Sosyoekonomi

2021, Vol. 29(50), 73-93

RESEARCH ARTICLE

ISSN: 1305-5577
DOI: 10.17233/sosyoekonomi.2021.04.04
Date Submitted: 10.02.2021
Date Revised: 21.07.2021
Date Accepted: 12.08.2021

Social Networks, Female Unemployment, and the Urban-Rural Divide in Turkey: Evidence from Tree-Based Machine Learning Algorithms

Mehmet Güney CELBİŞ *(https://orcid.org/0000-0002-2790-6035), Yeditepe University, Turkey; United Nations University, Maastricht Economic and Social Research Institute on Innovation and Technology (UNU-MERIT), The Netherlands; celbis@merit.unu.edu*

Sosyal Ağlar, Kadın İşsizliği ve Türkiye'de Kentsel-Kırsal Farklılaşmaları: Ağaç Bazlı Makine Öğrenmesi Algoritmalarından Bulgular

Abstract

This study takes a novel, algorithmic approach for understanding the underlying mechanisms related to the employment status of individuals. Using the data from the most recent survey of the International Social Survey Programme (ISSP) on Turkey, the present study examines how social connectivity and location play a role in the prediction of employment status through the use of two tree-based modern machine learning techniques, namely random forest, and extreme gradient boosting. We obtain a wide array of observations, with gender being the most prominent finding when periphery and rural locations are considered.

Keywords : Machine Learning, Unemployment, Turkey, Rural, Urban.

JEL Classification Codes : L26, R10, R20, R51, C40.

Öz

Bu çalışma kişi bazında işsizlik durumlarının nedenlerini anlamak amacıyla yeni ve algoritmik bir yaklaşım getirmektedir. Uluslararası Sosyal Anket Programı'nın (International Social Survey Programme, ISSP) Türkiye üzerine olan en güncel verilerini kullanarak kişilerin sosyal bağlantılarının ve bulundukları lokasyonların işsizlik statülerini tahmin etmede oynadıkları roller iki farklı modern makine öğrenmesi tekniği ile irdelenmektedir. Bu teknikler rassal orman ve ekstrem gradyan artırma modelleridir. Çalışmanın bulgularından yola çıkarak kırsal ve çevre bölgeler özelinde cinsiyet faktörünün rolünün en önde gözüktüğü bir dizi gözlem yapılmaktadır.

Anahtar Sözcükler : Makine Öğrenmesi, İşsizlik, Türkiye, Kırsal, Kentsel.

1. Introduction

Unemployment in rural areas is a complex issue that is challenging to understand, primarily due to the ever evolving and continuously transforming nature of the labour markets in these areas (Lindsay et al., 2003). Accordingly, policy implementations and proposals vary primarily based on the country, region, industry, or period in question. For instance, Gash (1935) associated rural unemployment with the technological advances in agriculture and accessibility to capital, which affects the labour market's seasonality dynamics. In more recent research, however, different mechanisms have been highlighted, such as network externalities, communication technology, firm size, being distanced from the large markets in urban areas, skills, and other personal employability attributes (Jones, 2004; Halden et al., 2005; McQuaid et al., 2004; McQuaid & Lindsay, 2003; Lindsay et al., 2003). The role of networks, social exclusion, social support, and social contacts have drawn particular interest within the context of labour markets. In this regard, evidence linking social networks and employment status have been frequently observed with varying prominence depending on location and gender, among other settings (Topa, 2001; Calvo-Armengol & Jackson, 2004; Russell, 1999).

While impeding the well-being of the affected persons, unemployment in rural areas has also been frequently linked to migration. Individuals who experience difficulties finding jobs in rural areas are often forced to relocate to urban centres (Lyu et al., 2019; Zenou, 2011; Jones, 2004; Cartmel & Furlong, 2000). Particularly in the case of Turkey, significant disparities in standards of living across regions have fuelled large migration flows from rural areas to cities since the 1950s. Consequently, Turkey has experienced a rapid urbanization process, leading to agglomeration economies that benefit both workers and firms in cities (Özgüzel, 2020).

Studies using traditional methods have produced noteworthy findings, particularly concerning female employment and earnings in Turkey. The role of distance to urban markets in rural unemployment has been observed for Turkey (Adanacıoğlu et al., 2012). In another mode, a recent study, Maru (2016) finds evidence that social customs may be a restraining factor limiting women's participation in the labour market in rural Turkey. In contrast, Ciğerci-Ulukan (2019) argues that the rise in rural poverty led to an increase in women's workload, forcing them to undertake extra -- and often unpaid -- jobs in agriculture, resulting in the "feminization of agriculture" in Turkey. Along similar lines, a 2011 UN report highlighted that most women in rural Turkey are unpaid workers with almost no job opportunities in sectors other than agriculture (Olhan, 2011). İlkaracan et al. (2011), on the other hand, observe a decline in agricultural labour participation by women and young individuals following the implementation of the Agricultural Reform Implementation Project (ARIP).

Nevertheless, the above outlined earlier findings on the obstructiveness for women, presented by the rural environment in Turkey, Gülümser et al. (2011) observed that Turkish

women in rural areas have a remarkably high motivation for self-employment[1]. Furthermore, regardless of location, it has been shown that a persisting gap has been existing between the earnings of men and women in Turkey in the period 2005-2017 (Türk, 2020). Therefore, it is evident in the literature that rural unemployment is subject to sophisticated interrelationships among numerous factors specific to the environment at focus.

Almost all techniques employed in the literature explore rural unemployment from a quantitative viewpoint involve traditional -- mostly linear -- models and descriptive statistics. On the other hand, flexible algorithmic approaches can help discover helpful information and clarify underlying frameworks of complex issues such as unemployment in rural areas. Machine learning (ML) approaches present many advantages to researchers against the challenges posed by complicated research questions (Mullainathan & Spiess, 2017; Harding & Hersh, 2018; Athey, 2018; Varian, 2014).

ML techniques are scarcely ever applied even in the analysis of the broader subject of employment. A handful of illustrations involving the use of ML algorithms on unemployment are the applications of support vector machine and neural network approaches for the prediction of US unemployment rates by Kreiner and Duca (2019); Cook and Hall (2017); and Xu et al. (2013). Taking a different perspective, we use personal level data in this study and apply random forest and extreme gradient boosting algorithms to predict individual employment status. The use of algorithmic selection and assessment techniques applied on individual features that come from a broad collection of potential predictors enables us to discover patterns in the data that would not be possible to achieve through traditional methods (such as econometric models that are often employed in labour market research). As a result, our approach also contributes to the currently sparse number of ML implementations about unemployment.

The remainder of the present study is organized as follows. Section 2 describes the survey data used in the present study, defines the sub-samples used in our analysis, and documents specific steps taken to prepare the raw data for analysis through ML models. Section 3 presents the details of the two ML algorithms used to predict employment status and the assessment of top predictors. The empirical results of the ML models are elaborating in Section 4. The concluding discussion and the relevance of the findings to future academic and policy-focused efforts are presented in Section 5.

2. The Data

In many cases, people working in rural jobs are not eligible for unemployment benefits used to evaluate unemployment rates. Therefore, unemployed individuals in rural areas are generally underestimated (Lasley & Korsching, 1984). The 2017 Turkey module

[1] *Aside of the studies in the rural context, the collection of studies by Günseli Berik on female employment in Turkey, and female labour in the rural carpet weaving sector should be noted (e. g. Berik 1987, 1989; Berik & Bilginsoy 2000).*

of the Social Networks and Social Resources survey conducted by the International Social Survey Programme (ISSP) presents elucidative data on personal unemployment status within a larger social network-oriented setting. The present study uses the most recent round of the ISSP Turkey survey published in 2020 (ISSP, 2017). The raw data is made of 1521 rows and 116 columns[2].

