National College of Ireland

Comparison of Crime Rates In Ireland And India Using Machine Learning Methodologies

MSc Research Project
MSc in Data Analytics

Chinmay Laxmikant Mukim
Student ID: x21145024

School of Computing
National College of Ireland

Supervisor: Prof. Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	**Chinmay Laxmikant Mukim**
Student ID:	**x21145024**
Programme:	**Masters of Science in Data Analytics**
Year:	**2022-2023**
Module:	**MSc Research Project**
Supervisor:	**Prof. Jorge Basilio**
Submission Due Date:	**15th December 2022**
Project Title:	Comparison of Crime Rates In Ireland And India Using Machine Learning Methodologies
Word Count:	**7214**
Page Count:	22

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:	
Date:	**15th December 2022**

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):	

Comparison of Crime Rates In Ireland And India Using Machine Learning Methodologies

Chinmay Laxmikant Mukim
x21145024

Abstract

A gross violation of law is one of the factors of the discussion in front of lots of regions. Various criminal activities nowadays are at their peak level. This research is about the major factor in front of the system which is the crime rate. When criminal activity happens in society, there are lots of factors responsible for it. This research is concentrated on two of the biggest factors for increasing or decreasing crime rates.

The machine learning methodologies are applied to the crime-related datasets for two countries namely Ireland and India. This paper investigates the machine learning methodologies to predict the crime rate with measuring units like the coefficient of determination, RMSE, MSE, and mean. The performance of the models applied on the dataset, in which the linear regression is the best model with the coefficient of determination of 92 percent for Ireland and 93 percent for India. The RMSE and MSE scores were also observed.

This research will help in predicting, and locating crime rates with the comparison of crime rates based on the factors and the models which are applied to the data respectively.

1 Introduction

1.1 Background

Increasing crime rates is one of the most important issues nowadays, based on this handling crime is the greatest challenge that the government is facing. As precaution is better than cure, dealing with a crime will be easy by predicting them and analyzing the factors such as population, and unemployment responsible for increasing or decreasing crime rates.

1.2 Importance

The aim of this work is a prediction of crime rates based on historical data, an analysis of all types of crime, and, a comparison of crime between Ireland and India based on population, and unemployment. The benefit of this is, we will come to know that how population, and unemployment affect crime rates. As it is the simplest way to get hands-on the criminal activities and maintain peace in the state.

1.3 Research question and objectives

Which factors such as population, and unemployment affect crime rates in Ireland and India? Using machine learning methodologies. The objectives of this research are, firstly, appropriate datasets will be gathered with respect to crime rates in both countries, datasets for population, and unemployment for both Ireland and India. The machine learning models used are ARIMA, Exponential Smoothing, Sequential Neural Network, and Linear Regression. Secondly, the datasets will be merged to build a connection and the machine learning models as mentioned above will be implemented. The respective results of these models will be compared to identify which model is best for crime prediction. And lastly, according to the results and to satisfy the research question the conclusion will be made as to which factor is responsible for the increase or decrease in crime rates for both countries.

1.4 Assumptions

the primary assumptions for this research are that factors such as population and unemployment will affect crime rates in both countries but mainly for India the factor affecting crime rate will be population and for Ireland, it might be unemployment. Even though the population rate is very high in India, unemployment might be a factor that is affecting crime rates more than the population. For Ireland, both factors can be responsible.

These assumptions are made before implementing the models and operating the data. In the conclusion and discussion section, these assumptions will be referred to and compared to the results as well.

1.5 Structure of the report

The flow of this research paper, as it consists of the previous work which will be mentioned in the related work, the methodology, design specifications, implementation, evaluation, discussion, conclusion, and future work.

2 Related Work

This section contains the past work done on crime rate prediction, which is related to this research. This section has the literature review which is based on the crime data, for separate research methodologies for both countries. Moreover, it also has details about which factor is responsible for the variation in the crime rate statistics. The time series methods as well as the statistical tests were referred to in this research from the previous papers. The related work section has both the previous work done with respect to Ireland and India as they have different levels of parameters with them. Therefore, this section is divided into subsections to get more clarity.

2.1 Previous Work on Indian Crime Rates

Agarwal et al. (2018) This research is published on the IEEE website in 2018. The crime prediction was done which is based on the statistical models. For this research, the dataset was obtained from the open government dataset for the crime rates. The crime statistics for India for each district. The dataset contains various types of crime types

such as murder, rape, theft, and so on. Statistical models such as WMA, FCR, and AGP where the AGP (Arithamatic Geometric Progression) model was the best one to predict the IPC crime. To increase the accuracy of the prediction it was suggested to use different machine learning models as for this research the SMAPE results were measured.

The crime data was analyzed, the patterns were detected and a prediction was done on the Indian crime rates. The paper was published in the year 2017. Yadav et al. (2017) The data was fetched from the online portal for Indian crimes. The weka tool was used as it has all the data mining and machine learning tools such as clustering, regression, and so on. For the crime pattern prediction and analysis in the paper, the Association mining, K-means clustering, and Naive Bayes classification technique, with correlation and regression respectively. Crime rate prediction with the help of a crime hotspot was suggested in this research. This factor was the main reason to focus on the region part in our research paper.

Moreover, the various machine learning algorithms were applied to Indian penal code data as the variables have their own dependencies. The research was published in July 2022. The author approached a data-driven approach to draw insightful knowledge from the crime rate data. The machine learning models like linear regression with single and multiple methods, random forests, and decision trees were implemented. The MYSQL workbench was used with the R programming. Among these models, the Random forest was the best model with the R square value of 0.96 and MAPE was 0.2. with the statistics of Andra Pradesh has the highest amount of crime rates. Aziz et al. (2022)

The multivariate time series approach was implemented on the crime statistics data to predict the crime trends was done in this research. This research was published IEEE in April 2007. Chandra et al. (2008) Similar to the main research, this paper also focuses on the crime location as the time series data to find the pattern. The sequential data pattern problem was solved in this research by using transformation techniques and time series clustering. As the dimensions have a different levels of weight, the time wrapping and parametric Minkowski model were proposed for finding similar crime trends.

The research paper on comprehensive comparative methods for crime rate prediction and analysis was implemented on Indian crime datasets. The research paper was published in 2018. A systematic crime analysis and prediction approach was implemented for measuring the trends and relationships in crime patterns. By using this the system predicts the region where a large amount of criminal activity happened. Vaidya et al. (2018) In this research, a good literature review was done with respect to the machine learning methods. The machine learning models such as K-means clustering, Fuzzy-c, hierarchical clustering, and, self-organizing map methods were compared in search of higher accuracy. In which the SOM method has the highest accuracy as the number of clusters increases.

Priya and Gupta (2015) In this research an autoregressive linear regression model was implemented by the author for predicting future crimes. The paper was published in 2015, where the time series sequence of the data was monitored with the effect of the population, economic conditions, and the effect of law enforcement agencies. the correlation was observed in the analysis for the dependent factors. The actual average crime rate was 10.389 whereas in years 2002 and 2010 has the most number of crimes with a prediction accuracy of 86.56 percent.

The research paper titled crime rate prediction was published in 2020. The linear regression method was used to predict the crime rates based on the historical data values. Mahendra et al. (2010) The multiple linear regression was used as the data has dates and the type of crime, and the population is a factor that was used to determine the crime

rates with the other model of logistic regression and KNN. The data cleaning and data transformation were also done before implementing the model. The types of crime are monitored but murder is the one that got the highest prediction rate by the model.

2.2 Previous Work on Irish Crime Rate

McKenna et al. (1997) This article is about suicide, homicide, and crime in Ireland. The article was published in 2007 and has the overall relationship in between the crime rate activities. The suicide and homicide rates were compared so to get the relationship. Not so technical work but it gave the overall review of the relationship between these two types of crime. In the conclusion, these two types of crime are related to each other. According to the hypothesis, there is a positive relationship in between suicide and homicide rates in Ireland.

