National College of Ireland

A Comprehensive study of applying machine learning algorithms for time series data prediction to the Irish Labour Market Unemployment Rate

MSc Research Project
MSc in Data Analytics

Shree Hari Krishnamurthy
Student ID: x21165441

School of Computing
National College of Ireland

Supervisor: Dr. Anh Duong Trinh

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Shree Hari Krishnamurthy
Student ID:	x21165441
Programme:	MSc in Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Anh Duong Trinh
Submission Due Date:	14/08/2023
Project Title:	A Comprehensive study of applying machine learning algorithms for time series data prediction to the Irish Labour Market Unemployment Rate
Word Count:	13394
Page Count:	33

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	*Shree Hari Krishnamurthy*
Date:	17th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Comprehensive study of applying machine learning algorithms for time series data prediction to the Irish Labour Market Unemployment Rate

Shree Hari Krishnamurthy
x21165441

Abstract

The utilisation of time series data provides significant insights into periodic patterns and the trends, making its study crucial across various fields, such as economic forecasting. This study seeks to examine prediction of the unemployment rate in the Ireland by utilising time series data. The study utilises a dual strategy, incorporating traditional statistical models as well as contemporary machine learning approaches in order to obtain precise predictions. Traditional statistical models like as ARIMA and SARIMA are commonly employed to capture the natural variations in time present in data, including seasonal oscillations. Simultaneously, advanced machine learning models such as Random Forest, Ridge Regression, K-Neighbors Regressor (KNN), and XGBoost are utilised to investigate their capabilities in improving the accuracy of predictions. The paper thoroughly assesses various models by considering their performance measures such as the RMSE, R-squared, and MAPE. It conducts an comparison analysis to determine effectiveness of each technique. Among the models studied, Ridge Regression outperformed others, demonstrating promising capabilities in conjunction with the time series data. This extensive research makes a valuable contribution to the existing body of knowledge by exploring the utilisation of statistical and machine learning models to enhance the precision and reliability of unemployment rate predictions.

Keywords:Time series, Ireland, Unemployment, ARIMA, SARIMA, XGBoost, KNN, Ridge, Random Forest

1 Introduction

Governments as well as companies can effectively plan for the future and allocate funds when a rise in unemployment is predicted. A strong economy with plenty of work possibilities is indicated by a low unemployment rate, as was the case in May when it reached a record low of 3.8% in Ireland. This stimulates investment growth, consumer consumption, and economic growth. However, when the Covid-19 epidemic was at its worst in March 2021, with a rate of 7.7% [1]. The impact of unemployment on the levels of psychological discomfort experienced by young adults is of utmost importance. In contrast to the adult population, social status and money do not serve as contributing variables. The social status of one's origin does really exert an influential impact Hannan et al. (1997). In

[1] `https://www.bbc.com/news/articles/c2v8670ejzpo`

the broader context, anticipating the likelihood of unemployment holds significant importance in influencing socioeconomic circumstances and directing endeavours towards a state of enhanced stability as well as wealth. Time series data is widely employed by organisations as a valuable tool for gaining insights into the underlying factors that drive trends or recurring patterns over a specific time period. [2] Through the utilisation of data visualisations, professionals in the business field are capable of observing recurring patterns that manifest during particular seasons and conducting deeper investigations into the fundamental factors contributing to these patterns. The introduction of modern analytics platforms has greatly broadened the range of visualisations available, going beyond the conventional use of line graphs.

The implementation of machine learning algorithms to the economic forecasting has been gained much interest in recent years as an effective method for improving the accuracy of forecasts and generating valuable insights. Among the several economic metrics, the unemployment rate is the important consideration that determines the country's overall economic health. Ireland is the country that is known for its dynamic economy and the labour market of the Ireland is one where the rate of unemployment is especially significant. Ireland, as the member of European Union and the active participant in global economy, faces the issue he efficiently regulated its labour market in order to maintain the economic growth. Predictions of the Irish Labour Markets Unemployment Rate that are timely and exact have the ability to help the policymakers to implement effective policies and alleviate the negative consequences of economic downturns. The past data and statistical models are frequently used in traditional ways of predicting unemployment rates. The development of machine learning, however, and its capacity to handle complicated time series data has led academics to resort to sophisticated algorithms to improve forecast accuracy.

1.1 Motivation

The forecasting of the time series data creates a significant obstacle, mainly due to the dynamic and unpredictable nature of economic trends and conditions. The economic environment experiences periodic swings that are influenced by a range of factors, including advances in technology, world events, modifications to policies, and unforeseen crises. The rapid and occasionally unprecedented fluctuations give rise to a significant degree of uncertainty and difficulty when trying to predict forthcoming values of economic and financial time series data. In spite of the inherent risk of economic markets, an additional interesting challenge in time series prediction pertains to the presence of incomplete information. Economic data is frequently obtained from a variety of sources, and instances of gaps or missing data points may arise as a result of reporting delays, challenges in data collection, or other factors. The absence of these values may introduce biases into the time series, thereby posing difficulties in attaining precise forecasts.

The objective of this research is to provide the thorough examination of the applying the various machine learning algorithms in the analysis of time series data, specifically in the prediction of the Irish Labour Market Unemployment Rate. Through the utilisation of historical data and the application of most recent machine learning algorithms, our

[2] `https://www.tableau.com/learn/articles/time-series-analysis`

objective is to investigate the capabilities of these machine learning algorithms in predicting the unemployment patterns with improved precision and detail.

1.2 Research Question

The aim of this study is to determine" How well Contemporary Machine-Learning models would improve the prediction of unemployment in Ireland when compared to traditional methods for time series data?"

- How do Modern machine-learning models perform for time series data set.
- In the context of forecasting unemployment in Ireland, how accurate are Modern machine learning models compared to Statistical models?

1.3 Report Structure

The present paper is structured into six distinct sections:

- **Section 2:** Presents a thorough examination of the existing literature of unemployment.
- **Section 3:** Outlines the methodology employed for data mining.
- **Section 4:** Discusses the implementation of various machine-learning models and statistical methods.
- **Section 5:** Presents the evaluation and comparison results of the statistical and Machine learning models using MAPE and R squared value metrics.
- **Section 6:** Gives concluding remarks and offers suggestions for future research endeavours.

2 Related Work

2.1 Traditional Approaches on Predicting Unemployment Rate

This study investigates the potential advantages of machine learning algorithms for forecasting the level of unemployment in the South Africa. In their study, Mulaudzi and Ajoodha (2020) employ a multivariate approach to examine the predictive accuracy of various machine learning models in forecasting instances of workforce reductions. The authors analyse the implications of their findings for policymakers and other stakeholders, as they have discovered the utility of machine learning models in predicting the rate of joblessness. In comparison to other traditional methods of statistical analysis, a perceptron system with three hidden layer demonstrated superior efficiency in predicting jobless rates. The model effectively captures the non-linear characteristics of the unemployment rate in South Africa. According to the findings of Mulaudzi and Ajoodha (2020), algorithms utilising machine learning techniques offer distinct advantages in accurately forecasting fluctuations in employment rates in Africa, surpassing the predictive capabilities of conventional statistical models such as ARIMA and SARIMA. The algorithms possess capability to the perform such tasks over the extended period of the time. Hence, it can be argued that algorithms based on machine learning are better suited for modelling and forecasting the unemployment rate in South Africa. The research results indicate that the multi-layer perceptron (MLP) with three hidden layers outperforms conventional statistical approaches in forecasting unemployment rates in South Africa. The advantage of the MLP neural network lies in its capacity to effectively capture the inherent non-linear structure of the unemployment rate in South Africa. This characteristic enables the MLP to handle intricate and ever-changing patterns in the data with greater proficiency. Moreover, ridge regression model exhibits a notable R-squared value, suggesting its proficiency in accurately capturing the robust associations between the unemployment rate in South Africa and its lagged values. The outcome underscores the significance of incorporating lagged data to enhance the precision of unemployment rate forecasts.

The study conducted by Ramli et al. (2018) aimed to forecast the unemployment rate using Holt's exponential smoothing and ARIMA models. Based on the findings derived from the regression analysis, it can be concluded that population growth and inflation exhibit statistically significant associations with unemployment. The ARIMA model (2,1,2) has been demonstrated as the most efficacious approach for forecasting the unemployment rate in Malaysia. The results of the study indicated a substantial rise in the unemployment rate within Malaysia. Based on certain projections, it is anticipated that Malaysia's unemployment rate will experience a modest increase during the period spanning from 2017 to 2026. However, Dumičić et al. (2015) selected several forecasting techniques that are appropriate for predicting time series data containing a trend component. One of the methods employed is double exponential smoothing, commonly known as Holt's method, which accounts for both pattern and seasonality. The present study has produced results that suggest the Holt-Winters' additive method to be the most efficacious model for forecasting the unemployment rate in Greece. Based on the MAPE analysis, it was concluded that the double exponential smoothing model is the most appropriate choice for Spain. Furthermore, the investigation revealed that the Holt-Winters' multiplicative model emerged as the most effective predictive model for estimating unemployment rates

in Croatia and Italy. On the other hand, in the case of Portugal, the double exponentially smoothed model was determined to be the most suitable model for forecasting the unemployment rate. Additionally, there is also a another research paper by Nkwatoh (2012), that aims to explore various univariate time series models that are employed for the purpose of predicting the rate of unemployment in Nigeria. The approaches under consideration in this study encompass trend regression analysis and ARIMA, GARCH, and the mixed ARIMA/GARCH algorithms. The comparative evaluation of these forecasting techniques is conducted using three essential criteria like RMSE, Mean Absolute Error, and also Mean Absolute percentage Error values. The findings of the research indicate that all the models under consideration possess the capacity to be utilised for forecasting future joblessness rates, dependent upon the importance of their variables and the level of accuracy in their fit. Nevertheless, the model selection parameters emphasise the advantage of the ARIMA/ARCH model in comparison to the pattern regression and ARIMA models in relation to the precision of forecasting. The research study proposes that the utilisation of an ARIMA [1,1,2] / ARCH[1] model could be a viable approach for accurately modelling and forecasting joblessness in Nigeria. It is important to highlight that the present findings are in opposition to the outcomes of a prior investigation carried out by Etuk et al. in the year 2012. The earlier study proposed the utilisation of an ARIMA [1,2,1] model to predict the rate of unemployment in Nigeria, employing monthly data spanning from 1999 to 2008. The differences in outcomes may be ascribed to variations in the time periods and the number of samples of the data, as both of these can have a substantial influence on the forecasting results.

