A Machine Learning Approach to Predicting Gross Domestic Product

MSc Research Project
Data Analytics

Ronan Flannery
Student ID: x19113269

School of Computing
National College of Ireland

Supervisor: Manaz Kaleel

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ronan Flannery
Student ID:	x19113269
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Manaz Kaleel
Submission Due Date:	17/08/2020
Project Title:	A Machine Learning Approach to Predicting Gross Domestic Product
Word Count:	8095
Page Count:	21

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:	16th August 2020

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 Machine Learning Approach to Predicting Gross Domestic Product

Ronan Flannery

x19113269

Abstract

There is an ever-increasing understanding of the benefits of using Machine Learning in the prediction of economic performance when compared to more traditional statistical methods. Machine Learning models can provide more accurate and timely forecasts of economic performance, which are very valuable for economists and policy advisors when making important decisions related to the future economic outlook. Gross Domestic Product (GDP) is a main indicator used in the evaluation of the performance of an economy. Traditionally, GDP was forecast using data available on a quarterly basis, such as imports, exports and government spending. Delays and issues with inaccuracy can arise with this quarterly data however, and the use of Machine Learning models has facilitated the evaluation of a wider range of economic indicators, available with more timely data, for predicting GDP. This research project creates Machine Learning models using Multilayer Perceptron - Artificial Neural Networks and Random Forests for predicting GDP. These methods have previously been successfully used for economic prediction models and this research project will build on the previous work in this area by applying the models to economic performance variables that impact GDP, for example Consumer Confidence and Business Confidence, Household Spending, Unemployment Rates, Interest Rates and Exchange Rates. The models built in this research project are evaluated in terms of accuracy and other statistical measures, and the results are promising in terms of the predicting GDP in Ireland and other countries.

1   Introduction

Gross Domestic Product (GDP) is an important measure of the economic performance of a country, in terms of both the overall market value of all goods, and the services produced by that country during the year. This statistical measure is valuable to many parties, including economists, politicians, analysts, banks and policy advisors when making important decisions with regards to the future economy. The prediction of GDP has always been important with regards to identifying possible trends and future impacts to the economy. Various models have been created for predicting GDP, however in recent years there has been a growing understanding of the importance of, and use of Machine Learning based models for predicting economic performance. Machine Learning has facilitated the identification of trends and patterns in greater depth than traditional statistical models and the power of Machine Learning algorithms and models has resulted in a greater accuracy of predictions. Many different Machine Learning models have been created for predicting GDP and various indicators of economic growth have been

examined by these models in the prediction of economic performance. A thorough examination of related work in the area found that Artificial Neural Networks (ANN) and Random Forests were positive in terms of their GDP predicting capabilities. Models have been created previously with an emphasis on specific countries and the indicators relevant to those countries. For example, Alamsyah and Permana (2018) created Multilayer Perceptron Artificial Neural Network models for predicting economic performance in Indonesia. Martin (2019) created Random Forests models for predicting GDP in South Africa and Behrens et al. (2018) also used Random Forests to examine the forecasting of macroeconomic data in Germany. Richardson et al (2018) used many different Machine Learning methods to predict GDP in New Zealand.

This research project creates Multilayer Perceptron ANN and Random Forest models for examining the predictive abilities of these Machine Learning models in terms of their accuracy at predicting GDP in Ireland and a number of other countries. The work conducted as part of this research project will add to the body of work in the area of creating Machine Learning models to predict GDP and economic performance. Economic Indicator data specific to Ireland as well as number of other countries is analysed, and the output of the models will provide an insight into the use of Machine Learning for predicting GDP in Ireland. This information will be useful for economic analysts and data analysts looking at models relevant to Ireland and the selected countries.

The Research Question that this research paper looks to examine is as follows: **'To what extent can machine learning algorithms predict Gross Domestic Product in Ireland and other countries based on various economic factors?** The objectives of this research are:

- Build two machine learning models (using Artificial Neural Networks and Random Forests) for predicting GDP

- Use selected economic indicators for predicting GDP with the models

- To evaluate the performance and accuracy of the models using selected statistical measures (RMSE, MAE, R2)

- To compare the GDP predictions for Ireland and other countries to the actual GDP numbers and analyse the potential value of the models

This research is done with the aim of building on previous work completed in the field of creating a machine learning approach to predicting GDP. The remaining sections of this research paper are structured as follows. The next section of the paper covers an analysis of the **Related Work** in the area of using Machine Learning for predicting economic performance. The motivation behind the work undertaken in this research project is covered, including the reasons why the prediction of economic performance is important and how GDP can be predicted using selected economic indicator data. An analysis of previous research that involves creating Machine Learning models for predicting GDP and also suitable evaluation metrics is also discussed in this section. In Section 3, the **Research Methodology** will be covered in detail, including the use and application of the CRISP-DM methodology as part of this project. Section 4 of the report will show the **Design Specification** which covers an overview of the frameworks that were created and used as part of this research project. Section 5 covers the **Implementation** of the models that are used to answer the research question and objectives of this project. Next, in Section 6 a thorough **Evaluation** of the results and main findings of the research

project will be provided. An analysis of the results will also be provided in this section along with a critical discussion of strengths and weaknesses of the implemented models. Finally, Section 7 will contain the Concluding remarks of this report and some possible Future Work that may be performed as continuation of this research project.

2 Related Work

This Research Project investigates the use of selected Machine Learning methods to predict economic performance. Gross Domestic Product (GDP) is the measure of economic performance predicted in this work. A Multi-Layer Perceptron (MLP) Artificial Neural Network is created to predict GDP from selected countries, and a Random Forest model is also created for predictive comparative purposes with the Artificial Neural Network. Selected economic indicator data is used as input to the models in order to make the GDP predictions. The data is taken from various countries, including Ireland, Germany, Sweden and the United Kingdom. Previous work has been conducted in relation to the prediction of economic performance and GDP, using both empirical methods and Machine Learning. In this section of the report, previous work in this area is analysed.

2.1 Economic Performance - Gross Domestic Product

Whilst there are various measures of the performance of an economy, Gross Domestic Product is one of the most important. GDP is a measure of the total value of goods and services produced in an economy over a time period, usually quarterly or yearly. GDP is a useful statistic for comparison and evaluation of the economic performance of countries and continents, and economic analysts are frequently considering improved methods and models for predicting GDP (Vrbka; 2016). The official calculation for GDP involves the use of variables including private consumption, government spending, government investment, imports and exports. One issue around the area of economic performance prediction is that often the data used for the predictions is incomplete or inaccurate (Richardson et al.; 2018). This is often due to delays that can occur with the publication of monthly and quarterly statistics in countries. As a result, the publication of economic performance measures such as GDP can be quite delayed, which in turn impacts the decisions made by economists and government policy makers.

These issues have led to a movement towards Machine Learning methods and models for predicting economic conditions, as part of the process of overcoming the challenges with traditional economic prediction methods. It has also led to the consideration of other variables in the prediction of GDP and the growth of economic 'nowcasting', where data is analysed in real-time or near real-time using Machine Learning models. Vrbka (2016) discusses the importance of Artificial Intelligence amongst the methods that have increased in importance due to their superior predictive power over conventional methods. The importance of accurate and timely economic predictions is also considered by Nik et al. (2016). The impact and disruption from financial crises can be severe. In their work, Nik et al. (2016) analyse the data from various countries that suffered a financial crisis and build an Artificial Neural Network model that could be used to predict the probability of future financial crises. Their work highlights some economic variables that are important for consideration with regards to the prediction of economic performance. Wu et al. (2019) also discuss the importance of economic forecasting, monitoring and early warning as part of macroeconomic decision making. In their work, they discuss

the model they have created for predicting GDP in China, and it is clear that such a model would be useful as part of any economic decision making process with regards to highlighting potential disruptions and issues in the economy. This section of the report discussed the reasons and benefits behind the prediction of economic performance, as well as some of the methods used in the prediction of GDP. The next section of the report looks at the selected economic indicators used for predicting Gross Domestic Product.