Adedokun (2020) This research paper was published in the national college of Ireland and referenced from the national college of Ireland library. The research paper was published in 2020 with the crime occurrence of crime in house price prediction for Ireland. The performance of the predictive and classification models was monitored and according to that the effect of crime was concluded. Comparative analysis was done by the data mining model with the help of the linear model, support vector machine, and, random forest.

The article titled Imprisonment and crime rate in Ireland is published in 2003 has an introduction about imprisonment in Ireland and the crime rates. O'Sullivan and O'Donnell (2003) The theoretical models with the criminal behavior were present with the prison system and the data of prisoners. Moving toward the crime rates, the results have a summary of the trends in crime and imprisonment. Basically the relationship between them and the crime and punishment statistics.

Therefore, while working on the crime statistics, to know the crime rates behavior these research papers and articles were referred to.

2.3 Previous Work Based on Models Implemented on Crime Data

The crime data is forecasted by using the ARIMA model. Property crime was monitored in a city in China with the given data of 50 weeks. The ARIMA model and the exponential smoothing models were compared. The ARIMA model has a higher fitting and forecasting accuracy than exponential smoothing. The research paper was published in 2008. Chen et al. (2008) The steps were carried out as identify the ARIMA. The SPSS was used for the forecasting purpose and the short-term forecasting was done using ARIMA. The fitting and forecasting results were compared with the other forecasting tools SES, and RES. The RMSE and MAPE values for the ARIMA model are 56.94 and 9.48 respectively.

Predicting and forecasting crime rates on a small scale is done using forecasting tools in this research. Gorr et al. (2003) The short term of forecasting the average crime of the instinct. A fixed effect of the regression model shows the absolute forecast for the crime rates. The Holt exponential smoothing with monthly estimated seasonality using the city-wide data. The data was on a monthly basis and as the data is less, an exponential smoothing model fits the data as its very user-friendly in use. The research paper states that it is necessary to have the crime count more than 30 percent with 20 percent of acceptable accuracy.

The Arima model and the exponential smoothing techniques were used for the forecasting in this research paper. Yadav and Nath (2017) The PM10 data was used for the particular matter and for the result were evaluated by the accuracy and the MAPE value. The burge method in AR is 14.23 percent and 10.20 percent. The MAPE values for Yule-Walker are 32.72 percent and 31.13 percent. The MAPE value for exponential smoothing is 5.81 percent. As per the results, the exponential smoothing model is the final model for forecasting crime rates in India.

The crime rates in the urban environment were predicted with the help of the social factors in this research. Ingilevich and Ivanov (2018) An aim of this research is to compare different approaches to the problem of forecasting the number of crimes in different cities. The linear regression, logistic regression, and gradient boosting models were implemented. The predictive factors which are used for the models are based on the feature selection techniques to increase accuracy and avoid overfitting. Among the three of them, gradient boosting is the best model with respect to the types of crimes and the task of predicting the crime rates.

The linear regression model was used to predict future crime in Bangladesh as the aim of the author is to forecast future trends. The research paper was published in 2016 on the IEEE portal. Awal et al. (2016) The linear regression model was fitted on the dataset and, the murder rates are predicted based on the historical data. The prime crime was selected as murder and a good literature review was provided. In the methodology section the data collection, dataset description, and model evaluation were explained. The crime rates for the metropolitan areas were also forecasted. The population in Bangladesh was the one factor that was taken as the main factor to predict crime rates.

The time series analysis for the time series-based crime data was done in this research paper. Moreover, the forecasting of the crime data was additionally done to get better results. The research was published in the year 2019. Devarakonda (2019) The research literature was maintained properly with the remarkable flow of the project. The data was of the crime data of Chicago and Los Angeles. To satisfy the aim, after analyzing the crime data, a few statsmodels were built. The data was weekly time based and dicky fuller statistical test was carried out on the data. Later on ARIMA, Auto ARIMA, Holt winter's forecasting, and, the prophet method was used and forecasting was done respectively. The MAPE was calculated and for Chicago, it was 31 and for Los Angeles, it was 6.5. The models were built in a good way as to gain a nice level of forecasting.

2.4 Previous Work on The Factors Affecting The Crime Rates

Phillips and Land (2012) This research is focusing on one of the most important factors affecting crime rates which are unemployment. The research paper contains an analysis of the fluctuation of the crime rate with respect to unemployment. The analysis of the research is on the state, county, and national levels. There are patterns in the data which are consistent as unemployment increases crime rate increases. However, to get the results the C and L model was implemented and has a relationship between unemployment and crime rates.

The statistical relationship between the population and UCR crime was represented in this research paper. It was published in the year 2004. Nolan III (2004) This paper established the relationship between these two factors in the increasing size of the population. The UCR data is recognizable in two formats crime volume and crime rate. The data is for the united states and according to it, there is a positive relationship between

population size and the crime rates as well as the average crime rate.

The paper investigates the relationship between unemployment and crime rates in Ireland during the time of financial crises. The summary statistics were calculated from the dataset of the garda division in that the minimum and maximum number of the statistics were focused on the unemployment roles and the types of crimes. The regression model was implemented on the data and changes in crime rates with respect to unemployment were monitored. The results were robust and there was an elasticity of 0.5 percent in crime rates with respect to unemployment. The paper was published in the year 2016. Hargaden (2016)

The results of the previous work are not that similar to this research as there is a variation in the results, and the aim of this research mentioned above.

3 Methodology

In the research methodology, various steps regarding the project development are carried out. From data analysis to the implementation and final results, the project procedure comes under all sections and are as follows: The data gathering, pre-processing, Data cleaning, transformation, modeling, Evaluation, and results.

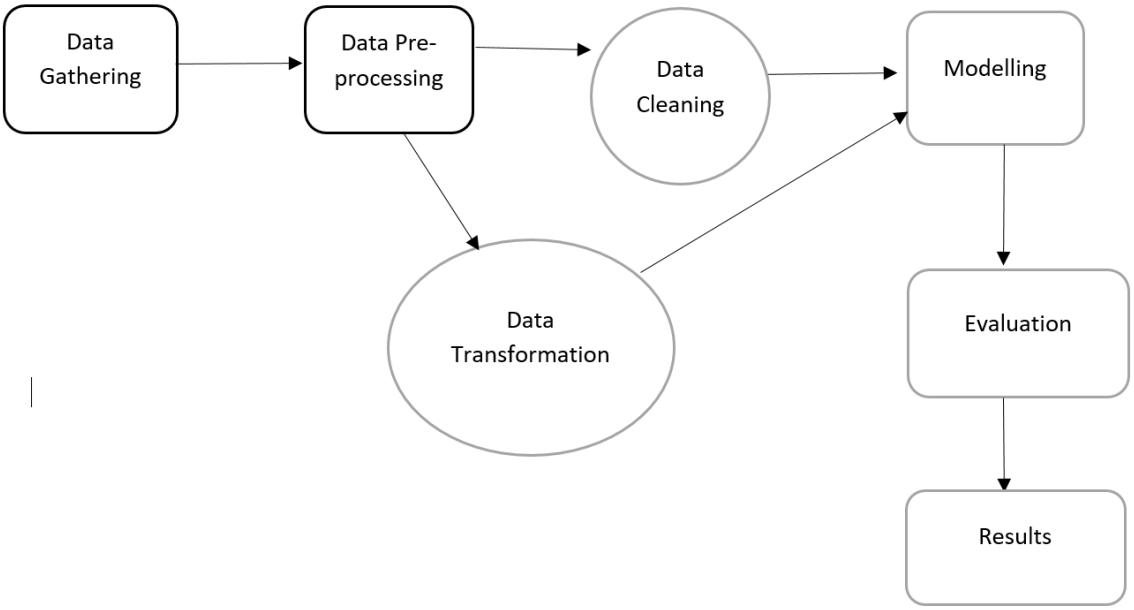

Figure 1: Project Procedure

3.1 Step 1: Data Gathering

For the crime rates in Ireland and India, the crime statistics and for India district-wise crime-related dataset was featured. Similarly in Ireland, the Ireland crime garda division dataset was used. Moreover, The world bank has its own database of crime statistics for each country. From the world bank database, the crime statistics for Ireland and India datasets were selected for implementation, meanwhile, the population and unemployment

statistics, the same database is used. The data fetched from the world bank database is in historical format and it's required performing some preprocessing on it.

