National College of Ireland

Evaluating the Prevalence and Effects of Disguised Unemployment in Ireland

MSc Research Project
MSc in Data Analytics

Sriram Rajgopalan
Student ID: x23213876

School of Computing
National College of Ireland

Supervisor: Jaswinder Singh

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: …Sriram Rajgopalan….

Student ID: …x23213876……

Programme: …Msc in Data Analytics …… **Year:** …2024……..

Module: …Msc Research Project……

Supervisor:……Jaswinder Singh……

Submission Due Date: ……12/12/2024……

Project Title: …Evaluating the Prevalence and Effects of Disguised Unemployment in Ireland……

Word Count: ……5931…… **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: …Sriram Rajgopalan……

Date: ……12/12/2024……

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):	

Evaluating the Prevalence and Effects of Disguised Unemployment in Ireland

Sriram Rajgopalan
x23213876

Abstract

"Unemployability" - a situational hazard in Ireland, that the majority of the population is experiencing currently. The idea of unemployment changes when the absolute factors that influence an individual's employability are considered. This is because the current employment situation of an individual in Ireland is mostly based on employable indices such as age, gender, skills, nativity, education and nature of work. When these factors are collectively analysed, it explains the fact that an individual's potential is not completely utilised but remains employed. This phenomenon is known as "Disguised Unemployment". This study is undertaken to evaluate various employable factors of an individual, using five different datasets, by understanding the distribution of unemployability due to each of the employable indices (One index per dataset). Once the trend of unemployment is understood, an ensemble technique "Stacking" using "Random Forest" and "SVM" as the base model and "Logistic regression" as a meta-model is implemented to identify how the state of unemployment is considered disguised. The model's outcome is evaluated via a confusion matrix by classifying every dataset individually and the key predictors are discovered using SHAP (SHapley Additive exPlanations). These key predictors are explained collectively in the end, to justify how unemployment is disguised among various factors considered for this study.

Keywords: Ireland, Disguised Unemployment, Stacking Classifier, Logistic regression, Random Forest, SVC, SHAP

1 Introduction

A demand to address the growing labour market inefficiencies in Ireland has spotted the limelight towards discussions on unemployment. This has led to a situation where the qualitative increase in employment is not helping the quantitative increase in unemployment. "This is known as Disguised Unemployment". This market situation starts prevailing when an individual being employed contributes very little or no productive work in comparison to the actual potential that they are employed for. This situation can be related by simply understanding of numbers of the Irish economy as provided by (CSO, 2020) where a 75.3 per cent growth in employment with 25,500 people by the end of 2023. Ironically there is a 4.5 percent increase, that is 117,700 people were unemployed in the same period. This could mean that either there is too much competition in the job market to fit in or it means that there is a huge discrepancy between the available employment opportunities and an individual's absolute employability, which is often overlooked.

1.1 Research Motivation

This research is motivated in a certain way such that even so there are lot of studies regarding unemployment, there are always lack of evidence to prove how exactly it is hidden even amongst the employed sectors. In today's research availability, a huge gap could be identified into creating practical experiments to understand the existence of disguised unemployment, as there is more of theoretical insights rather than practical interdisciplinary approaches. It becomes important to understand the multi dimension of this topic which could potentially intricate standard metrics into the study. Such situations are labour statistics, skill level statistics, migration statistics, education statistics and sector-based statistics. While the actual historical studies relating to this topic are independent, the main objective is to bring all the macro-economic factors that influences employment together and have them represented as a single case study.

1.2 Research Question

How exactly has unemployment disguised itself within a country's workforce in today's modern well-being and what are the factors that influence them?

1.3 Research Objectives

The main objective of this research is to bring together a group of factors that are relatable to unemployment and how exactly they are hidden in the concept of employment. As contradicting as it sounds, this is very important for sectors such as academia, Finance and Government organisations to know the actual quotient of the usable skilled labour force. While obviously addressing the gap in this pressing issue, it is notable that there are certain compelling works that have actually provided insights. (Gomis, R., Kapsos, S. and Kuhn, S., 2020) have a highlighted study on how there is a job quality decline by indicating lack of salary, benefits and limited opportunity in skill development. (Horne, R., 2024.) have an opinion on underemployment mainly on rural-urban dynamics where data analytics have been used in order to evaluate the extent of hidden unemployment. It is also to be noted, studies such as (Harris, J.R. and Todaro, M.P., 1970) clearly emphasises on role of technology, urban transition, employable structural inequalities and evolving job demands. An Ensemble machine-learning technique "Stacking Classifier" is used to study the topic and draw out conclusions on how the prevalence of "disguised unemployment which will be discussed in detail in the fourth coming sections.

1.4 Report Structure

With the above-mentioned objectives, the following of the research study is undertaken consisting for following distinct sections

- **Section 2, Literature Review:** Provide a thorough critic about the traditional studies undertaken, analysing the gaps involved and how a new study could be implemented for better traction of the existing literature on unemployment.
- **Section 3, Methodology:** With the insights from the literature review, a suitable data is selected to undertake the research and the same is prepared in order to make it experiment ready. An architectural framework is prepared in order to make the run experiments with the data
- **Section 4, Design Specification:** With the pre-processed data, the experiment is carried out on the selected machine learning technique by implementing them using various libraries and setting them up for evaluation of metrics.
- **Section 5, Evaluation:** A proper evaluation of the model is conducted, and a similar statistical comparison of results is done in order to measure the achievability of the objective. Apart from that, an honest discussion of the outcomes and missed elements of the experiments are also discussed in this section.
- **Section 6, Conclusion:** Final concluding remarks on the success of the experiments are quoted with recommendations for future research on the topic.

2 Related Work

A phenomenon where a country recognises a qualitative increase in employment but that does not stop the quantitative increase in the unemployment is considered disguised unemployment. Ireland having undergone various economic changes in the past decade have given room for study of vast labour market dynamics using machine learning models. This literature review explains how various studies have been implied to understand the concept of research study on disguised unemployment.

