European Productivity

Crispian Morris
Engineering Mathematics
University of Bristol
Bristol, UK
wg19671@bristol.ac.uk

Evgenia Karnavou
Computer Science
University of Bristol
Bristol, UK
dr19487@bristol.ac.uk

Isaac Evans
Engineering Mathematics
University of Bristol
Bristol, UK
jn19392@bristol.ac.uk

Katia Vasilonikolidaki
Computer Science
University of Bristol
Bristol, UK
bl19548@bristol.ac.uk

Nisa Bayraktar
Computer Science
University of Bristol
Bristol, UK
da19157@bristol.ac.uk

***Abstract*—An investigation is conducted into quarterly European productivity from 2008 to 2019 for fifteen European countries, by measuring the GDP per hour worked and the impact of different features on this. Productivity is measured by GDP per hour worked, which is impacted by many socio-economic features. The features investigated are education, depression, inflation and unemployment. To investigate productivity, dimensionality reduction, clustering algorithms, linear and penalised regression, a SARIMA model and a recurrent neural network (RNN) are applied to the data. Six clusters of countries are found in the datasets, which appear to match up historically similar and geographically close countries. This report discusses these approaches and the problems encountered when trying to source data at the resolutions required to train these machine learning techniques. The effectiveness of the methods are evaluated, and possible extensions to them are discussed further.**

I. Introduction

Productivity is the rate of output per resource. This rate of output can be measured at many levels, and at a global level, one way it is measured is Gross Domestic Product, which is the total value of all goods and services produced in a country. There are two main measures to compare GDP between countries: the nominal GDP, which uses the currency exchange rates and Purchasing Power Parity (PPP), which equates a basket of goods across currencies to the output of goods and services. The productivity is then the GDP per resource, and often per capita GDP is used. As this report considers European productivity all GDP values are measured in Euros.

However, due to varying labour markets from differences in age structure, culture and the economy, GDP per person employed and GDP per hours worked are used in this report. GDP per person employed gives an overview of the efficiency of the labour force in a country. This is only an average, so it does not provide a break-down of the distribution of productive labour among the labour force. On the other hand, GDP per hour worked breaks down per person employed again into the efficiency of every hour spent at work.

Productivity is investigated to track the output per employee or per hour worked and therefore the average efficiency per worker or hour worked. The factors that affect productivity are investigated to see how productivity can be increased. The factors that affect productivity can be broadly categorised into economic and social factors. Economic factors include inflation, which is the increase in the value of goods and services and unemployment. Economic factors can affect business confidence and the likelihood of investing capital, which lead to decreasing productivity. Social factors include the education levels in the population, as a more skilled work force will be able to work more productively. More personal social factors include mental health issues such as depression, which can significantly impact a persons productivity. Increasing productivity is increasing the efficiency of the labour force, improving the quality of life and encouraging human development and progression.

An often useful application of GDP is in evaluating the general economic state of the world, resulting in what is known as the world economic outlook. Here, GDP in countries around the globe is accounted for, and general trends are logged and investigated. This process allows for the economic effects of a diverse range of factors to be considered on a global scale, such as natural disasters or pandemics. However, in addition to logging these GDP statistics, historical data is used to predict the GDP for the next few years. Having an accurate GDP prediction is useful for estimating the impact on the economy by other events, so this report aims to use national socioeconomic characteristics to predict and analyse productivity in Europe. This is achieved using a combination of k-means clustering and several regression models, as well as a recurrent neural network.

II. Data Preparation

In order to conduct meaningful analysis, the amount of historical data needs to be large. The dataset provided for relative GDP in European countries is yearly, which is not a high enough time resolution for this analysis. Quarterly data for GDP, number of employees and hours worked per week for European countries is obtained from Eurostat. It is assumed that the EU provides reliable, accurate and trustworthy data. It is also assumed that the data made openly avaliable online by the EU is anonymous and ethical for use. The time and countries that overlap in these data frames are then matched. The GDP per employee data is calculated by dividing the GDP dataset element wise by the number of employees. The GDP per hour worked is calculated by dividing the GDP per employee dataset element wise by the number of hours worked and dividing by eleven to adjust for the number of weeks in a quarter minus some two weeks holiday. Overall, the quarterly sample rate is agreed upon since it offers a good compromise between dataset availability and resolution. Next, economic and social factors are selected because the general state of economy provides the conditions for a productive work force, and social factors ensure that a population has the ability to meet its maximum potential productivity. Unemployment is a measure of the people who are out of work and therefore unproductive, however one of the downsides of per employee GDP is that it does not account for this. On the other hand GDP per employee is then a more stable measure than GDP per capita as it is not directly affected by changes in the labour market. Social factors are useful to examine the overall ability and health of labour force, increased education would allow a worker to be more productive in a higher skilled job.

A. Economic Features

Inflation rates for Europe are sourced from the OECD database. This datas show the quarterly average percentages of the Consumer Price Index (CPI) in the European region between the years 2000 and 2021. CPI is a common measure of inflation which represents the price difference in the market of goods and services purchased by various urban consumers [1].

An unemployment dataset is also obtained from the OECD, and is similarly comprised of unemployment rate estimates for European countries between 2000 and 2021. The data is seasonally adjusted for each year, and the rate during a given quarter is calculated as the number of unemployed people as a percentage of the labour force. The labour force is the sum of employed and unemployed people, and the number of unemployed people is defined as the number of people of working age who are without work despite having taken specific steps to find employment. The data is originally sourced using Labour Force Surveys (LFS), and rates were estimated using Eurostat when LFS information is unavailable.

Data for wages is sourced from the ILOStat database, and includes the average income of citizens in European Countries from 2010 to 2019. This income is the ratio between the total labour income and gross domestic product, both provided unadjusted. The labour income includes both the earnings of employees and an estimation of income of the self-employed.

B. Social Features

The dataset of depression rates is from the Our World In Data website, it gives yearly depression rates in different age groups across European countries from 2000 to 2019. Since only members of the working population affect productivity, the only age range taken into consideration from this dataset was that of the labour force.

