Data-Driven Approach for Predicting and Explaining the Risk of Long-Term Unemployment

Linan (Frank) Zhao

Geelong Grammar School, Australia

Abstract—Long-term unemployment has significant societal impact and is of particular concerns for policymakers with regard to economic growth and public finances. This paper constructs advanced ensemble machine learning models to predict citizens' risks of becoming long-term unemployed using data collected from European public authorities for employment service. The proposed model achieves 81.2% accuracy on identifying citizens with high risks of long-term unemployment. This paper also examines how to dissect black-box machine learning models by offering explanations at both a local and global level using SHAP, a state-of-the-art model-agnostic approach to explain factors that contribute to long-term unemployment. Lastly, this paper addresses an under-explored question when applying machine learning in the public domain, that is, the inherent bias in model predictions. The results show that popular models such as gradient boosted trees may produce unfair predictions against senior age groups and immigrants. Overall, this paper sheds light on the recent increasing shift for governments to adopt machine learning models to profile and prioritize employment resources to reduce the detrimental effects of long-term unemployment and improve public welfare.

1 Introduction

Long-term unemployment (LTU), by definition of OECD, refers to unemployed people of working age who are actively looking for a job but remain unemployed for a span of over 12 months. [1] Long-term unemployment causes detrimental effects to the economy, including a lower aggregate demand and thus a lower GDP, a loss of tax revenue to the government, and an excess cost of unemployment benefits. On a personal level, the experience of unemployment can be damaging to physical and mental health, which ultimately affects social relationships. [2] Long-term unemployed people also tend to gain lower income even when they find a new job. Their families tend to be less stable and their children usually perform worse academically. Communities with high rates of LTU are often associated with higher crime rates and violence. [3] In essence, the underlying problem is the supply-demand asymmetry in the labor market.

To minimize the detrimental effects of long-term unemployment on individuals and the society, authorities like public employment services (PES) aim to help the long-term unemployed by connecting them with employers through information, placement and active support services. However, these programs are usually targeted at individuals who are already long-term unemployed and who may have already experienced some of the associated impacts of LTU. On the other end, it would be tremendously costly and inefficient to provide intensive aid to all unemployed workers. [4] Therefore, it is imperative to identify the features that attribute to LTU and use those features to identify people that are at high risk of LTU, such that PES could target those at high risk of LTU and design intervention programs to prevent them from becoming so.

However, there are two additional challenges for governments or public organizations to adopt machine learning-based models to automate the identification process. First, most popular machine learning models act like black boxes and can provide zero or little explanations of the predictions. In 2018, European Union implemented General Data Protection Regulation (GDPR) that mandates "the data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her." [5] Moreover, the growing use of machine learning models in public organizations has stirred a debate about bias and fairness embedded in the model. Cautions have to be made for policymakers before deploying models without understanding of inherent bias coming from the data. Thus, an accountable, transparent, and fair model is in critical need for automated decision making in areas such as tackling long-term unemployment.

In this paper, we aim to use advanced machine learning models (Logistic Regression, Random Forest, and XGBoost) to predict long-term unemployment risk in

frankz@outlook.com.au

one European country - Portugal - using national level data provided by its PES. In 2018, the country is among the top countries with highest rates of LTU in the European Union. The PES helps consulting unemployed people and provides them with career counseling, vocational training, and soft skills etc.; it plays a vital role to ensure the re-employment of the unemployed. The overall data set from PES includes 11 years of transactional history (2007-17) with 64 million interactions and 3.5 million registered individuals. The specific data sub-set used in this paper contains random samples from the overall data with stratified samples of different groups of people. There are 110,000 entries in total. Furthermore, the initial data contains 171 features, including micro personal features, transactional records with PES, and macroeconomic indicators. Through machine learning, we will be able to identify how each feature contributes to LTU and how they can predict if someone is going to become long-term unemployed or not. This will result in automated decision making with regards to someone's risk of becoming LTU, which would be a beneficial indication for the individual, the public employment services, and for the government as well. This machine learning model can become an assistive tool for PES within a human-in-the-loop system. Based on the decision, more resources can be prioritized to citizens that are more likely to become long-term unemployed.

However, unlike other widespread commercial automation systems such as YouTube recommendations and social media news-feeds, LTU predictions can be life-critical. Also, in accordance to GDPR, users should have a right to explanation, whereby they can ask what the algorithm does. [6] Thus, it is imperative that we should be able to explain the automated decisions made from machine learning models. We will use SHapley Additive exPlanations (SHAP) to explain the model decisions and investigate the major contributors to the decision.

More so, biases are systematic discrimination against certain individuals based on the inappropriate use of certain traits or characteristics. [7] Machine learning can either help to reduce bias or bake in and escalate bias. when machine learning models produce biased decisions, certain demographic groups may be discriminated against, which could lead to devastating effects on both an individual and a societal level. In our case, most of the bias comes from underlying data and reflects historical prejudices against certain demographics. Machine learning bias can also come from data labelling, data collection, the reductively represented nature of feature selection, and proxies etc. [8] Since predictions of LTU can be life-critical, bias identification becomes ever so important. We will use Aequitas to audit our final machine learning model for discrimination and bias.

Whilst GDPR may pose some challenges to industries, it opens new doors for researchers to design and evaluate algorithms that avoid discrimination and enable explanation. Although early technical progress has been made in this area, further research is required. This research project will contribute to the ongoing progress of explaining AI and deploying AI more fairly, particularly in economics. It is also surprising that China and the United States - two biggest countries in machine learning and AI - do not have such regulations. Therefore, this research project may serve as a pioneering pushing force to establish the importance of mitigating data discrimination in China.

Lastly, there has been extensive debates about AI automating certain jobs, which will lead to extensive unemployment of workers. Paradoxically enough, this research project using machine learning helps to resolve the issues of long-term unemployment. Thus, we may need to reconsider the societal position of machine learning and AI.

Research Question:

• Can we predict the risk of a citizen becoming long-term unemployed using advanced machine learning models and explain the model predictions?

• Does bias exist in the results from machine learning models that may systematically discriminate against certain disadvantageous groups?

1.1 Related Resrarch:

In 2000, Clive Payne and Joan Payne were commissioned by the British Government to work on LTU predictions. Their research is a preliminary feasibility test about whether it is possible to predict LTU with algorithms, which is representative of earlier developments in the field. The overall conclusion from their research is that "although the methodology to do this is available and can be applied to this purpose, in practice the pattern of errors that emerged presented a dilemma." [9] Their work had some limitations that required further development. Those limitations mainly include: 1) small data sample size (747) and limited number of predictors, 2) inability to use more sophisticated models due to lack of explainability, and 3) low model performance that leaves at least one of false positive and false negative too high.

Our research builds upon Payne's preliminary test and develops on most of their limitations. We use a larger data set with many more features; we use state-of-the-art machine learning models without losing the ability to explain predictions comprehensively; and we use these advanced models to yield better performances. Our research even extends further by identifying bias in machine learning models.

Recent researches have demonstrated progress as well. Íñigo Martínez de Rituerto de Troya *et al*. have shown the superior performance of XGBoost and how SHAP can be used for personalized explanations in a human-friendly way. [10] There are many other statistical or machine learning profiling researches in other counties over the years. [11] O'Connell, P. J. *et al*. detailed and summarized some countries' implementations of statistical profiling systems in PES, including the US, Australia, Denmark, Germany, and UK etc. However, most of these systems use limited variables or features and have limited complexity in their algorithms (usually Logistic Regression). [12,13] There remains more to be explored and developed in this field of research. This paper will extend from the continuing efforts made by past researchers and contribute to the wider goal of devising

accountable, transparent, and fair machine learning models for tackling economic problems.

2 Method

2.1 Data Processing and Feature Selection

As mentioned in the Introduction, the data set used contains a total of 110,000 entries with 171 features from PES. The data contains three major types of features. The first type is specific to the individual, including personal attributes such as age, sex, country, marital status, educational levels and so on. The second type includes interactions associated with PES such as intervention and job interview counts. The other type includes macroeconomics indicators, both at a national and administrative regional level, such as unemployment rate, GDP, poverty rate etc. Furthermore, the categorical features are already one-hot encoded.

First, three irrelevant features including "Unnamed: 0", "ute_id_anon" (used internally by PES), and "snapshot_date" (the date the information was recorded) were dropped by column. Furthermore, by plotting histograms of each feature's distribution, it was noticed that several features contained identical values across the whole data set. Therefore, these features would not affect the final result and were also dropped. They included "government_psd", "poverty_rate_after_social _transfers_total", "poverty_rate_before_social_transfers_total", and "recession". There was also a continuous numerical feature called "age" but the data set had already transformed the feature into one-hot encoded categorical features like "lessthan30". Therefore, the feature "age" was redundant and was dropped.

Then, the "target" column, which indicates whether an individual would become long-term unemployed or not in 12 months' time, was taken to be the label *y* set. We call the left-over set with only features and no target values the *X* set. After cleaning up the data set, we were only left with 162 features in the combined *X* set and binary classes in the combined *y* set. All the feature names are presented in Appendix 1.