2.1 Traditional approaches of research on unemployment

It is very important to understand the economic indices and its functions to predict the unemployment rate. There were traditional approaches such as ARIMA (Auto Regressive Integrated Moving Average) and SARIMA are commonly used in forecasting unemployment rates. These methods focus on the effectiveness of predicting cyclic trends of the past unemployment data and will predict the future trends. The focus on using these approaches is to have a decent accuracy which is one of the main factorials of any time series method to indulge in prediction. This was pretty much seen in the study of unemployment using SARIMA combing with Leading Economic Indicators (LEI) where the study showed immense improvements in the accuracy in comparison with simpler random walk models (Guerard et al., 2020). With seasonal trends being one of the main elements in prediction of unemployment rate the time series model approaches can be extended by introducing other specific or pointed predictors such as weekly unemployment rate, unemployment claims coincidental and lagging indicators like previous economic conditions, GDP growth, inflation, mortgages are all integrated to increase the prediction accuracy. Later these functions become base line for the usage of transfer function models that would include various economic factors to point towards the outcome (Xiao, H.,; Chen, R.,;Guerard Jr, J.B., 2022). With the above being said there are

other macroeconomic factors such as tax rates, industrial production, saving rates, and trade growth which corelates more with the unemployment. Also, labor market flexibility and tax policies are also influencing unemployment as per the study made by Berger and Everaert (2013) that only affected the European countries. The export and import growth are also macro-economic factors that influence the unemployment trends as discussed by Karanassou et al. (2008) and Wesselbaum (2015) where regression models are used in prediction in order to obtain interpretable results.

The main disadvantage of these traditional techniques is that there are lot of dependencies on the features that are involved in the model. For instance, time series models like ARIMA, SARIMA always assume that the relationship between the features are same considering different data is being used. However, there is a lot of possibility that the market can be extremely dynamic and unpredictable and so the relationships need not be linear (Box et al., 2015). This brings to the next topic of inconsistency in the modeling because of lack of handling of non-linearities among the features, for example GDP growth and inflation rates. Where neither of the features are necessarily measured in the same scale. Speaking of the scaling of the data, since the data recorded might be potentially inconsistent, proper preprocessed data is required for time series prediction. However, this becomes close to impossible because the data might be missing entries. This leads way to the lack of adaptability of the model since the data can have sudden changes when accumulated over a greater period (Mitchell, 1913). For example, the trend in unemployment saw a very steep increase due to the recession times recorded in the year 2008 and traditional models clearly fail to record these because of the drift in the labour market (Canale & Liotti, 2015). This is one of the reasons where any multidimensional aspect of unemployment is simplified around very limited resources such as GDP and inflation negating other dynamic labor factors that contribute to the core of unemployment.

2.2 Addressing gaps in traditional approaches

Capturing non-linearity is one of the most common gaps that could have been observed in the traditional approaches involving time series and regression models. ML models such as Random forests and Gradient boosting machines have good capability at capturing complex patterns and non-linearity of previously discussed macro-economic variables by detecting their feature interactions with each other (Gabrikova et al., 2023). Apart from the non-linearity, a wide range of datasets are expected to be involved in the analysis of unemployment because not all the economic indicators are relatable to each other and can be found on a single dataset. Understanding various data patterns club, them is integral and figuring out how exactly they make sense rather than relying on static information that usually the time-series models rely on. This dynamic nature of the dataset is found more suited for model training of techniques like random forest (Krishnamurthy, 2023) and also helps in imputing missing values which creates a huge deal in model training of datasets that becomes erratic with information over time (Berger & Everaert, 2013).

2.3 Ensemble Techniques for research on unemployment.

Stacking is one of the Ensemble Techniques which uses the advantages of multiple models to arrive at a conclusive prediction. A stacked model predictor of random forest, Support vector classifier and Logistic regression creates a framework that uses both classification and regression tasks, refer Figure 1. This way, there is a distinct influence in influencing both accuracy of unemployment prediction and understanding the feature importance in the analysis (Annals of CSIS 2023). The logistic regression reduces biases and understands the involvement of subtle patterns such as age, education, gender and skill level that are independently involved with the employment quotient of the dataset. It is also easily interpretable and usually takes its predictions from the base learners on the dataset by using models like RF and SVC in this context which provides adequate weights to reduce errors in predictions (Tarullo, G. and Desiere, S., 2023). There is another advantage of choosing the stacking approach with logistic regression because of the involvement of multiple datasets. It could be close to impossible to have all the datasets involved in a single prediction by, say for example merging them mainly since the macroeconomic factors on each of the datasets might not be measured with the same scale of measurement. Stacking allows computing results from different models and provides interpretability with techniques like SHAP (SHapley Additive exPlanations) which gives insights into actionable predictions on every model individually (AIFSG, 2024). This way it becomes much simpler to justify the evaluation of the models into bringing one conclusion of how unemployment behaves and is disguised with different economic factors making the study more reasonable.