In ISSP 2017, participants were classified based on their status of employment. The first two available categories were "unemployed and looking for a job" and "in paid work," alongside other classifications indicating being in education, domestic work, military service, retirement, etc. To make a proper comparison of the unemployed and employed persons possible, we reduced the number of categories into two classes by excluding individuals who do not fall into the first two classes. This subsetting step reduced the number of observations to 791. The data set does not include individuals younger than 19 years old. In other words, the requirement of being older than 15 years old for being included in the working-age population -- in line with the OECD (2020) definition designation -- is satisfied.

The ISSP 2017 Turkey survey also conveniently categorizes the participants by settlement hierarchy into the categories "a big city," "suburbs or outskirts of a big city," "town or a small city," "A country village," "farm or home in the country." It is important to note that the category "suburbs or outskirts of a big city" has been translated into Turkish in the Turkey module of the ISSP as "an outer neighbourhood, a ghetto." Therefore, Turkey's socioeconomic attributes pertaining to this category are different from European or North American countries, where the word "suburb" has a different meaning. Taking this definition into account, we regroup the categories in the data set into two classifications: "Periphery and Rural" and "Urban Centres" with 352 and 439 observations, respectively.

The complete ISSP - Turkey data includes a considerable number of variables with many missing values, significantly restricting the number of persons used in the ML models. We identified and dropped, one at a time, the variables or the combinations of variables that have the most significant number of missing values, causing the data to become unworkable[3]. The ISSP-Turkey data set also contains numerous identifier columns (e.g., study number, date of interview, etc.), which we removed. Subsequently, we encoded all classes belonging to categorical features into binary variables.

The measures above yielded a sample with 229 and 287 observations for periphery and rural areas and urban centres, respectively, with 281 predictors. Unsurprisingly, the unemployed persons are the minority in both samples (about 11\% of all individuals). Such a large disparity between the number of observations of each category in a two-class framework may lead to biased predictions; the models will reach about 88% accuracy even

---

[2] *The term "columns" is deliberately used instead of "variables" or "predictors," as at this stage, several of the columns were administrative identifiers which cannot be defined as actual variables.*

[3] *The code, written in the R language, used for the aforementioned steps, alongside with the script random forest and extreme gradient boosting algorithms can be shared with the reviewers if requested.*

if everyone is predicted as employed. While the ISSP - Turkey data set provides valuable and unique information, the samples are imbalanced and of mediocre size when the data is split based on settlement hierarchy, and only the individuals in the labour force are considered. A common remedy for dealing with this issue is to use the Synthetic Minority Over-Sampling Technique (SMOTE) developed by Chawla et al. (2002). SMOTE makes use of the *k*-nearest neighbours' algorithm to generate new synthetic minority instances for imbalanced data sets (Chawla et al., 2002). Using some nearest neighbours of 3 and creating new synthetic observations, we expanded our sample sizes to 415 and 542 for periphery and rural areas and urban centres, respectively. The unemployed individuals account for about 44%.

Even though we have taken various necessary steps to drop several variables, numerous predictors are available to the ML algorithms employed in this study. For this reason, we rename and list the definitions of only the variables selected by the random forest and extreme gradient boosting algorithms. The complete ISSP 2017 - Turkey data and the full variable documentation are available on the internet site of the ISSP. Our derived data set can also be downloaded for cross-checking[4].

3. Machine Learning Algorithms

We apply two tree-based algorithms on our rural/periphery and urban samples. These applications are the random forest method (Breiman, 2001) and the Extreme Gradient Boosting approach (XGBoost, Chen & Guestrin, 2016), which is an extension of the gradient boosting machine and the stochastic gradient boosting machine algorithms as established in Friedman et al. (2001) and Friedman (2002). The two approaches present different advantages. The random forest method aims to decorrelate the trees in the ensemble and generates a prediction through introducing randomizations to both the sampling and feature selection processes. On the other hand, Gradient boosting is a sequential learning procedure where each tree improves upon the previous tree's prediction. XGBoost, a highly acclaimed award-winning algorithm that has become increasingly popular, expands the gradient boosting method by introducing regularization, further randomization parameters, and increased computational efficiency.

Each single classification tree in both the random forest and gradient boosting ensembles follows the partitioning steps established in the Classification and Regression Trees (CART) algorithm of Breiman et al. (1984). In a classification framework, the impurity measure for a node k is $G_k = \sum_{z=1}^{Z} w_{kz}(1 - w_{kz})$ and $w_{kz} = \frac{1}{N_k}\sum_{i \in M_k} \mathbf{1}\,(y_i = z)$ where y_i is the observed employment outcome for the i'th person in the training data set $(i = 1, \ldots, N)$, z is the index for the observed class, and M_j the group of persons that fall into

[4] *The sample used in this present study is available on the link: <https://drive.google.com/file/d/1cDqr7IxLF62WR0eSRONKZn2r_SKAsenm/view?usp=sharing>. The ISSP 2017 - Turkey data and the survey and variable documentation is available on the link: <https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5521&db=e>.*

the k'th tree node, and N_k is the number of observations in k (James et al., 2013; Friedman, 2001)[5],[6]. A split feature x_c is selected, at each partition, from the feature space where $c = (1, \ldots, C)$ alongside with its splitter value v, minimizing the aggregate weighted Gini values:

$$\min_{c,v} \left[\frac{N_{k_1}(c,v)}{N} G_{k_1}(c,v) + \frac{N_{k_2}(c,v)}{N} G_{k_2}(c,v) \right] \quad (1)$$

where k_1 and k_2 are the two sub-nodes of k, and N is the total observation number (Friedman, 2001; Breiman et al., 1984; James et al., 2013). Our two ML methods diverge after the above step. While both techniques grow many trees using the above outlined recursive process, overfitting the data is common when the partitioning is allowed to run recursively until no more splits are possible. As remedies to overfitting, XGBoost has a variety of parameters that can be used for regularization. On the other hand, the random forest algorithm builds an ensemble of unpruned trees while allowing for generalization through accommodating stochasticity. Firstly, as in a bootstrapped aggregation model, each J tree in the random forest ensemble draws a random sample of individuals $j = (1, \ldots, J)$ of size N from the training data (Breiman, 1996). A high correlation between predictors may result in the undesirable exclusion of features which may be highly relevant to the research question (Athey & Imbens, 2019; James et al., 2013). Since every tree uses the same feature space, this shortcoming could apply to all trees in the ensemble, leading to a correlation between trees (Friedman, 2001; James et al., 2013). Therefore secondly, the random forest algorithm restricts the feature space to a random set of $\sqrt{C}$ predictors (Breiman, 2001; Friedman, 2001)[7],[8],[9]. Resulting from the combination of the J separate binary recursive partitioning processes, the random forest prediction for the employment status of person i is equal to the majority class vote of all J classification trees.

Another statistical learning technique that builds multiple trees is the earlier mentioned gradient boosting machine algorithm. After its development, gradient boosting has been promptly extended by incorporating stochasticity into the learning process, leading to the stochastic gradient boosting technique Friedman et al. (2001); Friedman (2002). The sequential boosting technique used in the present study, on the other hand, utilizes the XGBoost algorithm, which is not essentially algorithmic, but a computational (code-

---

[5] *Because we use only two types of employment status,* $G_k = 2w_{k1}(1 - w_{k1})$.

[6] *For both data sets (i.e., rural/periphery and urban centres) a random sample of 70% of the corresponding data set is used as the training data, and rest as the test data.*

[7] *The number* $\sqrt{C}$ *is a generally accepted rule of thumb value for classification models.*

[8] *The training data corresponds to the randomly sampled 70% of the complete data, and the test data consists of the remaining thirty percent.*

[9] *The present study utilized the following routines in the R software: randomForest written by Liaw and Wiener (2002) for the random forest model and the resulting proximity matrix, xgboost written by Chen et al. (2015) for the extreme gradient boosting procedure, pdp written by Greenwell (2017) for the individual conditional expectation and partial dependence plots, and ggplot2 written by Wickham (2011) for all visuals.*

specific) extension, for implementing Friedman's gradient boosting machine technique (Chen & Guestrin, 2016).