In this research Didiharyono and Syukri (2020), authors predicted the open unemployment rates in South Sulawesi by utilising the ARIMA model. In order to accomplish this objective, the study employs an applied research methodology, utilising numerical secondary data sourced from the Central Statistics Agency of South Sulawesi. The study utilises a range of research methodologies to achieve its objective. The primary objective of this study is to examine model identification, which is a fundamental and essential stage in ARIMA modelling. The method of identifying includes the examination and analysis of the autocorrelation function and PACF of the temporal data series. Such evaluations aid in determining the optimal arrangement of autoregressive and moving average terms, along with the necessary order of differencing [I] to achieve stationary behavior of in the data. Additionally, the researchers conduct data differencing, a technique used to convert non-stationary information into a stationary format by calculating the discrepancies between successive findings. The presence of stationarity is a crucial requirement for the ARIMA model to accurately capture the inherent trends and patterns present in the time series data. Subsequently, the process of parameter estimation is conducted, whereby the researchers make estimations of the coefficients of the ARIMA model through the utilisation of estimation techniques such as the highest probability or least squares. The researchers employed the ARIMA (1,2,1) model, which was identified as the most suitable model through rigorous analysis and evaluation conducted in the study. The selection of this model is justified by its accuracy, as evidenced by a low Mean Square score of 2.0474.

In order to conduct time series evaluation and prediction, this study Kyung and Nam (2019) utilises Holt's linearization method to effectively model the dynamic trends of two distinct datasets over a period of time. Particularly, the datasets under consideration are

the national jobless rate and the rate for increases in the federal debt. The precision of the Holt's the linearization algorithm varies across different datasets and when evaluating different return periods. When considering the overall unemployment rate, the Holt's model exhibits more preciseness compared to the assessment of the conceptual return period, although with relatively less precision. On the other hand, the Holt's linearization model exhibits less precision in that fits the data when applied to the national debt growth rate. However, it offers a more precise assessment of the conceptual return period of time. In general, the Holt's linearization approach demonstrates satisfactory performance as a forecasting tool with regard to the National Debt Growth Rate spanning from the year 1967 to year 2018. on the other hand the study Urrutia et al. (2017)is to utilise a time series model in order to provide the projection of the unemployment rate in Philippines. The model used to estimate and predict the unemployment rate in the Philippines is denoted as SARIMA [6, 1, 5] [0, 1, 1] 4. The projected values fall within range of six to eight percent and exhibit a close correlation of 72 percent with the actual values. The unemployment rate is influenced by two significant elements, namely the labour force rate and population. Moreover, it has been demonstrated that Population, GDP, and GNI exhibit Granger-causality with the dependent variable. Several variables have the potential to influence the fluctuations in the unemployment rate. An rise in such factors has the potential to either boost or decrease the unemployment rate.

2.2 Machine Learning Approaches for Predicting Unemployment Rate

The study Hatipoğlu et al. (2021) aims to estimate unemployment rates in the Turkey by employing the k-nearest neighbour regression (kNNR) algorithm. The rationale behind choosing this algorithm is rooted in its capacity to generate reliable forecasts by discerning resemblances among data points, taking into account the features that are likely to impact unemployment rates. The objective of this study is to improve the accuracy of unemployment forecasts by utilising machine learning techniques through the development of a new dataset that includes crucial factors influencing unemployment. In order to evaluate the efficacy of the kNNR algorithm, two additional machine learning algorithms, namely the ridge regression method and linear regression, are employed for comparative analysis. Ridge regression is selected due to its ability to address multicollinearity and high-dimensional datasets, whereas linear regression is employed as a benchmark for evaluating the efficacy of the kNNR algorithm.The findings of this study indicate that the kNNR method exhibits superior performance compared to the methods of ridge regression and the linear regression. Specifically, the kNNR method achieves a coefficient of determination R-squared value of 0.7498. The robust performance of the kNNR algorithm highlights its potential in accurately estimating unemployment rates in the country of Turkey. This has significant implications for both economists and policymakers as it provides valuable insights for the development of targeted actions and plans aimed at addressing the issue of unemployment.

In this study Sen et al. (2022) aims to identify the Indian states that are facing significant challenges in terms of job prospects. The application of Supervised Machine Learning techniques has been utilised to identify states characterised by the lowest levels of employment. The data visualisation provides a more comprehensive representation of the longitudinal patterns in the unemployment rate. The author applied Various popular

machine learning algorithms, such as Logistic Regression, Support Vector Machine and the K-nearest neighbours machine learning Algorithm, and Decision Tree, have been utilised. The researchers reached the conclusion that the Decision Tree algorithm exhibited outstanding performance compared to all other algorithms, demonstrating a high level of accuracy. In addition, the study Pavlyshenko (2016) examines different methodologies, such as linear models, the ARIMA algorithm, and the xgboost machine learning algorithm. The present study showcases the outcomes of different combinations of models, illustrating the capacity of these ensembles to attain improved accuracy in forecasting when compared to individual algorithms. The author introduces the copula approach as an important method for effectively addressing concerns related to risk assessment. The application of the copula methodology enables the depiction of probabilistic correlations between target variables and extreme factors. This methodology is especially beneficial in situations where the dependent variable exhibits a probability density function that deviates from the Gaussian distribution and is characterised by heavy tails. The capacity to effectively capture complex interdependencies greatly enhances the precision of risk assessment in real-world scenarios. The final outcome indicates that XGBoost exhibited better results in terms of achieving a lower Root Mean Square Error (RMSE) when compared to ARIMA.

The implementation of a case study framework for forecasting the Open Unemployment Rate in Indonesia, employing the Deep Learning methodology for prediction. In this study Rahmat et al. (2022), the prediction process consists of two fundamental stages the first is training stage and the testing stage. Deep Learning, a prominent machine learning technique, is employed in the training stage using the Keras framework. This process involves a series of sequential steps. The initial step involves the preparation of the dataset, which is subsequently followed by the definition of the Keras Model and its compilation to configure the optimisation settings. The Keras Model is subsequently modified and optimised according to the dataset. A comprehensive assessment of the Keras Model is undertaken to evaluate how it performs, and predictions are generated utilising the model that has been trained. The training procedure consists of multiple iterations, each comprising 500 epochs. This method of iteration that leads to a significant reduction in loss, as evaluated by the metric Mean Squared Error. During the training phase, the model demonstrate the impressive rate of an around the 97.93%, which serves as a strong indication of its ability to accurately capture the underlying patterns within the dataset. During the subsequent phase of testing, the model exhibits a commendable success rate of 92.10%, thereby reaffirming the model's durability and reliability in accurately predicting outcomes. On the other hand, the study Gogas et al. (2022) provided a directional forecast for the unemployment rate in the euro-area. To the best of current understanding, there are no existing studies that provide a comprehensive projection of the unemployment rate for the whole euro-area. The dataset comprises the unemployment rate and 36 explanatory factors, which have been selected based on theoretical considerations and relevant to the research. These variable cover the time period from 1998 to 2019, with data collected on an monthly basis. The aforementioned variables are utilised as the inputs for three distinct machine learning approaches, namely decision trees, random forests (RF), SVM. Additionally, an elastic-net logistic regression (logit) model, derived from the field of econometrics, is employed. The findings indicate that random forest model, which was determined to be the most effective, surpasses the other models by achieving an accuracy of 88.5% for train forecasting and 85.4% for test

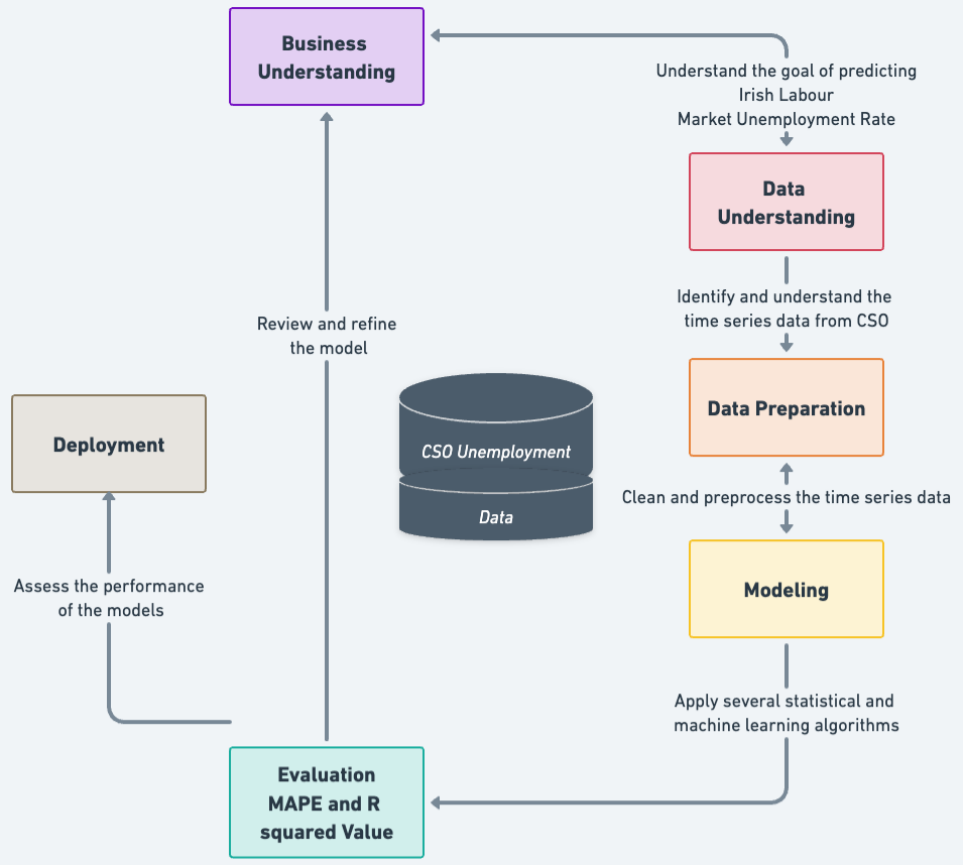

Figure 1: Adopted Methodology: CRISP-DM

forecasting.