For Ireland, the Ireland crime garda division dataset has columns namely, region, garda division, offense code, offense, type of offense, and crime rates in a yearly format. The dataset which is fetched from the world bank for Ireland has a date, per 100k population which is a crime per 100K population. Similarly for population and unemployment, the datasets have date and population and unemployment rate. The world bank datasets are in time series format.

For India, The district-wise crime committed dataset was used which has the columns state, year, murder, attempt to murder, district, rape, custodial rape, and so on. It has a total of 15 types of crime listed and also has the total IPC crimes and other IPC crimes. With respect to the world bank data for crime in India, it has date and crime per 100k population. however, for the population and unemployment rate, they both have the date and specific population, and unemployment rates. Similar to the Ireland dataset these datasets also are in a yearly time series format.

3.2 Step 2: Data Pre-processing

The data pre-processing is divided into two stages as shown in figure 1. The first stage in this research regarding data pre-processing is data cleaning.

In data cleaning, after importing the data into python using pandas, is there are any null values present in the dataset are checked. If there were any null values they need to transform as filling the blank spaces in the dataset. For checking the null values isnull command in pandas is used. Significantly, there are zero null values in the dataset. Similarly, the India crime-related dataset is also checked for the null values as well as the outliers. After testing there are no such issues in the dataset that has to be resolved at this stage.

The second stage in data pre-processing is Data transformation. For the garda division data and IPC crime data, an exploratory data analysis is done on these datasets and later on, for the crime, population, and, unemployment statistics for both countries, the specific datasets were fetched in data gathering. Therefore, Ireland's crime statistics has 3 datasets as historical crime data, population data, and unemployment data. Same as Ireland, country India, Indian historical crime statistics, Indian population, and Indian unemployment are 3 separate datasets used.

As per the research question, to perform the operations on Ireland's crime and Indian crime related to population and unemployment, for both countries, firstly crime and population and secondly crime and unemployment datasets are merged together. By using the .merge command in pandas merging these datasets become possible because they have a single common variable which is the date.

After merging the datasets two major datasets for both countries got available as a crime with respect to population and crime with respect to unemployment. On both datasets, an Augmented Dickey-Fuller test (ADF) is performed to check whether the data is stationary is not. The results of this test were observed and as per the results, the data is not stationary. Therefore, to make the data stationary, the logarithmic transformation is performed. Hence, the result of this transformation the data becomes stationary as its P value is less than 0.05 respectively.

3.3 Step 3: Modelling

As shown in figure 1, cleaned and transformed data is used for the modeling of the research project. On the merged datasets, a total of four machine-learning models are applied. The aim of the research is to predict the crime rate and to observe which factor among population and unemployment is affecting crime rates.

Firstly, on the merged dataset of crime rate with respect to population is transformed. after that, an Exponential smoothing model is applied to it. To observe the variations in the model results ARIMA, Sequential Neural Networks, and Linear regression are also tested. from all of them, the linear regression model is the final model according to the results.

In the same way for both countries, these four models are applied to crime with respect to the unemployment dataset.

3.4 Step 4: Evaluation

First of all, to get to know all of the datasets, the .describe and.info commands are used. After that, an exploratory data analysis is done on the datasets and the visualizations with the help of the matplotlib library.

The models which are applied to the datasets are evaluated by the model testing and measuring standards such as Accuracy, Mean, Mean absolute percentage error, Mean squared error and AIC.

3.5 Step 5: Results

Based on the nature of the models, results and test and measure standards the research is coming to the result which is Linear regression is the best-suited model to this environment. Moreover, the two most affecting factors to crime are monitored in this research, so based on the statistical and methodological results unemployment is affecting the crime rate more than the population in India. But for Ireland, there is a neutral graph of unemployment and population towards crime. The results are more distinctly explained in the evaluation section.

4 Design Specification

As mentioned in the methodology section, to satisfy the research question, machine learning methodologies are built on pre-processed datasets. This design and specification section includes the groundwork of this research, the architecture of the implementation procedure, a detailed description of the machine learning methodologies, the logic behind them, and the associated requirements for the implementation.

4.1 Groundwork of this research

As shown in figure 2, the groundwork of this research consists of the following three steps: The 1st and 2nd step is slightly different from each other but they both are connected to the final results.

Step 1 is an analysis of combined work done on the unemployment data and crime rates. The same procedure is followed for both Ireland and India. As mentioned in the

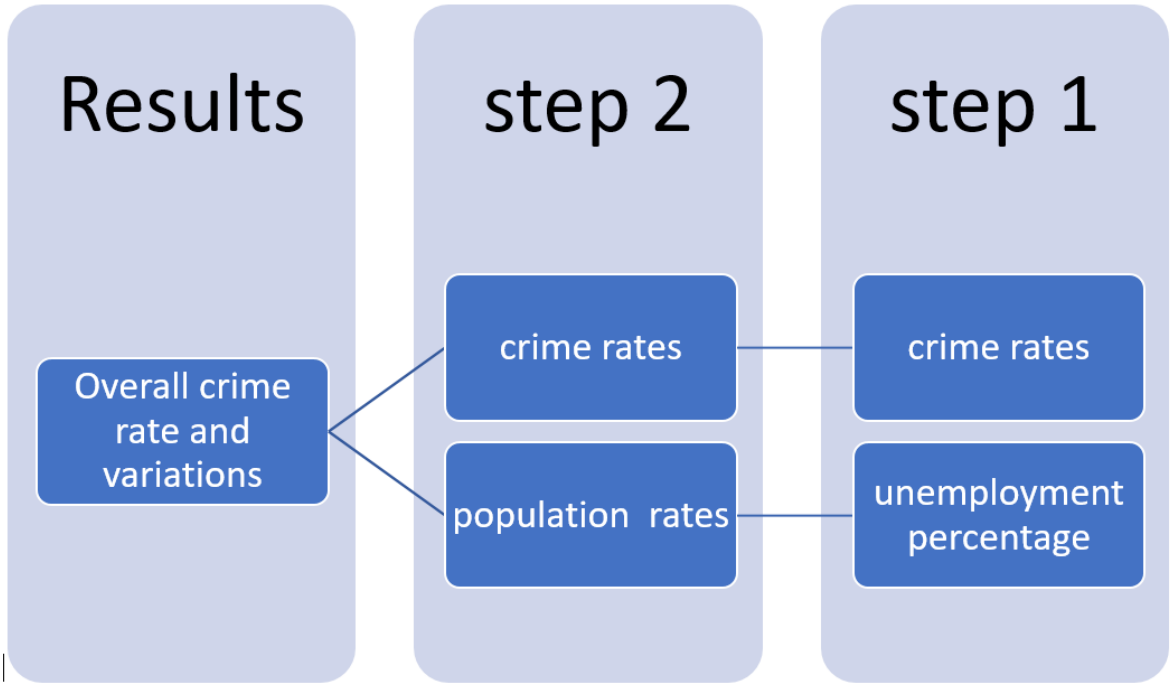

Figure 2: Project groundwork

methodology, the two datasets were merged to combine the data on unemployment and crime rates. Due to this a new CSV file is created separately of Ireland's crime rate with unemployment as well as population. The file pc.CSV is for Ireland's crime with population rates and the pc1.CSV file is for Ireland's crime rate with unemployment percentage. Similarly, for India, the ind1.CSV file is for crime rates with population rates, and the ind2.CSV file is for crime rates with unemployment percentages.

Moving towards step 2, for both countries the crime rates data is combined with the population rates and after implementing the models on the dataset prediction of crime rates with respect to the population rate is analyzed.

In step 3 which is the results, the overall crime rate and its variations with respect to population and unemployment are noted. Moreover, which factor of this two is more responsible for crime variations is also concluded.