In classification models, the gradient boosting algorithm is initialized by minimizing the negative log-likelihood function of the observed data concerning the predicted value in log-odds (Friedman, 2002; 2001; Friedman et al., 2001). This computation equals the logarithm of the odds that an individual is unemployed. Unlike the trees in the random forest, the next tree is not grown from scratch but is based on the initial prediction, and the third tree is based on the second, and so on. More specifically, the residuals -- which are the negative gradients of the negative log-likelihood (the loss function) -- of the initial prediction is fit into the first regression tree in the sequence $s = (1, \ldots, S)$[10]. At each iteration, only a portion of each residual, determined by the learning rate α is used fto improve the earlier prediction (i.e., higher weights are given on the persons misclassified by the preceding tree). Denoting terminal nodes of a tree j as $\bar{j} = (1, \ldots, \bar{J})$, the residuals $\epsilon_{\bar{j},j}$ of the predictions generated at each terminal region of all J trees ($M_{\bar{j},j}$) are used at a given iteration j to compute the new prediction $\hat{y}_{i,j+1}$. The new prediction is determined recursively such that $\hat{y}_i$, i.e., the prediction of the preceding tree j for the individual i, is improved by adding the residuals -- weighted by the learning rate α -- of the terminal region that the person i fell into at the j'th iteration (Friedman, 2002; 2001; Friedman et al., 2001):

$$\hat{y}_{i,j+1} = \hat{y}_{ij} + \alpha\epsilon_{\bar{j}j}\mathbf{1}(i \in M_{\bar{j}j}) \tag{2}$$

Therefore, the learning process which emphasizes previous errors is incremental and decelerated. The algorithm comes to a stop when further improvements on the prediction can no longer be made. The benefit of slowing down the process through the learning rate α is that it permits for new opportunities to also correct previous false improvements, i.e., the worsening of predictions, by revising the prediction for i and approaching the actual value by an increment $\alpha\epsilon_{\bar{j}j}$ (Schonlau, 2005; James et al., 2013).

XGBoost adds various opportunities to introduce regularization to the construction of the individual trees in the gradient boosting sequence, adding to the generalization capacity of the algorithm. This feature is beneficial as highly complex trees may lead to overfitting Friedman (2002). The XGBoost algorithm allows a flexible implementation of gradient boosting thanks to its computational speed and has even been shown to be used to discover the Higgs boson using the data obtained from the Large Hadron Collider (Chen & He, 2015; Adam-Bourdarios et al., 2015). Another advantage of XGBoost is that it allows for cross-validation. We have used 10-fold cross-validation to determine the model parameters, which are: the learning rate (α), the subsample of individuals to be considered

[10] *Since the data (i.e., residuals) in this intermediate step are no longer categorical values, the regression tree applies recursive binary partitioning by minimizing total squared error loss functions instead of the impurity levels (Breiman et al., 1984).*

at each iteration which is the defining feature of stochastic gradient boosting (Friedman, 2002), the feature subset to be used at each tree and for each split, and the minimum number of individuals in a node. The process was done for both the periphery/rural and the urban samples. The maximum number of iterations (J) is 35 for the periphery/rural and 52 for the urban sample. The algorithm was tuned to stop if predictions do not improve after 10 iterations.

Upon predicting employment status, the random forest and extreme gradient boosting models report a "variable importance" metric for each feature. The metric ranks the predictors based on their relative efficacy in improving the prediction process. For a given feature x_c, the variable importance score is calculated by aggregating, for each tree j, the reduction in nodal Gini impurity resulting from each instance where a node is split using x_c. The value is then averaged over all J trees, where the impurity decrease at any given node k is equal to $\Delta G = G_k - [\frac{N_{k1}}{N} G_{k1} + \frac{N_{k2}}{N} G_{k2}]$ (Breiman, 2001; James et al., 2013)[11].

4. Empirical Findings

The 500-tree random forest application and the XGBoost algorithm predict the employment status of the persons in the test samples with accuracies of about 94 for both the periphery/rural and urban centre samples. The variable importance levels for the top twenty predictors are represented in the first rows of Figures 1 and 4 for the random forest and XGBoost applications, respectively. The definitions of the algorithmically selected features and their summary statistics are presented in Tables 1 and 2. For binary variables, the percentages of each category are reported, while for continuous and ordinal variables, the mean, standard deviation, minimum and maximum values are shown. The results suggest that the set of variables that contribute to the correct prediction of unemployment status considerably differ from those selected in the prediction for the individuals who live in urban centres. In particular, in both sets of results, FEMALE appears as the strongest predictor of employment status for the periphery and rural areas while not being selected at all in the case of individuals living in urban centres. This finding is particularly relevant given that in the case of Turkish provinces, agglomeration gains on labour productivity seem to be larger for female workers than male workers (Özgüzel, 2020). Many factors drive these differences. However, one possible explanation often proposed in the literature is that cities provide more employment opportunities for women who are more likely to suffer from mobility restrictions imposed by family ties (Özgüzel, 2020).

In both models, frequency of face-to-face contact with other people, being isolated from others, and lack of companionship is among the top predictors for individuals who live in urban centres. The situations above can hinder access to social networks necessary for job search, leading to long-term unemployment (Lindsay, 2009). The remaining features generally pertain to the individuals' occupations that the respondents know (family, friends,

[11] *The outcome for each feature is scaled into a value between 1 and 100.*

acquaintances etc.). This outcome is not surprising; Topa (2001); Conley and Topa (2002) found that employed persons are more likely to transfer information about job opportunities given that they are within the individual's social network. Among the top predictors, we also observe variables indicating whether the respondent has anyone to ask for help under various undesirable conditions and the relationship of these persons to the respondent. In this regard, Jones (1991) has shown that social support and help are particularly relevant in a job loss and are a strong determinant of reemployment. Furthermore, we observe that the attributes of the individual and their household demographics have been effective in the predictions, in line with the findings of Adanacioglu et al. (2012) for Turkey.

The observed relationships are likely to be subject to considerable non-linearities and interactions. Therefore, it would be a reductionist strategy to interpret ML output similarly to elasticities resulting from econometric estimations. The directions of the associations between the predictors and employment status can certainly be examined in detail. However, many predictors necessitate us to focus on a specific variable of interest, FEMALE, which plays a role in one sample and not in the other.

The observed association between unemployment and gender is displayed in the individual conditional expectation plots (ICE) in the second rows of Figures 1 and 4[12]. Being a binary variable, the ICE lines for this predictor exhibit a kink right at the transition from zero to one (at 0.5). We display the urban centres ICE plot for this variable even though it is not selected as a top predictor for that sample. In all four plots, we observe that being female is associated with a lower probability of employment. Regarding the periphery/rural sample for which FEMALE is the top predictor by both algorithms, the drop in probability can be up to about 20% for some individuals. Except for the XGBoost model for periphery and rural areas, the ICE lines suggest heterogeneity in the relationship between employment status and gender. It is important to note that our ML techniques learn and automatically adapt to this heterogeneity in generating predictions with very high accuracy levels and robust across the two tree-based models.

In Figure 2, the two-way partial dependence plots visualize the role of FEMALE with the level of education and age, which are frequently found to be important determinants of job status and earnings (Cartmel & Furlong, 2000; Chandler, 1989; Unay-Gailhard, 2016). The lack of education and training has been earlier shown to harm the employment opportunities of specific women and young individuals Chandler (1989); Cartmel and Furlong (2000); Bock (2004).