3 Methodology

The primary objective of this comprehensive study is to examine the application of various machine learning algorithms for time series data in order to forecast the unemployment rate in the Irish Labour Market. The study aims comparative study of predicting unemployment rates in Ireland by utilising advanced machine learning techniques and Traditional Statistical models. The present study employs the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a widely recognised and extensively utilised framework, to provide guidance throughout the entirety of the research process. The CRISP-DM methodology offers a systematic framework for addressing data-driven challenges, encompassing various stages such as data comprehension and preprocessing, model development, assessment, and implementation. The Schröer et al. (2021) emphasises that the systematic literature review of recent studies utilising the CRISP-DM framework provides a solid basis for comprehending the research focus, identifying best practises, and exploring innovative methodologies.

This paper provides an overview of the six stages involved in a study, as illustrated in Figure 1, including their interconnections.

Table 1: Overview of Irish Unemployment Data

Column	Unique Values	Missing Values	Data Type
Statistic Label	1	0	object
Month	305	0	object
Age Group	1	0	object
Sex	1	0	object
UNIT	1	0	object
VALUE	281	0	float64

3.1 Business Understanding

The main aim of this comprehensive study is to conduct an analysis of time series data pertaining to the Irish Labour Market, with the goal of accurately predicting the unemployment rate. This research purpose is to utilise a diverse range of machine learning algorithms, encompassing statistical models and modern machine learning approaches with the purpose of offering significant insights to policymakers, businesses, and individuals. This will empower them to make more informed decisions and formulate efficient strategies considering the dynamic nature of the labour market in terms of unemployment. Comprehending and predicting the unemployment rate holds most importance in facilitating well-informed decision-making pertaining to economic policies, employment patterns, and workforce strategizing within the context of Ireland.

3.2 Data Understanding

The current study is based on data from Ireland's Central Statistics Office (CSO) regarding the labour market unemployment [3]. The dataset consists of six columns and around 5000 rows. This variable's statistical label designates the category of statistical information or indicator relating to Ireland's employment market. In this dataset, the unemployment data is from 1998 January to 2023 May, which is indicated by the month attribute. The dataset's inclusion of a time dimension makes it easier to analyze seasonal and cyclical changes in the unemployment rate. To facilitate analysis, the age group characteristic is used to categorise people into different age groups. The typical age ranges represented in the CSO unemployment dataset span from 15 to 74 years.The biological and social identities of people in the employment market are represented by this feature. Both genders are represented in the dataset that has been downloaded from CSO for all of the values.The attribute with the label "UNIT" is used to specify the unit of measurement used to report the data. The statistical indicator's precise numerical measurement of unemployment rate in Ireland, relating to a particular month, age bracket, and sex, is represented by the "value" variable.

3.3 Data Preparation

Ensuring the data is in the proper format and ready for analysis requires careful data preparation. Inconsistencies in the dataset, such as, missing values, or NA (Not Avail-

[3] https://www.cso.ie/en/statistics/labourmarket/monthlyunemployment/

able) values, must be absent in order to conduct valid analyses. Data integrity is crucial when working with time series data since any inconsistencies or absences might provide partial results and incorrect forecasts. Additionally, for time series models to be successful, stationarity is a key presumption. The term "stationarity" refers to the idea that the mean, autocorrelation, and variance of a time series are constant across the course of the data. In a stationary time series, it has been believed that the patterns and correlations seen in the past would continue to exist. It is necessary to transform the data into a stationary condition in order to make it appropriate for time series modelling.
For the study on the rise of unemployment in Ireland, the dataset was first put into a data processing system. Python with the pandas library was used for this. At first glance, there were several columns, including "Statistic Label", "Month", "Age Group", "Sex", "UNIT", and "VALUE". It was noticed that columns like "Statistic Label," "Age Group," "Sex," and "UNIT" all had the same value, so they might be able to be left out to make the information simpler. A careful check showed that there were no missing numbers, proving that the dataset was correct. Formatting Month column which was known to be an object data type, into a date-time style was an important step. This made sure that the data was in the right order and that it could be resampled, which is very important for time series analysis. Visualising the data set helped find clear trends, regular patterns, and possible outliers. The Augmented Dickey-Fuller test was used to make sure that the sample was stationary. When data isn't static, changes like differencing and log transformation are used. After these steps, the information was clean and stable enough to be used for modelling and more research.

3.3.1 Exploratory Data Analysis

The process of Exploratory Data Analysis (EDA) holds significant importance as it serves the fundamental step in understanding the fundamental characteristics of the dataset.The line plot depicted in Figure 2, illustrates a pattern of variability in the unemployment rate, characterised by discernible high points and low points. Certain periodic patterns indicate potential seasonality. There is an notable decline in the unemployment rate during the early 2000s, which may potentially indicate the period of economic expansion. Periodic increases in unemployment can be observed, which are likely to be associated with financial crises or the recessions. The latest data suggests the possibility of obvious trends that requires additional research.

The decomposition graph in Figure 3 facilitates knowledge of the basic components of a time series, including the trend, seasonal patterns, and residuals.The seasonal decomposition technique will be applied to an original unemployment data, using the additive model. Subsequently, the resulting component will be graphically represented. The additive model is employed in cases where the seasonal variations exhibit a somewhat consistent pattern across the course of time. The trend component characterised by additive effects exhibits the distinct and obvious increasing trajectory, particularly becoming greater following the midpoint of the series. This observation suggests a gradual rise in the overall unemployment rate as time progresses. The additive decomposition method uncovers regular the seasonal patterns, which may indicate the presence of monthly or yearly cycles. The constancy of the amplitude of the seasonal swings throughout time

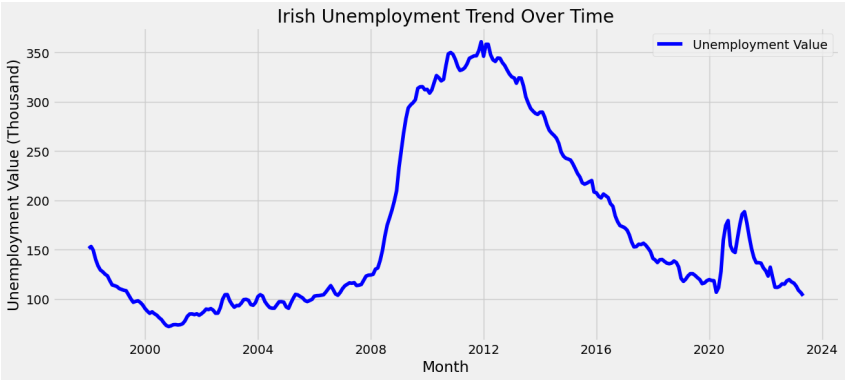

Figure 2: Irish Unemployment Trend over Time

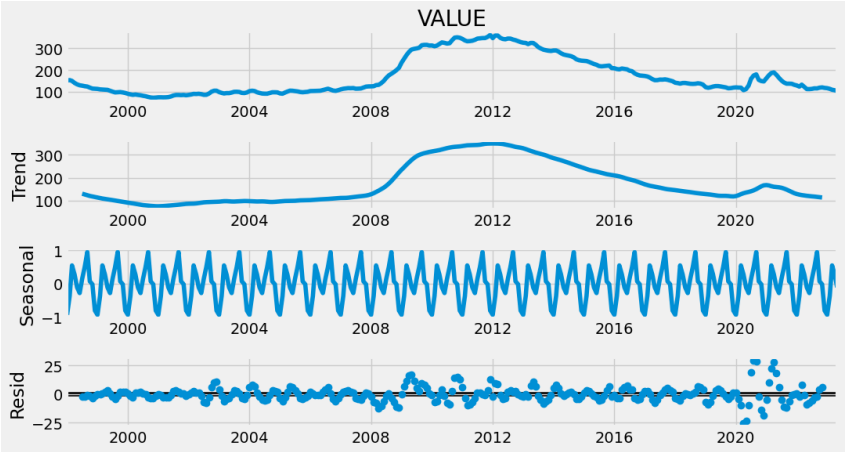

Figure 3: Seasonal Decomposition Using Additive Model

supports the appropriateness of using an additive model. The additive residuals exhibit stochastic fluctuations, although they tend to cluster around the mean of zero. It is possible that certain spikes may be observed, which could potentially be attributed to the outliers.

3.3.2 Data Pre-processing and Transformation

The assessment of the stationary holds significant importance in the study of the time series data, as numerous statistical techniques require the time series to exhibit stationarity. The Augmented Dickey-Fuller (ADF) test is a frequently employed technique for assessing the stationarity.

- Null Hypothesis: The data is non-stationary.
- Alternative Hypothesis : The data is stationary.

The ADF test is conducted using the adfuller function from the **statsmodels.tsa.stattools** package. Furthermore, the pandas library was employed for the purpose of manipulating the data, while the matplotlib library was utilised for purpose of visualising of the data. The outcome of the Augmented Dickey-Fuller (ADF) test for the Original data is depicted in Figure 4,The test statistic of -1.8041 and obtained p-value of 0.3784 exceeds commonly used significance values such as 0.05, suggests that there is insufficient evidence

Metric	Original Data	Transformed Data
Test Statistic	-1.8041	-3.4813
p-value	0.3784	0.0085
Critical Value (1%)	-3.4526	-3.4526
Critical Value (5%)	-2.8714	-2.8714
Critical Value (10%)	-2.5720	-2.5720

Figure 4: ADF Test: Original vs. Transformed Data

to reject the null hypothesis. The test statistic exceeds the critical values for all levels of significance, providing more evidence in favour of not rejecting the null hypothesis.The test incorporated a total of eight lags, so addressing the issue of autocorrelation within the series.