The education dataset was collected from the ILOSTAT database, and information is from 2008 to 2021. Here, data is segregated according to the highest level of education attained by the labour force in EU countries.

The levels of education included are compared with their corresponding European Qualifications Framework (EQF) level in Table I. The labels used during the analysis of education as a feature are also shown.

TABLE I
EDUCATION LEVELS

Eduction Level	Label	EQF Level
Less than Basic Education	0	Levels 0-2
Basic Education	1	Levels 3-8
Intermediate Education	2	Levels 3-4
Advanced Education	3	Levels 5-8

C. Wrangling

In the feature datasets, there is a wide array of time ranges, countries, and time granularity. For example, the wages dataset only includes data from 2010 to 2019, whereas the inflation data exists between 2000 and 2021. Therefore, the tightest range of years is used, meaning data before 2008-Q1 and after 2019-Q4 is removed. Additionally, some datasets only feature subsets of European countries, so the countries that complete data cannot be sourced for are omitted. The list of countries that a complete portfolio of data could be formed for ultimately summed only to 15. Finally, the granularity of time is considered. For instance, while the unemployment dataset contains quarterly data, the wages dataset is comprised of yearly records. The datasets with annually recorded values are interpolated to quarterly data by logging the same values for each quarter to ensure the minimal information is added to the dataset.

After the cropping the datasets into the overlapping countries and times, the data is fused together into a three-dimensional array, with the third dimension corresponding to the country. An overall average for the EU is then calculated, weighted by the hours worked by each country.

III. Data Exploration

A. Descriptive Statistics

The three-dimensional data frame comprises fifteen countries and ten separate features between the years 2008 and 2019.

Mutual information, F-test and Pearson's correlation coefficient are calculated for each feature, in comparison with the GDP per hour worked, as shown in Figure 1. The mutual information measure shows the information related to the GDP per hour worked data for each hour worked. The mutual information values for three of the four education categories are consistently between 0.77 to 0.78, however intermediate education is slightly lower, at 0.6 due to it's limited range of EQF levels, as shown in Table I. The F-test compares the distribution of the datasets that then are scaled using the maximum F-test score. The Pearson's correlation coefficient has also been calculated for all the features. Less than basic education is strongly negatively correlated with productivity. There is a high positive correlation for basic and advanced education, but shows almost no correlation with intermediate education at 0.051, this may be due to intermediate education including only two levels as shown in Table I. Inflation, unemployment, and depression rates are all slightly negatively correlated with scores between -0.2 and -0.27.

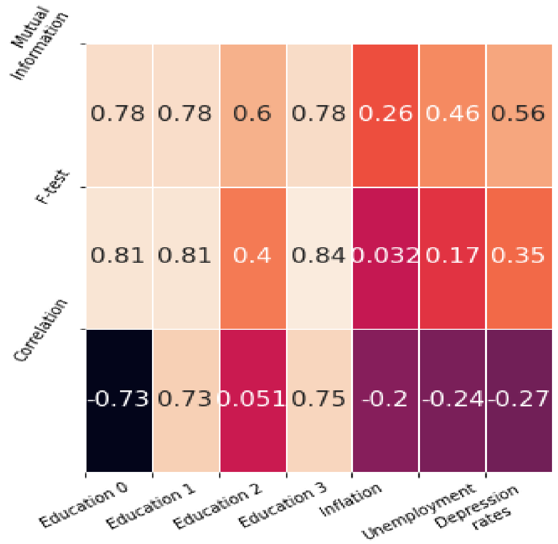

Fig. 1. This heat map shows three statistical measures mutual information, f-test and correlation for the seven features included in the data.

B. Dimensionality Reduction

Before the regression and cluster models are built, dimensionality reduction is used to project both the GDP and feature data into two dimensional space, so as to help visualise them. Principle component analysis (PCA), a commonly used linear dimensionality reduction technique which focuses on placing dissimilar points far apart in lower dimensional space, is used for this purpose. t-distributed stochastic neighbour embedding (t-SNE) is also used, which, unlike PCA, is a non-linear transformation and therefore does not preserve global structure. Instead, it converts high-dimensional distances between points into conditional probabilities that represent similarities [2]. By choosing an appropriate perplexity parameter value, which is a smooth measure of the effective number of neighbours, t-SNE can preserve local similarities, making comparison possible [3]. Before the application of these methods, the three-dimensional data frame is reshaped into a two-dimensional one, with the features as columns. This list of quarters for each country is normalised using sklearn's `StandardScaler`, and the reduction methods are applied.

Figure 2 show the effects of dimensionality reduction on the feature data. PCA is able to capture some information, but many of the countries' data points overlap significantly. t-SNE is able to separate out this bunching, and the differences between countries become much more visible.

Interestingly, some culturally similar countries appear to be grouped, such as Germany and Austria, as well as Spain and Portugal. This gives a visual indication that there exist potential clusters to examine.

C. Clustering

Clustering is applied so as to ascertain the existence of any groupings of similar countries in the data. Similarly to dimensionality reduction, both the GDP and feature data are investigated. Since no prediction is being done using these methods, this is allowed. If some clusters are determined, separate prediction models can then be tuned for prediction on each specific group of countries. This allows for increased accuracy by systematically taking into account subtle differences between the diverse range of countries being analysed.