4.2 Architecture of the implementation procedure

As per the research methodology, an architecture that is the flow of this research is taken place while working on the data. It starts with loading the datasets, and after that, the presence of null values is tested. Meanwhile, the outliers test is also carried out as a part of cleaning. There were separate datasets for population, unemployment, and crime rates, The major task was carried out to merge them as the population with crime rates and unemployment with the crime rates. The merging process was not that complex because there is a common factor of date which is available in every dataset. After combining the datasets, the Adfuller test was carried out on the dataset to gain the observation of whether the data is stationary or not. In addition to that, the transformation process is also carried out to convert the non-stationary data to stationary.

Therefore, the data cleaning and transformation process is completed and implementation is moved forward toward the model building.

4.3 Description of machine learning methodologies

Firstly, For Ireland, before the data transformation, an Exponential smoothing model is implemented on the population with crime dataset. The reason behind that is, at the primary stage the data is in a non-stationary format so for such kind of data, the exponential smoothing model is nearly suitable. After that, the transformation was done and the ARIMA model is applied to the population with crime data. Later on, the sequential neural network and linear regression models are implemented.

While dealing with the crime and unemployment dataset, firstly, the. CSV file is defined and on that particular data frame of crime and unemployment, the Sequential neural network, Exponential smoothing, and Linear regression are implemented.

For Indian crime rates, the same procedure was followed, after merging the dataset and creating a CSV file for the Indian crime rate with population, Exponential Smoothing, Sequential neural network, and Linear regression are implemented. Similarly, in the Indian crime rates with respect to unemployment, The flow of the implementation is Sequential neural network, Exponential Smoothing and Linear regression are implemented.

Due to the data structure and quantity of data, these four models are decided to be implemented on it, as not a huge amount of data is available for Ireland's crime rates.

For the associated requirements for the implementation, the required necessary libraries are imported into Jupyter notebook. The python environment is used for the implementation of the project, however, for the same pandas is a tool that is most useful. The other associated requirements like Keras and pmdarima which are used in this research are explained in the configuration manual.

5 Implementation

For implementation in this research, a few steps were followed which are mentioned in the design section. Therefore, according to the architecture proposed the procedure of implementation is carried out.

Before implementing the models, The datasets used for implementation are as follows,

The datasets are named as per convenience and aim to implement easily. The district-wise crime committed data is for Indian crime rates. The Ireland crime garda division-wise data is for Ireland crimes. Dataset ic is for Ireland's crime rate statistics, iu is for Ireland's unemployment rate statistics, and ip is for Ireland's population rate statistics. Similarly, The dataset c1 is of India's crime rate statistics, c2 is of India's population rate statistics, and c3 is of India's unemployment rate statistics. ind1 and ind2 are the datasets that are merged as Ind1 is of India's crime with respect to population and ind2 is of India's crime rate with respect to unemployment. For Ireland, The pc dataset is a merged dataset resulting in Ireland's crime with respect to population and pc1 is of Ireland's crime with respect to unemployment respectively.

5.1 Implementation on Ireland's data

Starting with Ireland's part, After importing the necessary libraries, the garda division dataset is loaded in pandas using jupyter notebook. With the help of the PD.read CSV

method. The data frame is created, and the data cleaning operations are performed on the data frame including testing the null values. To get the description and information about the dataset the.info and. describe methods are used. The described method provided the count, mean, std, minimum, and maximum values with 25, 50, and 75 percent of values for the dataset.

For exploratory data analysis, the matplotlib library is used to get to know the data and visualization. As shown in figure 3, the crime per quarter in Ireland is displayed. Moreover, figure 4 shows crime per year To demonstrate, which county region has the

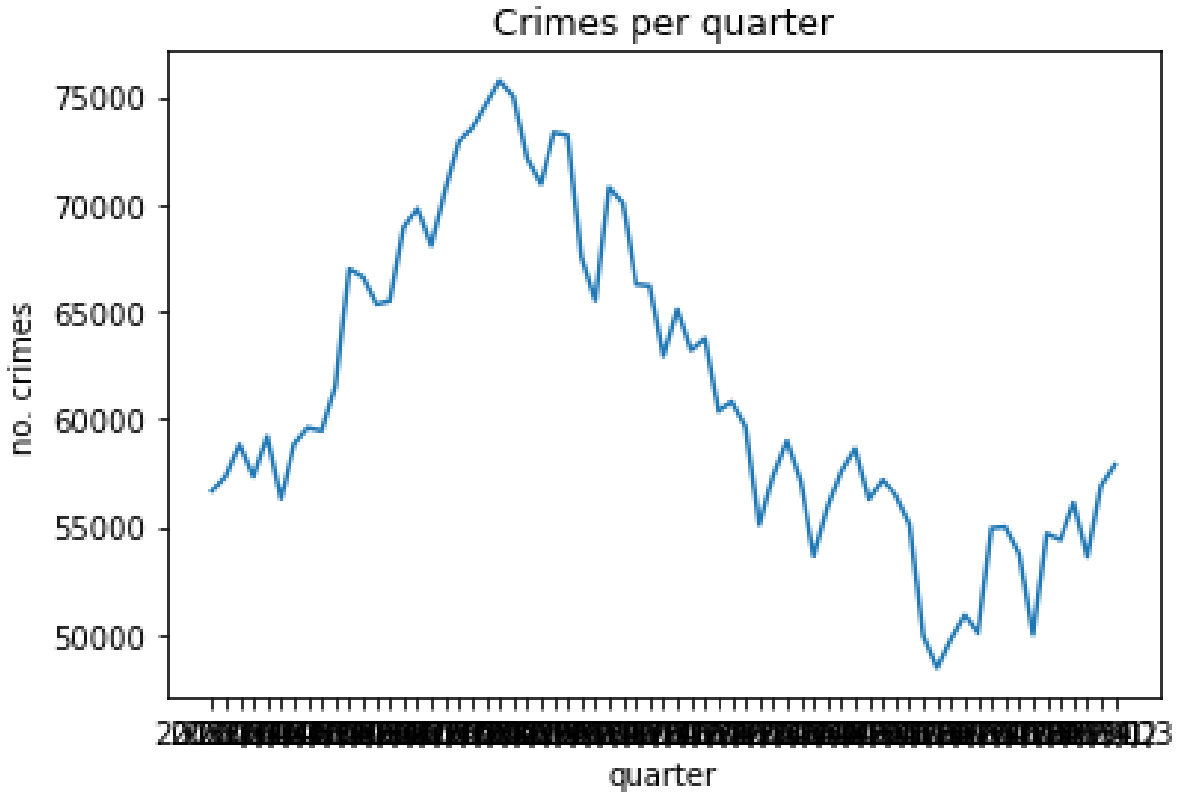

Figure 3: Crimes per quarter

most crimes is also calculated, and the Dublin metropolitan region has the most number of crimes of 1687832.

Additionally, figures 5, and 6 show the divisional crime ranking and the national offense counts.

Later on, The three data frames are created for loading three different datasets which are Ireland's population, Ireland's unemployment, and Ireland's crime rate. The crime rate dataset is in historical format. There are some null values present in those datasets so by using the dropna method the null values are removed and the datasets are ready to get merged.

The two datasets one for the crime rates and one for the population are merged together, similarly, one dataset on unemployment is merged with the dataset on crime rates by using the merge method. Therefore, the two main datasets are extracted one for Ireland's crime with Population and the second is for Ireland's crime with respect to unemployment. In this way, there is much reliability created to implement the models separately.