2.2 Predicting Gross Domestic Product with selected economic performance data

GDP is composed of various economic factors, including government and private expenditure along with imports and exports, however many other economic variables also impact on GDP. The data behind traditional economic indicators are often published quarterly, however as mentioned previously, this data is prone to error and incompletion. Other economic variable data is often available more frequently, for example monthly rather than quarterly, and recent work in the field of machine learning and economic performance prediction is examining the use of this data as part of creating models for economic predictions. In their paper, Stundziene (2013) examined many different monthly economic indicators and created models with this data for predicting GDP in Lithuania. It was found that turnover of retail trade, goods loading figures and passenger numbers at the airports provided data that could be used for forecasting GDP in Lithuania. The study provided an interesting insight into the fact that non-traditional variables can also be used for predicting GDP. The benefit of using these variables is that the data is available much more frequently than traditional GDP predictors. It is important to note that the impact of particular economic variables in predicting GDP can vary from country to country. Kuosmanen and Vataja (2017) performed an analysis on financial variables, both nominal and real in relation to their use in predicting economic performance. It was found in this work that the predictive ability of particular variables can be impacted during periods of economic uncertainty.

As part of the work carried out in this Research Project, certain economic indicators were selected for use in the prediction of GDP. The economic variables used in this project include: Various Business and Consumer Confidence Indicators, Imports, Exports, Exchange Rates, Interest Rates, Unemployment Rates and Political Factors. These factors were selected as they are a combination of traditional economic indicators, as well as monthly indicators that were considered to be relevant to the prediction of GDP in both Ireland and other countries. In their paper, Divya and Devi (2014) analysed various different economic indicators and their impact on GDP. An ANOVA and correlation analysis was performed on the indicators as part of the study and it was found that exchange rates were an important indicator in the prediction of GDP. Fiscal Deficit, Balance of Payments, Inflation and Foreign Exchange Reserves were also analysed as part of the study. Jiang et al. (2017) similarly found that indicator data with a monthly frequency was very useful in predicting GDP in China, particularly when used as part of a process that involves both dynamic predictors and mixed-frequency data, and when compared to traditional methods.

This section of the literature review covered the analysis and selection of economic variables as part of making economic performance predictions. In the next section, the report will look at the use of Machine Learning in the prediction of economic performance.

2.3 The use of Artificial Intelligence and Machine Learning in the Prediction of Economic Performance

Recent years have seen an increase in the use of Machine Learning models and algorithms for predicting GDP and economic performance. Machine Learning models are proving to often perform more accurately than traditional methods. In his paper, Kouziokas (2017) looked at the timeseries prediction of GDP and proposed a model based on a Feedforward Multilayer Perceptron that provided accurate predictions of GDP. The proposed model could be valuable as part of a financial management process. Cogoljević et al. (2018) created an Artificial Neural Network model to predict GDP based on energy resources. They compared their machine learning approach to a back-propagation algorithm and found, based on the results, that the Artificial Neural Network provided an improvement in accuracy. Richardson et al. (2018) investigate various different machine learning algorithms and evaluate their performance in predicting GDP in New Zealand. They look at the accuracy of the nowcasts provided by these Machine Learning models. Various economic indicator data is examined as part of the work carried out in the paper. Macroeconomic variables, confidence surveys and general domestic activity indicators are included. The study is useful as many different Machine Learning model are created, for example Neural Networks, Support Vector Machines, Boosted Trees and K-Nearest Neighbour models, which provide a good insight into the use of Machine Learning in the prediction of Economic Performance. Machine Learning has also been used to analyse economic data as part of gaining a retrospective insight into previous major financial occurrences, such as the 2008 Global economic crisis. Yu et al. (2019) use Machine Learning to examine the factors that influenced the global crash in 2008-2009.

Artificial Neural Networks have been found to be valuable for creating models for predicting economic growth. Alamsyah and Permana (2018) created a Multilayer Perceptron model for predicting economic growth based on macroeconomic indicators. The data used in their study is for Indonesia and is based on 37 years of data. A model with a 5-11-1 layer configuration was used and it provided an accuracy of over 95%. The economic indicator data used by Alamsyah and Permana included data on total investments, government expenditure, exports and imports. It is mentioned that future work in this area could look at additional economic indicator variables that effect economic growth. Vrbka (2016) also created a Multilayer Perceptron model for predicting future GDP, as well as a Radial Basic Function Neural Network. GDP related timeseries data was analysed, and the models interestingly predicted a drop in eurozone GDP over the coming years. Stevanovic et al. (2018) created an Artificial Neural Network model for predicting GDP based on electricity consumption. Their ANN models used both the extreme learning method and the back-propagation algorithm. The models created were positive in terms of their GDP prediction abilities using electricity consumption as an economic indicator variable. Jahn (2018) created a Single Hidden Layer Feedforward Artificial Neural Network for predicting GDP. Data was analysed for a number of European countries over a 20 year period and it was found that the ANN model produced much better results than a linear model in terms of prediction accuracy.

A thorough analysis of the Related Work has also found that Random Forest models have proved to be useful for predicting economic related performance. Martin (2019) used Random Forests along with other Machine Learning models to predict GDP in South Africa one quarter into the future. Fifteen different economic variables were used in the study including household income, consumer price index, gold price, oil price and

average wage. A process of creating multiple trees by bootstrapping the data, in order to improve the accuracy of the predictions is discussed. It was found that the Random Forest model performed strongly along with an Elastic-net model. Basuchoudhary et al. (2017) examined a number of algorithms for predicting economic growth, including Random Forests. The Random Forest model performed well in terms of its predictive qualities for predicting economic growth. Benefits of using tree-based algorithms were found, including their usefulness when multiple economic indicator variables affect economic growth in various countries differently. Overall, the Random Forest method was found to be a viable model for predicting economic growth, particularly in this instance where a large selection of economic indicators were examined. Random Forests were also used by Liang et al. (2020) for predicting GDP in China based on spatial distribution information. The Random Forest models performed better than traditional regression models, returning a higher accuracy level. Behrens et al. (2018) use multivariate Random Forests to determine the value of macroeconomic forecast data. The links between prediction and forecast errors were examined and the significance of the results were assessed. GDP related data and price index data from Germany was used in this work, and in concluding, the authors mention that it would be useful apply their work to macroeconomic data from other countries.

As part of the work carried out by this Research Project, data from a number of countries, including Germany, is examined using both ANN MLP and Random Forest models. Related work in the area of the use of Machine Learning in predicting economic performance was analysed in this section of the report. In the next section of the literature review, related work in the area of the evaluation of Machine Learning models in terms of their accuracy and prediction of economic performance will be reviewed.

2.4 Evaluating the Performance of Machine Learning Models in the prediction of Economic Growth

This research project created Multilayer Perceptron and Random Forest machine learning models for predicting GDP. The objectives of the research are to build and evaluate the models' predictive abilities based on the economic indicator variables selected and to determine the extent that these models can be used to predict GDP in Ireland and other countries. The models built in this research project are evaluated based on accuracy and some selected statistical based measures, including Root Mean Square Error (RMSE), Co-efficient of Determination (R2), Mean Absolute Error (Random Forest Model) and Mean Square Error (MLP Model). These statistical measures are part of the standard measures used in the evaluation of Machine Learning algorithms. Kordanuli et al (2017) created Artificial Neural Network models using extreme learning machine and back propagation for forecasting GDP. The predictive accuracy of the models that were created were evaluated using both the Root Means Square Error (RMSE) and Coefficient of Determination (R2) statistical measures. Tumer and Akkus (2018) also created an Artificial Neural Network for predicting GDP and evaluated it using Root Mean Square Error. Martin (2019) used Root Mean Square Error for evaluating the Random Forest models created for predicting GDP in South Africa and Behrens et al. (2018) used both Root Mean Square Error and Root Mean Absolute Error for evaluating their Random Forest models. This section of the Report has covered a review of the existing work related to GDP prediction and the use and evaluation of Machine Learning models as part of the prediction of economic performance.

