Estimation of Transition Probabilities for Turkey: An Upswing in Prediction Methods with Deep Learning

Yasin Kütük

Department of Economics, Altinbas University
Esentepe, İstanbul, Turkey
Email: yasin.kutuk@altinbas.edu.tr

Abstract: This article estimates the probability of transition to employment in the next period of unemployed individuals with high predictive power. We utilize HLFS questionnaires published by TURKSTAT between 2004 and 2016. Characteristics, qualifications, and experience information of the individuals are used to calculate transition probabilities. Any variable indicating employment status in the current period is excluded from the study. The logistic regression method is used from the econometric methodology, while classification algorithms and artificial neural networks are used from machine learning methodology. According to all results, multi-layered artificial neural networks are able to estimate 74% while XGBoost classifier and random forest follow, with performance rates remaining at 67%; this beats what Kütük and Güloğlu (2019) offered. The logistic regression has a performance rate of 63%. Researchers and academics from the Ministry of Labour, Turkish Employment Agency, private employment agencies, micro simulator modelers, forecasters, and TURKSTAT may use the deep learning method to predict the transition probability of any individual with relatively higher accuracy than other methods.

Keywords: Transition Probability, Employment, Deep Learning, Neural Networks, Classification

JEL Classification Numbers: J21, C38, C45

1. Introduction

Unemployment rates increase day by day among countries in the world. There is a similar rise for the Turkish economy, which nowadays is listed in the top 20 big economies in the world. Although Bulutay (1992) has calculated employment participation rates since 1923, unemployment rates have been officially announced since 1988, the date the first Household Labour Force Survey (HLFS hereafter) had been conducted in all over provinces of Turkey to represent the figures accurately. Accordingly, there is a natural or structural unemployment rate of around 6.5%. Above this, there is a significant upward trend in the cyclical unemployment rate, which needs to be tested and decomposed.

The reasons for this increase in Turkey depend on many factors. First, the young population rate in Turkey is quite high, therefore causing a rapid increase in participation in the labour force but available job opportunities for young people remains far from meeting the labour supply. Second, due to the strict labour market regulations and the tax

burden on employment, companies are reluctant to recruit, as employing people sometimes becomes too costly. Third, According to the report issued by ILO[1], overtime work in Turkey is ubiquitous, the ratio of overtime workers is 30%. However, this ratio is averaged around 15% in the world. Overtime work prevents the opening of new job vacancies. A rough estimate from over, like this, is the case; neither the employees can get as much productivity as required, and therefore companies cannot grow, nor the employment can increase sufficiently. Fourth, compared to developed and some developing countries, Turkey is lagging in flexible employment opportunities such as part-time jobs. We started to implement regulations to encourage flexible work. When the process is completed, job opportunities for women and young people will increase. Last, many countries in the second half of the 20th century began to increase women's labour force participation in the workforce. However, it accelerated after 2008 in Turkey.

Due to increased unemployment, Turkey has a significant problem in the labour market. One of the factors is severe inefficiency. These inefficiencies, i.e., unemployment rates, are relatively predictable on a macro scale. At least these estimates may be useful in making plans. However, it is challenging to predict at the micro-level. In order for an individual to get rid of unemployment, many factors must come together. The most important of these are undoubtedly individual developments. Moreover, personal characteristics, experience, and skills of individuals are significant internal factors. Also, the economy must continuously create new jobs.

This study aims to find out the probability of transition to employment with a high predictive power by considering the effects of individual abilities, skills, and characteristics on the transition to employment. Therefore, following the ILO definition above, individuals must be unemployed first. In the following period, they must move to the active labour force or stay their position. These transitional probabilities must be predictable at the individual level. It is planned to ensure that policies are created to prevent the accumulation of individuals in specific sectors, to direct them to new areas, and to increase job opportunities in more competitive sub-sectors. Previous studies are not sufficient, as policymakers or institutions need to calculate this transience, both because of limited predictability and small datasets.

For this purpose, firstly, the literature review, which includes the studies examining the probability of employment, will take place. Subsequently, the methods and data to be used in the study will be transferred. Then the results will be examined. The study will be concluded with an end-section, and a highly consistent estimation model will be presented with the methods used to predict individuals' transition to employment.

[1] Accessed: 28.08.2019, https://www.ilo.org/public/english/standards/relm/ilc/ilc93/pdf/rep-iii-1b.pdf

2. Literature Review

The first model that analyzes the participation of individuals in employment and the income they will receive is the human capital theory. The human capital theory put forward by Becker (1964) and Schultz (1961) argued that there is a relationship between the characteristics and experiences of individuals and their income. There is a correlation between factors such as education, experience, individual development, and income of individuals. People who know this want to go through a better and long-term education process. Later, Mincer (1958 and 1974) has recently established a more robust foundation of human capital theory by proposing a non-linear income model.