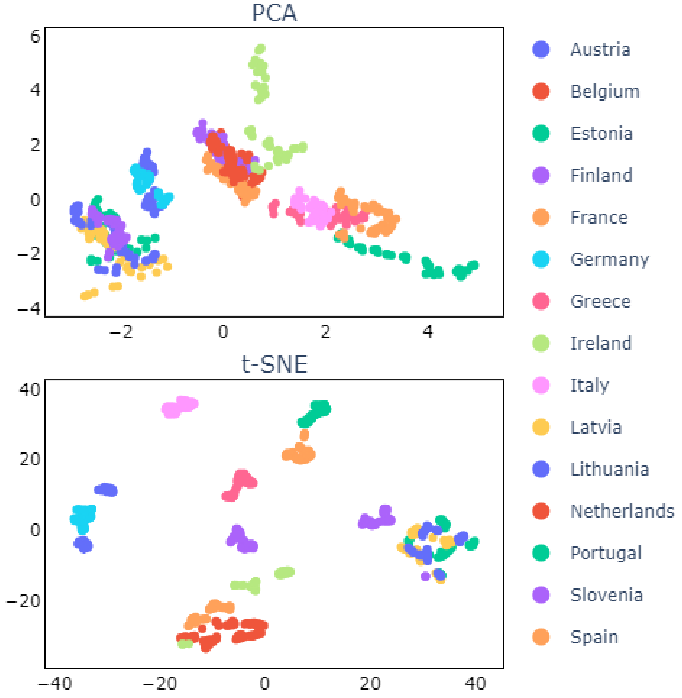

Fig. 2. Dimensionality reduction techniques applied to the features used in productivity analysis, coloured by country.

To determine if there are any useful clusters in the dataset, the k-means elbow method is employed. This heuristic uses the mean ratio of intra-cluster and nearest-cluster distance (silhouette) and the sum of squared distances of samples to their closest cluster centre (inertia) to evaluate the optimum number of clusters for a selection of features in a given dataset [4].

TABLE II
THE FIVE HIGHEST-SCORING CLUSTERS, DETECTED USING THE K-MEANS ELBOW METHOD.

Features	Score	Clusters
Education 3, GDP per employee, GDP per hour	0.641	5
Depression, Education 2, GDP per hour	0.615	6
Education 1, Education 2, Education 3	0.607	3
Depression, Education 3, GDP per employee	0.603	6
Depression, Education 1, Education 3	0.602	4

By looping through all possible combinations of the features, the method is used to find the optimum combination of three of them. The elbow method successfully found several high-scoring clusterable features, and the results are shown in Table II. Some feature combinations, such as Education 1, Education 2 and Education 3, are inherently redundant, and are omitted from consideration.

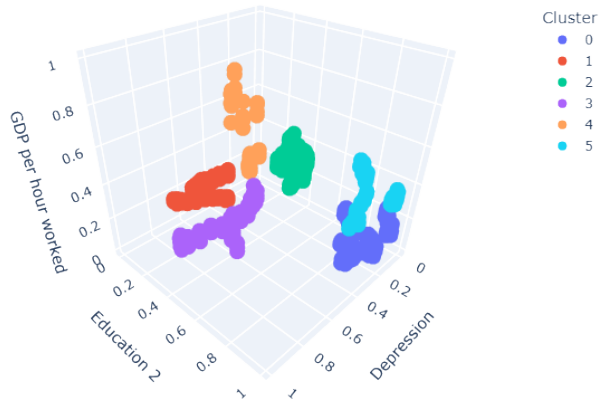

Fig. 3. Three dimensional plot showing the six clusters found by k-means

The highest scoring three-dimensional feature combination that is not subject to any redundancy is the combination of Education 2 (intermediate education), Depression, and GDP per Hour Worked. The clusters for this combination are visualised in Figure 3.

IV. Regression Approach

In order to determine the most important features in the acquired dataset, several regression models are tested and evaluated.

A. Linear Regression

Regression is a statistical technique which is used to estimate the nature and strength of a relationship between one dependent variable and a set of independent variables [5]. Thus, in this case, it is assumed that the found socio-economic features are all independent variables. Initially, these features will be used to predict GDP only for Austria, and afterwards, the best performing model will be fit to other countries individually. It is assumed that the best model for Austria will also be the best for any other European country, since the datasets are very similar.

Prior to the training process, the different ranges between the values of the features in the data are accounted for by standardising the dataset. Scaling the data is especially crucial for comparing the different features and deriving their importance, since it normalises the feature columns to a mean of zero and standard deviation of one [6]. Standardisation is chosen over other methods, since it preserves useful information about any potential outliers, making the algorithm less sensitive to them. This is in contrast to alternatives such as MinMax scaling, which scales the data to a limited range of values [7].

A linear regression is first trained on the Austrian dataset to calculate the regression coefficients of each feature for the country. The coefficients are then used as a measure for feature importance, as linear regression is a weighted sum of the feature inputs. In this type of regression, the independent variables x (the features), are mapped to the dependent variable y (GDP), to create a linear approximation of their relationship. This is expressed by

$$\hat{y} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon, \tag{1}$$

where p is the total number of features, β_j represents the learned weights (the coefficients), y is the predicted outcome of those features and ϵ represents the prediction error (following a Gaussian distribution). The weights are calculated using the ordinary least squares (OLS) distance measure, finding the coefficient which minimises the difference between the GDP per hour worked predicted using our features, and the actual value of the GDP per hour worked. The value of the weights of each feature β_x can be interpreted as a one unit increase in x_k that induces a β_k unit increase in the prediction of the independent variable $\hat{y}$ [8]. Therefore, we can conclude that the weight of a feature is proportional to the importance of each feature.

B. Penalized Regression

In order to avoid the overfitting introduced by multiple linear regression, three different shrinkage, or as they are more commonly known, regularisation regression techniques, are applied. Namely, Ridge, Lasso and ElasticNet regressions. The distinction between these methods and linear regression is that while the latter will continuously increase the weight of features it deems important during training, shrinkage methods will penalise the coefficient estimates, thereby shrinking them towards zero. Notably, shrinking the regression coefficients towards zero might introduce some biased estimates, but in turn these estimates will be characterised by lower variance [9]. As a result, these regression techniques have often been found to perform more favourably in comparison to the traditional linear method, enhancing the prediction accuracy of a given model [10].

1) Ridge Regression: Ridge regression (RR) is a $L2$ regularisation method which forces weights toward zero also known as coefficient shrinkage to simplify the model by penalising its large predictor coefficients [11]. Its weights are calculated using

$$\hat{\beta}_{Ridge} = (X^T X + \lambda I_p)^{-1} X^T y, \tag{2}$$

where I_p is the identity matrix, λ is the regularisation penalty and X is the matrix of the features. Besides preventing over-fitting and complexity, the RR also tackles the multicollinearity problem. Multicollinearity occurs when there is a correlation between the independent features of the dataset and may cause the inaccuracy of the regression coefficients [11]. The RR was implemented in the feature dataset to minimise over-fitting as well as avoid possible multicollinearity between the features.