3   Methodology

This research project utilised the Artificial Neural Networks and Random Forest machine learning methods for creating models used to predict GDP in Ireland and a number of other countries, based on economic indicator data gathered from the OECD (Organisation for Economic Co-operation and Development) data repository (OECD; 2020). The data mining methodology used in this research project is the Cross-Industry Standard Process for Data Mining (CRISP-DM). The steps in the CRISP-DM methodology provided the framework for the implementation of this research project and the basis for meeting the objectives of the project.

3.1   Business and Data Understanding

There has been an increase in the understanding of the benefits of using machine learning models for predicting economic performance. Previous work has focused on examining different machine learning models and their ability to predict GDP in various countries. After a thorough review of the related literature in this area, Multilayer Perceptron Artificial Neural Networks and Random Forest methods were selected for creating models to predict GDP in Ireland and other selected countries. After completing the business understanding step, the research question and objectives of this project were defined. The data for the economic indicators in this project are taken from the OECD data repository [1]. This data repository contains a wide variety of economic data, across many different areas. The data is freely available for download in various formats, and for the purposes of this research project the data was downloaded in csv format. The data was downloaded in individual files for a number of indicators including Current and Constant GDP, Unemployment Rate, Exchange Rate (Base Currency to USD), Long Term and Short Term Interest Rates, Government Expenditure, Private Consumption, Exports, Imports, Consumer Confidence Indicators and Business Confidence Indicators. The GDP figures are in the USD currency and most of the other indicators are in percentage format. Political data was also sourced [2] providing dates of when national elections were held in the various countries, selected as part of this research project. During the feature engineering part of the project, this election data was used to create a 'Political' variable indicating months when elections were held. This would be used as an additional indicator to determine if there was any additional impact on GDP during potential periods of political uncertainty or change of government.

The data from the OECD repository was gathered for ten countries (Ireland and nine other countries selected based on widespread availability of economic data over a prolonged period of time), for a period of 20 years (January 2000 – December 2019). Additional data for these countries was gathered for the period of January 2020 – May 2020, which was later used to perform further model testing and predictions. The files downloaded from the OECD data repository were of varying sizes depending on the data contained within the files. For example, the Business and Consumer Confidence data file had over 55,000 rows, the imports and exports file had 14,500 rows and the GDP data file had 1600 rows.

---

[1] https://stats.oecd.org/

[2] http://www.parties-and-elections.eu/countries.html, http://elections.uwa.edu.au/index.lasso and https://en.wikipedia.org/wiki/Elections_in_South_Korea

3.2 Preparation of the Data

An exploratory analysis of the data was carried out after all the data was sourced. Data was gathered for ten countries: Australia, Austria, Belgium, Denmark, Germany, Finland, Ireland, Korea, United Kingdom and Sweden. The relevant fields were extracted from the individual files and the data was combined into a master file. Figure 1 shows the Data Map for the project data and provides an overview of the source of the data used for the final GDP Prediction Dataset, as well as the fields contained in the Prediction Dataset. A further exploratory data analysis was carried out and it was found that there were some missing Business Confidence data for Australia and some missing Consumer and Business Confidence data for Korea. As a result, the decision was made to drop these two countries from the research, as the missing data would cause issues when training the models. This left a total of eight countries for analysis in the research. It was also found that there was one short-term interest rate value missing for Sweden from September 2001. As the values for this Interest Rate were almost identical from the months before and after September 2001 (3.7 and 3.71) it was decided to impute a value of 3.7 (mean for Aug and Oct 2001) for the short-term interest rate for Sweden for September 2001. No other relevant data was found to be missing from the data set. The GDP data sourced from the OECD data repository contained GDP figures in both Constant and Current GDP. As Constant GDP is inflation-adjusted it was determined that this would be more suitable to use for the GDP Predictor Variable in the research, as the inflation-adjusted value would provide a more accurate measure of the actual changes in GDP output. The Current GDP values (non-inflation adjusted) for previous quarter were also kept in the dataset and used as a variable as part of the feature engineering carried out. Other pre-processing that was carried out on the data included scaling of the independent variables (to standardise the values) and dropping of some columns that were not required for processing the data (for example Subject, Measure, Unit Code, Flag Code).

Figure 1: Data Map Overview
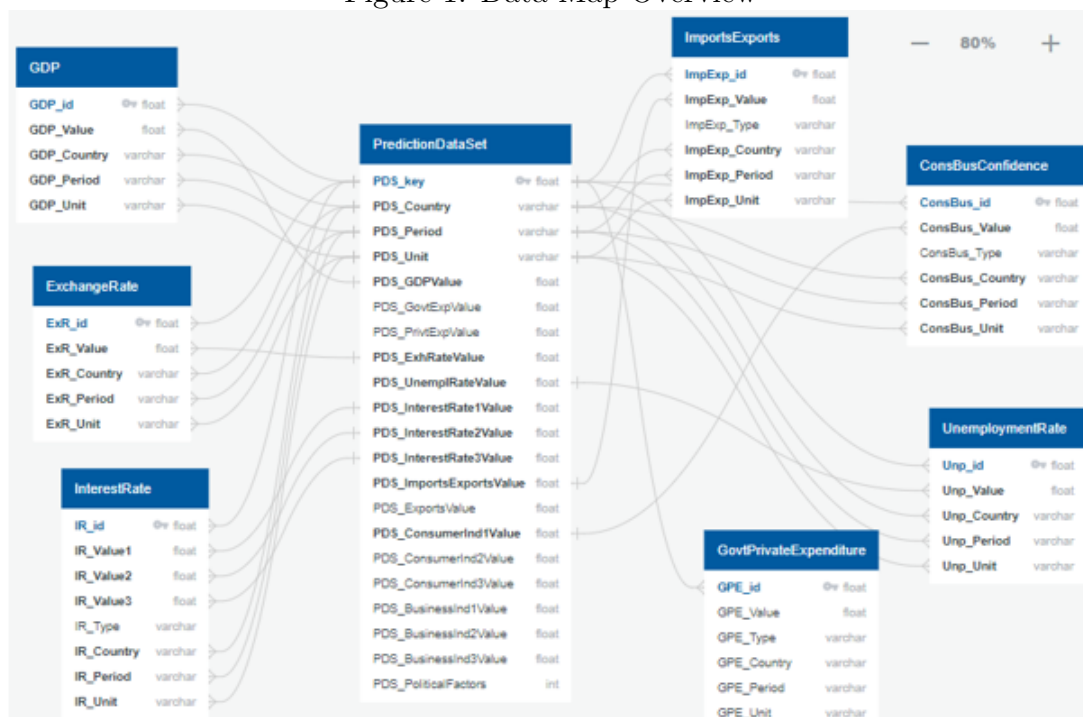
3.3 Modeling

In this research project, models for predicting GDP were created using both Multilayer Perceptron Artificial Neural Networks and Random Forests. The models were created and run in Google Colab with the Python coding language. Google Colab has benefits for creating Neural Network models as everything is run online without having to install or download any packages on a local machine. There is also less concern regarding resource constraints when working with Google Colab and when running models compared to on a local machine.