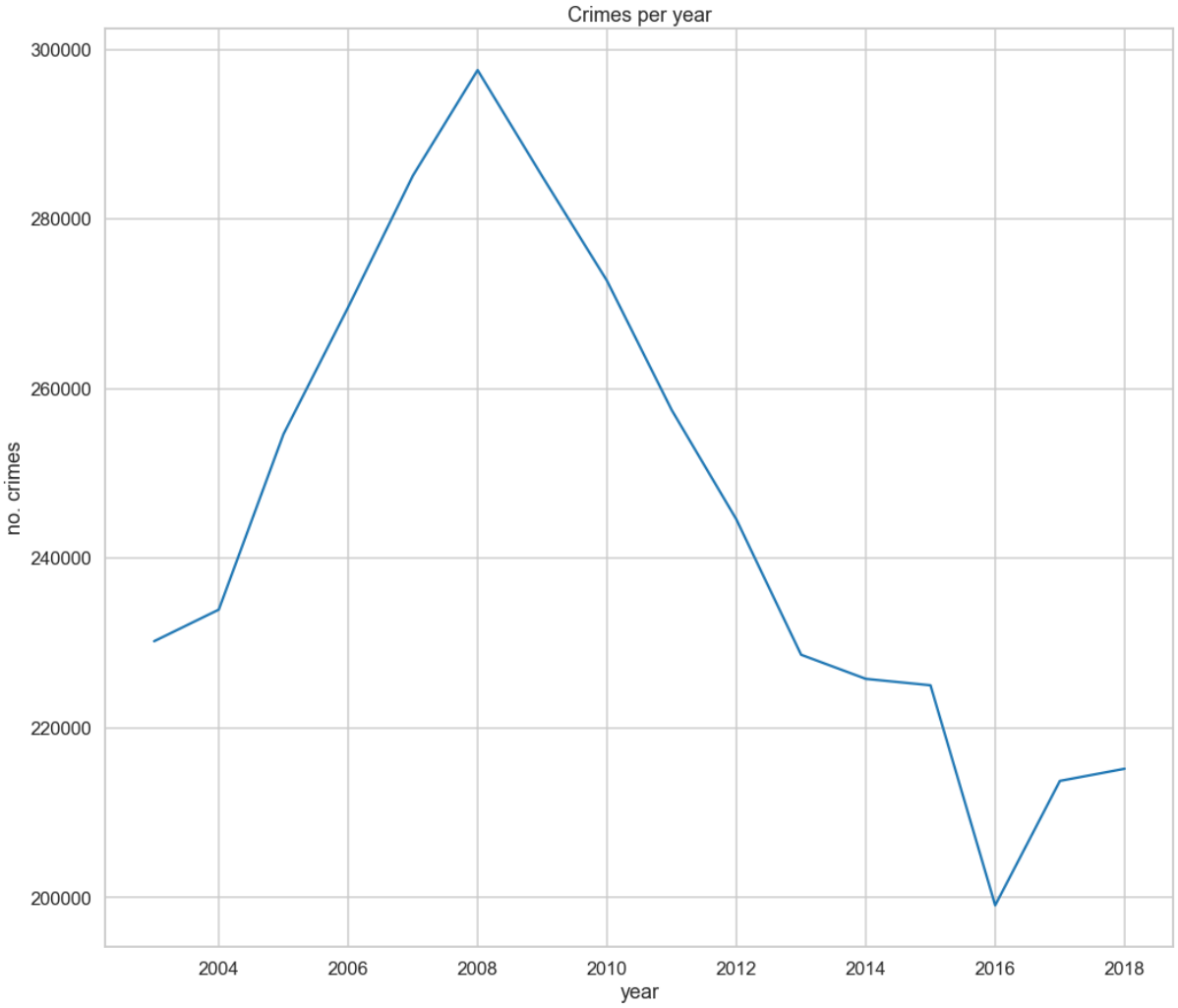

Figure 4: crime per year

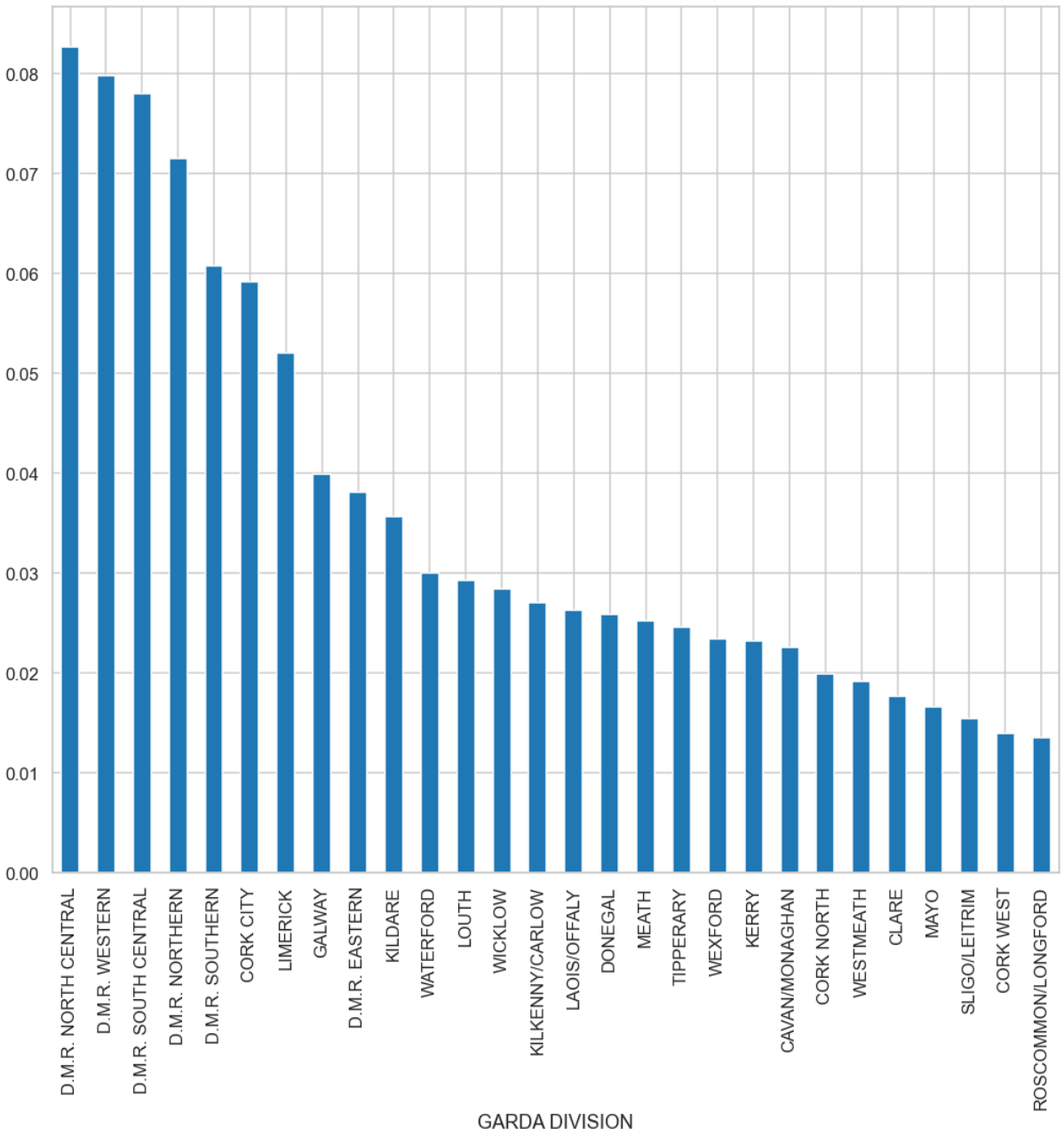

Figure 5: Divisional Crime Ranking

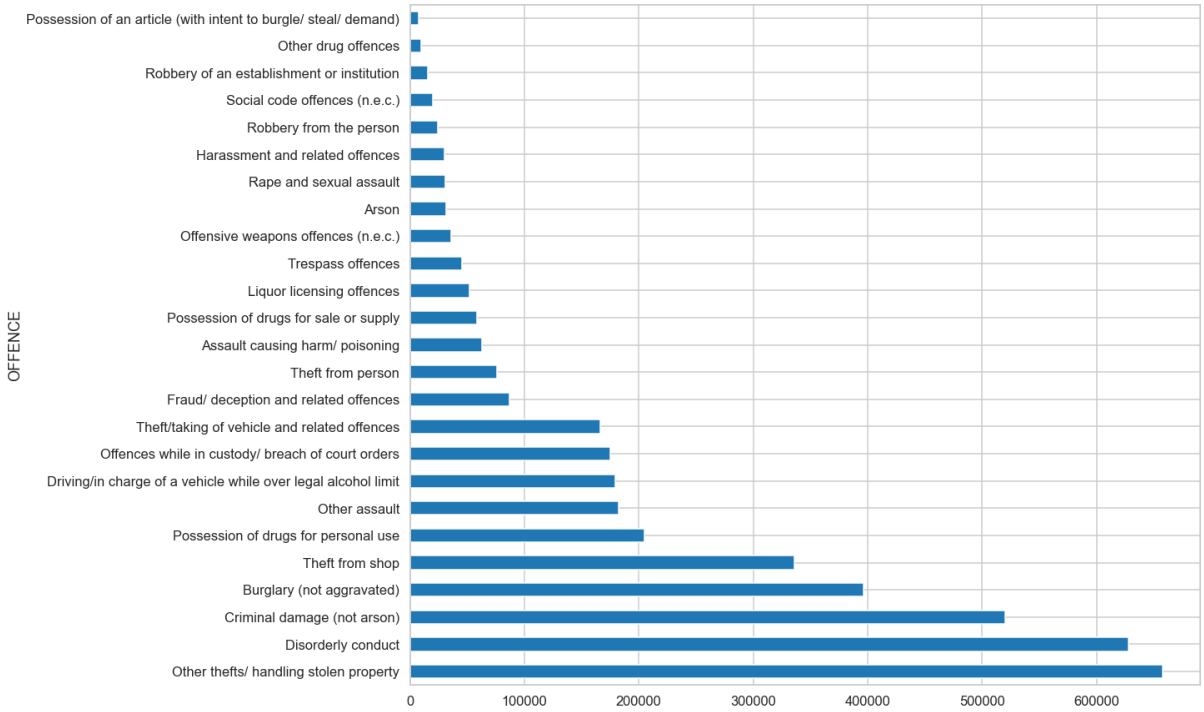

Figure 6: National Offense Counts

The dataset with Ireland's crime rate and Ireland population, on the dataset by creating a separate data frame firstly an exponential smoothing model is implemented. At this stage, the dataset is in its non-stationary format. This is the reason behind applying the exponential smoothing model on the dataset first. After getting the results, an ADfuller statistical test was carried out on the dataset. At the results of this test, the dataset was in its non-stationary format. Therefore, to make it stationary, a logarithmic transformation is performed on it. Moreover, the data is split into training data and testing data to make predictions as well as to avoid over fitting. Now as the dataset is in its stationary form, an Autoregressive Integrated Moving Average model is implemented. Along with ARIMA, the sequential neural network model is also fitted on the dataset using Keras. Finally, by using the sklearn.linear.model the Linear regression is implemented on the dataset.

Likewise the 1st dataset, the primary operations are performed on the dataset with Ireland's crime and Ireland's unemployment. Then the data split procedure is carried out in the training and testing set. The data is ready for model fit so, a sequential neural network model is implemented to it using Kerasmodel. To support the prediction, the Exponential Smoothing and Linear Regression models are also applied to the dataset. The Exponential Smoothing is implemented using the statsmodel in pandas.

Lastly, the garda station dataset is splitted into training and testing sets, and the ARIMA model is built on it. The offense code is the primary variable for this model.

5.2 Implementation on India's data