In addition, it was observed that some values were missing in the data set, marked as "nan". Hence, SimpleImputer from sklearn was used to fill in the missing values with the median value of existing ones. However, each feature still contained its original value, which would involve different orders of magnitudes. Some features went up into the hundreds whereas other one-hot encoded ones only contained 0s and 1s. "If a feature has a variance that is orders of magnitudes larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected." [14] Therefore, we needed to standardize all the features with StandardScaler from sklearn. StandardScaler scaled all the features' values to a normal distribution with mean=0 and variance=1. This step would generally help the machine learning models to perform better.

Further, it was noticed that only 33.1% of the target y set had a label value of 1; that is, "long-term unemployed". The imbalance of the data set could negatively influence models' performances. Therefore, SMOTE from imblearn was trained with *X* and *y* data sets to over-sample the minority - LTU - class, which gave equal number of samples for LTU and non-LTU.

Lastly, stratified sampling using Stratified Shuffle Split ("test size" =0.1, "iterations" =10) was used to split the training and testing sets. Stratified sampling would help ensure that the sample distribution is similar in the training and testing set. In fact, the stratified sampling yielded perfectly balanced y_{train} and y_{test} sets. There are 132,370 instances in the training set and 14,708 instances in the testing set.

This completes the feature selection and data processing work. We now have clean, processed, and valid data sets – X_{train}, X_{test}, y_{train}, and y_{test}.

2.2 Machine Learning Models and Hyper-parameter Tuning

Having a high-quality data set is the first pillar to ensure success in machine learning models. The other two pillars are high quality models and an optimum combination of hyper-parameters, which we will detail in this section.

2.2.1 Logistic Regression:

Logistic Regression is the baseline model we used in this research. It is a simple classi-fication model that predicts the risk of binary classes. If the predicted risk is more than 50%, the model classifies the instance into the positive LTU class and vice versa.

Normally, a vectorized form of linear regression is simply $h_\theta(x) = \theta^T x$. But a linear model would output a continuous range of values. To transform the continuous range into a boundary of (0,1) for classification problems, we can utilize the "S"-shaped sigmoid function: $\sigma(z) = \frac{1}{1+e^{-z}}$. By combining the sigmoid function with the linear regression form we can obtain the logistic function: $h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$.

Furthermore, there is a loss function associated with the logistic function, namely $J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)}\, log\left(h_\theta\left(x^{(i)}\right)\right) + (1-y^{(i)})\, log\left(1-h_\theta\left(x^{(i)}\right)\right)\right)$, where $y^{(i)}$ denotes the labeled value. The log function transforms a high value of $h_\theta\left(x^{(i)}\right)$ (close to 1) into a relatively lower absolute value. Therefore, when the y label is 1 and when the h_θ value is close to 1, the loss is small. Similarly, when the y label is 0 and when the h_θ value is close to 0, the loss is small. Since the cost function is convex, gradient descent can be used to find the global minimum, given the learning rate is small and the number of iterations is sufficient enough for the loss function to converge. This is done by continuously updating the coefficient $\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$ until convergence.

In sklearn, we can simply import LogisticRegression to employ the model.

2.2.2 Hyper-parameter Tuning for Logistic Regression:

Regularization is applied by default in the Logistic Regression model. However, we specified *L2* ridge regularization, which puts an additional hyper-parameter to restrict the value of $(\theta_j)^2$. The solver was set to "saga" because this specific solver can converge faster on large data sets and handles *L2* regularization. Additionally, "random state" was kept at 42 throughout the project for consistency and replication purposes.

The other three relevant hyper-parameters included "C", which is the inverse of regularization strength, "max iter", which is the maximum number of iterations, and "tol", which is the tolerance for stopping criteria. [15] Then, GridSearchCV from sklearn was used to evaluate different combinations of hyper-parameters through cross-validation and to find the best one. The hyper-parameter space is detailed in Appendix 2.

The reason we used GridSearchCV is to prevent data leakage and over-fitting. If we tuned the hyper-parameters with regards to performances on the testing set, it is likely that we would over-fit to the testing set data and the final model would not generalize well. GridSearchCV only uses the training set, thus preventing these problems.

GridSearchCV works by randomly dividing the training set into 5 (or some other number) equal-sized subsets. The algorithm then trains a model - Logistic Regression - on the first four data subsets combined and evaluates the model's accuracy on the other left-over data subset. GridSearchCV then trains another Logistic Regression on the data subsets without the fourth one and evaluates the new model's accuracy on the leftover fourth sub-set. This process will repeat five times so that each subset is tested exactly once. The accuracy score obtained from the 5 algorithms will be averaged to give the finalperformance score on the specific combination of hyper-parameters. This cross-validation method also prevents over-fitting, as the training sets and testing sets are different for each of the 5 models and the accuracy performance is averaged. This process was repeated for all combinations of hyper-parameters, from which we could compare the results.

By fitting GridSearchCV to X_{train} and y_{train}, we could call out "best params" to obtain the best set of hyper-parameters. After the first round of hyper-parameter tuning, we could shrink down the hyper-parameter space according to the best combination obtained previously. This process was repeated twice to make certain we have the best set of hyper-parameters. During the process, it was important to include the best combination from the previous round of tuning in the subsequent round to ensure that the new set of "best-_params_" always gave better results.

Once we had the optimum set of hyper-parameters, we trained the final Logistic Regression model on our whole training set. The tuned hyper-parameters are detailed in the Results section.

2.2.3 Random Forest:

Random Forest is our second machine learning model and has been quite popularly used in recent years. The ensemble model is bagging of Decision Trees and aggregates the independent weak learners together into a strong one.

Firstly, Decision Tree is a basic model that uses the tree structure to construct a predictor. It splits the data set into two subsets at every node, depending on a specific feature and it keeps splitting until it reaches the leaf node. One can predict the target value by following from the root down to a leaf node. A decision tree also has a loss function for a feature k and its threshold t_k:$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$. The m_{left} and m_{right} are simply the number of instances in the left/right subset. And G_{left}, G_{right} are the impurity of the left/right subset, i.e. the proportion of the misclassified instances. The decision tree greedily searches for the optimum split at the top level, one that reduces the loss function the most. Furthermore, Gini-index is $G_i = 1 - \sum_{k=1}^{n}({p_{i,k}}^2)$, where $p_{i,k}$ is the ratio of class k instances among all instances in the i^{th} node. The Gini-index is equal to 0 if the instances are perfectly classified. Similar to the process above, the decision tree algorithm selects a feature and threshold to split, such that the Gini-index is decreased the most. By default, sklearn uses Gini-index for the splitting criterion.

Although a Decision Tree is simple, intuitive, and easy-to-explain, it is not sophisticated enough to handle complex data sets and it is very unstable. That is why we use ensemble models to turn a weak learner into a strong one.

Random Forest trains several Decision Trees on sub-samples of the whole data set. It selects each sub-sample by random sampling with replacement - a process called bootstrap. And by aggregating each independent Decision Tree output together through averaging the predicted probability, we can usually get a higher predictive accuracy and reduce the possibility of over-fitting. Random Forest is a voting classifier that utilizes the principle of "wisdom of the crowd" to maximize model performance. Therefore, we would expect Random Forest classifiers to perform better than our baseline model. Also, Random Forest trains independent trees in parallel and can utilize all the CPU cores to make the process faster.

In sklearn, we simple import RandomForestClassifier to employ the model.

2.2.4 Hyper-parameter Tuning for Random Forest:

In RandomForestClassifier, there were more hyper-parameters that we needed to consider and hence more possible combinations. If we used GridSearchCV, it would take too long to run through all the possible combinations. Thus, we used RandomizedSearchCV from sklearn to find the optimum combination.

RandomizedSearchCV does not try out all hyper-parameter values, but only trains and tests on a fixed number of hyper-parameter settings sampled from the specified distributions. [16] It utilizes a cross-validation(CV) method similar to that of GridSearchCV

and outputs a CV average score for each hyper-parameter combination tested.

Several hyper-parameters specific to RandomForest-Classifier were tuned with RandomizedSearchCV. "n_estimators" dictates the number of independent decision trees trained in one model; "max depth" dictates the maximum levels in a specific tree, which is used to prevent over-fitting; "min samples split" is the minimum number of samples required to split an internal node and "min samples leaf" is the minimum number of samples required to be at a leaf node. Ideally, the minimum sample leaf should be exactly half or less than half of the amount of minimum sample split. In addition, "max features" is the number of features to consider when looking for the best split; "min impurity decrease" is a feature that determines whether or not a splitting occurs depending on the decrease in impurity. [17] The hyper-parameter space is detailed in Appendix 3.

Once the RandomizedSearchCV was trained on Xtrain and ytrain, we could obtain the optimum set of hyper-parameters. After the first round of tuning was finished, we shrank down the distributions according to the best combination obtained. We then ran RandomizedSearchCV for another 25 iterations to ensure that we obtained the best set of hyper-parameters.