2) Lasso Regression: The Lasso (Least Absolute Shrinkage and Selection Operator) regression was originally proposed by Robert Tibshirani for regularisation and feature selection of linear regression models [12], [13]. Lasso strives to identify the variables and corresponding regression coefficients which result in smaller prediction errors. This is achieved by setting a constraint on the model parameters, that 'shrinks' the regression coefficients towards zero by forcing the sum of their absolute values to be less that a predetermined value, λ. This type of regularisation, known as $L1$ regularisation, introduces sparseness to the model, and has led to the Lasso regression method receiving much attention in recent years [14]. The advantage of Lasso, when compared to the Ridge regression, is that it is able to perform feature selection, since the coefficients with the values that contribute the less to the minimisation of the loss function get set to zero. Essentially, Lasso regression penalises less important features of the dataset and makes their respective coefficients zero, thereby eliminating them. Therefore, Lasso regression also performs variable selection. As a result, models generated using Lasso are generally much easier to interpret than those produced by Ridge regression.

When performing Lasso regression on a country, eight coefficients $\hat{\beta}_{Lasso}$ were calculated each one corresponding to one of the eight features. Recapitulating subsubsection IV-B2, the magnitude of the coefficient, or weight, for one feature is proportional to the importance of that feature when calculating the $\hat{y}$ value of the regression, which is GDP per hours worked in our case. Hence, the higher the absolute value of the coefficient, the more the feature will contribute on the productivity of a country.

3) Elastic Net Regression: Elastic Net is also an extension of linear regression, calculated using a weighted combination of Lasso and Ridge, such that it uses both $L1$ penalty and $L2$ loss functions, respectively. It combines the properties of feature elimination from Lasso and the coefficient value reduction of less important features from Ridge to improve the model's prediction. The appropriate mix of Lasso and Ridge is determined by a hyperparameter α ($\alpha = 0$ for Ridge, $\alpha = 1$ for Lasso), estimated by cross-validation in order to optimise performance. Hyperparameter tuning using a GridSearch over a range of α values revealed the optimum to be $\alpha = 1$.

C. Feature Importance

Penalized regression is performed iteratively for all the countries in the combined feature dataset. Before applying each regression model, all the independent features are standardised using the `StandardScaler` from sklearn. The correlation coefficient, R^2, is recorded as well as the testing error, calculated using both Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Then, the cross-validation score for each regression is calculated to determine the model that minimises the test error.

TABLE III
AVERAGE TEST SCORES OF DIFFERENT PENALIZED REGRESSION MODELS TRAINED ON EACH COUNTRY.

Method	R^2 **Score**	**MAE Score**	**RMSE Score**
Linear	0.8163	1.2660	1.5464
Ridge	0.8086	1.3097	1.5949
Lasso	0.6721	1.5960	1.9985
ElasticNet	0.7102	1.6297	2.0380

This process is repeated for each of the countries in the dataset, and the average scores are shown in Table III. The linear regression method appears to be overfitting slightly, and observing the test error of each model, it is found that the regularisation method that performs the best is the Ridge Regression. For this reason, Ridge Regression was used to determine the importance of the features for each European country in the 2008 to 2019 timeframe.

V. Time Series Approach

A. Random Forest Regressor

To investigate the clusters found in the data further, one country is selected from each cluster. The countries selected are Austria, Finland, Ireland, Lithuania, The Netherlands, and Portugal. The EU average is also selected for investigation. A random forest regressor is chosen to predict the productivity of each of the selected countries as well as the EU as a whole from the features, instead of simply determining feature importance. Random forest regression uses a collection of randomly generated decision trees to create a regression from the data to a predicted GDP per hour worked. The average loss from this method across the six countries and the EU average using the feature data is 1.29%, while the average correlation coefficient between the actual and the predicted data is 0.974, indicating a strong positive correlation.

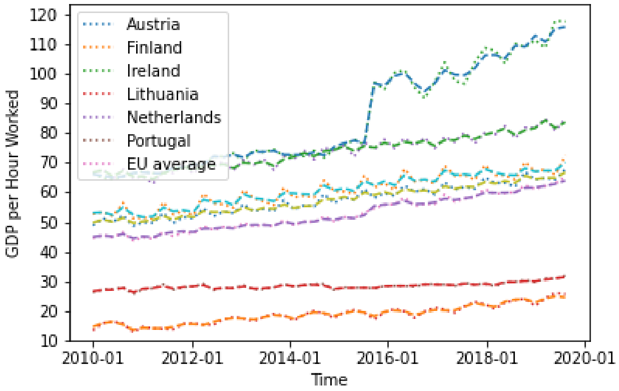

Fig. 4. A random forest regressor to fit the GDP per hours worked from the features matrix for one country per cluster. The real values are shown in solid colour for each country, whilst the predicted values are shown by the dotted line.

In order to investigate the significance of each feature, the random forest regressor is run iteratively. Each time, a selected feature is set to zero for the Austrian productivity data, and the loss is noted. The larger the loss is without a feature, the more important that feature is to the regression, and thus, the larger a factor it is in overall productivity. This allows for feature reduction for the random forest regressor to be carried out, reducing the risk of overfitting. When the depression rate is set to zero, the loss of the random forest is exactly the same as the loss with all the features included, at 0.012. This indicates that for random forest regression, the depression rate can be excluded from the feature data.

However, this feature reduction suffers from a significant shortcoming: as the number of features is reduced, the potential size of the decision trees is reduced. If all the features are removed, then the random forest regressor will not work, so this feature reduction is only carried out once. Additionally, for a given GDP estimation to be made, feature data at the selected time is required, suggesting that this method is not well-suited for use in predicting future values of productivity.