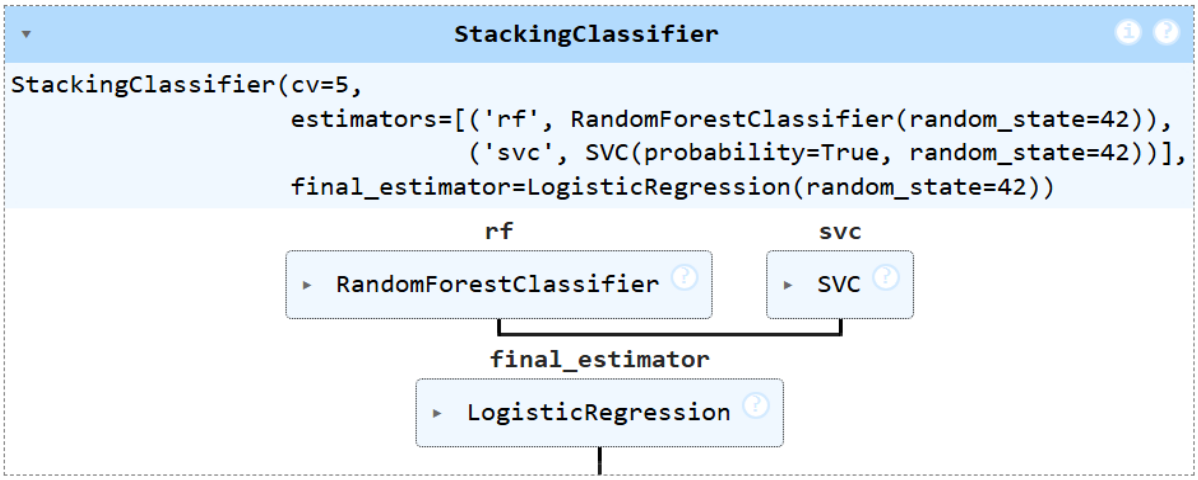

Figure 1: Pipeline of Stacking Classifier

3 Research Methodology

The primary goal of this objective is to individually understand how various employable factors are influencing unemployment and how can one collectively analyse the prevalence of disguised unemployment by collectively studying every result. This study involves an ensemble technique of three combined machine learning models which are random forest and support vector classifier used as base models which are used to capture non-linear relationships that are present in the employment pattern that are influenced by age, skills, education individually and collectively. The output of the base models is taken into consideration for the meta model which implies a base line to identify linear trends and key predictors by implying logistic regression. The meta model leverages the ability to identify outliers which help identify

subtle patterns by using random forest, the ability to map the data by reading into non-linear relationships which creates an overall effective marginal decision boundaries and logistic regression to explain the general patterns.

This research outlines an understanding of five different employable indices considered for study as illustrated in Figure 1, and how the individual outcome of each study helps in the collective understanding of the concept of disguised unemployment. The model design depicts the flow undertaken in this study, where 5 different datasets each considered for one employable index such as origin, skills, working age, gender, education level, nature of employment are collected, and they are feature engineered against the unemployment quotient such that all the datasets labelled in the same terms. The model design also gives a certain replica of CRISP-DM methodology where the modelling is done one by one and after reviewing the outcome of each and every model of the dataset, the outcome of the current model is compared with that of the previous model's outcome in order to understand how the unemployment is pitted against the selected employable indices using SHAP as the tool of measure for feature understanding as illustrated in Figure 2.

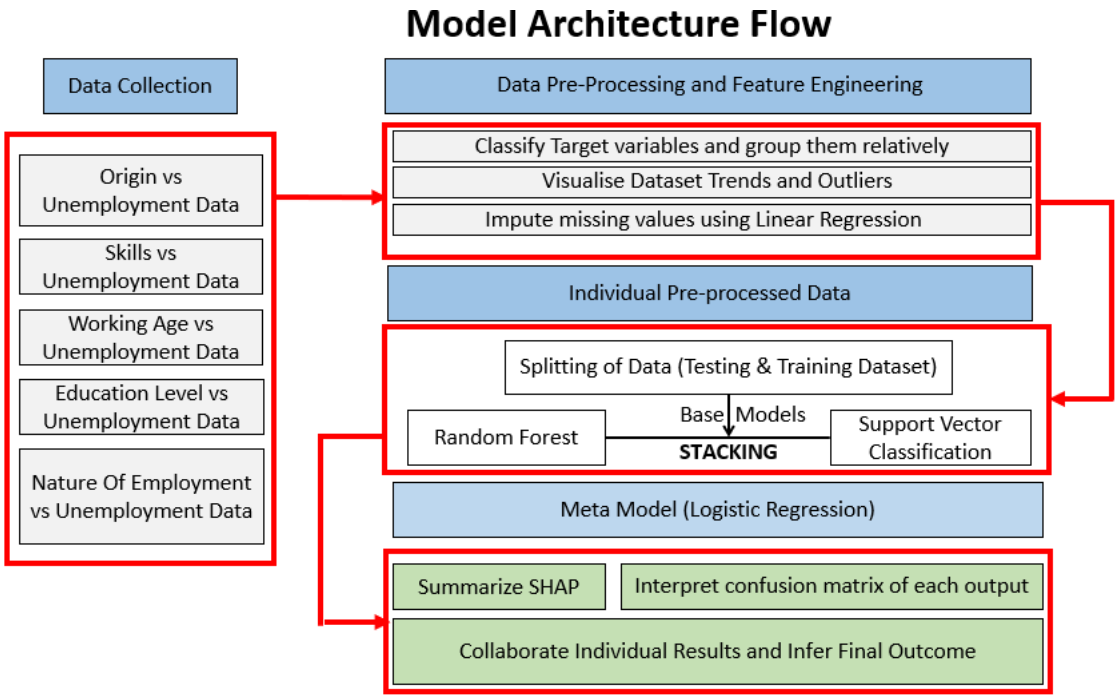

Figure 2: Model Architecture Flow

3.1 Introduction to Datasets

The study undertaken is based on the data from five different datasets taken from two different sites. The ILOSTAT[1] provides a piece of comprehensive information specific to unemployment trends that are influenced by a certain factor such as age, gender, nativity or skills,

[1] https://ilostat.ilo.org/data/#

Every dataset is discussed as follows.

3.1.1 Unemployment by Age and Gender

This dataset with 128364 x 12 records provides figures of unemployment segmented by sex consisting of male and female, age groups consisting of youth (15-24 years), Prime working age (25-54 years) and older workers (55+ years). This dataset is selected to understand the labour spread in Ireland pretty much involving the age bands.

3.1.2 Unemployment by Education and Gender

This dataset with 3168 x 12 records provides figures of unemployment segmented by sex consisting of male and female, education groups consisting of Basic education, Secondary education, and Tertiary education. This dataset is selected to understand the underutilisation of labour spread working in jobs which are not equivalent to their skill level.

3.1.3 Unemployment by Origin/Nativity

This dataset with 28082 x 12 records is one of the most important datasets that provides information on the utilisation of the skillset across areas and its relevance to unemployment. The skillset is segmented into figures of unemployment segmented by sex consisting of male and female, age groups and education levels similarly categorized as the sections 3.3.1 and 3.3.2. The dataset also contains segments depicting citizenship consisting of foreign or natives whose age and educational background are taken as factors to understand the likelihood of being unemployed or being available in low productive jobs.