The technique of transforming a non-stationary time series into a stationary one is a commonly used preliminary procedure in the field of time series analysis. A frequently employed approach to attain stationarity involves computing the initial difference of the series. This process involves the subtraction of the preceding observation from each subsequent observation.The rationale of the practise of the differencing is to eliminate pattern of trends and seasonality, resulting in a stationary time series. Based on the results of the Augmented Dickey-Fuller (ADF) test Figure 4, it can be observed that transformed series, which includes logarithmic, square root, and first-order differencing operations, exhibits characteristics of stationarity. Stationarity is the crucial need for the application of time series modelling.

In order to achieve the stationarity, log transformation was employed on the 'VALUE' column with the purpose of stabilising its variance.The data was subjected to a square root transformation after being log-transformed in order to enhance the stability of the series.First-order differencing was performed by calculating the difference between the consecutive data that had undergone square root log transformation, with the aim of eliminating any underlying behavioural patterns. Based on the results of the Augmented Dickey-Fuller (ADF) test Figure 4, it can be observed that transformed series, which includes logarithmic, square root, and first-order differencing operations, exhibits characteristics of stationarity. Stationarity is the crucial need for the application of time series modelling.

The plot 5 illustrates the stationary properties of the modified series, by displaying no obvious pattern or seasonality. The present stationary series has characteristics that make it appropriate for subsequent analysis and modelling.

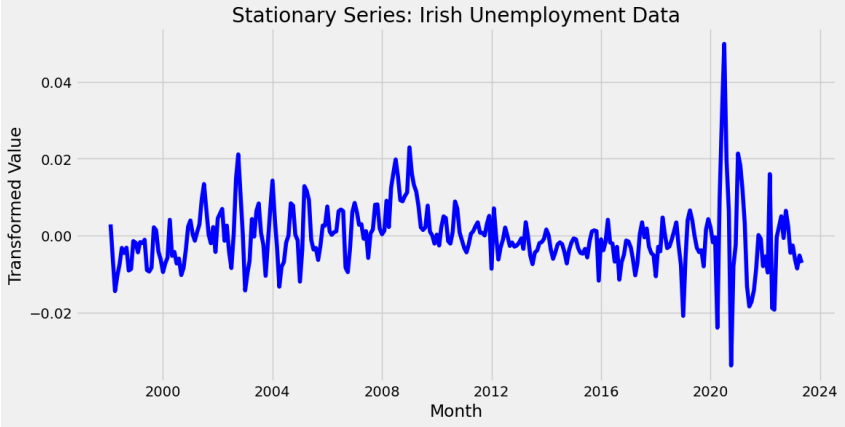

Figure 5: Stationary Transformation: Irish Unemployment Data

3.4 Modeling

This section presents an overview of the application of several machine learning algorithms in order to predict the Unemployment Rate in the Irish Labour Market. The models chosen for this research cover ARIMA and SARIMA for statistical modelling, as well as Random Forest, XGBoost, Ridge Regression, and K-Nearest Neighbours (KNN), which are modern machine learning models utilised for time series analysis. The implementation and evaluation of each model are carried out on a time series dataset, and their performances are thereafter compared using suitable evaluation measures.

3.4.1 Statistical Models

1. **ARIMA:** The ARIMA methodology, first introduced by Box and Jenkins Box et al. (2015), has become widely recognised and utilised as a key technique for forecasting time series data. It combines autoregressive, differencing, and moving average components to accurately represent the intrinsic patterns and connections within a particular time series. Stationary time series data, which exhibits statistical properties that remain constant throughout time, such as a consistent mean and variance, exhibits a high degree of efficacy in this particular scenario.

    - The autoregressive component, represented by the parameter "p," signifies the number of previous data points taken into account for predicting the present value.
    - The parameter "d" in time series analysis indicates the number of differencing operations performed in order to attain stationarity.
    - The parameter "q" in the moving average (MA) component signifies the number of previous error terms that are taken into account for forecasting the present value.

    The ARIMA model is characterized by 3 main parameters, that are p, d, and q. The proposed model integrates autoregressive which is p, differencing which is d, and moving average which is q elements in order to generate forecasts for forthcoming data points. The model known as ARIMA is commonly represented in a broader form as ARIMA(p, d, q).

2. **SARIMA:** The SARIMA model, which stands for Seasonal AutoRegressive Integrated Moving Average, is a variant of the ARIMA model that incorporates supplementary features to accommodate the presence of seasonal variations within time series data [4]. The inclusion of a seasonal element in the SARIMA model serves to capture the recurring patterns that manifest at consistent intervals within the time series. The seasonal aspect is characterised by 3 supplementary features, namely P, D, and Q. The parameter P, which stands for the seasonal autoregressive order, signifies the quantity of seasonal lags that are taken into account while forecasting the present value using prior information from the corresponding season. The seasonal differential order, represented as "D," bears resemblance to the non-seasonal differencing d in ARIMA. The seasonal MA order, represented as "Q," is comparable to the non-seasonal moving average order (q) in ARIMA. This parameter denotes the quantity of seasonal delays of the term errors that will be utilised for the purpose of forecasting.

3.4.2 Machine Learning Models

1. **Random Forest:** The Random Forest algorithm is a robust ensemble learning technique that is commonly employed for the purposes of both classification and regression. Random Forest is a machine learning approach that enhances forecast accuracy and mitigates overfitting by combining numerous decision trees. The Random Forest algorithm is extensively employed across several sectors owing to its durability, adaptability, and capacity to effectively handle intricate datasetsBreiman (2001). In addition to the process of randomly selecting data samples, the Random Forest algorithm incorporates an additional layer of unpredictability by exclusively examining a random subset of features at every decision tree node. In the context of tree-based models, the evaluation of splitting is performed on a limited collection of features, rather than encompassing all available data. This phenomenon contributes to the increased variability within the population of trees, so reducing the presence of correlation and resulting in predictions that are more precise and less influenced by bias.

2. **XGBoost:** XGBoost, also known as Extreme Gradient Boosting, is a cutting-edge gradient boosting method that has significantly transformed the domain of machine learning since its inception. XGBoost, which was created by Tianqi Chen and made available in 2016, has gained significant prominence and efficacy as a machine learning library. This may be attributed to its remarkable capabilities in terms of performance, scalability, and adaptability Chen and Guestrin (2016). The utilisation of this technology is prevalent throughout several fields, encompassing data science contests, industrial applications, and research endeavours. XGBoost has superior performance on classification and regression predictive modelling tasks involving organised or columnar datasets.

3. **Ridge Regression:** Ridge Regression, alternatively referred to as L2 regularisation, is a linear regression methodology that effectively tackles the problems of multicollinearity and the overfitting encountered in conventional linear regression. First proposed by Hoerl and Kennard (1970), Ridge Regression is a technique that incorporates a regularisation term into the conventional linear regression cost function.

[4] `https://rb.gy/otez8`

This additional term aims to promote the maintenance of modest and consistent model coefficients, often known as weights. Ridge Regression is particularly advantageous in scenarios involving datasets with a high number of dimensions or when there is a significant degree of correlation across variables.Ridge Regression is a linear regression model that has the capability of dealing with both numerical and categorical variables, provided that appropriate encoding techniques are applied. The algorithm demonstrates computing efficiency and may be effectively executed through the utilisation of matrix operations. Ridge Regression does not engage in the process of variable selection, as it retains all characteristics within the model while simultaneously reducing the magnitude of their correlations towards zero.

4. **KNN Regression:** The k-Nearest Neighbour algorithm (kNN) is a simple technique that is commonly used for classification or regression tasks Song et al. (2017). The K-Nearest Neighbour algorithm is a straightforward machine learning method that relies on the principles of supervised learning. In the training phase, the K-nearest neighbours (KNN) method retains the complete training dataset as a point of reference. When formulating forecasts, the algorithm computes the dissimilarity among the data that comes in point and each of the training instances, employing a designated distance measure such as Euclidean distance [5].In regression problems, the K-nearest neighbours (KNN) algorithm is utilised to estimate the target value of a new data point by taking the average of the target values of its K closest neighbouring data points. KNN is well-suited for problem domains characterised by continuous relationships in the data. The selection of a suitable value for K is imperative in the K-nearest neighbours algorithm.

   Hatipoğlu et al. (2021) employed a range of machine learning algorithms in their study to forecast the rate of joblessness in Turkey. A further dataset is constructed to incorporate the factors that are anticipated to influence the unemployment rate. Based on the findings of the study, it can be concluded that the kNNR technique exhibited superior performance compared to the alternative methods, as evidenced by an R2 value of 0.7498. The results indicate that the kNNR approach can be effectively utilised in this circumstance.

3.5 Evaluation

In the research project various machine learning algorithms (ARIMA, SARIMA, Random Forest, XGBoost, Ridge Regression, and KNN) in forecasting the Unemployment Rate of the Irish Labour Market, a comparison will be made based on the metrics of MAPE, R-Squared, and RMSE acquired for each model. Less Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) values, along with increased R-Squared values which is near to one, are indicative of enhanced prediction accuracy and a higher level of model reliability. By taking into account these three indicators collectively, one may make a well-informed judgement regarding the algorithm that exhibits superior performance for a certain forecasting assignment.

[5] `https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/`

- **R-squared/Adjusted R-squared**: The R-squared is statistic that quantifies the extent to which the variables that are independent in a model based on regression explain the variability in the dependent factor. On the other hand, the adjusted R-squared takes into consideration the proportion of unique variables included in the equation used for regression, aiming to offer an improved indication of the goodness of fit [6]. The highest R-squared number signifies that a notable percentage of variability in the dependent variable, implying that the good level of fit of the model. On the other hand, the lower R-squared value suggests that a model inadequately accounts for the variability observed in an target variable.