An implementation part for India's data is followed by the same procedure as Ireland's data. Firstly, all the necessary libraries are imported into jupyter notebook. By using PD. read method a CSV file namely district-wise crime committed IPC is loaded. The

dataset has various types of crimes with year, state, and district. The dataset has 9017 rows and 33 columns respectively. To get to know more about the dataset, the describe and info methods are used. As a part of data cleaning, whether the null values are present in the dataset or not is been tested. Moreover, as the dataset is big, are there any outliers present in the dataset or not is also tested by using grid and boxplot. To find the pairwise correlation in this data the corr method is used. This correlation is visualized by using a heatmap as shown in figure 7

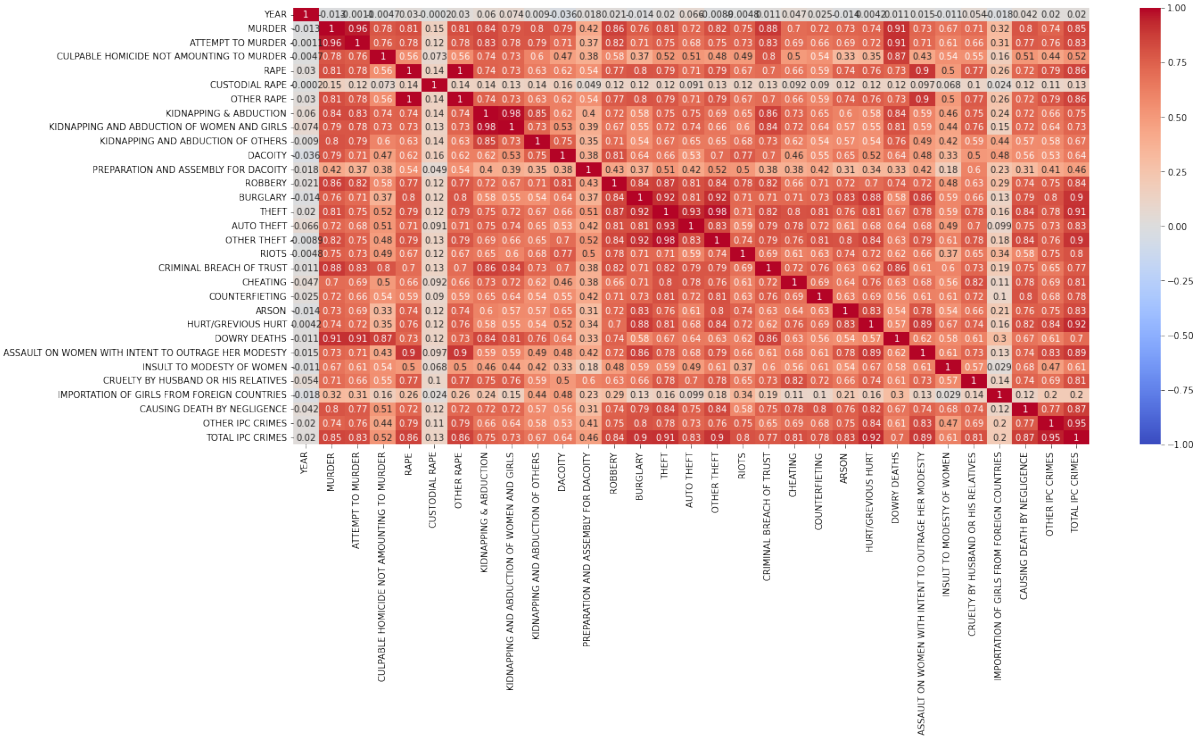

Figure 7: Pairwise Corrrelation

The same as previously done for Ireland's data, the three separate datasets are extracted from world bank data as India's crime rate, India's Population rate, and India's unemployment rate. Then India's population and India's crime rate data are merged to form a new dataset with a new data frame. The null values from this dataset are also dropped. This data frame's data is also divided into train and test split and by using statsmodel, an Exponential Smoothing is performed on it. The sequential neural network model is applied using Kerasmodel and Linear regression with the help of the sklearn.model.

As mentioned earlier, a new dataset is created for India's crime with respect to unemployment. By creating a separate data frame null values from the data which were replaced by NaN are removed and the data is divided into train and test sets. Although the data is ready for the model fit, firstly the Sequential Neural Network, secondly, Exponential Smoothing, and lastly, the Linear Regression are implemented on the dataset.

For evaluation of the model results, the AIC score, Accuracy, mean, RMSE, and MSE are a few parameters used.