3.1.4 Unemployment by Skills

This dataset with 41024 x 11 records is selected to understand the labour spread in Ireland pretty much involving the age bands, their corresponding skills and their average hourly earnings in relevance to that of corresponding skill set. A total of 9 different skillsets are represented and each of them are grouped/classified into 3 different skill sets. This dataset is very important to understand how skillset and it's corresponding pay scale need not necessarily relate, indicating one of the main factors of disguised unemployment.

3.1.5 Unemployment by Nature of Employment

This dataset with 4494 x 6 records provides information on categories in the society that are sub divided into sectors of unemployment based on the nature of their economic status. A typical segmentation of the data would be with respect to Employed, Unemployed and disguised unemployed. This dataset provides a critical dimension to the shift in the economic roles have led to underutilization of labours. It also helps in identifying the proportion of people who might be employed but are typically unemployed. This helps in measuring hidden unemployment.

3.2 Data Preparation and Transformation

Each of the datasets is analysed and ensured that they are in the proper format before running the model. Since different datasets are used it is important to clean the dataset into a similar scale that is used throughout, so that it is easy to compare. Each of the dataset features post-cleaning is discussed as follows.

The data taken consists of descriptor labels such as country information, data source information, and measurement labels such as currency information or percentage units that are mostly related to the absolute value of the data. These descriptor labels are repetitive and do not serve any purpose of the model and so they are removed as a part of preprocessing. A separate label matrix of classification is followed for the education, skillset and nature of employment dataset as per the ILOSTAT standards as seen in the Figure 3. This matrix helps in systemizing using the same labels used for classifiers across.

The features are checked for the datatypes and it can be understood that the data type is not stationary, i.e. string and object data types can be seen. Categorical variables cannot be used for SVC or random forest and so the feature variables are encoded. Label encoding is used for features like education, skills and age because the distribution of values among the features is considered ordinal.

Each dataset is identified for missing values and is handled according to the nature of the dataset. For instance, the datasets that relate to age, skills and nativity concerning employment do not have any missing values against the selected feature records. Still, the datasets relating to education and employment nature have certain missing values against the selected features for prediction. To handle the missing values, a linear regression is used to predict the missing values because it helps in capturing patterns in the missing data. Moreover, the features that are involved in the regression to impute missing data are typically age, gender and time across all three datasets. This way there is uniformity in imputing the missing data across all individual datasets.

Data outliers are a component of unemployment, because the employability is expected to be erratic. The intention of this research project is to understand the erratic nature and classify the grey section of unemployment without any influence. Hence, the trends in unemployment and nit-picking outliers are ignored as a part of data preparation.

Education Level Classifiers

S.No.	Label	Classified Keyword
1	Levels 1-2	Primary and lower secondary education
2	Levels 3 and 4	Upper secondary and post-secondary non-tertiary education
3	Levels 5-8 (Levels 0-8)	Tertiary education

Employment Skill Classifiers

S.No.	Label	Classified keyword
1	Managers	Skill levels 3 and 4 ~ high
2	Technicians and associate professionals	Skill levels 3 and 4 ~ high
3	Professionals	Skill levels 3 and 4 ~ high
4	Clerical support workers	Skill level 2 ~ medium
5	Service and sales workers	Skill level 2 ~ medium
6	Craft and related trades workers	Skill level 2 ~ medium
7	Plant and machine operators, and assemblers	Skill level 2 ~ medium
8	Elementary occupations	Skill level 1 ~ low
9	Armed Forces Occupations	Skill level 1 ~ low

Employment Type Classifiers

S.No.	Label	Classified Keyword
1	In employment part-time - underemployed	Disguised Unemployment
2	In employment part-time - not underemployed	Disguised Unemployment
3	Unemployed seeking full-time work/future job-starter	Disguised Unemployment
4	Unemployed seeking part-time work	Disguised Unemployment
5	Potential additional labour force	Disguised Unemployment
6	In labour force	Relatively Employed
7	All ILO economic status	Relatively Employed
8	In employment	Relatively Employed
9	In employment full-time	Relatively Employed
10	In employment part-time	Relatively Employed
11	Unemployed	Unemployed
12	ILO Economic Status	Unemployed
13	Not in labour force	Unemployed
14	Others not in labour force	Unemployed
15	Unemployed seeking work as self-employed	Unemployed

Figure 3: Label and Classified Keyword as per ILOSTAT

4 Design Specification

In order to understand the underlying question on the prevalence of disguised unemployment, an ensemble technique of stacking is used in order to create base models using, random forest

classifier and support vector classifier technique where the base model once evaluated is led into the meta-model to obtain the final result. This is carried out on all five different datasets each having a unique employable index as target variable such as age, skills, nativity, education and employment type individually and the model evaluation is verified by understanding the SHAP results. This is done as a technique to interpret the prediction against each, and every dataset in order to understand how the employable index is contributing to the final prediction. This is then compared with against outcomes of all other models in order to evaluate how unemployment prevails in a disguised manner.

The employable index that was discussed in the previous section is taken as the target variables for every dataset respectively and the common predictors were identified across the datasets to maintain uniformity for evaluation and also to provide a fair comparison. With the output from the SHAP values a percentage comparison is made to understand the impact of the features against the target variables.

4.1 Assumptions made based on Feature Engineering

Based on the model's systematic approach and the feature engineering done in the earlier sections the following assumptions are made as the framework of the design.

- The data is skewed, non-static and non-linear.
- The timeline used to capture data is independent to each dataset and does not have any co-dependence with the other data.
- There are no notable missing values, and the imputed values (imputed using linear regression) are handled in the feature engineering based on the analysed pattern of values that were missed.
- The target variables are correctly labelled, and the nomenclature of labelling carried across the datasets are same and the distinction between the features is unambiguous.
- No Categorical variables are employed in modelling and if at all present, they are encoded (label encoded) due to the ordinal nature.
- The model assumes that there are no external biases influencing the dataset apart from the features implied in the model learning and also assumes that the features implied are sufficient enough for predicting the likelihood of disguised unemployment.
- The design does not assume any balanced distribution of the data, since the original model is intended is an Ensemble and it is expected to handle the imbalance in the data.
- The usage of base models (Random Forest and SVC) encourages complementary points of view on the feature association with the data, balancing the complexity of non-linearity in the dataset and reducing overfitting. Hence evaluation technique like cross validation is not employed.
- The predictor variables in question are less and the same across all five datasets. Hence feature importance score and SHAP plot summary is considered the final evaluator to draw upon conclusions.