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

Where:
y_i is the observed value for the i^{th} observation.
$\hat{y}_i$ is the predicted value for the i^{th} observation.
$\bar{y}$ is the mean of the observed data.

- **Mean Absolute Percentage Error (MAPE):**

  The mean absolute percentage error (MAPE) is a commonly used measure for assessing the accuracy of forecasts. This measure is widely preferred because it employs percentage units, which enhances understanding of the scale of the variable [7]. The highest possible accuracy is attained when the dataset is devoid of outliers or zero values. The model with the lowest Mean Absolute Percentage Error (MAPE) is regarded as the most effective model, as it signifies that the model's forecasts are, on average, the most proximate to the actual values. The application of this function is widespread in regression analysis and model evaluation.In below formula, the variables A, F, and N represent the Actual, Forecast, and Number of observations, respectively.

$$\text{MAPE (Mean Absolute Percentage Error)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{2}$$

  Where:
  y_i is the actual value.
  $\hat{y}_i$ is the predicted value.
  n is the total number of observations.

- **Root Mean Square Error (RMSE):** The root mean squared error/RMSE is taking a square root of the average of the squared values of all the error. The utilisation of Root Mean Square Error (RMSE) is widely used and is regarded as a highly effective error metric for numerical predictions across various domains [8]. The model exhibiting the lowest root mean square error (RMSE) value is regarded as the most proficient model, as it signifies that the model's predictions exhibit the most proximity to the actual values on average.

[6] http://surl.li/jzvpa
[7] https://www.statisticshowto.com/mean-absolute-percentage-error-mape//
[8] https://www.sciencedirect.com/topics/engineering/root-mean-squared-error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

Where:
y_i is the actual value.
$\hat{y}_i$ is the predicted value.
n is the total number of observations.

4 Design Specification

The research project utilises a methodological strategy that includes the data acquisition, data preprocessing, exploratory analysis, modelling, and forecasting in order to examine and forecast patterns in Irish unemployment rate and Analysis of Time series data. The proposed framework incorporates many components such as time series decomposition, stationarity testing, data transformation, hyperparameter tuning, and visualisations. These components are implemented inside a Python programming environment, utilising popular libraries such Pandas, NumPy, Matplotlib, Seaborn, Statsmodels, and Development. The environment utilised for coding in Python Version 3.6 is often Jupyter Notebook or any other Integrated Development Environment (IDE) for Python.

5 Implementation

The primary aim of this implementation is to predict the unemployment rate of Ireland by employing a wide range of models. Additionally, it seeks to carry out a comparison analysis between statistical and machine learning methodologies for handling time series data. The models outlined in the methodology section are implemented and assessed in order to obtain a deeper understanding of their prediction abilities and effectiveness when applied to the unemployment rate dataset under consideration.

5.1 Statistical Models

5.1.1 ARIMA

After completing the preprocessing step, next step is to apply desired method or an technique. In order to achieve best model fit, it is an necessary to select optimal values for the p, d, and q parameters in the ARIMA model. The process is initiated by employing the 'auto arima' function from the 'pmdarima' package. The ARIMA model selection process is performed automatically on the converted series 'log sqrt df['shift log sqrt']', which has undergone logarithmic, square root, and differencing transformations. The code facilitates comprehensive progress tracking through the activation of the 'trace=True' setting, while also mitigating the display of warning messages by setting 'suppress warnings=True'. The significance of this stage lies in its ability to utilise computer resources to impartially and efficiently select the ideal model order based on preset criteria, such as AIC or BIC. Following the automated model selection process, the code proceeds to partition the updated data into two distinct sets, namely a training set and a test set.

The test set is composed of the final 30 observations, whereas the training set include the remaining 274 observations. The previously described division serves a pivotal function in facilitating the assessment of a model on data that has not been previously encountered. The purpose of this evaluation is to assess the model's ability to generalise and generate precise predictions. The incorporation of a test set is essential in order to simulate real-world forecasting scenarios, thereby ensuring that the analysis is firmly grounded in practical relevance.

The final segment of the code is devoted to the construction and implementation of a specific ARIMA model with an order of ((5, 0, 2)) that has been selected from the Auto Arima fuction, using the 'statsmodels' package. The model indicated above is applied to the training data, namely 'train['shift log sqrt']', in order to incorporate the selected autoregressive, differencing, and moving average components. The summary of the fitted model provides valuable insights into the coefficients, diagnostics, and the level of fit. The chosen sequence adeptly captures the intrinsic temporal patterns inherent in the unemployment data, tries to enable accurate forecasts for test data and a comprehensive understanding of the problem.

5.1.2 SARIMA

In the process of examining the unemployment rate in Ireland, the objective was to utilise the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, which is a robust statistical technique capable of effectively addressing both the trend and seasonality elements within a time series. Upon initial observation of the original series, it was noted that it exhibited non-stationary behaviour. To address this, a first-order differencing transformation was applied, resulting in a stationary series that could be effectively modelled using SARIMA. The model was subsequently selected with the special non-seasonal order (5, 0, 2) and for the seasonal part (0,0,0,12).

- The parameter (p = 5), denotes the number of autoregressive terms in the model. This quantity enables the model to effectively capture the associations between the observation and its five preceding values, thereby accounting for any potential delayed effects.
- The (d = 0), Given that the data had already undergone a transformation into a stationary form by the application of the first difference, there is no further differencing required.
- The parameter (q = 2), represents the number of moving average terms in the model. This parameter allows the model to account for the influence of past error terms, hence capturing the temporal dependencies present in the error structure.

The model did not incorporate any seasonal components, as this decision was made based on the distinct traits and patterns seen in the series of unemployment rates. The parameters were chosen based on a knowledge of the underlying dynamics of the data, exploratory analysis, and the theoretical principles of SARIMA modelling. The model fitting procedure, conducted using the SARIMA class from the 'statsmodels' package, resulted in a model that captures the fundamental temporal patterns of the Irish unemployment rate. This model may now be utilised as a foundation for subsequent analysis, prediction, and interpretation.

5.2 Machine Learning Models

5.2.1 Random Forest

In order to analyse the cyclic nature of the time series, we generated lagged features by moving the unemployment rate by intervals of one, two, and three months. The lagged values were utilised as the independent variables in our analysis for the Random forest,

while the current unemployment rate was considered as the dependent variable. After undergoing this process, the dataset was partitioned into two subsets: a training subset and a testing subset. The final 30 observations were specifically kept aside for the purpose of validation. Subsequently, we choose to employ the Random Forest Regressor, a robust ensemble learning technique, for the purpose of modelling the data.The model was initialised with predetermined parameters, consisting of 300 decision trees, a maximum of two features for each split, and a constant random state to ensure reproducibility.

The Random Forest model was subsequently trained using the lagged features and the related training data. The Random Forest's ensemble methodology facilitated the acquisition of knowledge from diverse viewpoints through the combination of many decision trees, each constructed using a random selection of attributes. The implementation of this particular approach effectively mitigated the potential issue of overfitting and facilitated the model's ability to accurately capture intricate non-linear patterns inherent in the dataset. Subsequently, the trained model was utilised to make predictions for the Irish unemployment rate by applying it to the test set.

5.2.2 Ridge Regression

In the field of time series analysis, it is widely recognised that past values, commonly referred to as lags, frequently possess significant informational content pertaining to future values. In this particular instance, three lagged variables were generated to represent the values of the unemployment rate from one, two, and three months before the current period like the other model mentioned above. The dataset is divided into train and test. The process of training the Ridge Regression model consisted of the selection of a suitable value for the alpha hyperparameter. The parameter alpha is responsible for regulating the intensity of the regularisation applied to the model. This regularisation serves the purpose of mitigating overfitting by imposing a penalty on the complexity of the model. The utilisation of an alpha value of 100 enables the model to strike a compromise between effectively capturing the patterns in the data and preserving a level of simplicity.

Once the model had been trained on the modified dataset, it was subsequently employed to make predictions regarding the unemployment rate for the test dataset. The predictions were generated by utilising the acquired coefficients and applying them to the lagged features within the test dataset. The prediction phase demonstrates the model's capacity to extrapolate from the training data to unfamiliar data, offering valuable insights on forthcoming developments in the unemployment rate.

5.2.3 KNN Regression

To enhance the original dataset on the unemployment rate in Ireland, three lagged features were generated to capture the values from the preceding one, two, and three months. The lagged values were utilised as predictor variables, including the temporal linkages present in the data. The implementation of a k-Nearest Neighbours (k-NN) regression model was subsequently carried out to forecast the unemployment rate based on altered dataset. A loop was executed to iterate over several values of k, ranging from 1 to 5, which corresponded to the number of neighbours taken into account for prediction. The model underwent the training on the training dataset for each value of k, taking into account the designated number of nearest neighbours. Subsequently, the k-Nearest Neighbours

(k-NN) model that had undergone training was employed to generate predictions on the test dataset, with each value of k being considered. The present study employed the k-NN regression model to effectively forecast an unemployment rate in Ireland, while also investigating the influence of varying numbers of neighbours on the accuracy of predictions. The analysis yielded significant findings regarding the ideal quantity of neighbours for this particular problem, indicating that the inclusion of four nearest neighbours resulted in the most accurate alignment with the data.

5.2.4 XGBoost

The XGBoost model was implemented to predict the unemployment rate in Ireland using the clearly defined methodology. To begin the data transformation process, three lagged features were generated to capture the one, two, and three-month changes of the unemployment rate within the original dataset. The lagged values were utilised as predictor variables in order to enable the model to comprehend the historical connections present in the data. Subsequently, the dataset was partitioned into distinct training and the testing subsets, employing an designated threshold date of January 1, 2015, in order to ascertain the evaluation of the model on previously unobserved data.