6 Evaluation

The evaluation section has an overall format of the results of the models and the required findings from them. This section is firstly focused on the results of Ireland's crime rates and secondly on India's crime rates. Lastly, there is a comparison of both results in a way related to the aim of this research.

6.1 Evaluation on Ireland's data

As mentioned in an earlier section is done on the data and along with the transformation, four machine learning methodologies are applied. Therefore, before model fitting, an ADFullar test is carried out in which for a population with crime rates data and unemployment with crime rates data has a p-value of 0.0621 and 0.0851 which clearly stated that the data is not stationary. To make it stationary a logarithmic transformation is carried out and after that, the p-values are 0.0005, and 0.0010 with the number of observations used for ADF regression being 29, and 28 respectively.

In total four models are applied to Ireland's crime with population data. Figure 8 shows the results for exponential smoothing. The AIC score for this model is 6.420. The

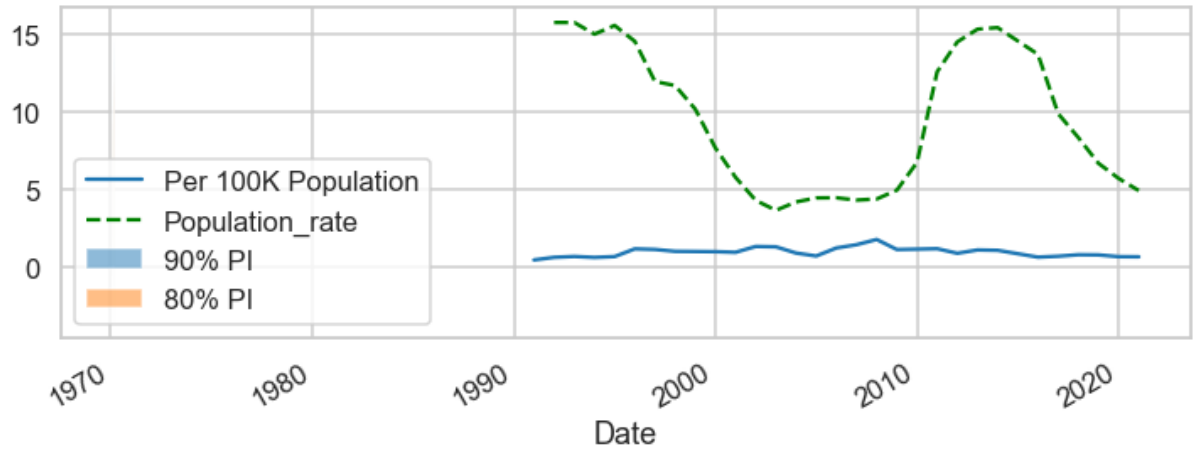

Figure 8: Crime with population exponential smoothing

BIC is 10.623. The smoothing level and initial level are 0.133 and 0.686. Figure 8 shows the results and relation in population rate and crime statistics, here per 100k population is the crime rate.

Secondly, The ARIMA model is applied to the stationary data, and for that, the best ARIMA model is (0,2,0) (0,0,0) with a model fit time of 0.472 sec. An AIC and BIC scores are 2484.8 and 2490.3 with a standard error of 746.25.

The Sequential neural network model is also used, by creating IP as an array. While dividing the data, for clarification the X train and X test variables are used as upper case letters same as test values. The Kerasmodel is used but the results are not so good, the epoch size is 10. This model is declared as a false model in this research.

Lastly, a Linear regression model is used for prediction. The coefficient of determination is 0.926. The RMSE and MSE values are also calculated as 1664, and 1290.

For the crime rates in Ireland with respect to Unemployment data, Similar to the population with crime rates the neural network is applied but things didn't go well with it. Later on, Exponential smoothing is implemented and the AIC score is 121.3 whereas

the BIC score is 125.4 with smoothing level and initial values are 0.173 and 4670.5. The

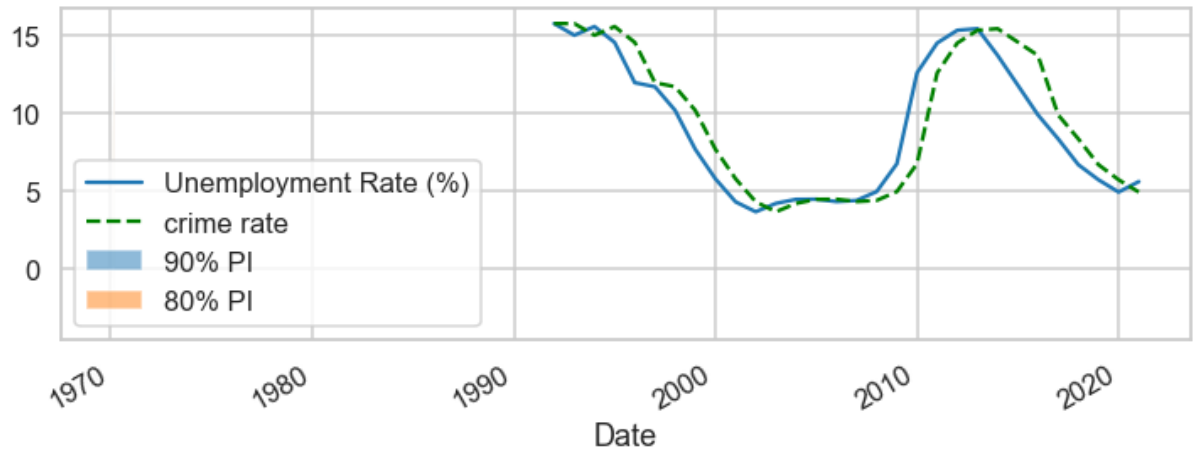

Figure 9: Crime with unemployment exponential smoothing

predictive analysis is done by using Linear regression. The MSE and RMSE values for it are 11.24 and 3.35.

Lastly, the Arima model is implemented on the garda station dataset with a target variable, an offense code. To minimize AIC the stepwise search was performed. The AIC and BIC scores are 22445 and 22488 with all the values of real imaginary modulus frequency. The best model is ARIMA(1,0,5)(0,0,0). Model fit time is around 49.04 sec and the mean is 501.6 and mean squared error is 441.8.

6.2 Evaluation on India's data

For India's crime with respect to the two factors population and unemployment, a similar process for implementation is followed as Ireland's crime data. First of all after EDA, and merging the data Indian population data with Indian crime statistics is operated. Totally the three models are implemented in this dataset. In the beginning stage, exponential smoothing is used. The AIC and BIC scores are -28.47 and -6.5 with the standard error of smoothing level and initial value are 1.180 and 50.11. The Sequential neural network is also implemented but the same as Ireland's data it is showing a bad form of results. Lastly, linear regression is applied to the data and its coefficient of determination is 0.934. The MSE and RMSE values are 0.17 and 0.13. The mean value of the prediction is 4.060.

Unemployment is a different factor than other factors affecting crime rates, whether the unemployment rate is high in both countries. Therefore, India's crime rates with unemployment data are collected in a different CSV file using data frames. The Sequential neural network is the first model for this dataset using relu and sigmoid as activations, but the results are as same as the model implemented before, Hence, the second model for this dataset is exponential smoothing with AIC and BIC score of 43.98 and 65.78. The coefficient values are 0.98 and 5.59 for the smoothing level and initial value. Moreover, the standard error values are 1.398 and 12.58. In the end, Linear Regression is implemented and the coefficient of determination value is 0.980 whereas, the MSE and RMSE values are 0.019 and 0.139 respectively.

In this research, a detailed description of the performance of the model and comparison is in the discussion section.

6.3 Comparison based on the results

Firstly, comparing the results for Ireland's crime statistics, among the four models which are implemented, the Exponential Smoothing gives a better clarification about the factors. As shown in Figures 8, and 9 the population factor is not that affecting crime rates, whereas unemployment is directly proportional to the crime rates. Where ever the unemployment is high crime rates are increased and vice versa.