Multilayer Perceptron Artificial Neural Network:

Artificial Neural networks can be useful for making predictions with non-linear independent variables. ANNs are easily adjustable based on the learning done by the model on the initial training data. They are also self-organised and have the ability to perform multiple real-time operations. Alamsyah and Permana (2018) discuss the benefits of ANNs in their paper. The Multilayer Perceptron trains using a backpropagation algorithm which is based on a gradient method. The training process is composed of an initialisation of the model followed by Forward Propagation, backpropagation and weight updates. Forward propagation is where the input is passed through the neural network and an output is calculated. The calculated output is compared to the actual output and a loss value is then calculated. This loss error is back propagated through the layers of the network and it allows updated weights to be calculated which can then be applied to the inputs and the process is repeated until a convergence is reached. As mentioned previously, Multilayer perceptron models have proven to be useful for predicting economic performance and GDP. In this research project, the ANN model has an input layer which takes in the various economic indicator input variables such as Unemployment rate, Interest Rates, Consumer and Business Confidence Indicators, Import and Exports. It also contains a number of hidden layers and an output layer.

Random Forest:

Random Forest models have also been successful when used to predict GDP previously. Martin (2019) used Random Forests for predicting GDP in South Africa. The random forest model uses a number of trees for making predictions, and from the input training data, samples are repeatedly taken from the input, using a bootstrapping process, to train the trees. The bootstrapping process allows multiple samples to be used from the same input training data. The Random Forest model uses a process to try and ensure that the output from each tree in the forest is uncorrelated with the other trees. It does this by using the bootstrapped random samples from the input training data. The number of estimators (trees) in the Random Forest model is a parameter that can be configured as part of the set up of the model. Different numbers of estimators can be tested in the model and an optimisation can be carried out to find a suitable number of trees to be used in the forest. Once the Random Forest model is run, each of the individual trees in the Random Forest produces a prediction and the model then takes the average or mean of each of the individual predictions to make an overall prediction result. The number of trees in the Random Forest has an impact on the model. There is a balance between having enough trees to make a good prediction and having too many trees which can lead to increased computational time whilst only providing minimal additional improvement to the final output prediction. Therefore, it is important to perform an optimisation of the number of estimators used in the model as part of the work done when creating the model. In the case of both the Multilayer Perceptron ANN and Random Forest Models

used in this project, the input data is split into a 70% - 15% - 15% ratio for Training, Validation and Test datasets. After the models are created, a process of evaluating and tuning of the models is conducted and any further tuning of the parameters in the models can be performed.

3.4   Evaluation

The output of the models created is evaluated for its performance and to determine the extent that the objectives of this research project have been met. Along with the accuracy of the models, a number of statistical measures will be used to evaluate the models. The statistical measures used in this research project are Root Mean Square Error (RMSE), Co-efficient of Determination (R2), Mean Absolute Error (Random Forest Model) and Mean Square Error (MLP Model).

- Root Mean Square Error (RMSE) is a measure of the square root of the mean square error. It is a standard metric for measuring the error in machine learning models.

- The mean square error (MSE) shows the difference between the predicted and actual values of the data points squared.

- The Co-efficient of determination (R2) shows how well the independent variables can predict the variance in the dependent variable, in this case the GDP. It is an indicator of how well the model fits the data.

- Mean Absolute Error (MAE) indicates an average value of difference between the predicted values and the actual values.

Some further evaluations are also carried out in the research project by comparing the predicted GDP values from the Artificial Neural Network model for 2020 to the actual GDP values for the first quarter of 2020. The results of the evaluation are discussed in detail in Section 6 below.

3.5   Deployment

In the deployment phase, the results and findings of the work conducted as part of this research project are analysed and compiled into the final project report. The Multilayer Perceptron and Random Forest models created in this project are evaluated for their ability to predict GDP in Ireland and other selected countries and the possible future work that can be done in this area is defined in this phase.

4   Design Specification

This research project uses the Python programming language, run in the Google Colab environment to create the models used in the project. The models are created based on Multilayer Perceptron Artificial Neural Networks and Random Forests algorithms. The data used in this project is downloaded in csv format from the OECD data repository and pre-processing of the data is performed in both Microsoft Excel and in Python. The data is fed into models created in Python, and the models are run in the Google Colab

environment. Google Colab is a free cloud-based environment that has a number of advantages, especially when building Neural Networks. Python is also used to analyse the output of the models. This is done by creating plots in Python using the matplotlib library. Figure 2 below shows an overview of the design framework/flow of this research project.

Figure 2: Design Flow Overview
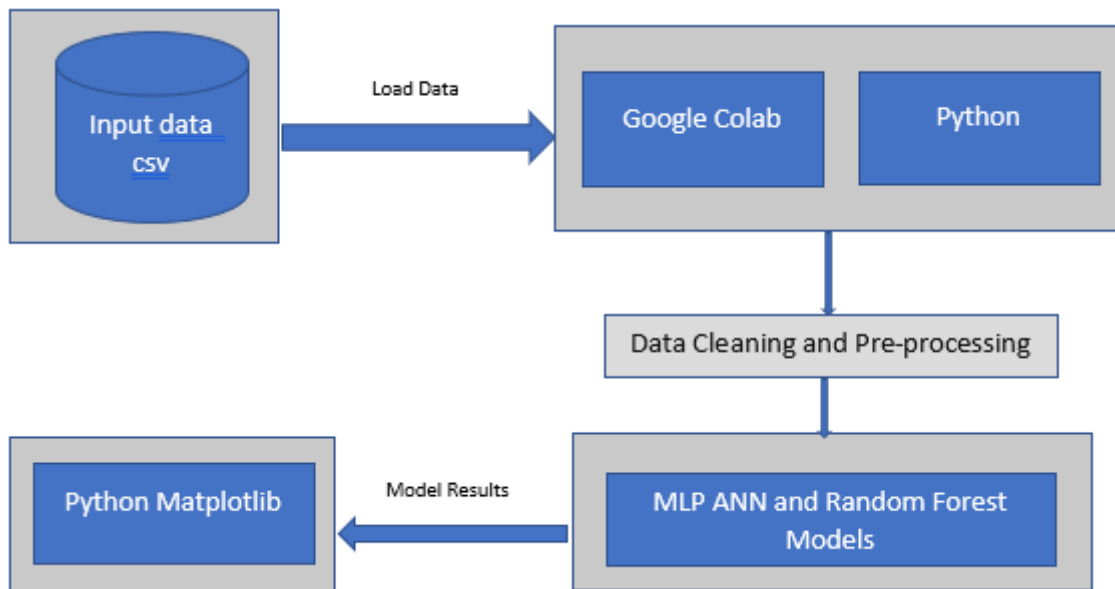
After the initial analysis of the input data is performed, the data files from the OECD website are combined into one input dataset in Microsoft Excel for use in the research project.

- This input data is then fed into the Google Colab cloud environment and additional data cleaning and pre-processing is performed.

- The Machine Learning models are created in the Google Colab Python environment and the data is input to the models.

- The output of the data is analysed using Python, and Matplotlib is used to create output graphs for further analysis and evaluation.

This section of the research report covered an overview of the Design Specification of the project. In the next section of the report, the Implementation will be discussed in detail.

5 Implementation

In this section of the report, the implementation of the final solution created in this research project will be discussed. The GDP and economic indicator data used in this project was downloaded in csv format from the OECD data repository. The data was analysed, combined, and cleaned using Microsoft Excel and Python. The machine learning models used in this project were created in Python, using the Google Colab cloud environment. Analysis and graphs of the output of the models was completed in Python with the Matplotlib library.