The lighter coloured pixels in Figure 2 indicate a higher probability of an individual being employed, and the darker colours represent higher unemployment probabilities. The random forest results highlight an apparent discrepancy between the two samples regarding the role of women's level of education. Women with higher degrees have a lower probability

[12] *The PDP is introduced by Friedman (2001), and the centred ICE and ICE graphs are based on the framework of Goldstein et al. (2015).*

of being employed in the periphery/rural areas, whereas the role of education is reversed in urban centres where higher education is associated with a higher probability of employment. In relation to age, both plots suggest that younger individuals have a higher chance of being unemployed. It is possible that his finding indicates that in periphery/rural locations, opportunities for women exist only for low-skilled jobs (and young age is a disadvantage). In contrast, in urban centres high-educated middle-aged and older women have better opportunities. This result implies a clear disadvantage for the women in the periphery/rural locations and highlights the lagging features of these local economies in Turkey.

The relationship between education level, age, and the probability of employment is not very different for males in urban centres compared to their female counterparts, as seen in the two plots in the second column of Figure 2. This observation is consistent with the fact that FEMALE was not selected as a top predictor in the urban centre samples by the random forest and XGBoost models. In other words, gender does not play a clear role in determining employment probabilities as it does in periphery/rural areas in Turkey. On the other hand, the PDP plot for males living in periphery/rural locations presents a different picture than the women in these locations. The bias towards low-skilled labour is still observable, albeit to a lesser degree, while being a middle-aged male seems to be associated with higher chances of employment. Middle-aged men with some but low education levels are predicted to have the highest probability of employment. The XGBoost prediction routine yields PDP plots with very similar patterns for the periphery/rural sample. The findings of XGBoost differ for the Urban Centres sample. While higher education is still associated with being employed for men in urban centers, the effect is not apparent for females.

In contrast, age stands out as a feature with a clear pattern, strongly highlighting that young people face higher chances of unemployment in urban centres. For older individuals, employment probability gradually becomes somewhat higher for females while becoming much higher for men with above-average levels of education. In other words, becoming more educated helps the chance of women to become employed in a lesser way than it does for men, underlining a further disadvantage for females in the labour market.

Lastly, as a means to illustrate the effectiveness of the random forest model in differentiating the two employment categories, the multidimensional scaling (MDS) plots of the proximity matrix resulting from the random forest predictions are presented in the subfigures of Figure 3, in two and three dimensions. The dark circles mark the unemployed persons in all proximity plots, while the light ones mark the employed individuals. The degree of proximity between any given two persons is given by their frequency of being assigned into the same terminal node at each iteration in which they are out-of-bag; in other words, when the random forest algorithm does not draw them in that particular iteration (Breiman & Cutler, 2020; Friedman, 2001). All MDS plots indicate that the random forest model has been relatively efficient in differentiating the unemployed persons from employed ones.

Table: 1
Predictors Selected by the Random Forest Application (out of 281 Variables)

Name		Definition	Summary
BARBER_NO		Predictor specifying whether the respondent does not know anyone who is a hairdresser/barber.	**0: Yes** (71.57%), **1: No** (28.43%)
CHILDREN		The number of children living in the respondent's household.	**Mean:** 0.6, **Min:** 0, **Max:** 4, **Sd:** 0.81
COMP_LACK		Predictor measuring the frequency of experiencing the lack of companionship by the respondent.	**Mean:** 2.04, **Min:** 1, **Max:** 5, **Sd:** 0.97
CONTACT		The number of people the respondent has a contact within a typical weekday.	**Mean:** 2.7, **Min:** 1, **Max:** 6, **Sd:** 1.3
CONTACT_INT		Predictor measuring how often the respondent communicates with close friends and family members over the internet.	**Mean:** 2.6, **Min:** 1, **Max:** 5, **Sd:** 1.08
CONTACT_FAM		Frequency of contact of the respondent with their most commonly contacted family member.	**Mean:** 1.41, **Min:** 1, **Max:** 3, **Sd:** 0.66
CONTACT_FRN		Frequency of contact of the respondent with their closest, most commonly contacted friend.	**0: No** (83.3%), **1: Yes** 16.7%)
CSIBLING		The respondent's frequency of contact with their most commonly contacted sibling.	**0: No** (66.7%), **1: Yes** (33.3%)
DEMAND		Predictor measuring the respondent's perception on whether their friends, relatives, and family makes too many demands on them.	**Mean:** 2.58, **Min:** 1, **Max:** 5, **Sd:** 1.08
DRIVER_OTH		Predictor specifying whether the respondent knows someone who works as a bus or truck driver (who is not a relative or close friend).	**0: No** (67%), **1: Yes** (33%)
EDUCYRS		The number of years of schooling.	**Mean:** 11.25, **Min:** 8, **Max:** 15, **Sd:** 2.66
ELEMENTARY		Predictor specifying whether the respondent's main occupation - regardless of employment status - is elementary (e.g., domestic helpers, window or laundry cleaners).	**0: No** (58.4%), **1: Yes** (41.6%)
FACETOCACE		The number of people with whom the respondent has face-to-face contact on a typical weekday.	**Mean:** 2.333, **Min:** 1, **Max:** 4, **Sd:** 0.98
FAIR		Predictor measuring the respondent's perception of how fair other people are.	**Mean:** 2.5, **Min:** 1, **Max:** 4, **Sd:** 1
FAMILYC_IHELP		Predictor indicating whether the respondent would ask for help from their family members in case of serious illness.	**0: No** (33.3%), **1: Yes** (66.6%)
FEMALE		Categorical predictor specifying whether the respondent is female.	**0: No** (66.6%), **1: Yes** (33.3%)
FRIENDS_HELP		Predictor measuring how supportive the respondent is of the idea that people who are better off should help friends who are worse off.	**0: No** (33%), **1: Yes** (67%)
GHETTO		Predictor indicating whether the type of the respondent's place of residence is a ghetto or slum area.[13]	**0: No** (75%), **1: Yes** (25%)
GO_OUT		Predictor measuring how often the respondent goes out to eat or drink with friends.	**Mean:** 3, **Min:** 1, **Max:** 8, **Sd:** 2.59
HEALTH_GOVT		Predictor categorizing the respondent's opinion that the government should provide health care for the sick.	**0: No** (17%), **1: Yes** (83%)
HOUSEPOP		The number of people who live in the respondent's household.	**Mean:** 4, **Min:** 3, **Max:** 5, **Sd:** 0.85
INCDIFF		Predictor representing the perception of the respondent regarding the income inequality in their country.	**Mean:** 1.7, **Min:** 1, **Max:** 5, **Sd:** 1.15
ISOLATED		Predictor measuring how often the respondent felt isolated from others in the past month.	**Mean:** 2.33, **Min:** 1, **Max:** 5, **Sd:** 1.3
LANGUAGES		The number of languages the respondent can speak.	**0: No** (33%), **1: Yes** (67%)
LEISURE		Over the past 12 months, the frequency of leisure activities that the respondent has taken part in.	**Mean:** 4, **Min:** 2, **Max:** 5, **Sd:** 1.04
LEFTOUT		Predictor measuring how often the respondent felt left out in the past month.	**Mean:** 2.587, **Min:** 1, **Max:** 5, **Sd:** 1.67
MECHANIC_OTH		Predictor specifying whether the respondent knows someone who works as a car mechanic (who is not a relative or close friend).	**0: No** (84%), **1: Yes** (16%)
NO_ONE		Predictor indicating whether the respondent has no one to look after them in case of serious illness.	**0: No** (50%), **1: Yes** (50%)
NURSE_OTH		Predictor specifying whether the respondent knows someone who works as a nurse (who is not a relative or close friend).	**0: No** (91.7%), **1: Yes** (8.3%)
OLD_GOVT		Predictor categorizing the respondent's opinion that the government should provide care for older people.	**0: No** (17%), **1: Yes** (83%)

[13] *This designation for Turkey differs from the definition for the other countries surveyed by the ISSP where the area is simply defined as "the suburbs or outskirts of a large city."*

POLICE_NO		Predictor specifying whether the respondent does not know anyone who is a police officer.	**0: No** (75%), **1: Yes** (25%)
PRESSURE		Predictor measuring the respondent's perception regarding how much they are subject to pressure from family members about their way of life.	**Mean:** 1.83, **Min:** 1, **Max:** 3, **Sd:** 0.57
TR_DEGR		Predictor specifying the highest level of education the respondent has attained.	**Mean:** 4, **Min:** 2, **Max:** 6, **Sd:** 1.7
TRUST_PRIV		Score measuring the degree of trust that the respondent has in major private firms.	**Mean:** 4.5, **Min:** 1, **Max:** 10, **Sd:** 2.54

Note: Variable definitions in the above table may be similar or identical to the explanations in the original ISSP 2017 documentation (ISSP, 2017).