Once we had the optimum set of hyper-parameters, we trained the final Random Forest model on our training sets. The tuned hyper-parameters are detailed in the Results section.

2.2.5 XGBoost:

XGBoost (Extreme Gradient Boosting) is another ensemble model that utilizes boosting to turn a Decision Tree base-model into a strong learner. Ever since its introduction in 2014, it has been known for its consistently extraordinary performance. It is the most robust model and will be our final machine learning model.

Contrary to Random Forest, a tree boosting model trains a new Decision Tree on the misclassified data points from the previous tree in a sequential manner. Or, it gives more weight for those instances that are misclassified by previous classifiers before training the subsequent one. Thus, each tree learns and improves from the previous one. Also, the more accurate classifiers are given more weight, and by aggregating the weighted average together, we usually obtain a better model and a better result. This process of sequential training on the residuals or errors, as commonly used by Gradient Boosting, is essentially using gradient descent to optimize the objective function and update the model.

Boosting models have better handling of data of mixed type, more flexibility in optimizing different objective functions, and have higher predictive power. However, the disadvantages include more careful hyper-parameter tuning, computationally expensiveness, and less interpretability.

But XGBoost is a better implementation of the normal Gradient Boosted Trees. XGBoost reduces execution speed through parallel and out-of-core computing and performs better than ordinary boosting models due to its scalability in all scenarios. It also introduces a regularization that controls the model's complexity and uses weighted quantile sketch to assign less weight to subsequent splits in a Decision Tree. [18] These improvements in XGBoost will help us avoid over-fitting and improve predictive power.

2.2.6 Hyper-parameter Tuning and Early Stopping for XGBoost:

The method used to tune hyper-parameters in XGBoost was similar to that of Random Forest. We utilized Randomized-SearchCV to run through random combinations of the hyper-parameters sampled from specified distributions.

There are more hyper-parameters associated with the XGBoost model. The "max_depth" and "n_estimators" hyper-parameters are similar to how they are used in Random Forest; "learning_rate" is the "step size shrinkage used in update to prevent over-fitting. After each boosting step, the learning rate shrinks the new feature weights to make the boosting process more conservative." These three hyper-parameters are the most important. In addition, "min_child_weight" is the minimum sum of instance weights needed in each child node, where a higher value prevents over-fitting. "subsample" is the proportion of training instances used in growing each tree, where a lower value avoids over-fitting but too low a value may lead to under-fitting. Similarly, the hyper-parameter "colsample_bytree" dictates the proportion of features used in constructing each tree. Lastly, "gamma" is the "minimum loss reduction required to make a further partition on a leaf node of the tree"; "alpha" and "lambda" are the L1 and L2 regularization terms respectively. [19] The hyper-parameter space is detailed in Appendix 4.

By training the RandomizedSearchCV on our training set, we could obtain the best combination of hyper-parameters. From these hyper-parameters, we could shrink down the possible distributions further in the second round of tuning. When we trained RandomizedSearchCV again, we obtained another optimum set of hyper-parameters. And it could be seen from the cv-score that the second round of tuning had higher average accuracy.

Furthermore, we can use the "cv" function in XGBoost to perform early-stopping. Early stopping is essential as a low number of iterations leads to under-fitting and a high number of iterations leads to over-fitting. There usually exists an optimum value that maximizes the cross-validation performance. By using 5-fold cross validation and setting early-stopping rounds to be 200, we could find in which iteration the mean test accuracy would be at its peak.

Once we had the optimum set of hyper-parameters, we trained the final XGBoost model on our training set, which needed to be transformed into a Dmatrix first. The tuned hyper-parameters are detailed in the Results section.

2.3 Evaluation of Model Performance

But XGBoost is a better implementation of the normal Gradient Boosted Trees. XGBoost reduces execution speed through parallel and out-of-core computing and performs better than ordinary boosting models due to its scalability in all scenarios. It also introduces a regularization that controls the model's complexity and uses weighted quantile sketch to assign less weight to subsequent splits in a Decision Tree.[18] These improvements in XGBoost will help us avoid over-fitting and improve predictive power.

Once all three models are trained with the optimum set of hyper-parameters, we can predict on the testing set. And by comparing the predictions with the actual y_{test} label values, we can evaluate each model's performance using different metrics.

Firstly, we need to use cross-validation to test performance on the training set, so that we can compare the performance between training and testing sets. The comparison is an indication of whether the model has over-fitted or not. Then, we may use some common metrics, including accuracy, precision, recall, F1 score, AUC, and precision at *k* to evaluate model performance on the testing set.

Table 1 Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Accuracy=$\frac{\#TP+\#TN}{\#Total}$, where # signifies "number of". Accuracy is a measure of the proportion of instances that are predicted correctly. It is the most straightforward indication of model performance. Accuracy shows how many people, regardless of LTU or not, are being predicted correctly.

Precision=$\frac{\#TP}{\#TP+\#FP}$. Precision shows the proportion of actual LTUs out of the instances that are predicted as LTU. This metric is especially important because people who are predicted as positive are likely to receive help from PES. And if the precision is too low, a large proportion of people who receive help are actually not going to be long-term unemployed, hence leading to the problem of resource misallocation. A high precision would mean that most of the resources such as new job opportunities and interventions are allocated to people that most need them.

Recall=$\frac{\#TP}{\#TP+\#FN}$. Recall is a measure of the percentage of people who are actually positive being identified as positive by the algorithm. This metric is also important in the case of LTU; it shows the proportion of identified LTU out of all people who are going to be long-term unemployed. The higher the value of recall, the more people who need help will receive help from PES.

F1 Score= $2 \times \frac{Precision \times Recall}{Precision+Recall}$. There is usually a trade-off between precision and recall, and F1 score is the harmonic mean of the two, which combines them into one metric. It will give a better indication than the two individual metrics.

AUC=Area under the Receiver Operating Characteristic (ROC) Curve. The ROC curve is plotted by True Positive Rate (Recall) against False Positive Rate ($\frac{\#FP}{\#FP+\#TN}$). We can plot the ROC curve using "roc_cruve" from sklearn. The performance is better when the curve is closer to the top-left corner, and therefore a larger Area Under the Curve (AUC) means a better performance.

We can simply import "accuracy_score", "precision_score", "recall_score", "f1_score", and "roc auc score" from sklearn. metrics. [20] By inputting the label values and the predicted values, these functions can easily output metrics performance values.

Precision at $k = \frac{\#TP_k}{\#TP_k+\#FP_k}$. This metric measures the precision for the top *k* individuals who are identified with the highest risk of LTU. Since public employment services have limited resources, they can only provide job opportunities or interventions to a limited number of people. Clearly, PES would want to give the resources to the ones that are most likely to become LTU. Therefore, precision at *k* shows a more realistic representation of the resource allocation efficiency. There are about 14,708 instances in the testing set and suppose that PES can only provide employment resources to 1000 of them. Then, it is crucial to know how many of the 1000 that receives help is actually going to need the help. [21] Since economics largely deals with the problem of allocating scarce resources, this metric is arguably the most important one.

For calculating precision at *k*, we can easily write a function in Python that first sorts the predicted probability values (usually using "predict_proba" function) from highest to lowest and then taking the $(k+1)^{th}$ value to be the threshold. A simple loop running through all the instances can count how many satisfies a) the predicted probability is larger than the threshold and b) the actual label is also positive (value=1 in our case). Dividing the total number counted by *k* will give us the precision at *k*. Further, by plotting precision at *k* versus the *k* value as it increases, we can obtain graphs that show how the precision at *k* drops off across all three models. By comparing the graphs, we can also compare between the three models' performances.

The performance results of all these metrics are detailed in the Results section.

2.4 Model Explanation using SHAP

Once we have trained the final XGBoost model, we can use SHAP to explain the predictive results, which would be incredibly important in the social sector. Although we will be using the XGBoost model trained on the re-balanced data set, we will only perform SHAP on the original data set, because new entries created by SMOTE is not representative of real-world individuals. For the feature interaction values, a smaller balanced data set with 5000 entries randomly taken from the original data set is used.

In the case of LTU, understanding why the model makes a certain prediction is crucial. If the model were to

be used in real life, individuals who go to PES will seek explanation– "why do I have such a high (or low) risk of LTU"? And in the case in which machine learning algorithms may be used to assist in decision making with regards to resource allocation, it is vital that we understand the process. However, the highest accuracy for large modern data sets is often achieved by complex models such as XGBoost, creating a trade-off between accuracy and interpretability. [22] And so far, SHAP provides the only unified framework for interpreting predictions and is the only tool that can satisfy the General Data Protection Regulation.

