as the optimum configuration.

**Ridge Regression:** The regularization parameter, lambda (λ), which controls the strength of the regularization, was tuned using cross-validation. Lambda values were chosen from a logarithmic scale ranging from 1e-4 to 1e4. The lambda value that minimized the cross-validated MSE was selected for the final model.

**Lasso Regression:** Like Ridge Regression, Lasso Regression likewise needs the selection of an ideal lambda value. A logarithmic scale of lambda values, same to that used for Ridge Regression, was utilised. Cross-validation confirmed that the chosen lambda struck the optimal balance between model fit and simplicity, avoiding both underfitting and overfitting.

**Support Vector Machine (SVM**): The essential hyperparameter for SVM, known as the Box Constraint (C), was tuned by cross-validating across a range of values from 1e-3 to 1e3. The C value that offered the best trade-off between bias and variance, as evidenced by the lowest cross-validated MSE, was selected for the final model configuration.

# **3. Empirical Results**

In this section, we report the outcomes of the forecasting models utilised to anticipate the U.S. unemployment rate. The performance of each model is assessed based on important metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R²). These metrics give insights into the correctness and dependability of each model in capturing the underlying patterns in the data.

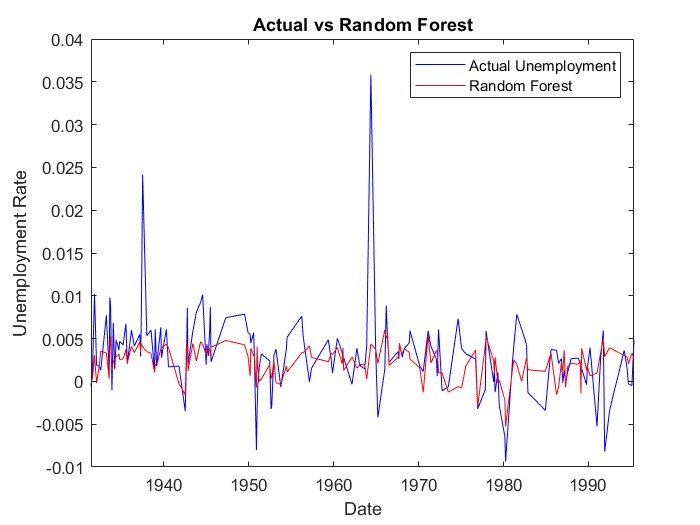
# **3.1 Model Performance Overview**

The table below summarizes the performance of each model:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Mean squared Error | Mean absolute Error | R-squared |
| Random Forest | 1.8348e-05 | 0.0026603 | 0.17487 |
| Gradient Boosting | 4.6136e-05 | 0.004885 | -1.0747 |
| Ridge Regression | 2.2482e-05 | 0.0029875 | -0.011012 |
| Lasso Regression | 2.2778e-05 | 0.003013 | -0.024351 |
| Support Vector Machine (SVM) | 2.213e-05 | 0.00296 | 0.004812 |

## **3.2 Analysis of Result**

### **3.2.1 Random Forest**

****

**Fig 1: Actual unemployment vs Random Forest**

The **Random Forest model** emerged as the most effective one for this analysis. It achieved the lowest Mean Squared Error (MSE) of 1.8348e-05 and the lowest Mean Absolute Error (MAE) of 0.0026603 among all the models. Furthermore, its R-squared value of 0.17487 indicates that this model was able to explain around 17.49% of the variance in the unemployment data, which is modest but notable. Random Forest's ensemble learning approach, which involves random sampling of features and bootstrap aggregating (bagging), helped reduce overfitting and enhanced the model's generalization to unseen data. These factors contributed to its ability to capture the complex, non-linear relationships present in the dataset, making it the most robust model in this analysis.

### **3.2.2 Gradient Boosting**

**A graph showing the actual and gradient boosting

Description automatically generated**

**Fig 2: Actual unemployment vs Gradient Boosting**

Gradient Boosting, typically known for its high accuracy, performed poorly in this case. It recorded the highest MSE of 4.6136e-05, the highest MAE of 0.004885, and a negative R-squared value of -1.0747, suggesting that its predictions were even worse than a simple mean-based model. The poor performance indicates that Gradient Boosting may have struggled with overfitting or suboptimal tuning of hyperparameters. While Gradient Boosting is usually effective at improving performance by reducing bias through sequential training of models, in this instance, it could not capture the non-linear patterns in the unemployment data effectively, resulting in its underperformance.

### **3.2.3 Ridge Regression**

**A graph showing the difference between unemployment and ridge regression

Description automatically generated**

**Fig 3: Actual unemployment vs Ridge Regression**

The **Ridge Regression model** provided a reasonable balance between prediction accuracy and error minimization. It achieved an MSE of 2.2482e-05 and an MAE of 0.0029875. However, its R-squared value of -0.011012 suggests that the model was not particularly adept at explaining the variability in the dataset. Ridge Regression’s use of L2 regularization helped in controlling overfitting by penalizing large coefficients, but its linear nature limited its ability to handle the more complex, non-linear relationships within the unemployment data, which hindered its overall performance.