5.1 Data Transformation

As mentioned previously, the data used in this project was downloaded from the OECD data repository. The data was downloaded in various individual files, all in the csv format. The raw data was split into individual files based on the different economic indicators, for example GDP, unemployment rate, business and consumer confidence, etc. An initial exploratory data analysis was carried out to determine the characteristics of the data sets and to note any potential issues with missing data. The individual data files were combined in to one master data set to be used as input for the models in the project. The 'TIME' column in the data set contained data in the format '2010-Q1'. As there was a similar column 'Period' already in the data set, it was decided to convert the 'TIME' column to a monthly data format which would be useful for analysis of the data. Data in the 'TIME' column was converted into the format '01/01/2000', '01/02/2000' etc.

Pre-processing of the data also found that some columns in the data would not be required so they were dropped from the Python dataframes that were created for storing the data used in the project. Columns such as 'Subject', 'Measure', 'Frequency' and 'Unit' were dropped as they were not deemed necessary for the analysis required in this project. It was found from the exploratory data analysis that some business and consumer confidence data was missing for the countries Australia and Korea. It was decided to drop the data for these countries from the study as the missing indicator data would cause issues in the training of the models. A missing interest rate for September 2001 for the country of Sweden was imputed from the preceding and following month as the value was almost identical (3.7 and 3.71). The data was examined in Python with the '.describe()' method which shows some basic statistics such as the mean, standard deviation and percentiles of the data. The correlations were also checked with the '.corr()' method in Python and histograms of the individual indicators were examined (for example, figure 3 below). After these checks and transformations were completed, the final dataset for use in the project was in place.

Figure 3: Data Histogram for Analysis
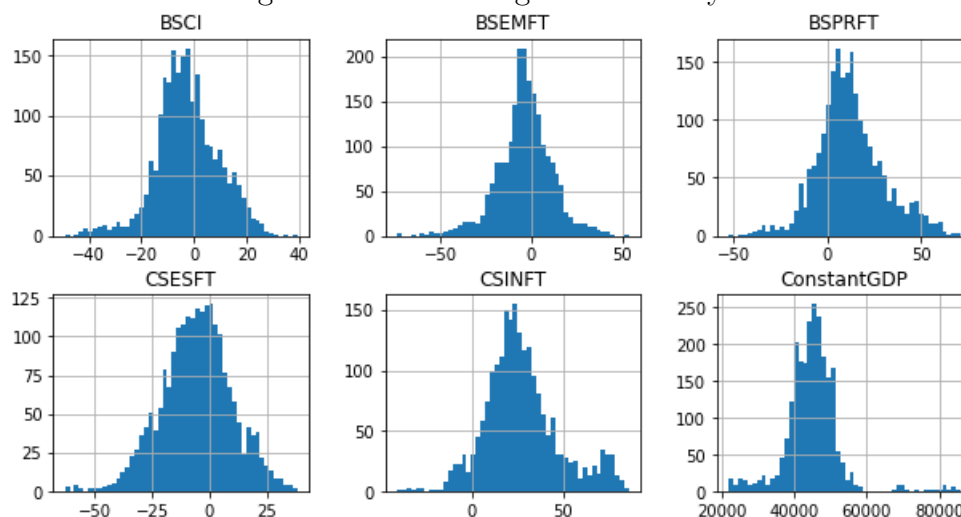
5.2 Model Development

In this research project, Artificial Neural Network and Random Forest Models were selected for creating models to predict GDP. These models were selected after an analysis of the related work in the area showed that they were suitable models for use in the prediction of economic performance. Google Colab was selected as the environment for creating the models in this work. The dataset used in this project was stored in Google Drive and accessed through Colab using an authentication module. The dataset is loaded using the Dataset.py module. In both models, the data is split in a 70% - 15% - 15% split for training, validation and testing. Various Python methods and libraries are used in these models and have been utilised previously for building models to predict GDP (Ghana and Singh; 2019) and (Thite; 2020). The Python train_test_split method from the sklearn.model_selection library is used for splitting the data. The data is scaled for the purpose of standardising using the sklearn.proprocessing library method StandardScalar. Scaling the data allows for better training in the cases where the variables have a wide range of values, which can impact the training of the data negatively.

5.2.1 Multilayer Perceptron Artificial Neural Network

The MLP model in this project is called via the Model.py module. This module contains the code for the MLP ANN model. As part of this module, various keras libraries are called for creating the Artificial Neural Network. This module links to the Options.py module which contains options for the Artificial Neural Network.

Figure 4: Artificial Neural Network Model
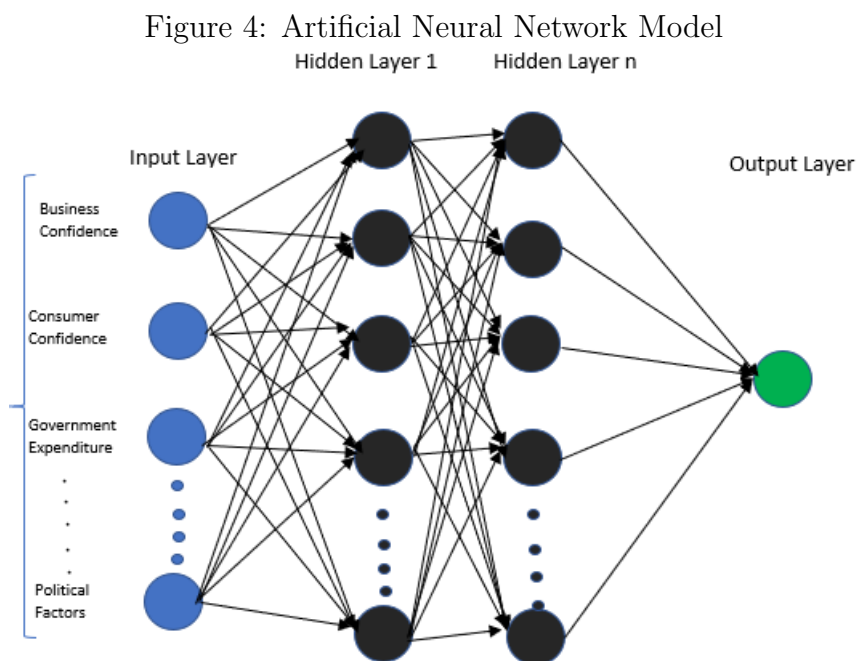
The MLP model is created using the 'MLPRegressor' which is part of the sklearn.neural network class in Python. Options such as the hidden layer sizes, validation fraction and max iterations are determined as part of creating the model. The output log of the model is also redirected to a log file, which can be used for further analysis of the model and its performance while the training epochs are running. In order to allow the Model to be saved and run separately if required, the 'pickle' and sklearn.externals library 'joblib' are

used. These allow the model to be saved and loaded separately. Using the pickle.load functionality then allows the model to be called in another environment if required. The plots of the output of the model are then called in the plots.py module. Figure 4 shows an overview of the framework of the Artificial Neural Network model. The input layer has a number of inputs based on the economic indicator data fed into the model. This then feeds into the Hidden Layers and eventually to the Output Layer.

5.2.2   Random Forest Model

The Random Forest Models use the same process for loading the data, and the train, validation and testing split of the data. The Random Forest model is called by the 'RandomForestRegressor' method which is part of the 'sklearn.ensemble' module. The Random Forest model has a number of options as well that can be modified and tuned if required such as the number of estimators (trees). In both the MLP and Random Forest models, the 'sklearn' method 'metrics' were used for calling evaluation statistics such as R2 score and RMSE. Figure 5 shows an overview of the Random Forest models framework.