In essence, SHAP calculates shapley values, which represents the feature importance by comparing what a model would predict with and without that feature. However, the order in which a model sees features can affect its predictions. SHAP thus calculates every possible coalitions and uses weighted sums to aggregate the feature importance. [23] Take the feature "female" as an example, we can create different coalitions for all the other features and keep "female" (0 or 1) the only independent variable. For each of these coalitions we compute the predicted long-term unemployment risk with and without the feature "female" and take the difference to obtain the marginal contribution. And the shapley value is the weighted average marginal contributions of a feature value across all possible coalitions.

By computing the shapley values for all possible features, we can obtain each feature's contribution to the final predictive decision. We would be able to produce a summary plot showing the most important predictors and how they contribute to the final decision. This would be important as we can explain which features are most likely to contribute to LTU and in which direction they contribute to the final decision.

Furthermore, since we can evaluate the effect of features on individual data points, we will also be able to produce dependence plots and visualize the interactions between different features. The dependence plots are simply made from shapley values. And the interaction plots are made from shap interaction values, which can be interpreted as the difference between the shap values for feature *i* when feature *j* is present and the shap values for feature *i* when feature *j* is absent.[24] From these plots, we can draw some social implications and begin to discuss the existence of bias in real-life labor markets.

Lastly, force plots can be made for explaining how an individual result was made. The plot explains how certain significant features increase the risk of LTU and by how much as well as how other features decrease the risk of LTU and by how much. Individual explanations are vitally important as LTU predictions are very individual-specific. Such plots would be beneficial for the implementation of this decision assistance tool as it will satisfy individual's queries and give him/her potential constructive feedback.

2.5 Bias-Identification using Aequitas

The last step in this research is bias identification. It is increasingly important to deploy AI fairly, especially in this case where a machine learning decision about LTU can be lifecritical. We use Aequitas, an open source bias audit toolkit for machine learning models, to audit for discrimination and bias and to make informed and equitable decisions around developing and deploying predictive risk-assessment tools. [25]

Similar to SHAP explainer, we only take the entries in the final test set that are from the original data set for use in Aequitas. We only leave the features we are interested in – "female", "age", "disability", and "country" – because Research has already shown that women, older workers, and people with disabilities are disproportionately affected by LTU, which renders them of particular concern. [26] We also need to transform the already one-hot encoded data back to the categorical data. Then, we would need to add a column of label values and another column of predicted values from XGBoost to complete the data transformation that is compatible with Aequitas.

Since we want to be fair on the basis of disparate errors and the machine learning algorithm is designed to be assistive, we are mainly interested in the False Omission Rate (FOR) Parity and False Negative Rate (FNR) Parity. The False Omission Rate is equal to $\frac{\#FN}{\#FN+\#TN}$. A high FOR value would indicate that a large proportion of people predicted as non-LTU are actually going to become LTU. This means that these people who do not receive necessary unemployment resources need the help. The False Negative Rate is equal to $\frac{\#FN}{\#FN+\#TP}$. A high FNR value would indicate that a large proportion of people who are actually going to be long-term unemployed are predicted as non-LTU. This again poses problems as those who are predicted as non-LTU may not receive the necessary resources for future re-employment. "A high rate of false negatives means we fail to give early help to people who may have benefited from it." [27]

Aequitas comes in three stages, Group, Bias, and Fairness. The Group() class enables us to evaluate biases across all population groups [28] in our transformed data set by assembling a confusion matrix for each group, as well as counts by group, and group prevalence among the sample population. The Bias() class is used to calculate disparities between groups based on the crosstab confusion metrics returned by the Group() class. Disparities are calculated as a ratio of a metric for a group of interest compared to a base group. For example, the FOR disparity for being female versus being male is calculated as $Disparity_{FOR} = \frac{FOR_{Female}}{FOR_{Male}}$. Finally, the Fairness () group evaluates the disparities produced by Bias () with many different types of parities and give the final audit for attribute-level and overall fairness. By default, it is considered to be fair for different sexes if the FOR disparity satisfies: $0.8 < Disparity_{FOR} = \frac{FOR_{Female}}{FOR_{Male}} < 1.25$. [29]

Through Aequitas, we can identify the extent of bias existence in different attributes and in the overall machine learning model. This bias audit will give PES an indication of some limitations of the algorithm's fairness, in order for it to be used appropriately.

3 Results

3.1 Best Hyper-parameters for 3 Models

After Grid Search and Randomized Search using cross validation, we obtained the following tuned hyper-parameters for the three models:

Logistic Regression Hyper-parameters:

penalty='l2', solver='saga', C=1, max_iter=1000, tol=-10^{-10}, n_jobs=-1, random state=42 (consistent for all three models).

Random Forest Hyper-parameters:

n_estimators=1708, max_depth=13, min_samples_split=17, min_samples_leaf=5, max_features=75, min_impurity decrease= 3×10^{-6} , oob_score='True', n_jobs=-1.

XGBoost Hyper-parameters:

num_boost_round=1034, learning_rate=0.009, max_depth=12, subsample=0.8, colsample_bytree=0.8, gamma=4, objective= 'binary:logistic', min_child_weight=4, alpha=0.3, lambda=50, seed=0, n_jobs=-1.

3.2 Evaluation of Results with Different Metrics

By comparing the different metrics detailed in the Methods section, we can observe and evaluate the improvements as the models become more complex. We can also draw some social implications from the improvements in metrics.

3.2.1 Metrics Summary:

Table 2 Model performance summary

Model	Cross _val	Accuracy	Precision	Recall	F1 Score	*P*1000	AUC
LR	74.0%	73.6%	72.6%	75.6%	0.741	83.0%	0.736
Random Forest	80.9%	80.5%	80.5%	80.6%	0.806	100%	0.805
XGBoost	81.6%	81.2%	82.8%	78.9%	0.808	100%	0.813

* Here, "LR" is Logistic Regression, "Cross val" represents the cross-validation score on the training set and "P1000" is the precision at 1000.

From the results table, we can see that there is a significant improvement from LogisticRegression to Random Forest and a smaller improvement from Random Forest to XGBoost. These improvements are not surprising as the machine learning models get moreand more complex and robust. For XGBoost, the higher performances signify a better prediction accuracy and a more efficient allocation of preventive resources if it were to be used in PES. For example, a higher precision means that out of all the individuals that are predicted as LTU, a higher percentage of them will become LTU and hence more right resources are given to the right people. Furthermore, it can be seen that the differences between cross-validation scores and accuracy on the testing set across different models are consistently marginal, suggesting that the models do not over-fit to the training data.

Overall, this performance summary is satisfactory and demonstrates the validity of our Methods. However, it may be possible to fine-tune the XGBoost model even more with Bayesian optimization algorithm to obtain better results.

The precision at 1000 value reaches perfection for Random Forest and XGBoost, hence we plot precision at k curves to compare between the three models.

3.2.2 Precision at k Curves:

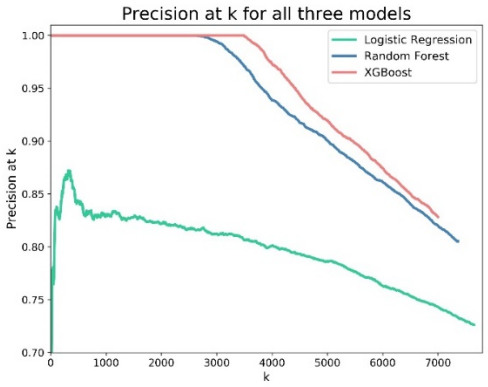

Figure 1 Precision at *k* for all three models

It can be seen that the precision at *k* curves for the three models also demonstrate consistent improvements. Firstly, there is a large improvement from Logistic Regression to Random Forest, with the peak value increasing by more than 10%. It can also be seen that Random Forest and XGBoost models perform much more consistently. Furthermore, although both Random Forest and XGBoost achieve perfect precision for predictions with the highest risk, it can be noticed that Random Forest's precision at k drops off earlier than XGBoost does, by almost 1000. Therefore, this demonstrates that XGBoost is the most robust model that will suit well to the public employment services.

3.3 Explanability Using SHAP

3.3.1 Feature Importance Summary Plot

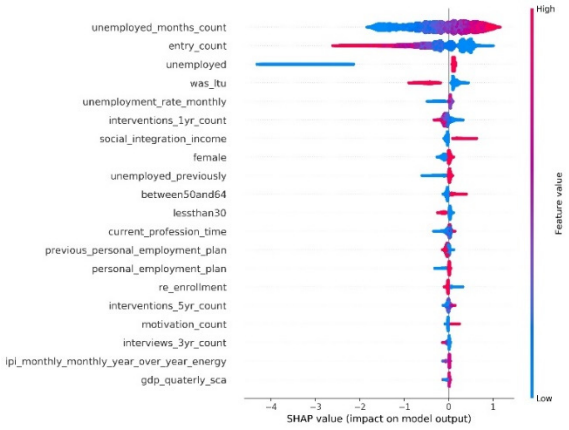

Figure 2 Feature Importance Summary Plot

From this summary plot, we notice that a wide array of features contributes to the long-term unemployment prediction. The most influential feature is "unemployed months count", which is the cumulative months of unemployment that a person experiences. It is obvious that this feature is positively correlated with LTU; if someone has been cumulatively unemployed for well over 12 months, it is very likely that the individual will either experience regular frictional unemployment or be trapped in long-term structural unemployment, which is directly correlated to LTU. On the other hand, if someone has a high "entry count" to PES, he/she is less likely to become long-term unemployed. This is perhaps an indication that the person is motivated to get a new job by checking in at the PES very often. As a result, he/she receives more employment information and opportunities and finds a new job more quickly. The feature is also an indication of the total number of unemployment spells. A higher number of spells indicates more regular frictional unemployment and hence a lesser risk of becoming LTU.