The XGBoost model underwent training using predetermined hyperparameters. These hyperparameters are included 100 estimators, which were used to regulate the number of boosting rounds. Additionally, a maximum depth of 5 was imposed to limit the complexity of the individual decision tree. Furthermore, a learning rate of 0.1 was employed to facilitate the convergence towards the best solution. The loss function was formally specified as the squared error, which is commonly employed in regression tasks. The model was applied to the training dataset and evaluated on the training and testing dataset, incorporating early stopping techniques to mitigate the risk of overfitting.After being trained, the XGBoost model was utilised to generate predictions on the test dataset, which served as a representation of the projected unemployment rate.

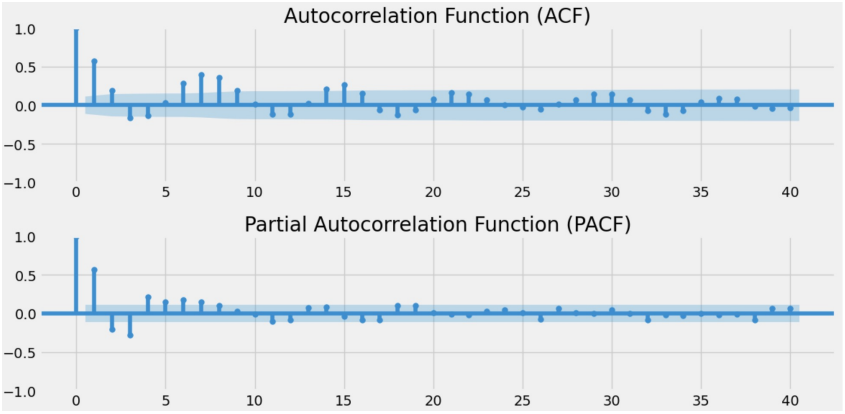

Figure 6: ACF and PACF Plots

Table 2: Evaluation Metrics for ARIMA

Order	RMSE	R-squared	MAPE (%)
(2, 1, 0)	0.012	-0.282	420.856
(2, 1, 1)	0.012	-0.248	405.358
(5, 0, 2)	0.010	0.141	268.453

6 Evaluation

6.1 Experiment 1 - ARIMA model

Accurately predicting time series data is of utmost importance in numerous fields of study, requiring the use of strong analytical techniques and careful model selection. The present study conducted the thorough analysis on a particular data series which has been converted to stationary with log shif values, with the aim of determining the most suitable Autoregressive Integrated Moving Average (ARIMA) model. The data set was initially subjected to the visual inspection of utilising the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) as shown in Figure 6. The partial autocorrelation function (PACF) indicated a statistically significant association at the first two lags values, which led the decision to choose an autoregressive order of p=2. Meanwhile, the Autocorrelation Function (ACF) demonstrated a significant decrease following the initial lag, prompting the selection of a moving average order of 1 (q=1). Given that the series had already undergone differencing, the differencing order of d=1 was applied, leading to the initial ARIMA order (2,1,1).

In order to assure the robustness of model selection, the study employed the Auto ARIMA function, which determined the optimal order as (5, 0, 2). The ordering of this model is indication of a sophisticated approach that incorporates five historical observations for the autoregressive (AR) component and two previous forecast mistakes for the moving average (MA) component. After applying ARIMA(5, 0, 2) model to the training dataset, a thorough evaluation of its performance was conducted using the test data. The evaluation metrics of the model are presented in Table 2, indicating an RMSE of 0.00960, an R-squared value of 0.1491, and a MAPE of 196.64%age. The results of the

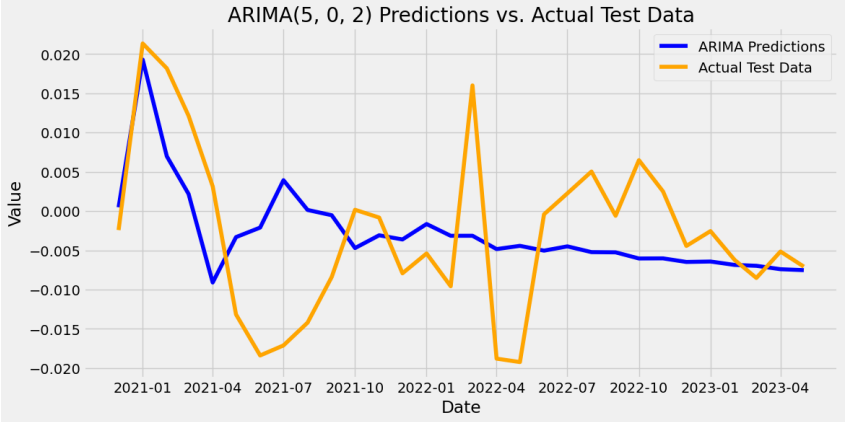

Figure 7: Predictions vs. Actual Test Data for ARIMA

analysis suggest that the ARIMA(5, 0, 2) model exhibited better results compared to other examined orders in accurately capturing the underlying patterns present in data.

Figure 7, depicts the graphical representation illustrating the model's predictions in comparison to the actual test data, so providing the visual evidence of the model's proficiency in forecasting. The similarity observed between the forecast and the observed values depicted in the graph helps to highlight the effectiveness of the model. The selected order indicates that the data might have short-term correlations. However, the high MAPE value suggests the model might not be capturing some nuances in the data, possibly non-linear patterns or external factors affecting unemployment.

6.2 Experiment 2 - SARIMA model

The study used the Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX) model, exploring the different orders for both non-seasonal and seasonal components. The seasonal order denotes the autoregressive (P), differencing (D), and moving average (Q) components at the seasonal frequency, in addition to the duration of seasonal cycle (S). The seasonal order employed in this study was maintained as (0, 0, 0, 12), denoting the absence of seasonal components and a seasonal cycle of 12 months.

Three SARIMAX models were assessed, specifically with the seasonal order indicated, namely (2, 1, 0), (2, 1, 1), and (5, 0, 2). The evaluation of the models was conducted by employing Root Mean Square Error (RMSE), R-squared, and Mean Absolute%age Error (MAPE) metrics. The SARIMAX model, specifically with an order of (5, 0, 2), demonstrated superior performance based on the evaluation metrics. The root mean squared error (RMSE) was calculated to be 0.011382, indicating the model's accuracy in predicting the target variable. The R-squared value, which measures the proportion of variance explained by the model, was found to be -0.196021. Additionally, the mean absolute%age error (MAPE) was determined to be 197.84%t, providing insight into the model's average prediction error as a%age of the actual values. Table 3 provides a detailed overview of the evaluation metrics related to various SARIMAX orders, encompassing RMSE, R-squared, and MAPE values. The table shown provides an comprehensive analysis of the models' performance, emphasising the superior fitting capabilities of SARIMAX(5, 0, 2) order.

Table 3: Evaluation Metrics for SARIMA

Order	Seasonal Order	RMSE	R-squared	MAPE (%)
(2, 1, 0)	(0, 0, 0, 12)	0.012	-0.282	420.856
(2, 1, 1)	(0, 0, 0, 12)	0.012	-0.248	405.358
(5, 0, 2)	(0, 0, 0, 12)	0.012	-0.407	301.793

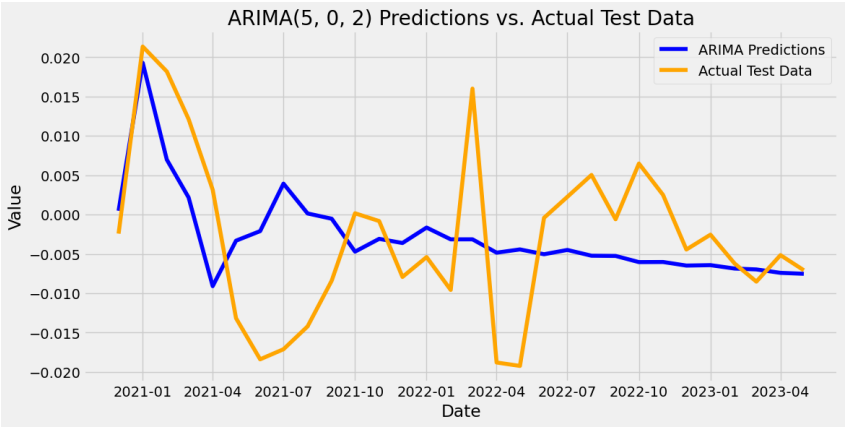

Figure 8: Predictions vs. Actual Test Data for SARIMA

After the selection of the most appropriate model, predictions were generated for the test segment of the dataset. The graph that is shown in Figure 8 provide an visual representation of the relationship between the predicted values and the actual test results, thereby demonstrating the effectiveness of the model. The SARIMAX predictions, shown by the blue line, exhibit a strong correspondence with the actual test results, as depicted by the orange line. The alignment seen in this study highlights the model's proficiency in accurately representing the fundamental dynamics of the modified unemployment data, thereby providing evidence of its ability to serve as an influential tool in economic forecasting and policy development. The data might not have significant seasonal patterns to exploit.

6.3 Experiment 3 - Random Forest

The purpose of this study was to utilise Random Forest algorithm, a robust ensemble learning technique, in order to model the time series data of Irish unemployment. The lagged variables representing one, two, and three-month shifts were derived from the raw data and utilised as features for model. Subsequently, dataset was partitioned into training and testing sets, wherein the final 30 observations were exclusively allocated for the purpose of validation. The process of hyperparameter tuning was undertaken in order to identify optimal configuration for Random Forest model. A total of four distinct hyperparameter combinations were evaluated, encompassing changes in both of the number of trees (n estimators) and the number of features taken into account for determining the optimal split (max features). The evaluation utilised the 3 metrics that is an RMSE, R-squared , and the MAPE (Mean Absolute%age Error), as shown in the Table 4.