- No Corelation matrix is used due to lack of features (only 4 employable index taken per dataset).

4.2 Modelling

Timeline, Age, Gender, Education and a relevant absolute value that are measured as percentages, average salary or numbers in thousands are considered common features across all the datasets. The data is split into training and testing 80:20 respectively for modelling.

4.2.1 Base Learners

Random forest classifier is used for to handle non-linear relationships and outliers possessed by the datasets by analysing feature importance in order to understand the primary factors of disguised unemployment. SVC balances the ensemble by handling the overfitting and interpretable linear trends. Overall, the base learners enable non-linear learning capabilities.

4.2.2 Meta Model

Logistic regression is used to predict combining predictions from RF and SVC as input features to the meta-model. The meta-model implies a weighted contribution of the base models optimizing final prediction. This is known as stacking.

4.2.3 Metrics

- **Accuracy** measures the overall proportion of correctness (predicting disguised unemployment) among the classifiers.
- **Precision** to measure the overall proportion of correctively identified positive cases (identifying disguised unemployment) among the classifiers out of all the categorised positive cases.
- **Recall** to measure the actual proportion of correctively identified positive cases (total number of disguised unemployed individuals) overall in the taken sample.
- **F1 score** to realise the trade-off between actually employed people to that of the actual people whose has disguised unemployment. This also helps in understanding overclassification if any, that is ranking of more appropriately employed individuals than disguised unemployed individuals leading to better prediction.
- **Confusion Matrix** to measure the classifiers that are not falsely classified as employed individuals (True Negatives) and falsely classified as unemployed who are actually employed (False Negatives). The above situation of classification is pretty much considered “missed disguised unemployed individuals”.
- **SHapley Additive exPlanations (SHAP)** to measure the factors of influence that contribute to the best of the model’s prediction concerning disguised unemployment in comparison among the datasets. SHAP serves as the final decision-making measurement to gauge the achievability of this research topic.

5 Implementation

Following the discussions on the data preprocessing from the previous sections, all the datasets are implemented in the same way to maintain uniformity in prediction and to ensure continuity in the overall interpretation of the model. The implementation is carried out in the following steps.

5.1 Tools and Technologies

The following requirements for the software and library resources were utilized in order to obtain the results.

- **System type:** 64-bit Windows operating system, x64-based processor
- **Processor:** 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz 2.70 GHz
- **Programming Language:** Python
- **Integrated Development Environment (IDE):** Jupyter Notebook
- **Python Libraries/Modules:** pandas, numpy, sklearn, matplotlib, seaborn, shap

5.2 Implementation of Model

The data is prepared by splitting for training and testing datasets with 80:20 split. The base models are initialised as estimators that are Random forests and SVC tuned for probability estimates as true. This is to not have just the class label predicted (not flexible) but to have a confident prediction. This is the framework of the base model of stacking classifier. The prediction from the base model is fed into the meta model. Logistic regression is used for the predicting the meta model completing stacking classifier, to make final predictions. The stacking predictions are trained on the training data and after training the predictions are implemented in the test data.

5.3 Implementation of SHAP

A kernel explainer is created with the training dataset consisting of 10 clusters as sampling data. This typically reduces the complexity. The SHAP values are then computed for the test set to understand how each of the features have contributed to the model prediction. Post execution the SHAP plots are summarised so that the weightage of distribution can be seen in a feature matrix.

5.4 Implementation of other metrices

The test set is used as the data sample and the final predicted values are employed in the matrix. The classes are decoded using inverse transform to understand the true labels of the classes present in the matrix. The predicted features are also run through a classification report furnishing accuracy, precision, recall and f1 scores. The ROC-AUC curve is drawn for

multi-classifiers by binarizing the true labels for each of the classifiers to plot the true positives against the true negatives from the predicted values.

6 Evaluation

The datasets are evaluated separately as discussed in the previous section and their metrics are compared with each other. This section consists of subsections that focus on the comparison of classification reports, SHAP values, Confusion matrix, ROC-AUC curve analysis of the individual datasets and compare them to understand how they behave with the predictors and one another.

6.1 Stacking Classifier Experiment

A stacking classifier model is run as an experiment across all the 5 datasets and a classification report is printed and evaluated side by side on the scores of factors influencing unemployment refer Figure 4.

Classification report for Unemployment Education Levels

Stacking Classifier Accuracy: 0.8501577287066246

	precision	recall	f1-score	support
Primary and lower secondary education	0.85	0.86	0.85	105
Tertiary education	0.84	0.79	0.81	221
Upper secondary and post-secondary education	0.86	0.89	0.87	308
accuracy			0.85	634
macro avg	0.85	0.85	0.85	634
weighted avg	0.85	0.85	0.85	634

Employment Type

Stacking Classifier Accuracy: 0.9210233592880979

	precision	recall	f1-score	support
Disguised Unemployment	0.92	0.92	0.92	316
Relatively Employed	0.97	0.96	0.96	313
Unemployed	0.87	0.88	0.87	270
accuracy			0.92	899
macro avg	0.92	0.92	0.92	899
weighted avg	0.92	0.92	0.92	899

Nativity

Stacking Classifier Accuracy: 0.8793103448275862

	precision	recall	f1-score	support
Citizenship: Foreign citizens	0.79	0.90	0.84	21
Citizenship: National citizens	0.94	0.86	0.90	37
accuracy			0.88	58
macro avg	0.87	0.88	0.87	58
weighted avg	0.89	0.88	0.88	58