Furthermore, the feature "unemployed" is a clear indication for someone who is not going to be LTU. If an individual is not considered to be unemployed, it remains obvious that the individual is unlikely to be LTU. However, it is surprising that people who were "LTU" before have a lesser risk of becoming LTU again, as suggested by the graph. Studies have shown that long-term unemployment will deteriorate an individual's skill level [30] and the long-term unemployed usually lack in specific skills that the employers are looking for. [31] This "skill gap" theory illustrates that people who were LTU before are likely to become LTU again. Therefore, further research needs to be conducted to explain how previous long-term unemployment can reduce someone's risk of becoming LTU again.

In addition, some macroeconomic indicators play a role in the decision. Macroeconomic indicators can signal cyclical unemployment, which occurs at the downturns of the business cycle. It is clear that the economic conditions are less favorable when the monthly unemployment rate is high, hence contributing to a higher likelihood of LTU. Indeed, studies have shown that a very significant predictor of unemployment length is "the state of the economy when he or she loses his or her job". [32]

More so, living on social integration income is positively correlated with long-term unemployment. To be eligible to apply for social integration income, an individual must meet a number of requirements. His or her household must not have movable assets or goods that are subject to registration worth more than 25,734 euros. The individual must also be in serious financial need and is registered in the PES. [33] Although social integration income is designed to provide financial support for households to integrate more successfully into society, it is simply not effective for the long-term unemployed. The summary plot demonstrates that living on a social integration income significantly increases someone's risk of becoming LTU. We may postulate that people simply depend on social integration income without making an effort to find a job. It is also coincidental that an individual can receive social integration income for a maximum of 12 months - the exact same threshold for LTU - before renewal. Therefore, individuals can live off social integration income for 12 months without a job, hence contributing to LTU. The government should aim to resolve this issue by providing incentives or more resources for people living on social integration income to go and find a steady occupation.

Lastly, some more personal features such as sex, gender, disability, and intervention or interview counts contribute to the LTU prediction. These features will be discussed in detail in the following subsection.

3.3.2 Dependence Plots for Different Features

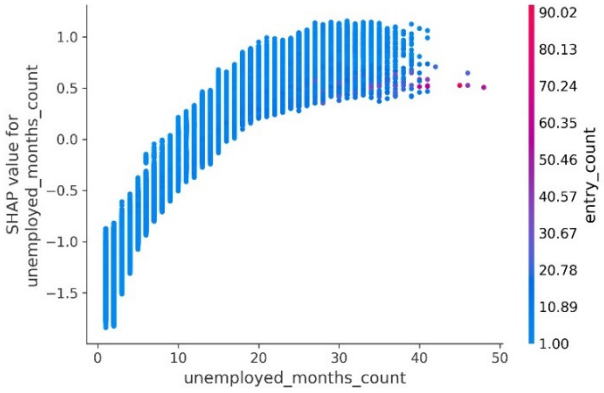

Figure 3 Dependence plot for "unemployed months count"

As discussed before, having a higher "unemployed month count" is a direct indication of long-term unemployment - a lower value indicates a lower risk whereas a higher value indicates a very high risk. Researches have demonstrated that, the greater the number of previous spells or the longer the duration of unemployment, the more likely is the individual going to be unemployed again in the future, which can lead to LTU. The first reason is rooted in the economic theory, such as lower work experiences, skill asymmetry, and deteriorating skill levels, which worsens the problem and increases future risks. The second interpretation, however, suggests that it

is the other variables in which individuals differ that influence someone's risk of unemployment, and the experience of previous unemployment serves only as a proxy for the actual causes. High risk of future unemployment is simply due to factors that contributed to previous unemployment, which are not corrected overtime, but is not necessarily caused by previous unemployment itself. [34]

But recent studies have also revealed new theories. A 2012 study conducted by Dr. Rand Ghayad discovered that employer discrimination plays a significant role in keeping the long-term unemployed from receiving fair consideration in the job market. [35] There is in fact prejudice against people with higher unemployment experiences even though their other attributes like work experiences and level of education stay the same. And since employers hold certain biases against the long-term unemployed, it is suggested that people who are long-term unemployed may be permanently pushed out of the job market. Thus, the high cumulative unemployed months count may not only be an indication but also a likely cause of LTU.

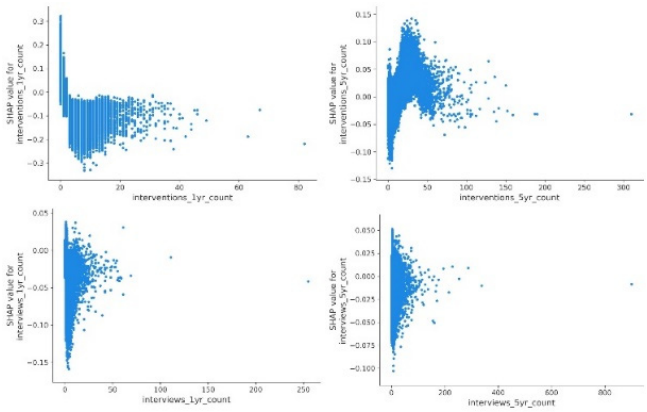

Figure 4 Summary dependence plot for interventions and interviews

By looking at interventions or interviews count in a year, it remains clear that the more of them an individual goes to, the less likely he/she will become long-term unemployed. However, if we compare them to the 5-year count, the negative shap values are no longer that clear. The mean value stays roughly in the middle, and some high interventions or interviews count even contributes to a higher risk of LTU. This illustrates that going to PES activities with high frequency will be beneficial in the short run but less so in the long run. Over long periods of time, they are clearly less helpful and serve as a mere indication that someone has not yet found a job.

This observation derived from the underlying data may be pointing out a problem in PES's programs. It can be reasonably inferred that the service PES provides is relatively ineffective in long periods; it clearly has not helped those people who keep attending interventions because they never manage to get a job. These expenditures made by public employment services constitute a large proportion of governments' welfare budgets and it is important to note that taxpayers are essentially paying for the costs associated with running these interventions. If the data suggests that the training programs are not that efficient in certain circumstances, we should start to look at ways in which PES or the government can improve this efficiency. For example, PES can devise a long-term program that aims specifically at the long-term unemployed, which helps them to obtain the necessary skills and resources. Furthermore, the government might work towards eliminating employer bias in the long run.

This could further lead to the argument around government transparency. Without the data, no one can be certain if the interventions are effective or not. But since the data points out this problem, it can become more transparent to the public, thus helping the government to fix the underlying issues.

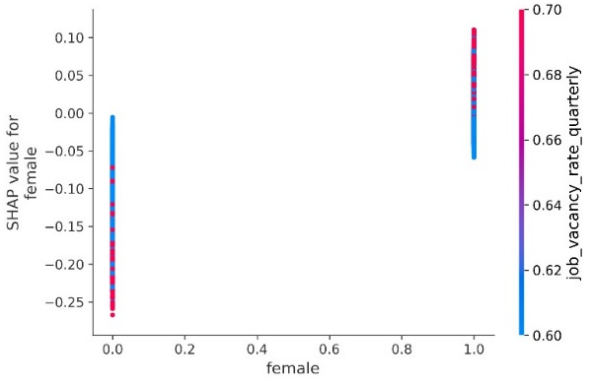

Figure 5 Dependence plot for "Female"

As the model suggests, being female increases someone's risk of becoming LTU, which may be a reflection of the underlying social bias in the data. There has been some research about the female-male differential in unemployment rates. According to Niemi's research, the three reasons for women's higher unemployment rates are: "a higher level of frictional unemployment in and out of the labor force", "a relative lack of specific training" that leads to "susceptibility to cyclical layoffs", and occupational or geographic immobility, which leads to more structural unemployment. [36] The society most females grow up in is conditioned to expect that they spend a relatively large portion of their adult lives outside the work force. Females may be pregnant or have to take care of children and therefore are less likely to be employed by companies. For all we know, "most hiring processes are intuitive and ineffective" and "women are often evaluated more negatively by others even when there are few granular behavioral differences between women and men." [37] This discrimination against females in the labor market reinforces the societal roles of men and women and concretizes the societal expectations of women. This creates a self-fulling prophecy that traps females in unfavorable conditions in the labor market, hence leading to higher risks of long-term unemployment.