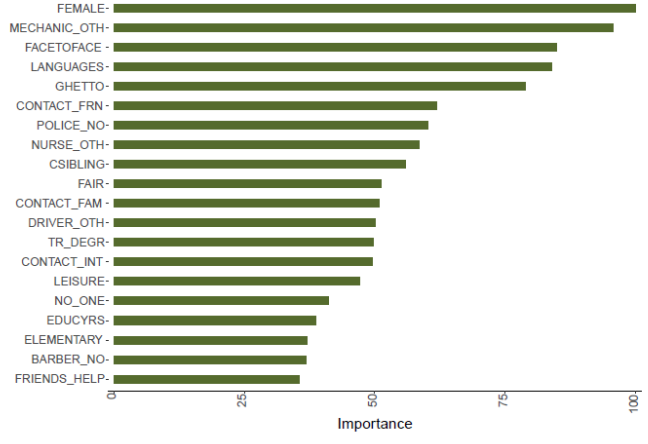
Table: 2
Predictors Selected by the XGBoost Application
(out of 281 Variables, not including previously defined features)

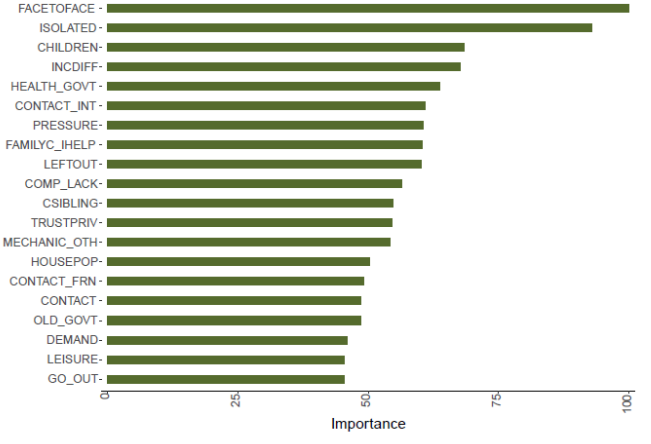
Name		Definition	Summary
BARBER_FRN		Predictor indicating whether the respondent has a friend who works as a hairdresser/barber.	**0: No** (59.33%), **1: Yes** (41.67%)
DECISION		Predictor specifying that the respondent believes they have no say about what the government does.	**Mean:** 2.583, **Min:** 1, **Max:** 5, **Sd:** 1.67
FAMILY_AHELP		Predictor indicating whether the respondent would ask for help from their family members regarding administrative problems.	**0: No** (50%), **1: Yes** (50%)
POLITICS		Predictor indicating whether the respondent has taken part in activities of political parties.	**Mean:** 4.417, **Min:** 1, **Max:** 5, **Sd:** 1.64
MECHANIC_FRN		Predictor specifying whether the respondent has a friend who works as a car mechanic.	**0: No** (66.7%), **1: Yes** (33.3%)
PARENTS		Predictor measuring how supportive the respondent is of the idea that adult children have a responsibility to look after their elderly parents.	**0: No** (83.3%), **1: Yes** 16.7%)

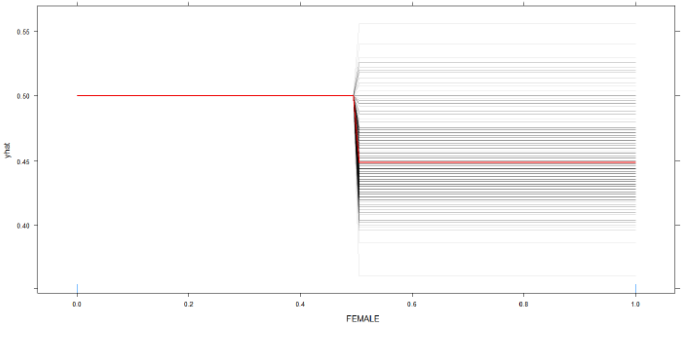
Note: Variable definitions in the above table may be similar or identical to the explanations in the original ISSP 2017 documentation (ISSP, 2017).