Table 4: Evaluation Metrics for Random Forest

n_estimators	max_features	random_state	RMSE	R-squared	MAPE (%)
100	2	1	7.935	0.8937	19.721
200	3	1	7.710	0.899	19.690
300	2	1	7.987	0.892	19.678
400	auto	1	7.642	0.901	19.724

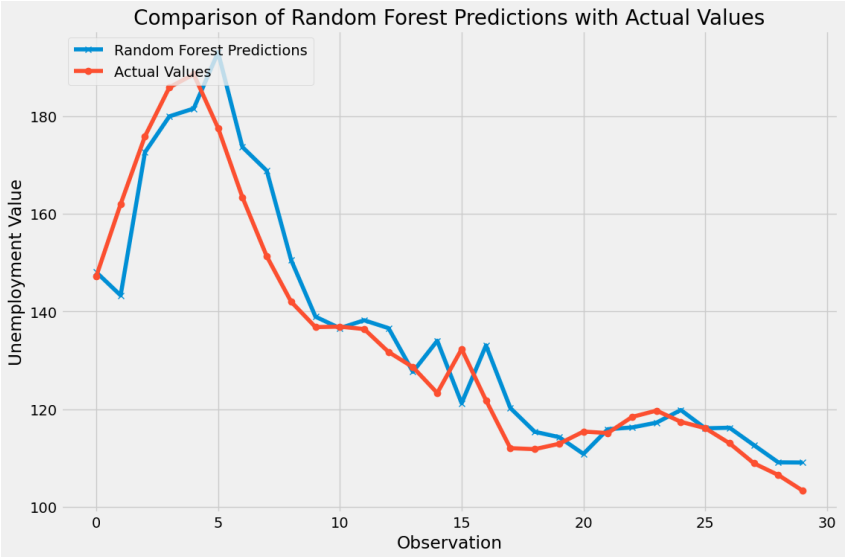

Figure 9: Predictions vs. Actual Test Data for Random Forest

The model that exhibited an highest performance, utilising hyperparameters 'n estimators: 150', 'max features': 'auto', 'random state': 1, attained a root mean squared error value of 7.64, R-squared value as 0.90, and a mean absolute%age error (MAPE) value of 19.72%. The previously mentioned optimal model was utilised to generate subsequent predictions, and a visual representation of the comparison between predicted and actual values can be seen in Figure 9. Random Forest can capture intricate relationships in data, making it apt for datasets with complex underlying patterns. The lagged variables (representing previous months) likely provided valuable temporal information for the model.

6.4 Experiment 4 - Ridge Regression

The study used the Ridge regression as the machine learning technique to predict future values in a time series and predict the unemployment rate. The approach incorporates an hyperparameter, denoted as alpha, which serves as the regularisation strength. This hyperparameter governs the balance between precisely fitting the data and preserving simplicity (reducing variance. Various values of alpha were examined, spanning from 0.001 to 10, in order to ascertain the influence on the model. Based on the results obtained, the ideal value of alpha was determined to be 0.001 (refer to Table 5).

Table 5: Evaluation Metrics for Ridge Regression

Alpha	RMSE	R-squared	MAPE (%)
0.001	5.859	0.942	2.897
0.010	5.883	0.941	2.951
0.100	6.179	0.935	3.278
1.000	7.529	0.904	4.064
10.000	10.846	0.801	6.236

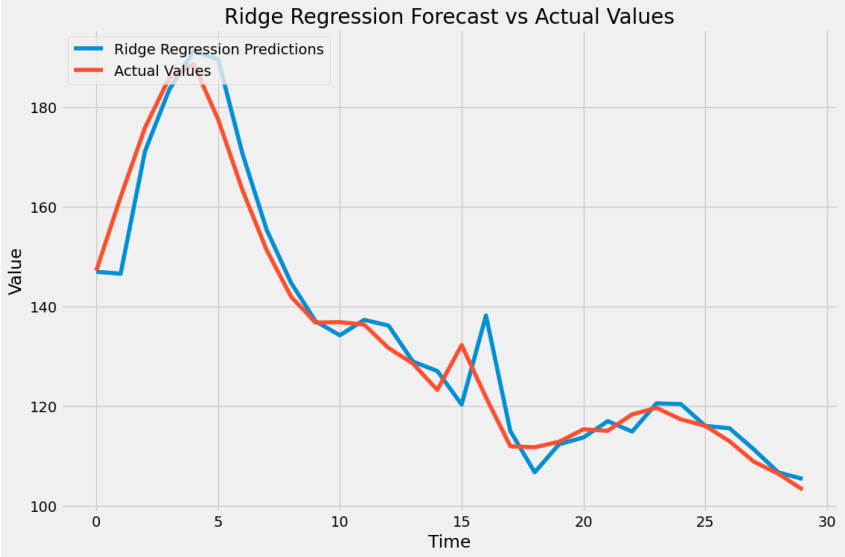

Figure 10: Predictions vs. Actual Test Data for Ridge Regression

The model that corresponds to the given data produced a Root Mean Square Error (RMSE) of 5.8594, a coefficient of determination (R-squared) value of 0.9421, and a Mean Absolute%age Error (MAPE) of 2.87%. These results indicate a high level of predicted accuracy. The comparison between the anticipated and actual values, as depicted in Figure 10, provides more evidence supporting the model's capacity to accurately represent an underlying patterns and variations. The methodical methodology employed for selection and evaluation of the alpha parameter showcases the robustness of Ridge regression as a forecasting technique within the specified context. The data might have a largely linear relationship with the unemployment rate, allowing Ridge to excel. The regularization ensured that the model didn't overfit, leading to better generalization on test data.

6.5 Experiment 5 - KNN Regression

The purpose of this study was to analyse Irish unemployment data in order to effectively predict future trends. In this context, the K-Neighbors Regressor was utilised, which is a non-parametric technique that operates based on the concept of similarity. It leverages the k nearest observations from the training set to make predictions for unknown data points. The data underwent preprocessing by generating features that capture one, two, and three-month shifts. Additionally, a partition was established to reserve the final 30 observations for testing purposes. The choice of the neighbourhood size, denoted as k,

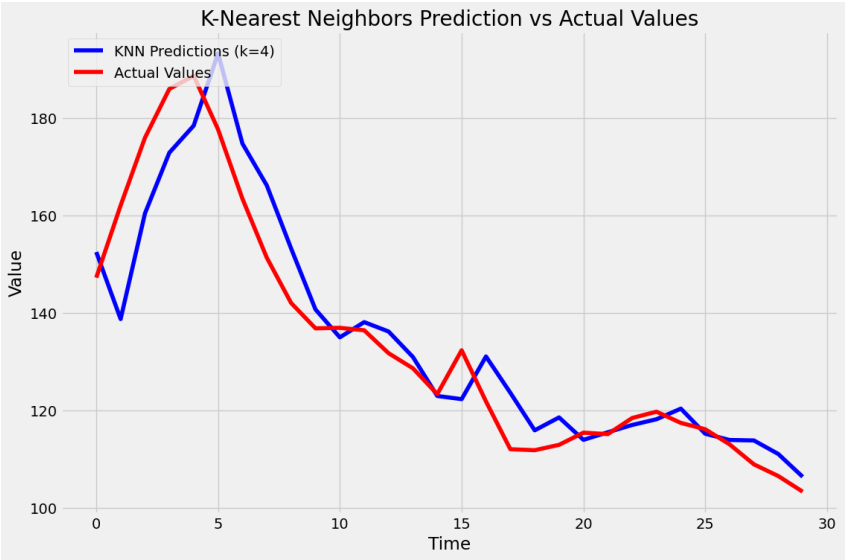

Figure 11: Predictions vs. Actual Test Data for KNN Regression

Table 6: Evaluation Metrics for Ridge Regression

K	RMSE	R-squared	MAPE (%)
1	9.527	0.846	4.934
2	10.246	0.822	5.077
3	8.943	0.865	4.657
4	8.733	0.871	4.585
5	8.850	0.867	4.781

was a crucial factor in determining the effectiveness of model. A number of test were done to assess the performance using various values of k, ranging from 1 to 5. The primary evaluation of all the measures employed were RMSE, Mean MAPE, and R-squared. The findings revealed a complex association between the size of the neighbourhood and the predicted accuracy of the model. As an example, when the neighbourhood size was set to k=4, the performance of the model was found to be as the most balanced. The root mean square error (RMSE) was measured to be 8.73, the mean absolute%age error (MAPE) was 4.585%, and the coefficient of determination ($R\hat{2}$) value was 0.87. The observed arrangement showed an ideal balance between overfitting and the underfitting, guaranteeing that the model accurately captured the inherent patterns in the dataset without excessively increasing its complexity.

Table 6 presents a comprehensive overview of the performance metrics associated with the various hyperparameter configurations. This resource functions as a thorough reference for the comprehending the trade-offs associated with the selection of neighbourhood size, hence providing guidance in determining the most appropriate configuration for the given task. The inclusion of a graphical depiction comparing the observed values to anticipated values offered further understanding into performance of the model as shown in the Figure 11. The analysis demonstrated a strong correspondence between the projected outcomes and the observed values, underscoring the model's capacity to capture the

Table 7: Evaluation Metrics for XGBoost

n_estimators	RMSE	R-squared	MAPE (%)
10	139.436	-15.202	90.07
100	55.252	-1.544	36.42
200	17.256	0.751	10.77
1000	9.338	0.927	4.68

sequential relationships within the unemployment dataset. The data might have short-term correlations, which KNN can leverage. The chosen 'k' value of 4 indicates the model found a balance between capturing these correlations without overfitting.

6.6 Experiment 6 - XGBoost

The objective of this study was to employ the XGBoost algorithm, a widely utilised gradient boosting framework, for purpose of the modelling and predicting the target variable. The number of estimators (n estimators) that is a crucial hyperparameter in XGBoost, as it governs the number of boosting rounds and hence impacts the complexity and performance of model. In order to determine most suitable value for the parameter n estimators, an experiment was undertaken. This involved training the XGBoost model using various values, specifically 10, 100, 200, and 1000. Every model underwent training with a learning rate of 0.01, a maximum depth of 10, and early stopping rounds set to 100 in order to mitigate the risk of overfitting.