Furthermore, Niemi's research mentions that the discrepancy in unemployment rate between men and women historically had been greatest at the peak of the business cycle. This may explain why during high job vacancy rates, women have an even higher risk of LTU. We can postulate that at the peak of the business cycle, most of the new structural shifts that open new employment opportunities are taken by men.

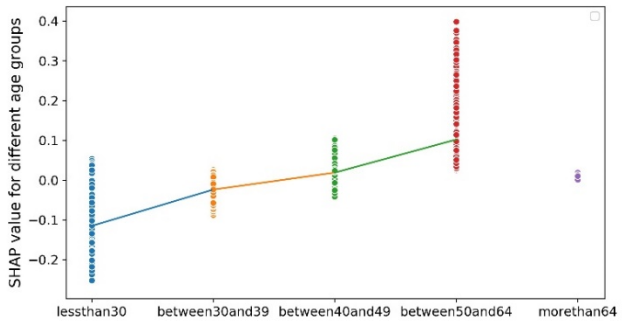

Figure 6 Dependence plot for Age

The lines connect the mean values in each age group together ("morethan64" has a small sample size and may be considered as an outlier).

By looking at the summary graph for different age groups, we can see a consistent increase in shap value as age increases. For young people, the shap values are generally negative, indicating a lesser risk of becoming LTU. For older people, however, the shap values become positive and older ages seem to contribute positively to the prediction of LTU. This implies that being older, especially between 50 and 64 years old, increases someone's risk of becoming LTU. As the structure of the labor market changes in the modern era, older people in general are beginning to lack the necessary skills demanded in the modern market. "From the point of view of aggregate productivity, the aging of the pool of unemployed workers may aggravate the persistence in the mismatch between the set of skills that employers demand and the skills of available workers. Arguably, older workers may find it more difficult - and less rewarding - to upgrade their skills and catch up with innovations in technology." [38] Also, there may exist bias towards different age groups in the job market, as older people may be less likely to receive a new job.

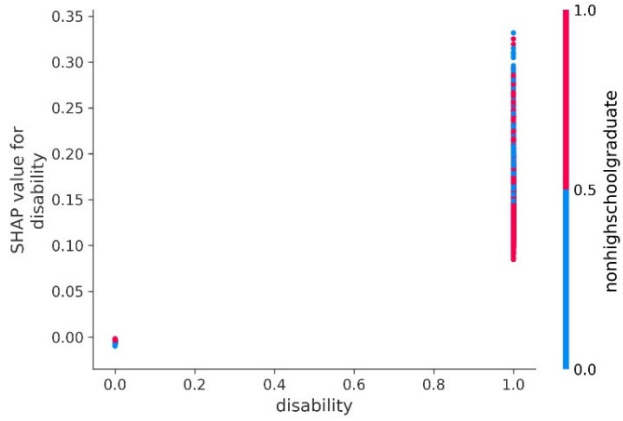

Figure 7 Dependence plot for "disability"

From this graph, it becomes apparent that people with disabilities are much more likely to become long-term unemployed – a reflection from the underlying data. This finding can be confirmed by Josh Mitchell's research. [39] "In Europe, about one eighth people of working age report having a disability; that is, the presence of a long-term limiting health condition. Despite the introduction of a wide array of legislative efforts and policy initiatives aimed to mitigate discrimination and enable retention of and entry into work, disability is still associated with substantial and enduring employment disadvantages." [40] It is thus likely that there is bias towards people with disabilities in the job market.

3.3.3 Interaction Plots for Interrelated Features:

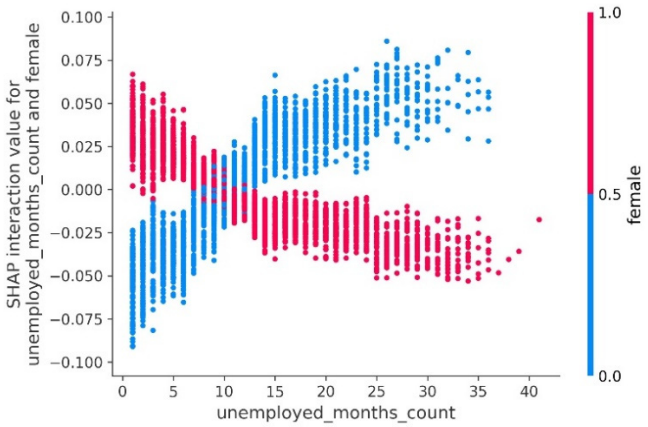

Figure 8 Interaction plot for "female" and "unemployed months count"

It would seem that for lower unemployed months counts, females have a higher risk of becoming LTU, which is reflective of the disadvantages society poses on them. However, for higher unemployed months counts, females actually have a lower risk of becoming LTU. One explanation of this is that females are more subject to frictional unemployment, as they move in and out of the labor market temporarily. Females may take maternal leave - 6 months in this country - whenever they have a new child. Therefore, it is possible that when females have 2 or more children, they are cumulatively taking more than 12 months out of the labor market. Although the unemployed months count is still increasing, those females are simply experiencing periods of frictional unemployment, which is not classified as LTU. That is why even though unemployed months' count is increasing, females' risks of long-term unemployment are in fact relatively decreasing. Another useful feature, therefore, would be the total number of unemployment spells an individual experience. It would be interesting to see how the total number of unemployment spells interacts with the overall cumulative months count and how they collectively contribute to LTU. This graph may also show that females are more resilient in the long run, as they can quickly find re-employment.

In the case of males, however, the trend is much more intuitive. As the unemployed months count increases, so does the individual's risk of long-term unemployment increase.

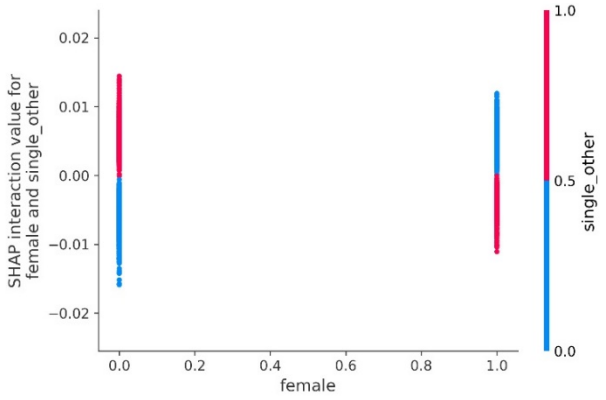

Figure 9 Interaction plot for "female" and "single other"

This also exhibits a clear difference. If someone is female and single, her risk of becoming LTU decreases. It is likely that the marital status variable - being single - is proxying for the influences of factors related to higher levels of labor market mobility among single females and a lower reservation wage due to the absence of dependent children. [41] If someone is female and married, her risk of LTU increases, perhaps due to family factors and even the social roles of men and women. Maternal leave and time spent for looking after children may potentially deteriorate the mother's skill levels and employability, hence contributing to future LTU. It also links to the possibility that married females are disadvantaged or prejudiced against in the job market. Employers may fear that they will take maternal leaves or need to spend time looking after their children. The underlying data exposes this problem.

As for a man, if he is single, his risk of LTU is usually more positive and higher; if he is married, his risk of becoming LTU is lower. This is possibly due to the fact that most men marry after they have a steady job and that they will need to support their family. Chiodo, A. J. and Owyang, M. T. proposed three reasons for married men's higher wages, which are equally applicable to the case of LTU: (1) employers discriminate in favor of married men, as they believe married men to be more stable and responsible; (2) marriage makes men more productive via specialization; or (3) more-productive men are more likely to be married. [42] Due to these reasons, married men have a lower risk of becoming LTU. On the other hand, single men do not perform as well, but the causal relationship is not clear if being single reduces employment probability. Therefore, it is most likely that a male stays single because he does not have a steady job or income, which correlates positively with LTU.

3.3.4 Individual Explanations

If the algorithm were to be used in actual PES, it is very likely that individuals will seek explanations to the machine learning algorithm's decisions. Why does the algorithm predict that I am not likely to become LTU and therefore refuse to offer me training opportunities? Why does the algorithm predict that I will become long-term unemployed? What potential practical things can I improve on? If it predicts that I have a low risk of LTU, what am I doing right and what can I take advantage of? Thus, individual plots will not only answer an individual's queries, but also give them constructive feedback and potential advises about the future. "Crucially, explanations may provide individuals with actionable recourse for changing their prospects in light of their profiling." [43]

Figure 10 Entry with highest predicted risk of LTU

For this individual with the highest risk of long-term unemployment, we can see that some contributing factors include a relatively old age, living on social integration income, low entry counts, and high unemployed months count. Although the macroeconomic indicator shows a favorable time for re-employment, the person is still likely to be long-term unemployed because he/she has already been unemployed for 24 months and is still unemployed currently. The individual depends on social integration income without making an effort to check in at PES and ask for employment resources.

One suggestion the social workers could give to the individual is to come to PES more often to obtain more employment resources, which will help the person to find a job.