Figure 5: Random Forest Model
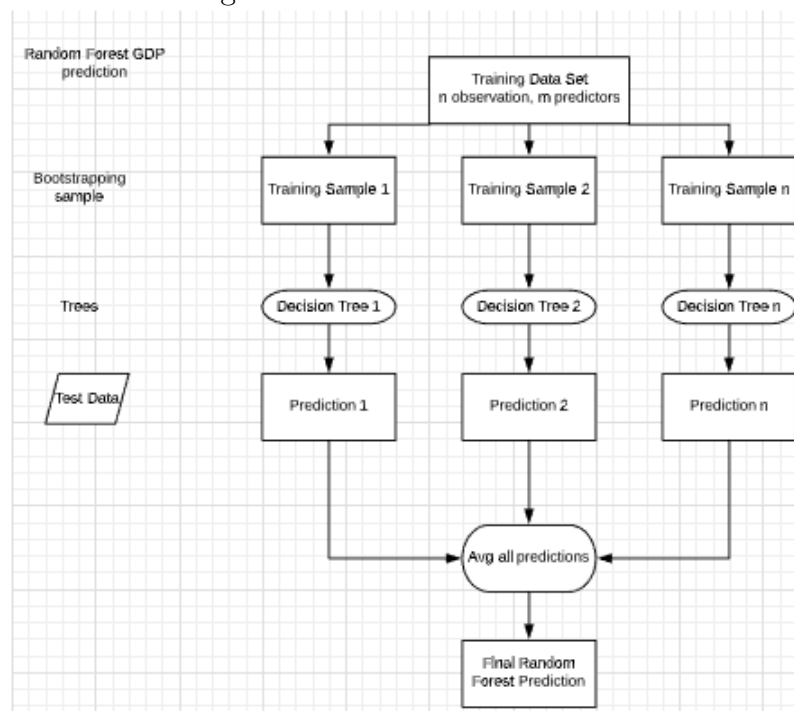
5.3   Output Analysis

After the Multilayer Perceptron and Random Forest models were created and run in Google Colab, the output of the models was analysed with Python and graphs of the output were created for further evaluation of the models. The Python library matplotlib was used to create the graphs of the output. Various graphs were created, for example graphs comparing actual to predicted GDP (Overall) and differences in GDP for Ireland (Actual v Predicted). Graphs were also created examining some predictions specifically for the year 2020 in some of the countries.

6 Evaluation

In this section of the report, the results of the implemented models are discussed. The objective of this research project was to create machine learning models for predicting GDP in Ireland and other selected countries. Multi-layer Perceptron Artificial Neural Network and Random Forest models were used for predicting economic performance, based on selected economic indicator data sourced from the OECD data repository. These machine learning methods were selected after an extensive review of related work in the area of predicting economic performance using machine learning. Both Artificial Neural Networks and Random Forest models have been used previously for predicting GDP. In this project, economic indicators were also selected for predicting GDP, based on a study of previous work conducted in the prediction of GDP. These indicators were composed of data with a monthly frequency (Consumer and Business confidence Indicators, Unemployment rates, Interest Rates etc) as well traditional indicators of GDP (for example Import and Exports). The results of the models are discussed in detail in this section.

6.1 Model Evaluation

Figure 6 shows the results of the evaluation of the models. The Multilayer Perceptron model had an R2 score of 0.94, while the Random Forest model had a higher score of 0.98.

Figure 6: Model Evaluation Scores

MLP ANN			Random Forest	
R2	0.94		R2	0.98
MSE	2946980.9		MSE	929508.09
RMSE	1716.67		RMSE	964.11
			MAE	426.23

Figure 7 below shows a selected comparison of GDP predictions for Ireland in 2019 from both the MLP and Random Forest models.

Figure 7: Selected Predictions of GDP Ireland 2019

2019 Selected Data for Ireland Actual vs Pred for MLP and RF			
Date	Actual GDP	MLP Predicted	RF Predicted
01/01/2019	83785.34812	80940.34419	83166.08808
01/05/2019	83104.43988	83392.41446	82030.91513
01/09/2019	84600.3956	78513.82205	84640.41719
01/10/2019	85901.24537	78816.46421	85028.2341
01/12/2019	85901.24537	80819.82747	84451.59843

6.2 Multilayer Perceptron Predictions

Figure 8 below shows the actual GDP values vs the predicted GDP values for Ireland for the period January 2000 – December 2019. The predicted values from the model are relatively good apart from a couple of periods of divergence.

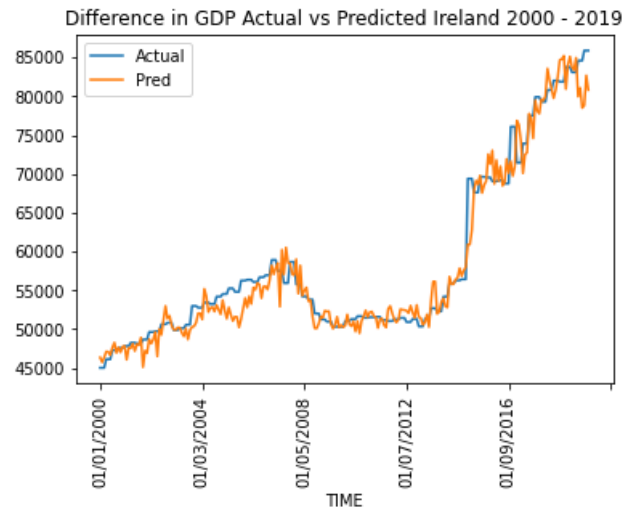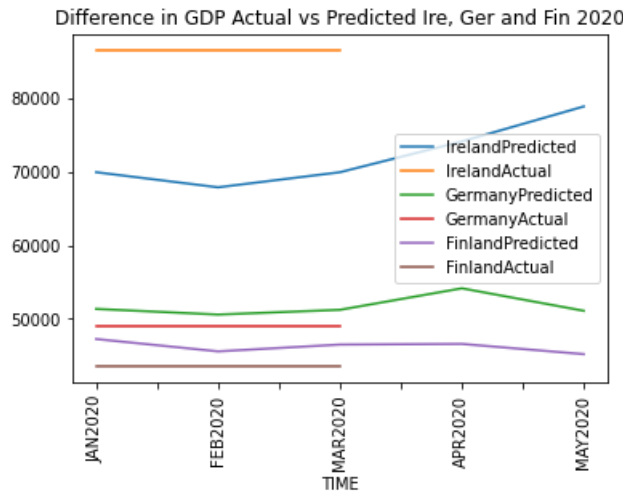Figure 8: Actual vs Predicted GDP Ireland 2000 - 2019 MLP ANN

Figure 9: Actual vs Predicted GDP 2020 - Ireland, Germany, Finland

In this study, actual GDP values were available up to the first quarter of 2020. The graph above shows actual and predicted values for some selected countries (Ireland, Germany, and Finland) for quarter one of 2020 along with predicted values for the first couple of months of quarter two (April and May of 2020). There is a noticeable difference between the actual and predicted values for quarter one 2020 for Ireland, whereas the figures for Germany and Finland are closer. Ireland's GDP figures are high in comparison to other countries as the GDP figures in Ireland are somewhat distorted by the activity of multinational companies. Many multinational companies use Ireland as a base for taxation purposes and earnings of the companies are often declared in Ireland. Sales and exports related to intellectual property and economic activities are declared in Ireland and so contribute to Irish GDP rather than the country where activity may have taken place. These factors can have a distorting impact on Irish GDP and so make it trickier to predict in comparison to other countries.