B. SARIMA

In order to predict future productivity, a SARIMA model is implemented. SARIMA stands for Seasonal, AutoRegressive, Integrated, Moving Average, and each of these sections decode the time series data in order to analyse it. Seasonality is a periodic part of the signal which is separated from the trend, in order to predict future values separately. AutoRegressive refers to the use of time lagged values in the model, which are used as the inputs to a regression calculation for the next time step. These are weighted using the Integrated section for the errors in the previous time steps, so as to calculate a Moving Average.

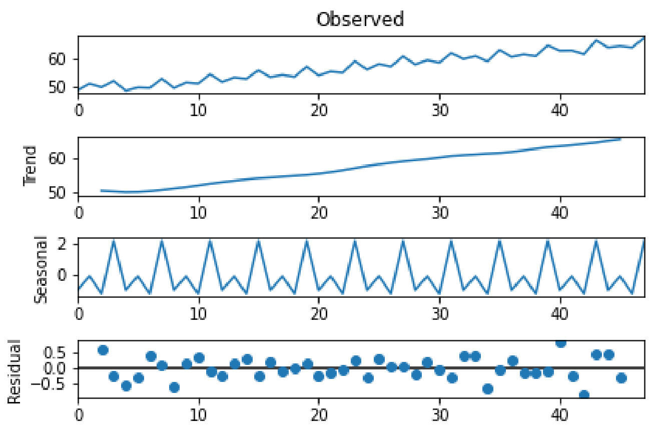

Fig. 5. Seasonal decomposition of the Austrian data using a SARIMA model. The first figure shows the observed GDP per hour worked. The second figure shows the trend in the data. The third figure shows the seasonal data. The fourth figure shows the residuals of the model.

In this case, seasonality is the section of the data that is affected by the changing seasons. The obtained productivity data is quarterly, so the season is four periods long. The data follows a repeating cyclic pattern shown here in the third section of Figure 5. Once the seasonality is removed, the trend data in the second section of shows much less periodicity.

C. Recurrent Neural Network

However, the SARIMA model implemented only uses the productivity data for future predictions. Therefore, this method may fail to detect trends in the feature data that have affected productivity data and will continue to affect productivity in the future. In order to assess the impact of the feature data as well, a Recurrent Neural Network (RNN) is implemented, which uses the features dataset in its prediction of productivity. The RNN is made up of three layers, the first two of which are Gated Recurrent Units (GRUs), with 16 units, hyperbolic tangent as the activation, and dropout and recurrent dropout both set to 0.2 to reduce overfitting. The third layer is a dense layer with one unit and linear activation, to reduce the dimensions of the output to just the prediction of GDP per Hour Worked. A GRU detects trends in the data by passing through the previous time steps with some removed at the recurrent dropout rate. The RNN uses the previous 12 data points to predict four data points, or one year, into the future.

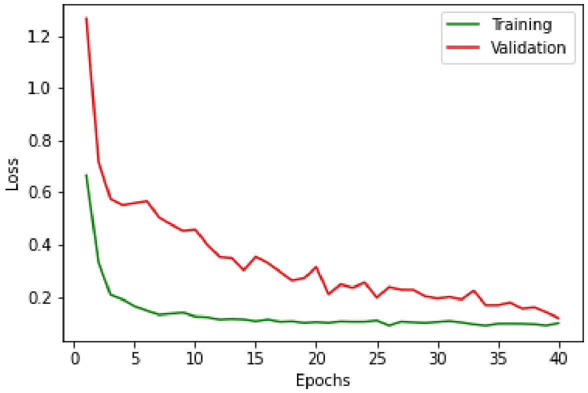

Fig. 6. Learning curves showing the loss on the training and cross-validation sets for the recurrent neural network model.

The RNN is trained for forty epochs, and in order to asses it's accuracy, the training loss and validation loss are calculated. The values throughout the training process are displayed in Figure 6. A training curve indicates a good fit for the model when the training and validation curves converge close to each other. Figure 6 shows the training curves converging, with a final training loss of 9% and a final validation loss of 12.6%. The relatively small dataset used to train the model may have affected the time it takes for the RNN to fully converge, so more data may be necessary to fully optimise the model.

VI. Results

In general, the approaches produced results that closely match reality. Historically similar countries are grouped roughly together using the clustering methods, such as Germany & Austria and Spain & Portugal.

The countries present in each cluster are shown by Figure 7. A violin plot shows the distribution shapes of each of the three features in each cluster, with the width held constant.

In order to quantify the effect of all the features on productivity, the computed coefficients are converted into a ratio corresponding to the contribution of the feature on the GDP per hour worked of each particular country. To achieve this, for each separate country the magnitude of the coefficient was calculated as a function of the sum of absolute values

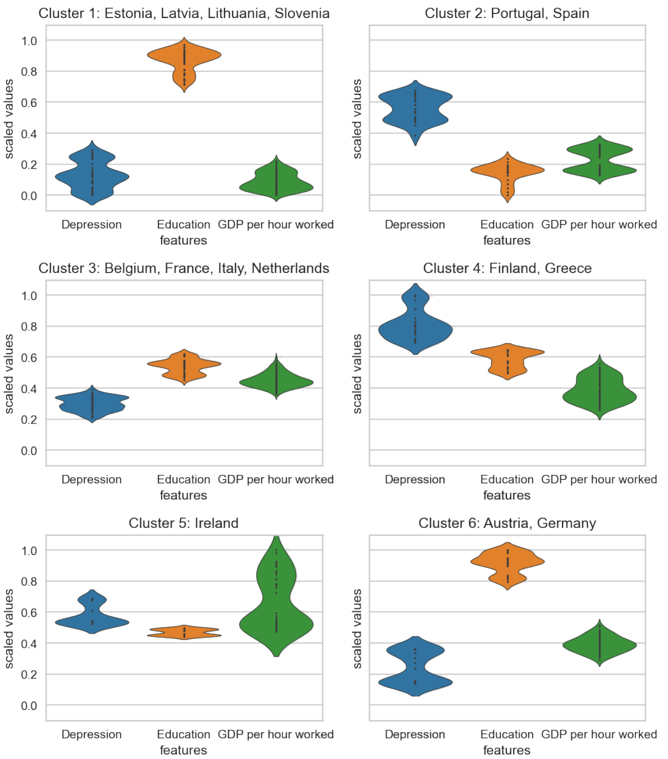

Fig. 7. Violin plots of the six clusters found by k-means using the three features found using the elbow method.

of all coefficients for that country. This procedure permitted the transformation of the coefficient of each feature into a percentage contribution towards calculating the dependent variable, productivity, specifically for that country.