Figure 11 Entry with lowest predicted risk of LTU

For this individual with the lowest risk of long-term unemployment, we can see that the person is not currently unemployed and makes an effort to check in at the PES very often. We can deduce that although the person has been cumulatively unemployed for 27 months, it has been mostly frictional unemployment and has not led to serious issues. The individual might have learned from the previous LTU experience and have recognized the benefits of coming to public employment services for help.

One suggestion social workers could give to the person is to keep checking in at PES and to request any help when needed.

To sum up, we can employ SHAP to make explanations both on a global and an individual level. The global explanations give us insights into the contributing factors of LTU, from which we can find social implications and explore the underlying social mechanism; the individual explanations are necessary for the model's implementation in PES and can aid the social workers to give constructive feedback for the individuals.

3.4 Bias-Identification using Aequitas:

Since long-term unemployment predictions can assist employment services to determine which individuals to give resources to, the machine learning algorithm is considered to be life-critical. Therefore, it is imperative that we identify the existence of any biases.

Table 3 FOR and FNR values and disparities for different groups

Attribute	Population Group	FOR	FNR	FOR disparity	FNR disparity
Sex	Male	0.18	0.41	1	1
Sex	Female	0.21	0.42	1.19	1.04
Age	Less than 30	0.17	0.58	1	1
Age	Between 30 and 39	0.18	0.50	1.07	0.85
Age	Between 40 and 49	0.22	0.38	1.32	0.66
Age	Between 50 and 64	0.26	0.29	1.53	0.50
Age	More than 64	0.12	0.23	0.71	0.40
Disability	People without disabilities	0.20	0.42	1	1
Disability	People with disabilities	0.22	0.31	1.12	0.74
Country	Portuguese	0.20	0.41	1	1
Country	Brazilian	0.125	0.65	0.62	1.60
Country	Other EU countries	0.17	0.61	0.86	1.49
Country	non-EU countries	0.19	0.6	0.93	1.47

Green signifies that the particular group passes the bias audit for FOR or FNR whereas red signifies that the particular group does not pass the bias audit.

For FOR: It can be seen that both Disability and Sex introduce a small discrepancy, but it is not significant enough to be considered unfair. However, there clearly exists bias towards senior age groups such as between 40 and 49 years old as well as between 50 and 64 years old. The senior age groups are discriminated against because a larger proportion of them is identified as non-LTU but will become LTU in the future; proportionally more people who do not receive additional assistance due to the prediction in fact require it.

For FNR: It can be noticed that only Sex passes this bias audit for FNR. However, it is surprising that people with disabilities are favored for as they have a lower FNR value, which may reflect the model's ability to recognize disability as a significant predictor. Furthermore, all the country groups other than Portugal are prejudiced against with their higher FNR values, signifying a discrimination against immigrants.

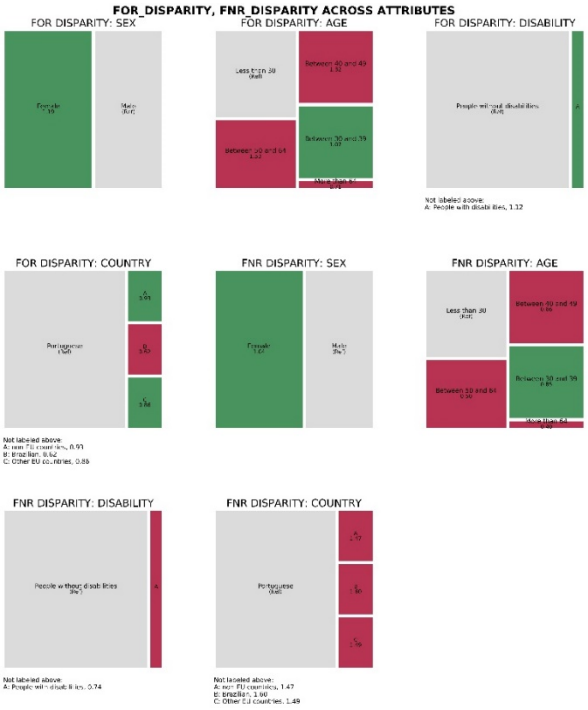

Figure 12 FOR_Disparity, FNR_ Disparity, across attributes

This graph shows the FOR and FNR disparities of different groups in each attribute. We can see that although Sex passes the bias audit, there is still an unfav0.20orable disadvantage for females as they have a higher FOR and FNR value. For people with disabilities, on the other hand, they have a lower FNR value compared to people without a disability, meaning those with disabilities and high risk of LTU are less likely to be misclassified as low risk. In other words, people without disabilities and with a high risk of LTU are 1.35 times more likely to be wrongly predicted as having a low risk, thus giving those with disabilities and high risk of LTU disproportionally more assistance. For Age and Country, however, the disparities are much more significant. Compared to Portuguese, all the other nations will be prejudiced against in terms of FNR parity if this machine learning model were to be implemented in real life. This is clearly unfair for the immigrant population groups, as those who require additional assistance are discriminated against by the system. For different age groups, on the other hand, the model prediction is disadvantageous for people between 40 and 49 years old and for people between 50 and 64 years old in terms of FOR parity. The fact that the algorithm favors people more than 64 years old is of less significance, as the sample size is very small.

It is reassuring that both the Sex and Disability attributes are not subject to significant bias, but we are still able to surface the specific demographics such as older ages and immigrants for which the model is imposing bias on. Hence, we have a clearer lens when evaluating models and making recommendations for individuals who come to PES. For instance, when deciding about resource allocation, individuals may be prejudiced against on the basis of sex, either intentionally or subconsciously by the PES staff. In that case, implementing an assistive machine learning system may actually help to reduce bias.

The machine learning algorithm's overall fairness is as follows:

Table 4 The machine learning algorithm's overall fairness

Unsupervised Fairness	Supervised Fairness	Overall Fairness
False	False	False

4 Discussion

As shown in the performance summary, we can use advanced machine learning models such as gradient boosted trees (XGBoost) to predict long-term unemployment with 81.2% accuracy and this represents 10% better performance than the baseline Logistic Regression model that most PES currently adopt. In particular, the model can help social workers in PES to make decisions. For example, when they decide to open a new training program that can only accommodate a limited number of people, as such they can utilize machine learning algorithms to rank all the individuals by the predicted risk. Since the precision at *k* performance is very high, PES can achieve more efficient resource allocation by placing less to those with low risk. This helps the citizens who are most in need and minimizes the negative personal and social effects of LTU.

Furthermore, by using SHAP explainer we can also explain how each feature, or a pair of features contributes to the prediction. This gives us a more comprehensive understanding of the risk of long-term unemployment and what contributes most to that risk. A wide myriad of features on a personal, employment service, and macroeconomics level are important for the final prediction. By plotting dependence plots, we can also draw conclusions from a wider social context - including employer bias towards the long-term unemployed, women, and people with disability etc. These observations and analysis are confirmed by other researches and also give us more concrete evidence for the existence of some prejudices in the labor market. These explanations also expose some problems associated with the current job market; they can also offer useful suggestions for governments and PES to further ameliorate the status-quo of long-term unemployment.

SHAP explainer also provides an explanation for each individual decision, detailing which features contribute positively towards LTU and which ones contribute against the risk of LTU. This provides the individual with a simple explanation, which complies with the General Data Protection Regulation. Furthermore, social workers and individuals can gain insights from these explanations and see what can be improved on an individual-specific level. Individual explanations allow a far more tailored approach to support the long-term unemployed.

Lastly, it is imperative that we identify the existence of bias in machine learning models. As discussed in the Introduction, it is increasingly more important to deploy AI fairly, especially in a life-critical situation such as long-term unemployment. By using Aequitas, we find that there does exist some bias in the final XGBoost model, especially for senior age groups and countries. Although the algorithm fails the bias audit, we can see that it is reasonably fair towards sex and disability attributes. The identification of bias suggests that further researches should aim to mitigate bias before implementing the assistive machine learning algorithm in the real world.

It is very likely that machine learning models can be trained to reduce bias in the real world because we can increase the data quality and carefully design the machine learning system. The system design should consider to avoid model bias towards the features related to legally protected classes. However, this is still an active area of research and the exact method is not clear.

In the modern era of digitalization, many private companies are implementing advanced machine learning models in order to make a profit. People have more incentive to work for private firms as they receive higher wages. This causes the utilization of automation systems in public service sectors to lag behind. This research project aims to assist with closing that "gap". It provides a complete and comprehensive solution for public employment services when predicting an individual's risk of becoming LTU. Machine learning will help public employment services improve efficiency as most processes will be assisted with automation systems. Not only can the machine learning model help make better decisions, but also can we provide extensive and human-friendly explanations as well as possible limitations of the model.