6.3   Random Forest Output

The Random Forest models in this project were created for comparison with the MLP ANN models. It can be seen that the Random Forest model has a higher accuracy than the MLP model. Figure 10 below shows a comparison of the Actual vs Predicted values of GDP for the Random Forest model. It can be seen that the predictions from this model are quite accurate.
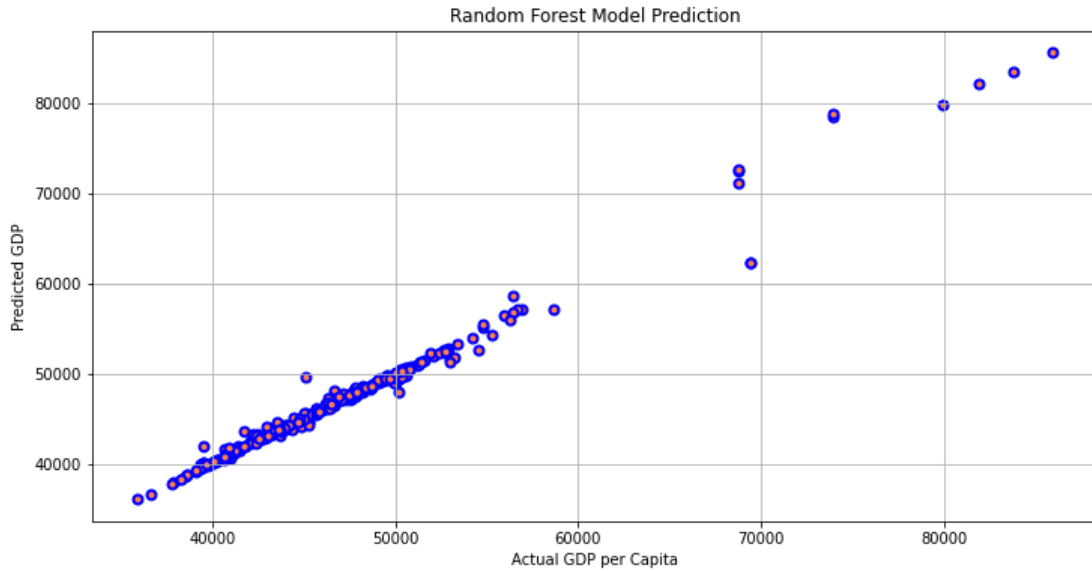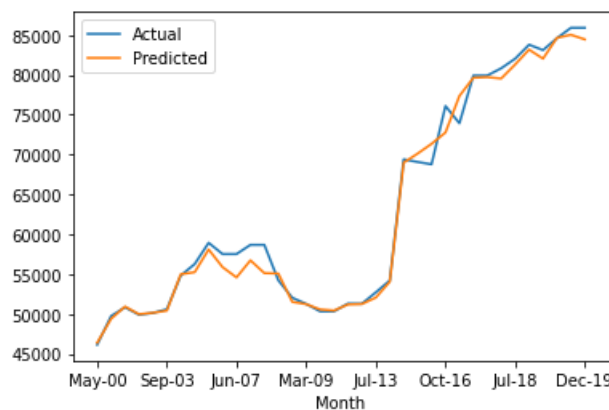
Figure 10: Random Forest Model Predictions

Figure 11 below shows the Actual vs Predicted values from the Random Forest model for Ireland. Other than a few divergences it shows that the predictions were generally in line with the actual values. Figure 12 below shows an example of the actual vs predicted values of GDP in Denmark. The predicted values were also generally in line with the actual values of Danish GDP.
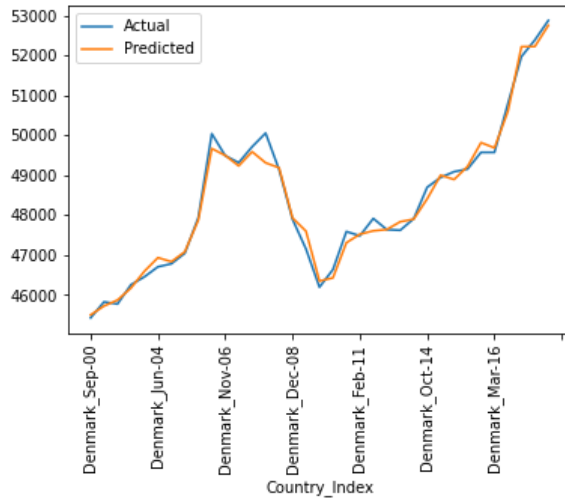
Figure 11: Random Forest Model Ireland 2000 - 2019

6.4   Discussion

The MLP ANN and Random Forest models created in this project had a good level of accuracy. The economic indicators selected appear to be positive in terms of their GDP

Figure 12: Random Forest Model Denmark

predictive abilities. The MLP model had an R2 value of around 0.94 which was similar to the ANN model created by Alamsyah and Permana (2018) which had an accuracy of approximately 95%. Based on the data from twenty years (2000 – 2019), the models predicted values for GDP relatively in line with the actual GDP values. When looking at the predictions from 2020 and the prediction of future GDP with the MLP model, some interesting points can be noted. It can be seen that the model is predicting an increase in GDP in Ireland in the second quarter of 2020, which due to issues in the global economy caused by the COVID-19 virus will not be the case. However, this increase in the GDP prediction in the model is likely caused by a large increase in pharmaceutical/chemical exports in Ireland which took place in March 2020. This large increase in exports is believed to be linked to global demand due to the COVID-19 pandemic [3].

As mentioned above, Ireland is quite unusual in terms of its GDP figures in comparison to other countries, due to the large presence of multinational companies in Ireland which distort the Irish GDP figure. This distortion is due to the declaration and routing of economic activity thorough Ireland and the use of Ireland as a taxation base, resulting in economic activity being declared in Ireland that may not have taken place in the country, whilst still contributing to Irish GDP [4]. The actual values for GDP in 2020 were generally lower than the predicted values from the MLP model, except for Ireland. These lower GDP values were likely at least partially due to the impact of the COVID-19 pandemic. As part of the initial work carried out in this research project, an LSTM RNN model was built as well for predicting GDP, however the RNN model was having issues with inconsistent validations, so the MLP and Random Forest models were used for making economic predictions in this research project. Both models proved to be adequate at making GDP predictions based on the data used in the project. A larger dataset with more countries and for a greater period of time would be useful for further testing of these models, however it should be noted that the availability of economic data can sometimes present difficulties, especially historically, as found in this research project (for Korea and Australia) where data can be missing or unavailable for periods of time for different countries, thereby making it difficult to accurately test some countries in the models.

---

[3]https://www.irishtimes.com/business/economy/irish-exports-hit-record-15-7bn-in-march-as-demand-for-medicines-surges-1.4254319

[4]http://www.oecd.org/sdd/na/Irish-GDP-up-in-2015-OECD.pdf

7 Conclusion and Future Work

This research project examined the question 'To what extent can machine learning algorithms predict Gross Domestic Product in Ireland and other countries based on various economic factors?'. The objectives of the research were to create two machine learning models for predicting GDP based on selected economic indicators, to evaluate the performance and accuracy of the created models and to study the GDP predictions from the models in order to determine their potential value for use in future economic analysis. The idea behind this project was to also examine some alternate economic indicators with data that is available more frequently to traditional economic indicators for predicting GDP. Traditional economic indicator data is usually available only quarterly at best, so the use of more frequent data would be beneficial for economic analysis purposes. Based on the review of related literature in this area, and from the work carried out in this research project, it can be seen that machine learning models can be used to predict GDP in Ireland and other countries based on the economic indicators selected. A number of models were considered for use in this research project, however after a thorough analysis, Artificial Neural Networks and Random Forest methods were chosen for use and comparison in this project as they were found to be highly proficient for predicting economic performance. A multilayer perceptron ANN model was created in this project. Initially an LTSM RNN model was also considered, however after some preliminary work the MLP model was proving to be more suitable for use with the selected data. The Random Forest model was also proficient, and it provided a good comparison to the Artificial Neural Network model created.