Figure 8 illustrates the percentage importance of each feature in predicting the GDP per hours worked of each country as a result of performing Ridge regression on that country. The total length of each bar represents the average GDP per hours worked between 2008 and 2019 for that country. The colour-coded sub-bars in each country's row represent the extent to which the aforementioned key variables contribute to GDP per hours worked, and therefore productivity.

The random forest regression demonstrated that productivity could accurately be predicted from the selected features as shown in Figure 4.

The importance of different features for the random forest regression is shown in Figure 9. The graph shows the varying ability of reduced feature data sets to predict productivity.

The importance of each feature in the random forest regression is shown in Table IV by the reduction in accuracy. Therefore, for random forest regressions, less than basic education is the most important feature, since without this the loss is 4.5%. The least important feature is the depression rate as without this the loss is 1.2%, which is exactly the same as the performance of the random forest regressor on all the data.

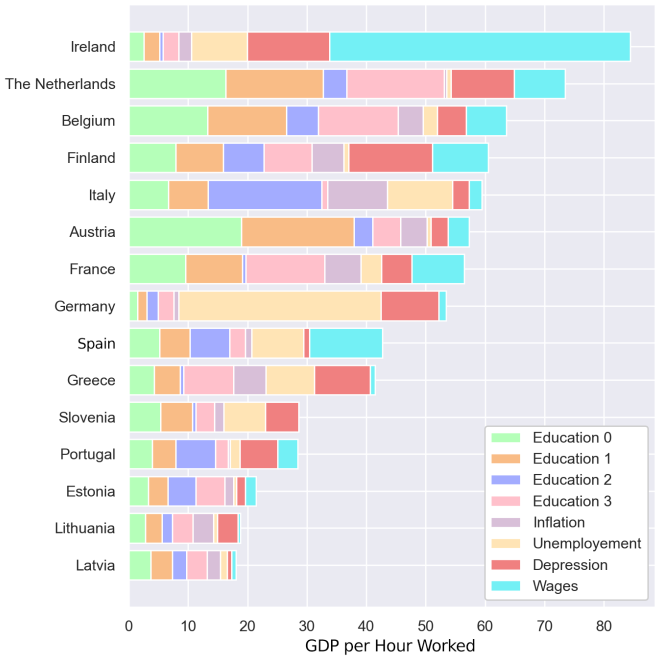

Fig. 8. Graph demonstrating feature importance with Ridge for each country analysed between 2008 and 2019.

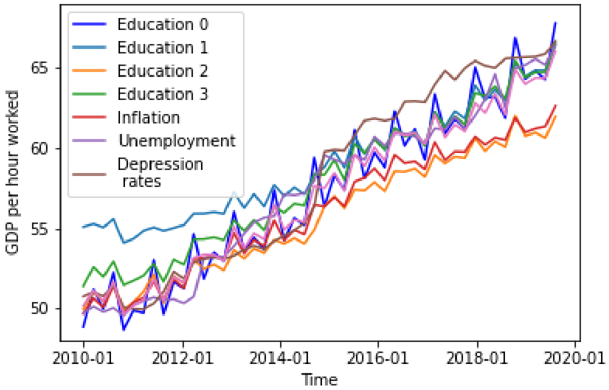

Fig. 9. The prediction of a random forest regressor when each variable is iteratively set to zero. As well as plotting the actual data for Austrian GDP per hour worked in Euros.

In order to predict future values of productivity, a SARIMA model is trained on the productivity data, as seen in Figure 10. The SARIMA model predicts an error bound which can be seen in Figure 11, where the error bound starts low, at only ±1.6% of the predicted value. However, the errors compound, giving a larger potential range of future values. At three years, it is ±6.4% of the predicted value, and after twelve years, this rises to ±24.6%.

The results of training the RNN on the Austrian data are shown in Figure 12. The predictions capture the trend in productivity data well for the training set. However the predictions do not capture the seasonality of the data as well. The future predictions show a levelling off in productivity

TABLE IV
THE LOSS FOR EACH FEATURE FROM THE PREDICTION OF A RANDOM FOREST REGRESSOR WHEN THE VARIABLE IS SET TO ZERO WHEN TRAINED ON AUSTRIAN DATA.

Feature	Loss
All features	0.012
Education 0	0.045
Education 1	0.031
Education 2	0.021
Education 3	0.023
Inflation	0.022
Unemployment	0.029
Depression	0.012

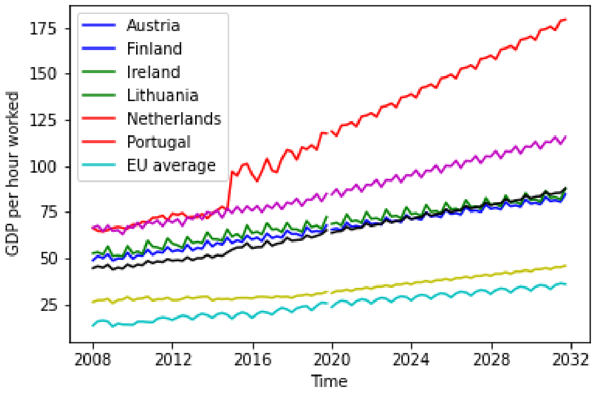

Fig. 10. A SARIMA model for time series prediction on the GDP per hours worked data for the selected countries and the EU average. The actual data is displayed from 2008 to the end of 2019 and the predicted data is for the 12 years following till the end of 2031.

rather than continuing on the past trend.