4.1 Further Research

In the hyper-parameter tuning stage of XGBoost, we can only run a limited number of iterations based solely on probabilities. The tuning techniques used in this paper do not pay attention to past results, which makes them somewhat ineffective. To further improve the model performance, we can use a Bayesian optimization algorithm, which is a dynamic search algorithm that updates on prior information on the hyper-parameter combinations it has seen thus far when choosing the hyper-parameter set to evaluate next. [44] The algorithm focuses on those areas of the hyper-parameter space that it believes will bring the most promising performances by evaluating past results. This will generally reduce the number of iterations required to obtain the optimum set of hyper-parameters. Python has a "hyperopt" library that can implement this optimization process. The effectiveness of using Bayesian optimization could be an area of further research.

Lastly, we can run trials in selected PES institutions with the final machine learning model. We can track and document its performance in real-life and compare the performance with theoretical values and with skilled human staff members. We will also be able to identify if certain bias exists in the real-world implementation and investigate how to mitigate that bias and how the automation system can best assist social workers. The effectiveness of the machine learning models is only as good as the effectiveness of the interventions and the way in which PES uses it.

5 Conclusion

In conclusion, the methodology of this research remains an effective one, through which we have substantially answered the research questions. We can use advanced machine learning models to assist public employment services in order to improve the allocation of employment resources as well as to do so in a responsible and ethical way. We can offer a global explanation of the machine learning model and individual explanations for the long-term unemployed. Through those explanations, we can draw conclusions that extend to the wider economic and social context, including the efficiency of PES's interventions and social integration income, the potential bias in the labor market exerted upon individuals' attributes such as sex, age, and disability, as well as the role of macroeconomic indicators in the determination of long-term unemployment. Explaining the machine learning model offers us evidence that enrich our economic and social understanding of long-term unemployment. Furthermore, a certain level of bias is identified in the machine learning algorithm, which contributes to the continuing effort of researching how to deploy automation systems more fairly.

References

1. OECD. Long-Term Unemployment.https://stats.oecd.org/glossary/ detail.asp?ID=3586.
2. OECD. A Broken Social Elevator? How to Promote Social Mobility. doi:https: / /doi.org /https://doi.org/10.1787/9789264301085-en (OECD Publishing, Paris, France, 2018).
3. Nichols, A. et al. Consequences of Long-Term Unemployment. Urban Institute.https://www.urban.org/sites/default/files/publication/23921/412887 - Consequences - of - Long - Term - Unemployment.PDF(Aug. 2013).
4. Payne, C. & Payne, J. Early Identification of the Long-Term Unemployed. Policy Studies Institute.http://www.psi.org.uk/publications/Research% 20Discussion%20Series/pdffiles/Research DiscussionPaper 4.pdf (2000).
5. Regulation (EU) 2016/679 Of The European Parliament And Of The Council (General Data Protection Regulation). Official Journal of the European Union Article 22. https://eur-lex.europa.eu/eli/reg/2016/679/oj (Apr. 2016).
6. Goodman, B. & Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine 38. https://arxiv.org /pdf/1606.08813.pdf (2017).
7. Silberg, J. & Manyika, J. Notes from the AI frontier: Tackling bias in AI (and in humans). McKinsey Global Institute (June 2019).
8. Barocas, S. & Selbst, A. D. Big Data's Disparate Impact. California Law Review 104. http ://dx.doi.org/10.2139/ssrn.2477899 (Sept. 2016).
9. Payne, C. & Payne, J. op. cit.
10. De Rituerto de Troya, ´I. M. et al. Predicting, explaining, and understanding risk of long-term unemployment in 32nd Conference on Neural Information Processing Systems (Montr´eal, Canada, 2018).https://aiforsocialgood.github.io/2018/pdfs/track1/97 aisg neurips 2018.pdf.
11. Denmark: Rosholm et al., 2006; Ireland: O'Connell, P. J. el al., 2010; Portugal: ´I˜nigo Mart´ınez deRituerto de Troya et al., 2018 etc.
12. O'Connell, P. J. et al. National Profiling of the Unemployed in Ireland. Research Series RS10.https://ideas.repec.org/b/esr/resser/rs010.html (Economic and Social Research Institute (ESRI), 2009).
13. O'Connell, P. J. et al. A Statistical Profiling Model of Long-Term Unemployment Risk in Ireland.Papers WP345 (Economic and Social Research Institute (ESRI), May 2010). https://ideas.repec.org/p /esr/wpaper/wp345.html.
14. Scikit-learn. StandardScaler. https: / / scikit - learn. org / stable / modules / generated / sklearn .preprocessing.StandardScaler.html.
15. Scikit-learn. LogisticRegression. https://scikit-learn.org/stable/modules /generated/sklearn.linearmodel.LogisticRegression.html.
16. Scikit-learn. RandomizedSearchCV. https: / / scikit - learn. org / stable /modules / generated / sklearn. model selection. Randomized SearchCV.html.
17. Scikit-learn. RandomForestClassifier. https://scikitlearn.org/stable/ modules/generated/sklearn.ensemble.RandomForest Classifier.html.
18. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, San Francisco, California, USA, 2016), 785–794. isbn: 978-1-4503-4232-2. http: / / doi. acm. org / 10. 1145 / 2939672.2939785.
19. Chen, T. & Guestrin, C. XGBoost Parameters. https://xgboost. readthedocs.io/en/latest/parameter.html.
20. Scikit-learn. Model evaluation: quantifying the quality of predictions. https://scikit-learn.org/stable/ modules/model evaluation.html#model-evaluation.
21. Note that the "k" value is ultimately dependent on the PES's objectives and resources.
22. Lundberg, S. & Lee, S. A unified approach to interpreting model predictions. CoRR abs/1705.07874. arXiv: 1705.07874. http://arxiv.org/abs/1705.07874 (2017).
23. Tseng, G. Interpreting complex models with SHAP values. https: / /medium. com /@gabrieltseng /interpreting-complex-models-with-shap-values-1c187db6ec83 (2018).

24. Lundberg, S. M. et al. Consistent Individualized Feature Attribution for Tree Ensembles. ArXiv abs/1802.03888. https://arxiv.org /pdf/1802.03888.pdf (2018).
25. Saleiro, P. et al. Aequitas: A Bias and Fairness Audit Toolkit. CoRR abs/1811.05577. arXiv:1811.05577. http://arxiv.org/abs/1811.05577 (2018).
26. OECD. A Broken Social Elevator? How to Promote Social Mobility. op. cit.
27. Payne, C. & Payne, J. op. cit.
28. To clarify, e.g. ”Female” is referred to as a group whereas ”Sex” is referred to as an attribute.
29. COMPAS Analysis using Aequitas. https://dssg.github.io/ aequitas/examples/compas demo.html(2018).
30. Nichols, A. et al. Consequences of Long-Term Unemployment. op. cit.
31. Louie, K. Long-Term Unemployment: A Destructive and Persistent Social Issue. www.onlinemswprograms.com/resources/social-issues /long-term-unemployment/.
32. Casselman, B. The Biggest Predictor of How Long You’ll Be Unemployed Is When You Lose Your Job. https://fivethirtyeight.com/features/the-biggest-predictor- of-how-long- youll-be-unemployed-is when-you-lose-your-job/ (2014).
33. Portugal - Social Integration Income. https://ec.europa.eu/social /main.jsp?catId=1125&langId=en&intPageId=4742.
34. Heckman, J. J. & Borjas, G. J. Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence. Economica 47, 247–283 (Aug. 1980).
35. Louie, K. op. cit.
36. Niemi, B. The Female-Male Differential in Unemployment Rates. Industrial and Labor Relations Review 27, 331–350. issn: 00197939, 2162271X (1974).
37. Chamorro-Premuzic, T. Will AI Reduce Gender Bias in Hiring? https://hbr.org/2019/06/will-aireduce-gender-bias-in-hiring (2019).
38. Monge-Naranjo, A. & Sohail, F. The Composition of Long-term Unemployment Is Changing Toward Older Workers. The Regional Economist. https://www.stlouisfed.org/∼/media/publications/regionaleconomist/2015/october/unemployment.pdf (Oct. 2015).
39. Mitchell, J. Who Are the Long-Term Unemployed? Urban Institute. https://www.urban.org/sites/default/files/publication /23911/412885-Who-Are-the-Long-Term-Unemployed-.PDF (Aug. 2013).
40. Jones, M. Disability and labor market outcomes. IZA World of Labor. doi:10. 15185 /izawol. 253(2016).
41. O’Connell, P. J. et al. National Profiling of the Unemployed in Ireland. op. cit.
42. Chiodo, A. J. & Owyang, M. T. Marriage, Motherhood and Money: How Do Women’s Life Decisions Influence Their Wages? The Regional Economist. https://www.stlouisfed.org/ ∼/media/files/pdfs/publications/pub assets/pdf/re/2003/b/marriage.pdf (Apr. 2003).
43. De Rituerto de Troya, ´I. M. et al. op. cit.
44. Kraus, M. Using Bayesian Optimization to reduce the time spent on hyperparameter tuning. https://medium.com/vantageai/bringing - back - the - time - spent - on - hyperparameter - tuning - with - bayesian -optimisation-2e21a3198afb (2019).