The evaluation criteria employed for assessing the models encompassed RMSE, R-squared, and Mean Absolute%age Error (MAPE). The findings are clearly presented in Table 7, indicating a positive correlation between the number of estimators and the performance enhancement. The model that was configured with n-estimators set to 1000 demonstrated lowest RMSE of 9.34, the highest R-squared value of 0.93, and a mean absolute%age error of 4.68%. Based on the results obtained, we have chosen the model with n-estimators = 1000 as the most effective model. Subsequently, the model mentioned earlier was employed to forecast the target variable. The ensuing comparison between observed and projected values is visually depicted in Figure 12. The strong correspondence observed between the observed and forecasted values provides evidence of the model's resilience and ability to accurately predict outcomes. With a high number of estimators (1000), the model was able to progressively refine its predictions. The nature of gradient boosting allows XGBoost to improve iteratively, capturing both linear and non-linear patterns.

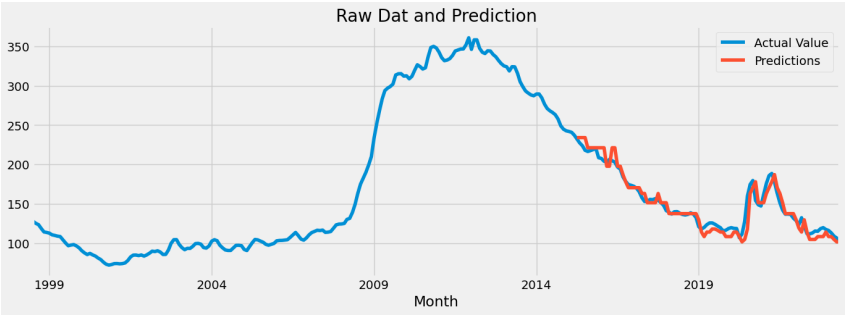

Figure 12: Predictions vs. Actual Test Data for XGboodt Regression

6.7 Discussion

This study undertook the comprehensive investigation into a prediction of the unemployment rate in Ireland, employing a combination of traditional statistical techniques and sophisticated machine learning models. During initial stage, the Autoregressive Integrated Moving Average (ARIMA) model was utilised, considering several orders to effectively capture the intrinsic patterns present in the time series data. SARIMA, which is an extension of ARIMA (Autoregressive Integrated Moving Average), was utilised to incorporate seasonal components in order to improve the model's ability to capture periodic changes in the unemployment rate.

In addition to utilising statistical models, the study also incorporated modern machine learning techniques such as Random Forest, Ridge Regression, K-Neighbors Regressor (KNN), and XGBoost. The Random Forest algorithm, which is an ensemble learning technique, offers a reliable approach for modelling complex data by optimising hyperparameters. Ridge Regression, renowned for its regularisation capability, was carefully optimised over a range of alpha values in order to achieve an optimal balance between the accurately fitting the data and preserving simplicity. The K-Neighbors Regressor (KNN), which is an non-parametric method, is evaluated using thevarious sizes in order to determine the optimal value that is between overfitting and the underfitting. Finally, the research conducted experiments with XGBoost, a framework for gradient boosting, using varying numbers of estimators in order to assess its robustness and prediction accuracy.

The Ridge Regression model exhibits an superior R-squared value in the comparison to the ARIMA model. The R-squared number, sometimes referred to as the coefficient of determination and also goodness of fit, provides insight into the extent to which the variability in the dependent variable can be explained by the independent variable(s). A greater R-squared value signify a better degree of explanation provided by the model for the variability observed in the response data relative to its mean. In this particular scenario, it can be observed that the Ridge Regression model exhibit an higher degree of explanatory power in comparison to the ARIMA model. This observation may suggest that Ridge Regression model possesses better forecasting capabilities. In addition, it should be noted that the Ridge Regression model has a comparatively reduced the Mean Absolute%age Error (MAPE) in comparison to the ARIMA and SARIMA statistical model. The Mean Absolute%age Error (MAPE) is an indicator of statistics used to assess the accuracy of a forecasting approach. A lower MAPE value is indicative of higher prediction accuracy.

Table 8: Evaluation Metrics for Various Models

Model	RMSE	R-squared	MAPE (%)
ARIMA(5,0,2)	0.0096	0.1491	196.64
SARIMA(5,0,2)	0.011	-0.1960	197.84
Random Forest	7.64	0.90	19.72
Ridge Regression	5.859	0.9421	2.89
KNN(k=4)	8.733	0.871	4.585
XGBoost(1000)	9.338	0.927	4.68

Hence, based on an evaluation of both R-squared and MAPE metrics, it can be concluded that Ridge Regression model outperforms the ARIMA model, despite the latter exhibiting a lower RMSE value. This highlight the importance of taking into account many factor in the process of defining the optimal model, compared to the only relying on a singular statistic. A comprehensive analysis of these models yielded intricate observations regarding their respective performance as show in table 8. Within the field of statistical models, the ARIMA model with an order of (5, 0, 2) demonstrated a robust framework, showcasing a refined methodology that incorporated five past observations for the autoregressive component and two preceding forecast errors for moving average component. In this specific situation, the performance of the SARIMA model, despite its inclusion of seasonal components, did not exhibit a significant improvement over ARIMA model. On the oher hand, Ridge Regression, an machine learning model, demonstrated superior performance compared to other models. It achieved an RMSE (Root Mean Square Error) of 5.8594, an R-squared value of 0.9421, and a MAPE (Mean Absolute percentage age Error) of 2.89%. The aforementioned measures highlight the capacity of Ridge Regression to effectively capture and depict fundamental patterns and fluctuations. This study discovered the potential of machine learning as compared to traditional statistical models in the field of economic forecasting, particularly with regard to the prediction of unemployment time series data in Ireland.

7 Conclusion and Future Work

The current investigation has conducted a comprehensive analysis of Traditional statistical models, namely ARIMA and SARIMA, as well as contemporary machine learning methods including Random Forest, Ridge Regression, KNN, and XGBoost, in order to forecast the unemployment rate in Ireland. The results demonstrate the merits and limitations of each approach, highlighting the notable performance of the Ridge Regression and XGBoost with 1000 estimators in terms of their capacity to strike a favourable balance between the accuracy and interpretability as mentioned in table 7. Although traditional models have offered fundamental insights, machine learning models have shown an distinct advantage in capturing intricate patterns and providing a sophisticated comprehension of the underlying dynamics of unemployment rates in the region. There exist numerous potential in the areas for future research that show promise. The inclusion of further macroeconomic factors has the potential to offer a more intricate understanding of the unemployment rate. The utilisation of deep learning techniques like as Long Short-Term Memory (LSTM) has the potential to record the intricate patterns of even greater complexity. The extension of the research to more economies would enable the examination of practicality of the results in a broader context. The practical implications of implementing real-time prediction system are noteworthy for policymakers and economists. By placing emphasis on the interpretability of models, it is possible to enhance accessibility of these tools to a wider range of individuals. These prospective undertakings have the potential to yield the more accurate, comprehensible, and flexible models, so augmenting the domain of economic forecasting.

References

Box, G. E., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. (2015). *Time series analysis: forecasting and control*, John Wiley & Sons.

Breiman, L. (2001). Random forests, *Machine learning* **45**: 5–32.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Didiharyono, D. and Syukri, M. (2020). Forecasting with arima model in anticipating open unemployment rates in south sulawesi, *Int. J. Sci. Technol. Res* **9**(3): 3838–3841.

Dumičić, K., Čeh Časni, A. and Žmuk, B. (2015). Forecasting unemployment rate in selected european countries using smoothing methods, *World Academy of Science, Engineering and Technology: International Journal of Social, Education, Economics and Management Engineering* **9**(4): 867–872.

Gogas, P., Papadimitriou, T. and Sofianos, E. (2022). Forecasting unemployment in the euro area with machine learning, *Journal of Forecasting* **41**(3): 551–566.

Hannan, D. F., Ó Riain, S. and Whelan, C. T. (1997). Youth unemployment and psychological distress in the republic of ireland, *Journal of Adolescence* **20**(3): 307–320.

Hatipoğlu, , Belgrat, M. A., Degirmenci, A. and Karal, (2021). Prediction of unemployment rates in turkey by k-nearest neighbor regression analysis, *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–5.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**(1): 55–67.

Kyung, A. and Nam, S. (2019). Study on unemployment rate in usa using computational and statistical methods, *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE, pp. 0987–0993.

Mulaudzi, R. and Ajoodha, R. (2020). An exploration of machine learning models to forecast the unemployment rate of south africa: a univariate approach, *2020 2nd international multidisciplinary information technology and engineering conference (IMITEC)*, IEEE, pp. 1–7.

Nkwatoh, L. S. (2012). Forecasting unemployment rates in nigeria using univariate time series models, *International Journal of Business and Commerce* **1**(12): 33–46.

Pavlyshenko, B. M. (2016). Linear, machine learning and probabilistic approaches for time series analysis, *2016 IEEE First International Conference on Data Stream Mining Processing (DSMP)*, pp. 377–381.

Rahmat, B., Purbasari, I. Y., Sari, N. K., Widiwurjani, Nugroho, B. and Widyantara, H. (2022). Indonesia's open unemployment rate prediction system using deep learning, *2022 IEEE 8th Information Technology International Seminar (ITIS)*, pp. 7–10.

Ramli, S. F., Firdaus, M., Uzair, H., Khairi, M. and Zharif, A. (2018). Prediction of the unemployment rate in malaysia, *Int. J. Mod. Trends Soc. Sci* **1**(4): 38–44.

Schröer, C., Kruse, F. and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model, *Procedia Computer Science* **181**: 526–534.

Sen, M., Basu, S., Chatterjee, A., Banerjee, A., Pal, S., Mukhopadhyay, P. K., Dutta, S. and Tarafdar, A. (2022). Prediction of unemployment using machine learning approach, *2022 OITS International Conference on Information Technology (OCIT)*, pp. 1–5.

Song, Y., Liang, J., Lu, J. and Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression, *Neurocomputing* **251**: 26–34.

Urrutia, J., Tampis, R. and Atienza, J. E. (2017). An analysis on the unemployment rate in the philippines: A time series data approach, *Journal of Physics: Conference Series*, Vol. 820, IOP Publishing, p. 012008.