Figure: 1
Variable Importance and ICE Plots - Random Forest

A) Periphery and Rural **B) Urban Centres**







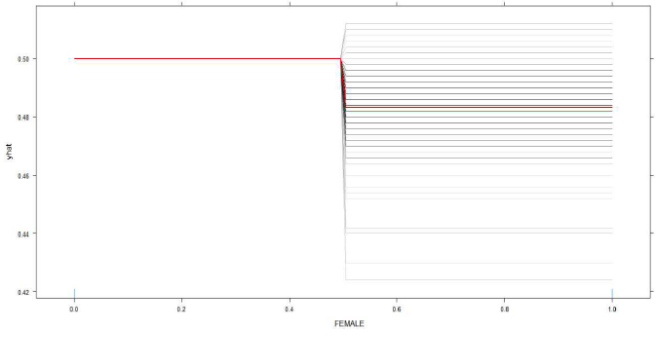


Figure: 2

Two-Way Random Forest Partial Dependence Plots

A) Periphery and Rural **B) Urban Centres**

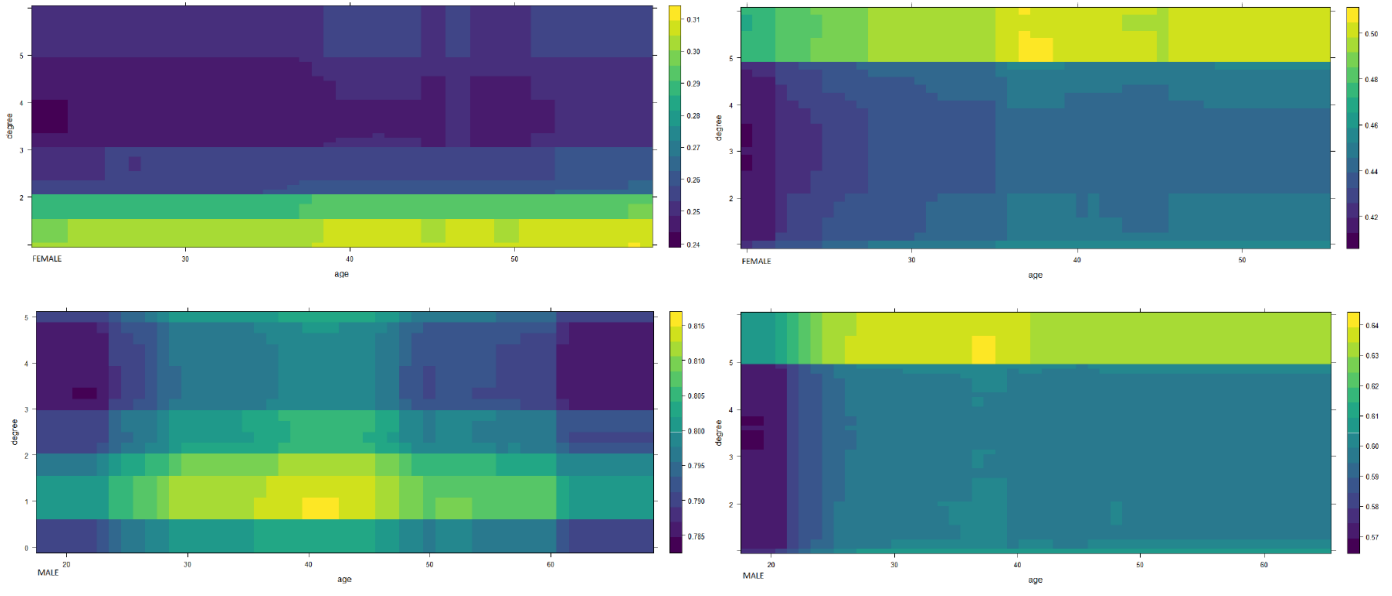


Figure: 3
MDS Proximity Plots Random Forest

A) Periphery and Rural **B) Urban Centres**

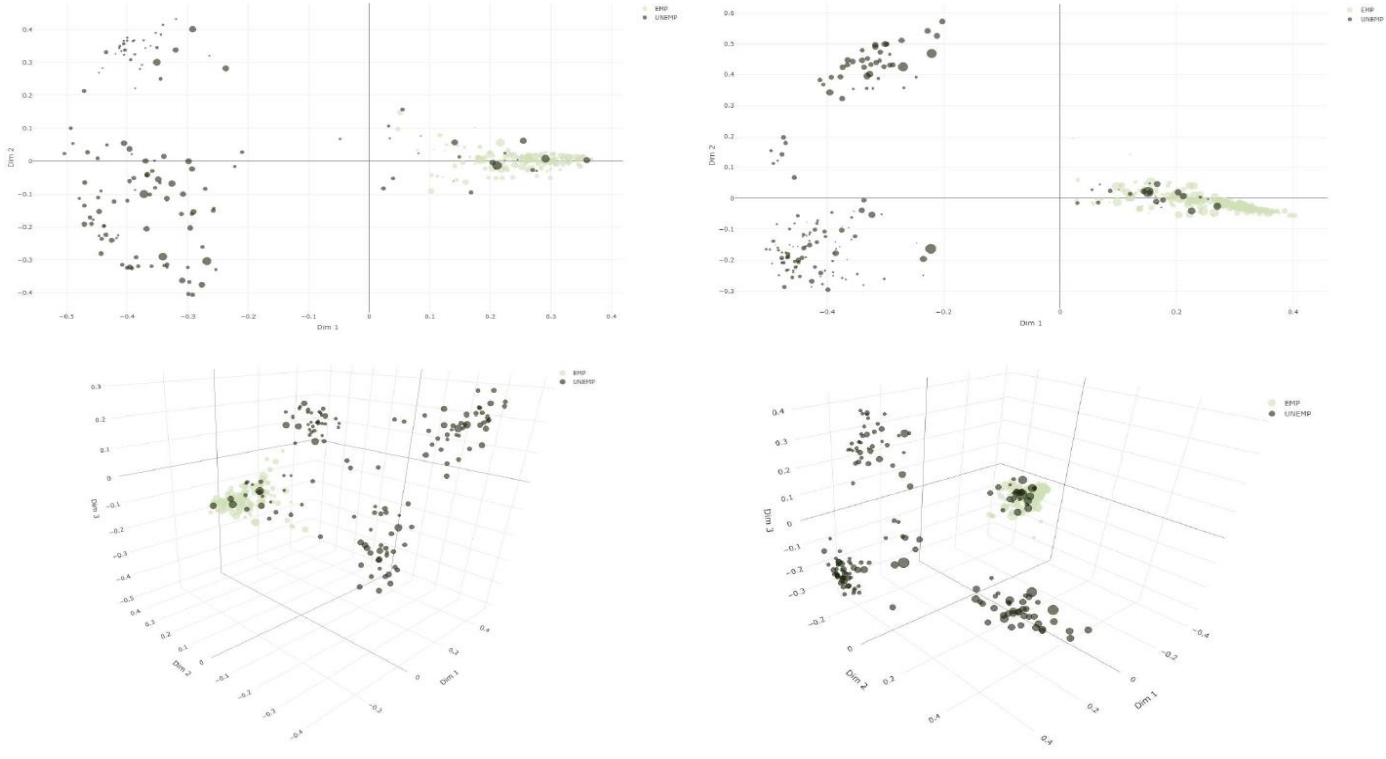


Figure: 4
Variable Importance and ICE Plots - XGBOOST

A) Periphery and Rural **B) Urban Centres**

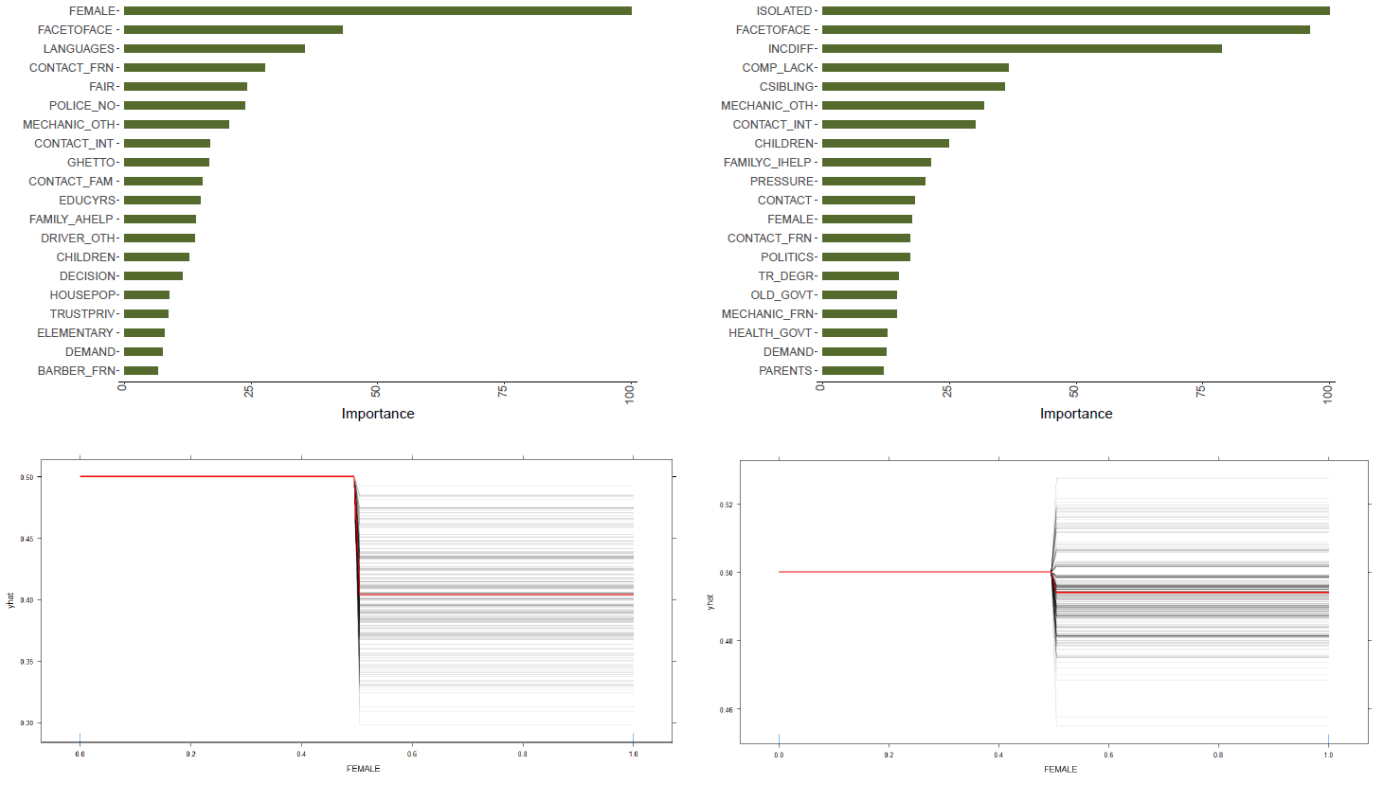
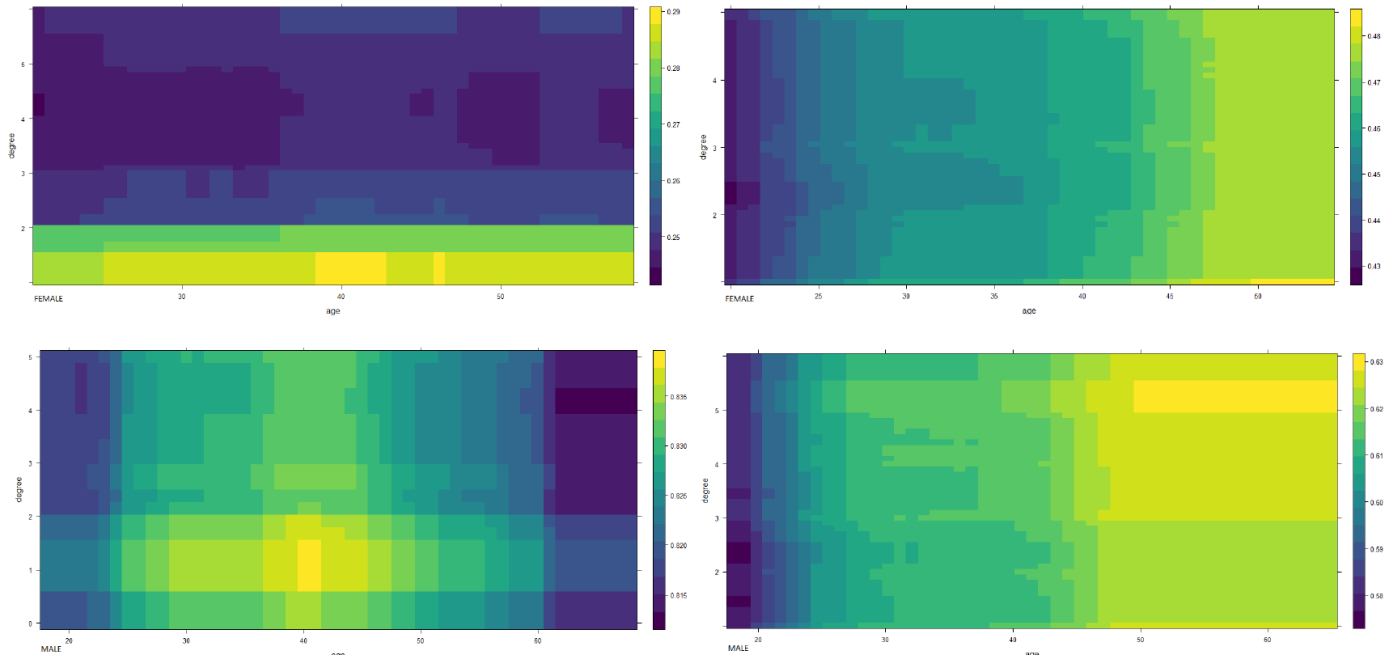


Figure: 5
Two-Way XGBoost Partial Dependence Plots

A) Periphery and Rural **B) Urban Centres**



5. Concluding Remarks

Being largely an outcome of macroeconomic effects, an individual's employment status depends on social, environmental, and personal characteristics. Given the same macroeconomic conditions, certain characteristics of individuals may be associated with their employment status in differing ways, based on the environment and social structures where they live. In the present study, gender was selected and assessed by two separate machine learning algorithms as one of such characteristics. This finding applies particularly to individuals who live in locations other than urban centres. Alongside the role of gender, we have observed that features representing social connections, the manner of contacting people, the employment status of an individual's friends, acquaintances, and family members, have been algorithmically selected and used as top predictors by our ML procedures.

It is reasonable to expect that the social and individual characteristics mentioned above may have highly non-linear and interactive relationships with employment status. It follows that a manual selection of features and theory-driven modelling of each of those features (e.g., deciding whether their relationships are linear or not) is infeasible, given the large number of variables in our data set. Such complicated mechanisms are very hard to capture using traditional techniques (Mullainathan & Spiess, 2017). Being still firmly grounded in the theoretical foundations, particularly concerning the social capital, we have used modern statistical, algorithmic techniques instead of traditional approaches to understanding particular dynamics of not general unemployment levels and rates, but individual unemployment. We have based our empirical analysis on samples divided by two main attributes; employment status individuals living in urbanised locations in Turkey were compared and contrasted to those who do not live in these locations. Results were illuminating and novel, underlining differences mainly related to gender and providing machine learning-based explanations and confirmations related to