### **3.2.4 Lasso Regression**

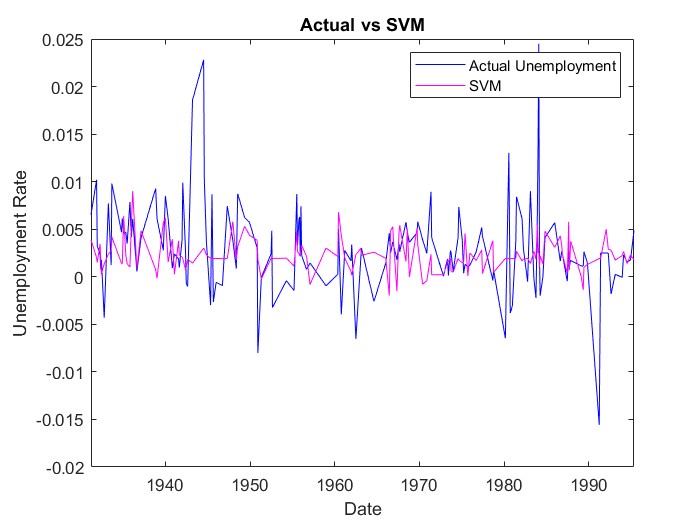
**A graph showing the time line

Description automatically generated**

**Fig 4: Actual unemployment vs Lasso Regression**

Lasso Regression, which uses L1 regularization to shrink some coefficients to zero, performed similarly to Ridge Regression, with an MSE of 2.2778e-05 and an MAE of 0.003013. Its R-squared value of -0.024351 suggests that it also struggled to explain the variance in the data. The feature selection ability of Lasso may have caused the model to omit important variables that were crucial for accurate unemployment forecasting. This limitation, combined with the model's linear approach, contributed to its inability to match the performance of more complex models like Random Forest.

### **3.2.5 Support Vector Machines (SVM)**

****

**Fig 5: Actual unemployment vs SVM**

The Support Vector Machine (SVM) model, while capable of handling non-linearities via kernel functions, did not perform as well as expected. It had an MSE of 2.213e-05, an MAE of 0.00296, and an R-squared value of 0.004812, suggesting that the model captured some variability in the data but was not very effective overall. SVM is generally powerful for classification and regression tasks in high-dimensional spaces, but its performance in this case indicates that further hyperparameter tuning and selection of an appropriate kernel function might have been necessary to fully leverage its potential for unemployment rate forecasting.

## **3.3 Implications of Findings**

The empirical findings imply that machine learning techniques, especially ensemble approaches like Random Forest, may provide considerable gains over standard econometric models in predicting unemployment rates. The Random Forest model displayed solid performance across all measures, making it a powerful tool for economic forecasting. However, the underperformance of Gradient Boosting underscores the significance of rigorous model tweaking and validation, particularly when working with complicated datasets.

The study's results underline the promise of machine learning approaches in economic forecasting, especially in detecting non-linear linkages that conventional models may overlook. These findings further underline the necessity for a balanced approach in model selection and tweaking, including both accuracy and computing economy.

## **3.4 Random Forest Feature Importance**

A graph of blue bars

Description automatically generated

**Fig 6: Feature importance, Gradient Boosting**

Random Forest Feature shows the feature priority scores as calculated by the Random Forest model. The x-axis indicates the economic indicators used as features, while the y-axis reflects their value in generating forecasts. A higher score implies a larger influence of the characteristic on the model's predictions.

As seen in Figure, the 'RPI' (Retail Price Index) stands out as the most crucial element, with a substantially higher relevance score compared to other indicators. This shows that inflation, as reflected by the RPI, plays a major role in projecting unemployment. The second most significant characteristic is 'W875RX1', an employment-related indicator, followed by 'Retail' (Retail Sales) and 'INDPRO' (Industrial Production). These findings are consistent with economic theory, which holds that consumer prices, job levels, and industrial activity are primary causes of unemployment.

The prevalence of 'RPI' in the feature relevance rankings highlights the strong link between rising inflation and unemployment. Policymakers might use this understanding to watch inflation carefully while adopting policies to decrease unemployment. Similarly, the relevance of 'W875RX1' and 'INDPRO' implies that maintaining strong employment and industrial development might be essential in keeping low unemployment rates.

### **3.4.2 Gradient Boosting Feature Importance**

A graph of a number of forest

Description automatically generated with medium confidence

**Fig 7: Feature importance, Random Forest**

Figure depicts the feature significance scores produced from the Gradient Boosting model. Similar to the Random Forest analysis, the x-axis identifies the economic indicators, while the y-axis shows their relative significance.

The Gradient Boosting model likewise indicates 'RPI' as the most important predictor, comparable with the results from the Random Forest model. However, the relevance rankings for the other aspects show considerable fluctuation. Notably, 'W875RX1' and 'RETAIL' remain among the top contributors, while 'INDPRO' still has substantial relevance. These changes show that although both models agree on the fundamental indications, they may assess the relevance of less dominating traits differently.