Both models created in this project performed well using the selected economic indicators, however the Random Forest model performed slightly better than the MLP model. The MLP model had an R2 score of 0.94 compared to the Random Forest model with an 0.98 R2 score. The predicted GDP output from the models was compared to the Actual GDP over a period of 20 years (2000 – 2019) and with the exception of some divergences, the GDP output from the models was relatively in line with the actual figures. The models were used to make predictions for GDP in Q1 and Q2 of 2020 and the predictions looked reasonable for most countries. A known issue related to Irish GDP was visible in this research project. Irish GDP figures are generally distorted due to the impact of multinational companies using Ireland as their headquarters for taxation purposes. These companies filter much of their economic activity and intellectual property profits through Ireland, with a result that production and export figures in Ireland appear significantly higher than actual production rates, resulting in a high GDP figure. This can make prediction of Irish GDP trickier than other countries. However, with the selection of relevant economic indicators, these concerns can be mitigated.

The output from these models could be useful as part of economic research and analysis done on specific countries economic performance and future GDP outlook. This research also contributes to the area of machine learning based predictions of GDP and economic performance, particularly with regards to the predictive ability of the economic indicators used in this research project. Future work in this area would look at a more in-depth analysis of the economic indicators and their specific impact on various countries, as well as the impact of other factors that can impact GDP. For example, the COVID-19 pandemic caused a large drop in GDP around the world and it would be interesting to link a machine learning based analysis of the impact of COVID-19 to this work in order to further enhance the GDP and economic performance predictions.

References

Alamsyah, A. and Permana, M. (2018). Artificial Neural Network for Predicting Indonesian Economic Growth Using Macroeconomics Indicators, *2018 International Symposium on Advanced Intelligent Informatics (SAIN)*, Yogyakarta, Indonesia, 22-23 August, pp. 15–19. IEEE Xplore. doi: 10.1109/SAIN.2018.8673347.

Basuchoudhary, A., Bang, J. and Sen, T. (2017). Machine-learning Techniques in Economics New Tools for Predicting Economic Growth, *Springer* . Available at: `https://doi.org/10.1007/978-3-319-69014-8` [Accessed 22 May 2020].

Behrens, C., Pierdzioch, C. and Risse, M. (2018). A test of the joint efficiency of macroeconomic forecasts using multivariate random forests, *Journal of Forecasting* **37**(5): 560–572. doi: 10.1002/for.2520.

Cogoljevic, D., Alizamir, M., Piljan, I., Piljan, T., Prljic, K. and Zimonjic, S. (2018). A machine learning approach for predicting the relationship between energy resources and economic development, *Physica A: Statistical Mechanics and its Applications* (495): 211–214. doi: 10.1016/j.physa.2017.12.082.

Divya, K. and Devi, V. (2014). A Study on Predictors of GDP: Early Signals, *Procedia Economics and Finance* (11): 375–382. doi: 10.1016/S2212-5671(14)00205-6.

Ghana, S. and Singh, S. (2019). India GDP Prediction, *Github* . Available at: `https://github.com/satyajitghana/IndiaGDPPrediction` (Accessed: 25 May 2020).

Jahn, M. (2018). Artificial neural network regression models: Predicting GDP growth, *HWWI Research Papers 185, Hamburg Institute of International Economics (HWWI)* . Available at: `https://ideas.repec.org/p/zbw/hwwirp/185.html` (Accessed: 26 May 2020).

Jiang, Y., Guo, Y. and Zhang, Y. (2017). Forecasting China's GDP growth using dynamic factors and mixed-frequency data, *Economic Modelling* (66): 132–138. doi: 10.1016/j.econmod.2017.06.005.

Kordanuli, B. (2017). Appraisal of artificial neural network for forecasting of economic parameters, *Physica A: Statistical Mechanics and its Applications* (465): 515–519. doi: 10.1016/j.physa.2016.08.062.

Kouziokas, G. (2017). Machine Learning Technique in Time Series Prediction of Gross Domestic Product, *Informatics. (ACM International Conference Proceeding Series)* pp. 1–2. doi: 10.1145/3139367.3139443.

Kuosmanen, P. and Vataja, J. (2017). The return of financial variables in forecasting gdp growth in the g-7, *Economic Change and Restructuring* **50**(3): 259–277. doi: 10.1007/s10644-017-9212-7.

Liang, H. (2020). Gdp spatialization in Ningbo City based on NPP/VIIRS night-time light and auxiliary data using random forest regression, *Advances in Space Research* **65**(1): 481–493. doi: 10.1016/j.asr.2019.09.035.

Martin, L. (2019). Machine Learning vs Traditional Forecasting Methods: An Application to South African GDP, *Working Papers 12/2019, Stellenbosch University, Department of Economics* . Available at: `https://ideas.repec.org/p/sza/wpaper/wpapers326.html` (Accessed: 26 May 2020).

Nik, P., Jusoh, A. M. and Sarmdi, T. (2016). Predicting the Probability of Financial Crisis in Emerging Countries Using an Early Warning System: Artificial Neural Network, *Journal of Economic Cooperation and Development* **37**(1): 25. Available at: `http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=edo&AN=116321159&site=eds-live&scope=site` (Accessed: 25 May 2020).

OECD (2020). Oecd Data Repository, *OECD.Stat* . Available at: `:https://stats.oecd.org/` (Accessed: 25 May 2020).

Richardson, A., van Florenstein Mulder, T. and Vehbi, T. (2018). Nowcasting New Zealand GDP using machine learning algorithms, *CAMA Working Papers 2018-47, Centre for Applied Macroeconomic Analysis, Crawford School of Public Policy, The Australian National University* . Available at: `https://ideas.repec.org/p/een/camaaa/2018-47.html`.

Stevanovic, M., Vujicic, S. and Gajic, A. (2018). Gross domestic product estimation based on electricity utilization by artificial neural network, *Physica A: Statistical Mechanics and its Applications* **489**: 28–31. doi: 10.1016/j.physa.2017.07.023.

Stundziene, A. (2013). Prediction of Lithuanian GDP: Are Regression Models or Time Series Models Better?, *Economics and Management* **18**(4): 721–734. doi: 10.5755/j01.em.18.4.5041.

Thite, A. (2020). Gdp Prediction, *Github* . Available at: `:https://github.com/thite-amol/gdp-prediction` (Accessed: 15 June 2020).

Tumer, A. and Akkus, A. (2018). Forecasting Gross Domestic Product per Capita Using Artificial Neural Networks with Non-Economical Parameters, *Physica A: Statistical Mechanics and its Applications* (512): 468–473. doi: 10.1016/j.physa.2018.08.047.

Vrbka, J. (2016). Predicting Future GDP Development by Means of Artificial Intelligence, *Littera Scripta* (3): 154. Available at: `http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=edsdoj&AN=edsdoj.775645272b841ce907dc34e9b6a0022&site=eds-live&scope=site` (Accessed: 25 May 2020).

Wu, W., Zhang, T. and Zheng, C. (2019). A Novel Optimized Nonlinear Grey Bernoulli Model for Forecasting China's GDP, *Complexity* . doi: 10.1155/2019/1731262.

Yu, J., Zhang, J., Shin, H. and Kong, J. (2019). Revisiting the economic crisis after a decade: Statistical and machine learning perspectives, *Annals of Dunarea de Jos University. Fascicle I: Economics and Applied Informatics* (2): 14. Available at: `http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=edsdoj&AN=edsdoj.9976a30f0a074b5f9f3ccabed9fc7a6e&site=eds-live&scope=site` (Accessed: 23 May 2020).