earlier literature claims. The results have implications on many dimensions of regional and national policies ranging from local commuting/infrastructure and safety policies to nationwide educational endeavours to support female employment in rural areas. In particular, the habits and established general approaches towards the division of labour between women and men need to be addressed on many levels, from manual labour positions to top management within the context of rural locations.

Lastly, by being the only application of ML algorithms on the topic, particularly for the case of Turkey, the present study brings new techniques under focus that can significantly help understand the issue in question and help the generation of policies for tackling unemployment in Turkey.

References

Adam-Bourdarios, C. & G. Cowan & C. Germain & I. Guyon & B. Kegl & D. Rousseau (2015), "The Higgs Boson Machine Learning Challenge", in: *NIPS 2014 Workshop on High-energy Physics and Machine Learning*, 19-55.

Adanacıoğlu, H. & S.G. Gümüş & F.A. Olgun (2012), "Rural Unemployment: The Problems which it Generates and Strategies to Reduce it: A Case-Study from Rural Turkey", *New Medit*, 11(2), 50-57.

Athey, S. & G.W. Imbens (2019), "Machine Learning Methods that Economists Should Know About", *Annual Review of Economics*, 11.

Athey, S. (2018), "The Impact of Machine Learning on Economics", in: *The economics of artificial intelligence: An agenda*, University of Chicago Press, 507-547.

Berik, G. & C. Bilginsoy (2000), "Type of Work Matters: Women's Labor Force Participation and the Child Sex Ratio in Turkey", *World Development*, 28(5), 861-878.

Berik, G. (1987), "Women Carpet Weavers in Rural Turkey: Patterns of Employment", *Earnings and Status (Geneva: International Labour Office, 1987)*, 13-15.

Berik, G. (1989), "Born Factories: Women's Labor in Carpet Workshops in Rural Turkey", *International Studies Notes*, 14(3), 62.

Bock, B. (2004), "It Still Matters Where You Live: Rural Women's Employment Throughout Europe", in: H. Buller & K. Hoggart (eds.), *Women in the European Countryside*, Ashgate Publishing Limited, 14-41.

Breiman, L. & A. Cutler, *Random Forests*, <https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>, 02.01.2020.

Breiman, L. & J.H. Friedman & R.A. Olshen & C.J. Stone (1984), *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA.

Breiman, L. (1996), "Bagging Predictors", *Machine Learning*, 24(2), 123-140.

Breiman, L. (2001), "Random Forests", *Machine Learning*, 45(1), 5-32.

Calvo-Armengol, A. & M.O. Jackson (2004), "The Effects of Social Networks on Employment and Inequality", *American Economic Review*, 94(3), 426-454.

Çarkoğlu, A. & E. Kalaycıoğlu (2020), *International Social Survey Programme 2017: Social Networks and Social Resources* - ISSP 2017 (Turkey).