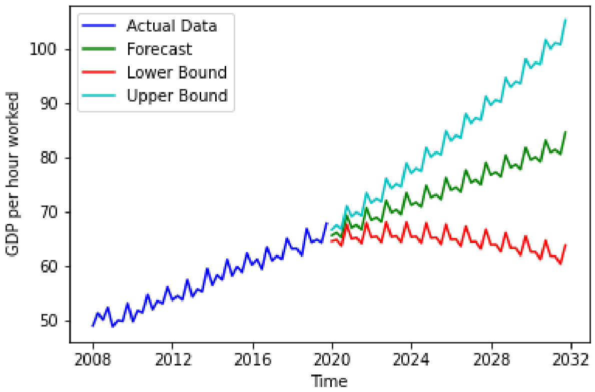

Fig. 11. The SARIMA model ran on the Austrian productivity data to predict 12 years into the future. The figure displays the actual data the forecast data and then the upper and lower bounds given by the SARIMA model.

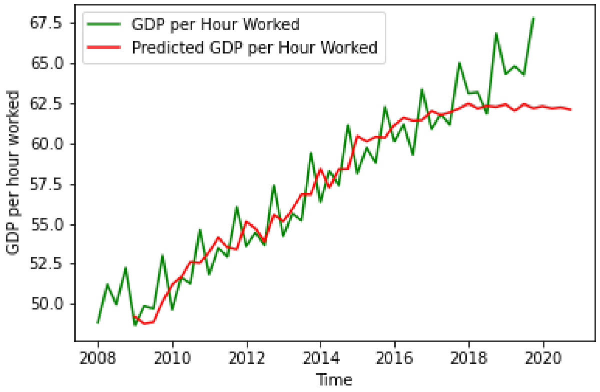

Fig. 12. A RNN predicting the GDP per hours worked in Austria from 2009 to 2019 and the actual GDP per Hours Worked in Austria from 2008-2019

VII. DISCUSSION AND CONCLUSION

The clusters found using GDP and the education and depression features seem to correlate strongly with real life. Germany and Austria are grouped together, and so are Spain and Portugal. Estonia, Latvia, Lithuania and Slovenia are put together, and these countries also have many similarities. The third cluster of Belgium, France, Italy, and the Netherlands may appear slightly unusual culturally, yet these countries are all similarly located geographically, and all each share a border with another.

An unusual cluster is that of Finland and Greece. There don't appear to be many similarities between these two countries broadly speaking, so it is reasonable to believe that this cluster is comprised of countries which didn't fit into other clusters. Further work could be done to see how the clusters would be affected by adding more countries to the dataset.

The final cluster of, however, consists exclusively of Ireland. This could be due to the fact that Ireland is home to many multi-national corporations as a tax haven, artificially inflating its GDP above what would be considered 'normal' by the clustering algorithm for its average education levels and depression rates.

These observations can be affirmed by the Ridge Regression results in Figure 8, where wages are the major contributor towards Irish GDP. Additionally, the factors affecting productivity in Estonia, Latvia and Lithuania demonstrate very similar composition, with Slovenia not far off. These four countries are countries with weak economies in which, as illustrated, education is almost the only factor affecting productivity out of all the features. According to Eurostat publications, these countries have a below average educational level, relative to other EU countries, which could explain their relatively low productivity [15], [16].

Also interesting is the geo-spatial distribution of countries with features contributing in a similar way to productivity. It appears that in countries that are close together, feature importance is similar. Take Latvia, Lithuania, and Estonia for example again. These countries have similarly low levels of productivity, and almost the exact same features affect productivity, and all of them share borders. The same pattern is observable with Belgium and the Netherlands, as they have a relatively high value of productivity and features of education, depression and wages contribute similarly.

Other insights include that in Germany, which is an advanced economy with relatively high GDP, unemployment is a feature largely affecting its productivity. Germany has kept one of the lowest levels of unemployment throughout the years, signifying that one of the key factors driving its productivity is utilising almost all of its available workforce resources.

An important distinction to be made is that the percentage contribution of the features in Figure 8 does not convey if the effect of that particular feature is negative or positive on the productivity. Instead, it only discloses how much the feature has affected the GDP in comparison to the other features, which could either increase it or decrease it. This is likely why in Figure 8 for most countries the magnitude of one level of education is much larger than the others. Nevertheless, the effect of the labour force education level on the productivity of a country is evident, as it comprises a significant percentage of each country's productivity. This finding is supported by prior research stating that education is a key human capital indicator and that an educated workforce is a highly valuable factor that increases productivity [17], [18]. Therefore, educational attainment of the workforce is one of the most important contributors in productivity growth.

With regards to the methods developed to predict future productivity, SARIMA excels at predicting the periodic and predictable information in the Austrian productivity dataset. The results show that European productivity is on an upwards trend for each of the clusters identified. However, there is scope in the error bounds for productivity to fall or grow at almost double the rate predicted. The potential for faster growth puts an impetus on the European Union to improve the underlying factors of productivity growth that are outlined in this report. However, the implementation of SARIMA only uses the existing productivity data to predict future productivity data, and so may not be capturing any underlying trends or events in the feature data. However, it is known that events like natural disasters or pandemics severely affect the socio-economic status of civilians, and these are nearly impossible to predict. The upper bound and lower bound given by the SARIMA model, as shown in Figure 11, provide upper and lower estimates for productivity growth in the European Union. The factors that affect the trajectory that productivity continues on are then determined by the features.

The RNN is implemented to predict the future productivity using the features, as shown in Figure 12. The RNN captures the trend in the training data as seen in Figure 12, however it fails to capture the seasonality. In order to improve the RNN the model could be trained on the decomposed data or more seasonal features could be obtained or the data could be standardised for hours worked in different seasons. Therefore the RNN does not fit the data as well as the SARIMA model, however the RNN has the capability to detect the future impact of the features selected on the productivity. The future impact of different features on productivity will be of vital importance in working out areas to improve to improve overall productivity. The RNN is only training on 336 data points, which is a very small data set for a neural network. To improve the RNN it could be trained on all the countries in the data set at once to give 5040 data points to train the model on.