Skill Levels

Stacking Classifier Accuracy: 0.875

	precision	recall	f1-score	support
Skill level 1 ~ low	0.00	0.00	0.00	1
Skill level 2 ~ medium	0.84	1.00	0.91	16
Skill levels 3 and 4 ~ high	1.00	0.71	0.83	7
accuracy			0.88	24
macro avg	0.61	0.57	0.58	24
weighted avg	0.85	0.88	0.85	24

Age

Stacking Classifier Accuracy: 0.5872576177285319

	precision	recall	f1-score	support
Older workers	0.56	0.78	0.65	130
Prime working age	0.61	0.74	0.67	147
Youth	1.00	0.01	0.02	84
accuracy			0.59	361
macro avg	0.72	0.51	0.45	361
weighted avg	0.68	0.59	0.51	361

Figure 4: Classification summary

It can be noted that the accuracy of 4 of the predictors (Education, Skillset, Nativity and Employment Type) are considerably above 85 per cent overall which can account to the fact that the model is correct 85% of the time on average. It can also be noted that the Precision and recall of the classes of the above four predictors are complementing and are substantially over 75 percent.

Among the classes, the predictor “Age” seems to have lower accuracy (58.7%), as highlighted in Figure 1. It is to be noted, as highlighted in Figure 1, that the class “Prime working age” shows better classification balance than others with a 67% F1 Score. This means that the individuals usually in their prime age are the most productive, and so the disguised unemployment is expected to be minimal.

On the flip side, it is to be noted that there is a huge imbalance between the f1-score and the precision where all the classifications of “Youth” are technically correct by a recall of 0.01 is immensely misleading. This means that the “Youth” class of the working section are heavily affected by disguised unemployment.

6.2 Evaluation of Confusion Matrix

A confusion matrix is plotted to understand the correctly classified instances of each group. From the previous observation of the classification report the predictor “Age” is scrutinised

because of its classification of “False Negatives”. Along with “Age”, from Figure 5 it is also to be noted that Education Level and Employment type also provide a different insight on the classification of classes by “True negatives”, which seems to be a good sign. The following insights are made from the confusion matrices.

6.2.1 Predictor Age

Most of the youth (102) and older people (109) are correctly classified but close to (28) and (38) are classified as older workers and youths respectively. This could be related to the previous section’s precision-recall imbalance. This could mean that the youth and the older segment are sharing characteristics possibly being employed on the same job where skill is independent.

Surprisingly the prime age group also is classified (42) as youth and (41) as older workers with only 1 classification as the correct one. This means that almost majority of the classification of the prime age group is confused with that of the older or younger (Close to 519 False negatives). Hence the model is reluctant to distinguish causing heavy misclassification.

6.2.2 Predictor Education Levels

The “Upper Secondary Education” and the “Tertiary Education” are having majority of the classes being correctly classified (92) and (172) respectively but however (25) “Upper Secondary Education” are classified as Tertiary Education indicating overestimation of educational distinctions. Similarly, (39) “Tertiary Education” are classified as “Upper Secondary Education” indicating an underestimation of educational distinctions. Highter percent of in “True Negatives” can lead to realisation of blurry educational requirements satisfied for an apt employability.

6.2.3 Predictor Employment Type

The class “Disguised Unemployment” is correctly classified 291 times and 21 times as “Unemployed” but never “Relatively Employed”. This classification is spot on because the model is satisfied with the input features to understand clear difference between “Unemployed” and “Disguised unemployment”.

However, among 234 times the “Unemployed” is classified 29 instances are “Disguised Employment” and 11 instances are “Relatively employed”. This means that the misclassification does not indicate that the individuals are not employed at all but means an overlapping of scenarios where employed means is not exactly employed (Disguised Unemployment).

Also, along with the above observation, the model also exhibits low false positive rate meaning, any negative instance (unemployed) is not classified as (Relatively Employed) which is a huge plus.