Spence (1978) indicates that the job-finding probability is shaped as a conditional probability problem given the information about indices and signals. Cohn et al. (1987), Groot and Oosterbeek (1994) and Kroch and Sjoblom (1994) show that human capital theory generally explains the earnings and job finding probability while Acemoglu and Autor (2011) support signaling theory when unobservable factors mostly surpass the obvious ones.

The second theory examining the possibility of finding a job is search theory. The search theory, first introduced by Stigler (1961), emerged in the same years as the human capital theory. Exploration theory designs the market as an open economy model in which price and quantity are uncertain. Price and quantity are only information in this model (Reynolds, 1951). Search theory refers to the search action that a seller or buyer must spend to find the addressee.

The last framework finally examined is matching theory in terms of job finding the probability. For the first time, college applications and partner matching problems were analyzed, and the most effective matching solution was provided by an algorithm developed by Gale and Shapley (1962). Generalized solutions, on the other hand, always match individuals within the game theory to a solution and are valid nowadays (Abdulkadiroglu and Sönmez, 2013).

Roth (1984) applied the algorithm for matching theory for the first time to the economy. Accordingly, it was aimed to produce a mechanism that enables the most effective placement of college graduates who wish to do an internship in the hospitals considering their characteristics, experiences, and successes. Accordingly, a stable solution was created that no one should stay outside. Hall (1979) examined how individuals transition from unemployment to employment or vice versa during cyclic fluctuation periods. Hall (1979) found that recessions could have lasting effects on the labour market. Kaitz (1970), on the other hand, explained this situation with the duration of unemployment.

Studies examining the probability of finding a job for Turkey are minimal. After 1923, a macro-scale but very comprehensive study of labour, employment, and wages were carried out by Bulutay (1995). Bulutay (1992) studies by examining the labour market dynamics between 1923 and 1989. Şenses (1996) finds the disadvantage of being less-educated; unfortunately, this is a negative factor in the transition to knowledge-based employment such as science, innovation, and entrepreneurship.

The most comprehensive study of the probability of transition to employment was conducted by Tansel and Tasci (2004) for the first time. It was found that men, urbanites, and married individuals were more likely to get employment. Tunali and Ercan (2003) state that graduates of vocational high schools are more easily employed than other graduates. Ilkkaracan (2012), besides, finds that the chances of finding a job are sexist and that women have lower chances than men. A study was done by Kütük and Güloğlu (2019) in where the same research question was examined; however, it was found that XGBoost and Random Forest applications on estimating transition probabilities were stuck at about 67%.

3. Data

All data used in this study are HLFSs data published between 2004-2016 issued by officially TURKSTAT[2]. Although there are annual HLFSs data between 2000 and 2003, they cannot be used because individuals' statuses in the previous period cannot be differentiated. In total, the data size consists of 203,891 people and its span is 13 years. All data are filtered vertically (in terms of rows, the survey data) and horizontally (in terms of variables/features). Initially, vertical filtering was applied in the data of individuals as follows:

- The individual must be unemployed in the previous year[3] (according to the ILO definition).
- The next period should be employed or continue to be unemployed.
- The individual must be 15 years old or more.
- The individual must not be counted in the institutional population.

In addition to the filters mentioned above, horizontal filtering was also applied as follows:

- In the current period, it should not contain any variables that indicate whether the individual is employed or not.
- Variables/features that are not (directly or indirectly) related to the individual should be excluded from the set.

[2] Turkish Statistical Institute.
[3] JOBLEFT YEAR

Afterwards, certain variables are generated from original variables, for instance, the duration of inactivity [4] in terms of years.

The data obtained in the light of all these consist of 203,891 individuals. There are 43 variables in this dataset, including derived data. The whole dataset was then divided into the train set and test set. Logit binary choice estimation and ML algorithms, which are accepted as econometric methods, will be run primarily on the train set, which is approximately 85 % of the dataset. Again, 15 % of the data is reserved as a test set that will be used to determine and compare the performance rates.

The tabulation matrix, Table 1, is given below by year: The data have numerous NAs in both qualitative and quantitative features. To be able to run an ML classifier that is so sensitive to NAs, the NAs in this data were replaced as follows:

- Qualitative features/variables are encoded as factors. In order not to drop NAs, they are regarded as another level which is unknown in each qualitative features.
- In quantitative features, NAs are filled with medians of that feature for the current survey. Results do not show significant changes.