Cartmel, F. & A. Furlong (2000), *Youth Unemployment in Rural Areas*, Number 18, York Publishing Services for the Joseph Rowntree Foundation York.

Chandler, J. (1989), "Youth Unemployment in Rural Areas: Local Government and Training Agencies", *Local Government Studies*, 15(3), 59-73.

Chawla, N.V. & K.W. Bowyer & L.O. Hall & W.P. Kegelmeyer (2002), "SMOTE: Synthetic Minority Over-Sampling Technique", *Journal of Artificial Intelligence Research*, 16, 321-357.

Chen, T. & C. Guestrin (2016), "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785-794.

Chen, T. & T. He & M. Benesty & V. Khotilovich & Y. Tang (2015), "XGBboost: Extreme Gradient Boosting", *R package version 0.4-2*, 1-4.

Chen, T. & T. He (2015), "Higgs Boson Discovery with Boosted Trees", *NIPS 2014 workshop on high-energy physics and machine learning*, 69-80.

Conley, T.G. & G. Topa (2002), "Socio-Economic Distance and Spatial Patterns in Unemployment", *Journal of Applied Econometrics*, 17(4), 303-327.

Cook, T. & A.S. Hall (2017), "Macroeconomic Indicator Forecasting with Deep Neural Networks", Federal Reserve Bank of Kansas City, *Research Working Paper*, (17-11).

Friedman, J. & T. Hastie & R. Tibshirani (2001), *The Elements of Statistical Learning*, Volume 1. Springer Series in Statistics, New York.

Friedman, J.H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine", *Annals of Statistics*, 5, 1189-1232.

Friedman, J.H. (2002), "Stochastic Gradient Boosting", *Computational Statistics & Data Analysis*, 38(4), 367-378.

Gash, N. (1935), "Rural Unemployment, 1815-34", *The Economic History Review*, 6(1), 90.

Goldstein, A. & A. Kapelner & J. Bleich & E. Pitkin (2015), "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation", *Journal of Computational and Graphical Statistics*, 24(1), 44-65.

Greenwell, B.M. (2017), "pdp: An R Package for Constructing Partial Dependence Plots", *The R Journal*, 9(1), 421-436.

Gülümser, A.A. & T. Baycan-Levent & P. Nijkamp (2011), "Changing Trends in Rural Self-Employment in Europe and Turkey", in: A. Torre & J. Traversac (eds.), *Territorial Governance*, Physica-Verlag HD, 3-25.

Halden, D. & R. McQuaid & M. Greig (2005), *Relationships Between Transport and the Rural Economies*, The Countryside Agency, Derek Halden Consultancy and Employment Research Institute, <http://www.dhc1.co.uk/projects/transport_rural_economies.html>, 07.12.2020.

Harding, M. & J. Hersh (2018), "Big Data in Economics", *IZA World of Labor*, (451).

İlkaracan, I. & I. Tunalı & B. Karapınar & F. Adaman & G. Özertan (2011), "Agricultural Transformation and the Rural Labor Market in Turkey", *Rethinking Structural Reform in Turkish Agriculture: Beyond the World Bank's Strategy*, 105-48.

James, G. & D. Witten & T. Hastie & R. Tibshirani (2013), *An Introduction to Statistical Learning*, Volume 112, Springer, New York.

Jones, L.P. (1991), "Unemployment: The Effect on Social Networks, Depression, and Reemployment", *Journal of Social Service Research*, 15(1-2), 1-22.

Jones, M.K. (2004), "Rural Labour Markets: The Welsh Example", *Local Economy*: *The Journal of the Local Economy Policy Unit*, 19(3), 226-248.

Kreiner, A. & J.V. Duca (2019), "Can Machine Learning on Economic Data Better Forecast the Unemployment Rate?", *Applied Economics Letters*, 1-4.

Lasley, P. & P.F. Korsching (1984), "Examining Rural Unemployment", *Journal of Extension*, 22(5), 32-36.

Liaw, A. & M. Wiener (2002), "Classification and Regression by Randomforest", *R News*, 2(3), 18-22.

Lindsay, C. & M. McCracken & R.W. McQuaid (2003), "Unemployment Duration and Employability in Remote Rural Labour Markets", *Journal of Rural Studies*, 19(2), 187-200.

Lindsay, C. (2009), "In a Lonely Place? Social Networks, Job Seeking and the Experience of Long-Term Unemployment", *Social Policy and Society*, 9(1), 25-37.

Lyu, H. & Z. Dong & M. Roobavannan & J. Kandasamy & S. Pande (2019), *Rural Unemployment Pushes Migrants to Urban Areas in Jiangsu Province, China*, Palgrave Communications, 5(1).

Maru, T. (2016), "How Social Customs Restrict EU Accession Effects on Female Labor Participation in Agricultural Production in Rural Adana, Turkey: A Simulation Analysis", *The Japanese Journal of Rural Economics*, 18(0), 17-31.

McQuaid, R. & C. Lindsay & M. Greig (2004), "'Reconnecting' The Unemployed information and Communication Technology and Services for Jobseekers in Rural Areas", *Information, Communication & Society*, 7(3), 364-388.

McQuaid, R.W. & C. Lindsay (2003), "Delivering Job Search Services for Unemployed People in Rural Areas: The Role of ICT", *43rd Congress of the European Regional Science Association: "Peripheries, Centres, and Spatial Development in the New Europe"*, 27th - 30th August 2003, Jyväskylä, Finland.

Mullainathan, S. & J. Spiess (2017), "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives*, 31(2), 87-106.

OECD (2020), Working Age Population (Indicator), <https://data.oecd.org/pop/working-age-population.htm>, 07.04.2020.

Olhan, E. (2011), "The Structure of Rural Employment in Turkey", *FAO Turkey Report 2011*.

Özgüzel, C. (2020), "Agglomeration Effects in a Developing Economy: Evidence from Turkey", *PSE Working Papers*, 2020-41.

Russell, H. (1999), "Friends in Low Places: Gender, Unemployment and Sociability", *Work, Employment and Society*, 13(2), 205-224.

Schonlau, M. (2005), "Boosted Regression (Boosting): An Introductory Tutorial and A Stata Plugin", *The Stata Journal*, 5(3), 330-354.

Topa, G. (2001), "Social Interactions, Local Spillovers and Unemployment", *The Review of Economic Studies*, 68(2), 261-295.

Türk, U. (2020), "Gelir Dağılımında Fırsat Eşitsizliği ve Alt Kırılımları: Türkiye Üzerine Bir Araştırma", *Alternatif Politika*, 12(2), 311-335.

Ulukan, U. & N. Ciğerci-Ulukan (2019), "Is Agriculture Feminized? Female Labor in Contemporary Rural Turkey", in: M. Meciar & K. Gökten & A.A. Eren (eds.), *Economic & Business Issues in Retrospect & Prospect*, IJOPEC Publication Limited, London.

Unay-Gailhard, I. (2016), "Job Access After Leaving Education: A Comparative Analysis of Young Women and Men in Rural Germany", *Journal of Youth Studies*, 19(10), 1355-1381.

Varian, H.R. (2014), "Big Data: New Tricks for Econometrics", *Journal of Economic Perspectives*, 28(2), 3-28.

Wickham, H. (2011), "ggplot2", *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2), 180-185.

Xu, W. & Z. Li & C. Cheng & T. Zheng (2013), *Data Mining for Unemployment Rate Prediction Using Search Engine Query Data, Volume 7*, Springer, 33-42.

Zenou, Y. (2011), "Rural-Urban Migration and Unemployment: Theory and Policy Implications", *Journal of Regional Science*, 51(1), 65-82.

Celbiş, M.G. (2021), "Social Networks, Female Unemployment, and the Urban-Rural Divide in Turkey: Evidence from Tree-Based Machine Learning Algorithms", *Sosyoekonomi*, 29(50), 73-93.