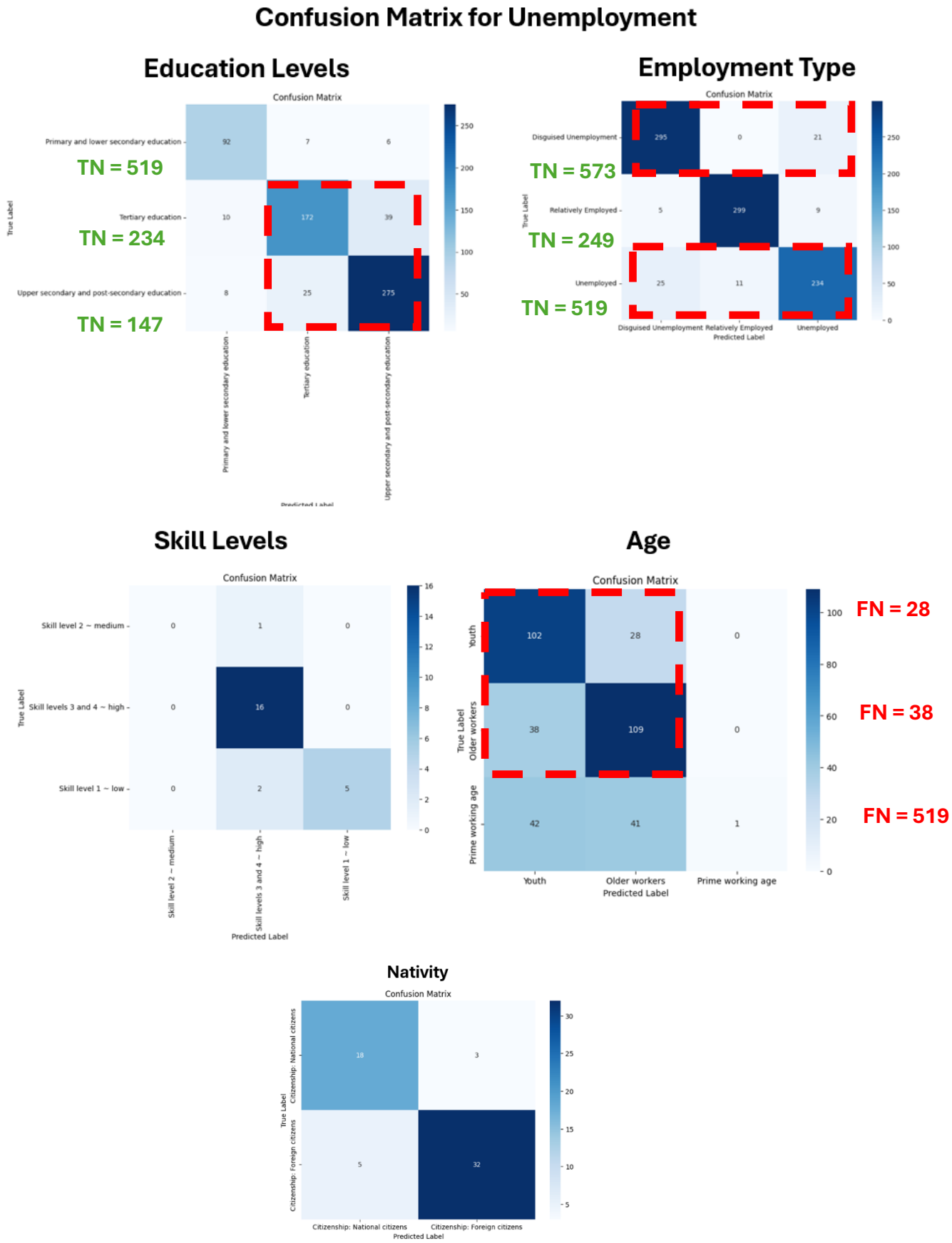

Figure 5: Confusion Matrix

6.3 Interpretation of SHAP

From the previous experiments it can been seen that “Age”, “Education levels” and “Employment type” are most intrusive employable indices. This can be confirmed from the comparison of the feature importance score of all the five predictors refer Figure 6. The feature contribution is uniform for these three predictor variables. On the flip side that the model is dependent only on the absolute value for “skills” and “nativity” predictors, which tells the tale that the relation is directly proportional and irrespective of the other features in question, confirming that there is a linear relationship.

Feature Importance for Unemployment

Education Levels

	Feature	Importance
3	VALUE	0.516562
2	Age Group	0.391009
0	Quarter	0.103628
1	Sex	0.054574

Employment Type

	Feature	Importance
3	value_thousands	0.582202
2	Sex	0.283871
1	Quarter	0.173081
0	Age	0.000000

Skill Levels

	Feature	Importance
2	value_average	0.275000
1	time	-0.004167
0	Sex	-0.012500

Age

	Feature	Importance
3	obs_value	0.199169
1	Nativity	0.154294
2	time	-0.004709
0	sex.label	-0.014404

Nativity

	Feature	Importance
3	obs_value	0.401724
1	Education Level	0.146552
0	sex.label	0.063793
2	time	0.048276

Figure 6: Feature Importance Score

It is a different scenario when the other 3 predictors feature importance score is observed. Sinc the same predictors are used across all the target variables, so the SHAP technique refer Figure 7 is used to understand the interaction of features among each other among the predictors.

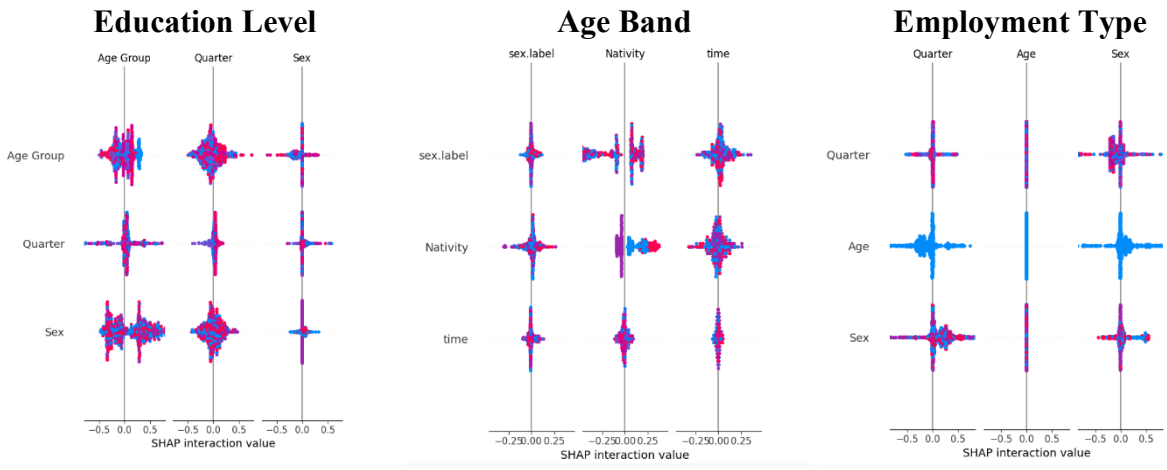

Figure 7: Interpretation of SHAP

6.3.1 Education Level Classification

Age vs Education Level: There is a strong positive interaction with age bands, implying that younger and older individuals with lower/moderate education levels are more exposed to disguised unemployment.

Quarter vs Education Level: The interaction is almost the same on both the positive and negative side indicating fluctuations or trend in the Education level affecting employment with respect to time.

Sex vs Education Level: There is a huge positive interaction of sex with education level, indicating a certain gender (red dots) face more unemployment than other.

6.3.2 Employment Type Classification

Quarter vs Employment type: There is an even distribution of positive and negative interaction, implying that there is a seasonal drift in employment in coherence with the education level.

Age vs Employment type: The interaction seems to be biased to only one section of the community (blue), in comparison with education levels meaning that unemployment prevails only to a certain age band.

Sex vs Employment type: A similar pattern of a certain gender is being provided with more employment opportunities which we can see (blue dots) with the drift of the feature towards the positive interaction.

6.3.3 Age Band Classification

It can be observed that, there is a huge negative interaction, with respect to age and nativity and age and Sex. This means that, the unemployment quotient is very high for a section of natives and also a section of the gender

6.4 Discussion

Post the experiment conducted on the "Prevalence of Disguised unemployment", the following interpretations are made. The model has predicted the parity of each of the predictors which has similar features among them. Notably, the same features behave differently when across different employable indices.