In conclusion, the productivity of the EU is at a crucial point in history, with a potential for stagnation or a period of rapid sustained growth shown by the upper and lower bounds for the SARIMA model in Figure 11. The random forest regression predicts the GDP per hour worked with an average loss of 1.29%. The SARIMA model provides upper and lower estimates for productivity growth. The RNN achieves a training loss of 9 % and a validation loss of 12.6 %. In order to unlock this growth the EU must tackle the problems associated with the underlying features affecting productivity detailed in this report.

References

[1] N. Patel and A. Villar. Measuring inflation. *Inflation Mechanisms, Expectations and Monetary Policy BIS Paper No. 89*, 2016.

[2] Andre Violante. An introduction to t-sne with python example, Aug 2018.

[3] Binu Melit Devassy and Sony George. Dimensionality reduction and visualisation of hyperspectral ink data using t-sne. *Forensic Science International*, 311:110194, 2020.

[4] Diana Lin and Sampath Jayarathna. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6):441–458, 1996.

[5] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[6] Sebastian Raschka, Yuxi Liu, and Vahid Mirjalili. Machine learning with pytorch and scikit-learn, 2022.

[7] Sebastian Raschka. *Python machine learning*. Packt publishing ltd, 2015.

[8] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[9] Patrik Waldmann, Gábor Mészáros, Birgit Gredler, Christian Fuerst, and Johann Sölkner. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics*, 4:270, 2013.

[10] SE Lazic. The elements of statistical learning: Data mining, inference, and prediction, 2nd edn.

[11] A. M. E. Saleh, M. Arashi, and B. G. Kibria. *Theory of ridge regression estimation with applications (Vol. 285)*. John Wiley & Sons, 2019.

[12] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[13] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[14] Hong-Kun Xu. Properties and iterative methods for the lasso and its variants. *Chinese Annals of Mathematics, Series B*, 35(3):501–518, 2014.

[15] Lene Mejer, Paolo Turchetti, and Eric Gere. Trends in european education during the last decade. *Statistics in focus*, 54:8, 2011.

[16] Stanislav Ranguelov, Isabelle De Coster, Sogol Norani, and Giulia Paolini. *Key Data on Education in Europe 2012*. ERIC, 2012.

[17] Alistair Dieppe. *Global productivity: Trends, drivers, and policies*. World Bank Publications, 2021.

[18] Wolfgang Modery, Paloma Lopez-Garcia, Maria Albani, Claudio Baccianti, Rodrigo Barrela, Katalin Bodnár, Jan De Mulder, Beatriz Lopez, Vincent Labhard, Julien Le Roux, et al. Key factors behind productivity trends in eu countries. 2021.

Reflective Discussion

Nisa Bayraktar
da19157

I. Contribution To The Project

Throughout the project, I worked on multiple parts and stages. Our aim in this project was to find the most accurate GDP prediction using socioeconomic and national features. Therefore, initially, we did data collection to investigate features that have an impact on productivity and check their correlations with the GDP dataset which was our productivity measure. The features I found the correlations with productivity were inflation and depression. I collected these features' datasets from Our World In Data and OECD databases. After determining the correlation, I did data wrangling to prepare the inflation and depression datasets for the next step of the data preparation which was combining all features into one data frame. The data wrangling process was matching the attributes of the datasets, countries, and years, then putting them into a new data frame. Once the data was prepared, we decided to implement regression methods to establish the importance of each feature on productivity and overcome the overfitting in our combined dataset. I implemented the Ridge Regression on the data after normalising it and found its accuracy to compare with the other regularisation methods (Lasso and ElasticNet) and choose the most accurate method's coefficients. We visualise the feature's importance by creating a bar chart that is inspired by World Happiness Report using the Ridge regression's coefficients which was the chosen regression. When doing this bar chart, I helped with finding the best alpha value of the regression which represents the best coefficient results. Then, I adjusted the bar chart's appearance to make it tidier and more understandable. In the end, I explained all the parts that I mentioned above in the report and presentation as well as edited our presentation video for the submission.

II. Strengths & Weaknesses of The Project

A. Strengths

The most crucial part of our project was coming up with a topic to investigate the GDP dataset and its trend. Unfortunately, the GDP dataset that we initially have was not enough to find a trend to explore the data and do research about it. Therefore, finding more datasets to consolidate the given GDP dataset and doing a deep exploration using many techniques such as regression, clustering, and machine learning methods were the key strengths of our project. To apply all these methods, we needed to do a data wrangling to match the different feature datasets that we found and combine them in a short-time period. This process was essential to make a start on our project and we managed to do it in a short-time period which can be considered a success. Additionally, we did various visualisation of the results of these methods to make it easier to explain the trends and find a relation between them (finding similar patterns). Once we visualised the results, we also did deep research to compare them with the recent research and discuss the relationship between the findings and the results.

B. Weaknesses

Even though we found extra datasets our data points were still not enough to implement some methods to get more realistic and accurate results. Therefore, we put the ideas which could have been done to improve the results in the conclusion section of our report. Not many data points arose from the difficulty of finding datasets about socioeconomic and national data that might have an impact on GDP. Throughout the project, we were searching very unique datasets with specific time-period and countries to match our GDP dataset. Thus, these datasets were too exclusive to find in the database sources even though I contacted the database websites to ask about these data. The lack of source limited our time to improve our results and apply more methods because searching these datasets were time-consuming.

III. Project Relation With The Unit Content

In the Data Science unit, we learned about how to explore and analyse the data with diverse methods. The European Productivity project allowed me to apply these techniques such as data wrangling, exploration, preparation, and visualisation to the real-world problem following the data science pipeline.

IV. Achievements & Project Experience

In this project, we worked as a group with students who have different academic majors. As a computer science student, this collaboration helped me to widen my knowledge about the mathematical side of data science by discussing and implementing ideas with my teammates who do Engineering Math. Moreover, working as a team also improved my communication skills because I needed to express my thoughts clearly to my teammates and my supervisor. In my opinion, we implemented many techniques to explore our data and find patterns, by combining completely different datasets. Furthermore, we linked our results together and explained them as clear as possible by doing many visualisations to end up with a sensible conclusion. We could be able to do all this in a short-term period even though we spent time finding all our datasets by ourselves, this may be considered as our achievement.