Although both models agree on the relevance of 'RPI' and 'W875RX1', the Gradient Boosting model gives somewhat different priority ratings to the remaining attributes. This might be owing to the varied methods in which the two models handle interactions between features and mistakes in predictions. The Gradient Boosting model's stronger sensitivity to indicators, such as 'CMRMTSPL' (Consumer Metrics), demonstrates its ability to capture short-term variations in consumer behaviour more effectively than the Random Forest model.

## **3.5 Implications for Economic Forecasting**

The feature relevance analysis gives vital insights into the economic factors that most substantially effect unemployment projections. The constant identification of 'RPI' as an important characteristic across models implies that inflation management is crucial for managing unemployment. Additionally, the relevance of 'W875RX1' (employment levels) and 'INDPRO' (industrial production) corresponds with classic economic theories that highlight the role of employment and industrial activity in determining labor markets.  
  
By knowing which qualities are most significant, policymakers may better focus actions to stabilize the economy and minimise unemployment. This investigation also demonstrates the potential for machine learning algorithms to identify complicated, non-linear correlations between economic variables and unemployment that conventional models could overlook.

# **4. Conclusion:**

This Research paper has explored the application of advanced machine learning techniques to enhance the accuracy of unemployment rate forecasting, particularly by utilizing the FRED-MD dataset, which includes a wide range of U.S. economic variables. Traditional econometric models such as ARIMA and VAR, while foundational, have limitations in capturing the intricate and non-linear dynamics of unemployment, particularly in periods of economic volatility. The research demonstrated that machine learning models, particularly ensemble methods like Random Forest, outperformed these conventional models by providing more robust, accurate forecasts. These methods excelled at capturing complex interactions between multiple economic factors, offering more precise predictions than standard approaches. Additionally, the application of Ridge and Lasso Regression showcased the effectiveness of regularization techniques in managing high-dimensional data and mitigating overfitting. Despite the promising results, the study also acknowledged the computational challenges and the need for meticulous hyperparameter tuning in machine learning models. This dissertation provides valuable insights for policymakers, suggesting that the adoption of machine learning methods can lead to more informed economic decisions and improved labor market policies. Moving forward, the integration of real-time data sources could further refine the accuracy and timeliness of unemployment forecasts, positioning machine learning as a key tool in the future of macroeconomic forecasting.

# **References**

1. McCracken, M. W., & Ng, S. (2015). FRED-MD: A monthly database for macroeconomic research. Federal Reserve Bank of St. Louis Working Paper, 2015-012B. [https://doi.org/10.20955/wp.2015.012]
2. Gogas, Periklis, Theophilos Papadimitriou, and Emmanouil Sofianos. "Forecasting unemployment in the euro area with machine learning." *Journal of Forecasting* 41.3 (2022): 551-566.
3. Dobre, Ion, and A. A. Alexandru. "Modelling unemployment rate using Box-Jenkins procedure." Journal of applied quantitative methods 3.2 (2008): 156-166.
4. Floros, Christos. "Forecasting the UK unemployment rate: model comparisons." International Journal of Applied Econometrics and Quantitative Studies 2.4 (2005): 57-72.
5. Jelena, Mladenovic, Ilic Ivana, and Kostic Zorana. "Modeling the unemployment rate at the EU level by using box-jenkins methodology." KnE Social Sciences (2017): 1-13
6. Kurita, Takamitsu. "A Forecasting Model for Japan's Unemployment Rate." Eurasian Journal of Business and Economics 3.5 (2010): 127-134.
7. Maehashi, Kohei, and Mototsugu Shintani. "Macroeconomic forecasting using factor models and machine learning: an application to Japan." Journal of the Japanese and International Economies 58 (2020): 101104.
8. Cicceri, Giovanni, Giuseppe Inserra, and Michele Limosani. "A machine learning approach to forecast economic recessions—an italian case study." Mathematics 8.2 (2020): 241.
9. Kuusisto, Niklas. Forecasting Methods Applied to Macroeconomic Variables. BS thesis. 2021.
10. Yoon, Jaehyun. "Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach." Computational Economics 57.1 (2021): 247-265.
11. Tattikota, Sri Rajitha, and Naveen Srinivasan. Integration of econometric models and machine learning-study on us inflation and unemployment. Madras School of Economics, 2021.
12. Gabrikova, B.; Svabova, L.; Kramarova, K. Machine Learning Ensemble Modelling for Predicting Unemployment Duration. *Appl. Sci.* **2023**, *13*, 10146. [<https://doi.org/10.3390/app131810146>]
13. Güler, Mehmet, et al. "Forecasting of the Unemployment Rate in Turkey: Comparison of the Machine Learning Models." *Sustainability* 16.15 (2024): 6509.
14. Celbiş, Mehmet Güney. "Unemployment in rural Europe: A machine learning perspective." Applied Spatial Analysis and Policy 16.3 (2023): 1071-1095.
15. Vinaya, Nareddy, et al. "Unemployment Rate Future Forecasting Using Supervised Machine Learning Models." RICE. 2023.
16. Yurtsever, Mustafa. "Unemployment rate forecasting: LSTM-GRU hybrid approach." *Journal for Labour Market Research* 57.1 (2023): 18.

# Appendix