- There are significant misclassifications between the education levels and employment types which point towards systematic unclear need for job roles in terms of actual qualification – A huge point on Disguised Unemployment.
- When Employment type is considered, the classification is confused between "Disguised Unemployment" and "Unemployment" which means that the model is inclined towards agreeing that there is an outright lack of employment among those who are considered employed – Another base characteristic of Disguised Unemployment.

- The model is excellent in avoiding false alarms that is, it does not label an individual as unemployed when he isn't. This is a major boost for the overall success, since the concept of disguised unemployment is more hidden on an individual who is employed rather than an unemployed person.
- The prime working age seems to be misclassified and the interaction of the age band with other features such as nativity, sex and timeline means that there is a huge underutilisation of individuals in this age band holding a certain education and skill level – A major effect due to Disguised Unemployment.

6.4.1 Room for Improvement

- Among all the predictors used, features such as the measure of productivity and work hours could have improved the model, because right now the model is pretty straightforward, and a lot of feature overlaps are observed during the experiment.
- Systematic issues such as overstaffing/understaffing/lack of job opportunities are not deduced in this experiment which as significant vectors of "Disguised Unemployment"
- The experiment is based only on the perspective of availability of the classes with respect to unemployment. The experiment does not signify specific requirements for model improvement. The addition of more features with granularity, instead of random trial and error experiments can provide better specifications for disguised unemployment on the requirement.
- There is more focus on the correlation of features employed than causation where a high probability of disguised unemployment is exhibited but fails to say why such characteristic is even present. This means that a better dynamic analysis is required.

7 Conclusion and Future Work

The main objective of answering the question "How exactly has unemployment disguised itself within a country's workforce in today's modern well-being and what are all the factors that influence them?" is to focus on the factual reasoning behind why unemployment is hidden with the society's employable indices and evaluate them with the available data, concluding from the previous section it is fair enough to say that this research study has achieved its intended purpose. The usage of "Stacking Classifier" has helped in identifying disguised unemployment in demographic patterns, skill-based patterns, educational and employment categories as well. For example, an individual applying for a loan in a bank that utilizes this model would consider loaning for those not at repayment risk. This means that the experiment has successfully identified the "Prevalence of Disguised Unemployment" in real-world conditions and the experiments have thrown light on the effect it has on certain factors (say age, skills, nativity) when interacting with other factors (say employability, education). The visual interpretability with the help of the confusion matrix and Feature importance scores just provides an easier understanding of the situation of employability much better than the traditional time-series approaches as discussed in the literature review section.

With that being said, the concept of disguised unemployment does not have any sure-cut definitions to have a better understanding of what the experiment is searching for. There is

a very huge possibility of biased interpretation of labels or lack of clear thresholds to qualitatively develop the outcome. In other words, more relevant data collection and creating standards on the feature specifications are a whole area that is to be left explored for future works rather than focusing on modelling to overcome the lack of relevant data collection.

References

Gomis, R., Kapsos, S. and Kuhn, S., 2020. World employment and social outlook: trends 2020. *ILO: Geneva, Switzerland*, *127*.

Horne, R., 2024. Employment and social trends by region. *World Employment and Social Outlook*, *2024*(1), pp.37-60.

Harris, J.R. and Todaro, M.P., 1970. Migration, unemployment and development: a two-sector analysis. The American economic review, 60(1), pp.126-142.

Parida, J.K., 2023. Harnessing India's demographic dividend: The way forward. In *Higher Education, Employment, and Economic Development in India* (pp. 131-150). Routledge India.

Mishel, L., 2022. The Persistent Absence of Full Employment: A Critical Flaw in the Legal "Freedom of Contract" Framework. *Journal of Law and Political Economy*, *3*(1).

Murshid, K.A.S., Mahmood, T. and Shashi, N.A., 2019. Employment and unemployment amongst educated youth in Bangladesh. *The Bangladesh Development Studies*, *42*(4), pp.1-49.

Banerjee, A.V. and Duflo, E., 2019. *Good economics for hard times*. PublicAffairs.

Celbiş, M.G., 2023. Unemployment in rural Europe: A machine learning perspective. *Applied Spatial Analysis and Policy*, *16*(3), pp.1071-1095.

Zhou, Y., Zou, M., Woods, S.A. and Wu, C.H., 2019. The restorative effect of work after unemployment: An intraindividual analysis of subjective well-being recovery through reemployment. *Journal of Applied Psychology*, *104*(9), p.1195.

Guerard, J., Thomakos, D. and Kyriazi, F., 2020. Automatic time series modelling and forecasting: A replication case study of forecasting real GDP, the unemployment rate and the impact of leading economic indicators. Cogent Economics & Finance, 8(1), p.1759483

Xiao, H., Chen, R. and Guerard Jr, J.B., 2022. Forecasting the US Unemployment Rate: Another Look. Wilmott Magazine, November, pp.20-31.

Gabrikova, B., Svabova, L. and Kramarova, K., 2023. Machine learning ensemble modelling for predicting unemployment duration. Applied Sciences, 13(18), p.10146.

Enhancing binary classification: A new stacking method via leveraging computational geometry, 2024.

IEEE, 2024. Using Stacking Approaches for Machine Learning Models. IEEE Xplore

Box, G. E., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. (2015). Time series analysis: forecasting and control, John Wiley & Sons. Breiman, L. (2001). Random forests, Machine learning 45: 5–32. Chen, T. and Guestrin, C. (2016).

Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Central Statistics Office (CSO) (2020) *Labour Force Survey*. Available at*: https://www.cso.ie* (Accessed: 1 August 2024).

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

O'Higgins, N. (2012) 'This Time It's Different? Youth Labour Markets During 'The Great Recession'', *Comparative Economic Studies*, 54(2), pp. 395-412.

Tarullo, G. and Desiere, S., 2023. Short-Time Work or Unemployment? On the role of individual traits and job-related features during COVID-19. In *8th European User Conference for EU-Microdata. Microdata from Eurostat SILC, LFS, AES, SES, CIS, CSIS, EHIS, HBS and TUS*.