Linear regression is able to provide a better quality of results as compared to the other models, based on the coefficient of determination, MSE, and RMSE values.

Lastly, in the ARIMA model which is implemented on the garda dataset, The MSE value is nearly less than the mean value so it is well treating the variable offense code for division data.

From the perspective of Indian crime statistics, the linear regression model is the best model as compared to the rest of the two. And similarly to the Irish crimes unemployment is the factor which is affecting Indian crime rates more than the population.

6.4 Discussion

Based on the results which are obtained from this research, the training and testing set of data is avoiding overfitting of the models. According to the model results, even after making the data stationary, the ARIMA model is not able to provide appropriate results. In terms of the Sequential Neural Network model, this model gave bad-quality results. The main reason behind this is that maybe there is a very less amount of crime data available to use. Additionally, the number of samples available to train this model is not equal so this might be another reason which is responsible for the model's failure. Therefore, due to the less amount of data model results are unsatisfactory. Moreover, the crime rate is such a sensitive topic to work on, however, the rest of the models and the exploratory data analysis helped in increasing good level of knowledge of this topic. And therefore it helped in answering the research question. Hence, unemployment is a factor that is affecting the crime rates most.

Phillips and Land (2012) As per the results of this research paper, this research has the same level of results that there is a bit of fluctuation in crime rates and unemployment is responsible for increasing and decreasing crime rates.

Hargaden (2016) In this research paper, the focus is on the crime type therefore the regression results are quite different from this research. Moreover, working on the social factors this research gives diversity in the results as one thing is in common which is unemployment.

Nolan III (2004) as mentioned earlier, there is a relationship shown in this research in crime rates and population size. The fact that both countries have a high count of the population but by comparing, an observation comes out that in India the population is high so unemployment is high. Therefore, there is a relationship between population and unemployment. To deny the fact that our research results are different from this research results as more than population, unemployment is affecting crime rates.

7 Conclusion

According to the evaluation of this research, to answer the research question that which factors from population and unemployment affect the crime rates in Ireland and India

is possible to use machine learning methodologies. As mentioned earlier, unemployment is a factor affecting crime rates the most compared to the population. Also, based on the results of the machine learning models and an exploratory data analysis, Linear Regression is the best model in this research with the results as shown in figure 10. Additionally, for Ireland the Dublin region, and in India the Andra Pradesh region has the most number of crime rates. But to compare the crime rates, India has more rate of crime annually than Ireland.

In conclusion of this research, the results also satisfy the assumptions stated in the introduction section and the results are beyond expectations. There are some limitations in this research, such as the data is not that big and the data was not organized concerning the sample size. The regional crime rate records are low in Ireland, which is one reason for the model's failure.

8 Future work

The results which are mentioned above and the limitations, there might be some future work to do in this research. Firstly, if possible, a good dataset with a great level of organized crime can be obtained. Secondly, some more machine learning models can be implemented on this dataset such as Recurrent Neural Network and Long Short Term Memory models. After implementing them the results might be improved. Also, a good level of tuning can be done on the Sequential Neural Network model to avoid model failure. A more effective exploratory data analysis can be done and if the data is fetched in a large amount for Ireland the Box Jenkins method can also be used.

To analyze the crime rates more significantly, can be done by selecting a single type of crime for example murder or homicide rates, and work on them can be processed. The comparison of crime rates is done in this research for two countries based on which country is safest to live in. this factor can be included in the research question. Not just depend on the time series data, by selecting one target variable every type of crime can be noticed.

Acknowledgement

I would like to thank Prof. Jorge Basilio for his guidance, suggestions, and motivation in the research project module. Completion of this research project would not be possible without his help and guidance. I would also like to thank the National College of Ireland for giving me this opportunity to work on my thesis. And lastly, I am very thankful to my father, Adv. Laxmikant Mukim for his great support and motivation. Thank you.

References

Adedokun, S. A. (2020). *Housing Price Prediction and Classification Based on Crime Occurrence using Machine Learning Algorithms: Ireland*, PhD thesis, Dublin, National College of Ireland.

Agarwal, S., Yadav, L. and Thakur, M. K. (2018). Crime prediction based on statistical models, *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pp. 1–3.

Machine Learning Models	Ireland	
	Crime Rates with Population	Crime Rates with Unemployment
ARIMA	AIC = 2484.8 BIC = 2490.3	
Exponential Smoothing	AIC = 6.420 BIC = 10.623	AIC = 121.3 BIC = 125.4
Linear Regression	COD = 0.92 MSE = 1290 RMSE = 1664	RMSE = 3.35 MSE = 11.24

Machine Learning Models	India Crime Rates with Population	Crime Rates with Unemployment
Exponential Smoothing	AIC = -28.47 BIC = -6.5	AIC = 43.98 BIC = 65.78
Linear Regression	COD = 0.93 MSE = 0.17 RMSE = 0.13 Mean = 4.060	COD = 0.98 MSE = 0.019 RMSE = 0.139

Figure 10: Results

Awal, M. A., Rabbi, J., Hossain, S. I. and Hashem, M. (2016). Using linear regression to forecast future trends in crime of bangladesh, *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, IEEE, pp. 333–338.

Aziz, R. M., Sharma, P. and Hussain, A. (2022). Machine learning algorithms for crime prediction under indian penal code, *Annals of Data Science* pp. 1–32.

Chandra, B., Gupta, M. and Gupta, M. (2008). A multivariate time series clustering approach for crime trends prediction, *2008 IEEE International Conference on Systems, Man and Cybernetics*, IEEE, pp. 892–896.

Chen, P., Yuan, H. and Shu, X. (2008). Forecasting crime using the arima model, *2008 fifth international conference on fuzzy systems and knowledge discovery*, Vol. 5, IEEE, pp. 627–630.

Devarakonda, D. S. (2019). Time series analysis and forecasting of crime data, *California State University, Sacramento* .

Gorr, W., Olligschlaeger, A. and Thompson, Y. (2003). Short-term forecasting of crime, *International Journal of Forecasting* **19**(4): 579–594.

Hargaden, E. (2016). Crime and unemployment in ireland, 2003-2016, *online] Hargaden. com. Available at: http://www. hargaden. com/enda/hargaden_crime. pdf [Accessed 21 Nov. 2017]* .

Ingilevich, V. and Ivanov, S. (2018). Crime rate prediction in the urban environment using social factors, *Procedia Computer Science* **136**: 472–478.

Mahendra, C., Babu, G. N., Chandra, G. B. N., Avinash, A. and Aditya, Y. (2010). Crime rate prediction, *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)* **2**.

McKenna, C., Kelleher, M. J. and Corcoran, P. (1997). Suicide, homicide and crime in ireland: What are the relationships?, *Archives of Suicide Research* **3**(1): 53–64.

Nolan III, J. J. (2004). Establishing the statistical relationship between population size and ucr crime rate: Its impact and implications, *Journal of Criminal Justice* **32**(6): 547–555.

O'Sullivan, E. and O'Donnell, I. (2003). Imprisonment and the crime rate in ireland, *Vol. XX, No. XX, Issue, Year* .

Phillips, J. and Land, K. C. (2012). The link between unemployment and crime rate fluctuations: An analysis at the county, state, and national levels, *Social science research* **41**(3): 681–694.

Priya, S. S. and Gupta, L. (2015). Predicting the future in time series using auto regressive linear regression modeling, *2015 Twelfth International Conference on Wireless and Optical Communications Networks (WOCN)*, IEEE, pp. 1–4.

Vaidya, O., Mitra, S., Kumbhar, R., Chavan, S. and Patil, R. (2018). Comprehensive comparative analysis of methods for crime rate prediction, *Int. Res. J. Eng. Technol* **5**(02): 715–718.

Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R. and Yadav, N. (2017). Crime pattern detection, analysis prediction, *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, Vol. 1, pp. 225–230.

Yadav, V. and Nath, S. (2017). Forecasting of pm 10 using autoregressive models and exponential smoothing technique, *Asian Journal of Water, Environment and Pollution* **14**(4): 109–113.