Table 1: Tabulation of Transitions by Year

Years	t_0 Situtation		t_1 Situation		Transition Ratio
			Unemp	13426	
2004	Unemp	21148			0.365141
			Emp	7722	
			Unemp	15136	
2005	Unemp	23275			0.349689
			Emp	8139	
			Unemp	14110	
2006	Unemp	21669			0.348839
			Emp	7559	
			Unemp	13685	
2007	Unemp	20972			0.347463
			Emp	7287	
			Unemp	14024	
2008	Unemp	21428			0.345529
			Emp	7404	
			Unemp	18625	
2009	Unemp	28869			0.354844
			Emp	10244	
			Unemp	20097	

[4] Inactive=Survey Year - Jobleft Year

Table 1 continued

2010	Unemp	29930			0.328533
			Emp	9833	
			Unemp	18223	
2011	Unemp	26200			0.304466
			Emp	7977	
			Unemp	15941	
2012	Unemp	22782			0.300281
			Emp	6841	
			Unemp	14997	
2013	Unemp	21753			0.310578
			Emp	6756	
			Unemp	15064	
2014	Unemp	22131			0.319326
			Emp	7067	
			Unemp	15497	
2015	Unemp	22836			0.321379
			Emp	7339	
			Unemp	15066	
2016	Unemp	22449			0.328879
			Emp	7383	
			Σ	203,891	

4. Methodology

The main purpose of the study is to estimate the transition probabilities of individuals with a good prediction power. However, an individual must be unemployed in the past; another side of transition probabilities are not examined in this study when currently working. If an individual is employed in the following period, it becomes 1, and if the unemployment situation continues, it becomes 0. Therefore, the dependent variable set is discrete data consisting of y = 0,1. Such discrete variable estimates are referred to as a classification problem in the machine learning literature. Classification problems are estimated by binary choice solutions called probit, logit, and Tobit family in econometric literature.

In machine learning literature, classification problems can be solved with numerous methodologies. Nevertheless, here, the models which are shallow are preferred. The whole econometric process, therefore, can be called as a shallow neural network which consists of one perceptron. It also allows comparing both estimation methodologies easily. Hence, the algorithms used in this study are given in Table 2.

Table 2: Machine Learning Algorithms

Function *F*	**Methods**
Decision Trees	Extra Trees Classifier
	XGBoost Classifier
Ensemble Methods	Random Forest
	Bagging
Neural Networks	Shallow Artifical Neural Network
	Multilayer (Deep) Artifical Neural Network

Note: Impurity metrics for top four algorithms which were able to allow was provided in Appendix 3.

The solutions in the Scikit-Learn (Pedregosa et al., 2011) solve Extra Tree Classifier, XGBoost Classifier, Random Forest, and Bagging listed in the machine learning algorithms. Shallow Artifical Neural Network (ANN) model, one of the artificial neural network models in the last two stages, is solved in Keras (Chollet, 2015) and Multilayer (Deep) Artifical Neural Network is solved in Tensorflow (Abadi et al., 2015). All solutions are implemented in Python 3.7 (Van Rossum and Drake Jr, 1995).

In order to use the above machine learning algorithms counted in Table 2 and logistic regression, the dataset is divided into two parts, train set, and test set. Receiving the dataset sequentially over the years will produce biased estimates since it also includes trends and being a member of family information. The dataset was stratified by Stratified K-Fold (James et al., 2013) method on a yearly basis. Therefore, the number of layers will be 13 that was parallel to the length of the year. Repeated observations are also prevented by random shuffling. Finally, train and test data have dimensions [188208,43] and [15683,43] respectively[5].

Due to that ANN-based models shallow ANN and Deep ANN which is as difficult as to calculate machine learning algorithms for both the train set and the test set, min-max normalizer is used for scale variables.

In addition, the Deep ANN infrastructure needs specific interest, which should be mentioned. For this architecture, the Deep ANN infrastructure was established with four hidden layers and a total of 43 variables considered as inputs. There are a total of 25 neurons in the first hidden layer after the input layer. Softmax is used as the activation function in that one. The second hidden layer contained a total of 40 neurons, and the activation function was again selected as Softmax. The third hidden layer contains ten neurons, but in here, this time using the RELU activation function. In the last hidden layer,

[5] [number of rows (observations), number of variables/features]

there is a total of 10 neurons. The sigmoid function was selected as the activation function to estimate the transition probability.

Finally, after all the results obtained, the decision rule for the transition probabilities was decided as follows:

- If the probability of an individual is equal or below 0.5000, an individual is accepted as unemployed in the current period. That is to say; individual stays her/his position as unemployed.
- If the probability of an individual is above 0.5001, an individual has a chance to be employed in the current period, namely, individual transit to be employed from being unemployed.

5. Results

As previously mentioned, all variables used in the dataset were divided into two parts. These are quantitative and qualitative variables/features. Descriptive statistics for quantitative variables can be extracted and interpreted; these statistics are presented in Appendix 1. However, for qualitative variables, it is necessary to use tabulation instead of descriptive statistics. These tabulations are provided in Appendix 2. In addition, weights are included in the variable set since it carries valuable information from the NUTS2 level. Similarly, the deflator is also added in the variable set as it contains a basic level of information about the economic conditions of the country for the current year of HLFS.

Then, the probability of transition was estimated with the algorithms in Table 2 and logistic regression as the econometric approach. Accordingly, the estimated probabilities [Prob(ˆy)] are classified by applying the decision rule mentioned above. In the algorithms based on the learning rate, for instance, λ is determined as 0.1 for XGBoost while λ is 0.01 for Shallow and Deep ANN. In addition to these, for Deep ANN, the Dropout technique is used to get rid of vanishing or exploding gradient decent problem. In order to avoid overfitting problems for Shallow ANN, L2 type, namely Ridge Regression regularization, was performed in Keras framework. Extra Tree Classifier algorithm could not get rid of the overfitting problem in the training dataset since it could not be optimized in all parameters.

When the results were examined, ANN-based on deep learning showed a performance that would surpass all other algorithms. Accordingly, by using the 43 variables in the table given in addition to the deep learning, individuals' probability of transition to employment can be estimated with 74% accuracy. Moreover, no exploding or vanishing gradient descent problems were encountered in this 5-layer deep learning network. The other two

algorithms that followed were very close to each other. Accordingly, XGBoost Gradient Descent Classifier and Random Forest Classifier estimate 67%, 66%, respectively. Extra Tree Classifier, one of its closest followers, and logistic regression, an econometric method, remained at around 63%. All results are presented in detail in Table 3.

When the results are evaluated, it has been seen that the methods based on machine learning give better results than the econometric method. As stated in Mincer (1974)'s study, the main reason for this success is the feeding of the model with non-linear combinations of variables. Moreover, it can easily and automatically determine the effects of prediction by giving more weights to those who affect it.

Table 3: **Best Classifiers Results**

Classifier	**Classes**	**Precision**	**Recall**	**F1-score**	**Support**	**Confusion Matrix**		**Accuracy**	**AUC**
Extra Tree	0	1	1	1	93740	93740	0	OVERFIT	OVERFIT
Train	1	1	1	1	94468	26	94442		
	Avg/Total	1	1	1	188208			0.9999	1.0000
Extra Tree	0	0.62	0.69	0.65	7811	5409	2402		
Test	1	0.65	0.58	0.61	7872	3324	4548		
	Avg/Total	0.64	0.63	0.63	15683			0.6349	0.7073
XGBoost	0	0.73	0.73	0.73	93740	68501	25239		
Train	1	0.73	0.74	0.73	94468	24955	69513	0.7333	0.8243
	AVG/Total	0.73	0.73	0.73	188208				
XGBoost	0	0.67	0.66	0.67	7811	5176	2635		
Test	1	0.67	0.68	0.67	7872	2534	5338		
	AVG/Total	0.67	0.67	0.67	15683			0.6704	0.7492
Random Forest	0	0.76	0.75	0.75	93740	70549	23191		
Train	1	0.75	0.76	0.76	94468	22852	71616	0.7554	0.8486
	AVG/Total	0.76	0.76	0.76	188208				
Random Forest	0	0.66	0.66	0.66	7811	5146	2665		
Test	1	0.66	0.66	0.66	7872	2653	5219		
	AVG/Total	0.66	0.66	0.66	15683			0.6609	0.7404
Bagging	0	0.65	0.6	0.62	93740	56347	37393		
Train	1	0.63	0.67	0.65	94468	30916	63552	0.6371	0.6948
	Avg/Total	0.63	0.63	0.63	188208				
Bagging	0	0.64	0.57	0.6	7811	4661	3150		
Test	1	0.62	0.68	0.65	7872	2559	5313		
	Avg/Total	0.63	0.63	0.63	15683			0.6360	0.6958

Table 3: Best Classifiers Results (cont'd)

Classifier	Classes	Precision	Recall	F1-score	Support	Confusion Matrix		Accuracy	AUC
Extra Tree	0	1	1	1	93740	93740	0	OVERFIT	OVERFIT
Train	1	1	1	1	94468	26	94442		
	Avg/Total	1	1	1	188208			0.9999	1.0000
Extra Tree	0	0.62	0.69	0.65	7811	5409	2402		
Test	1	0.65	0.58	0.61	7872	3324	4548		
	Avg/Total	0.64	0.63	0.63	15683			0.6349	0.7073
Shallow NN								0.5018	0.5000
Train	Avg/Total								
Shallow NN									
Test	Avg/Total							0.5019	0.5000
Deep NN								0.7614	0.7597
Train	Avg/Total								
Deep NN									
Test	Avg/Total							0.7461	0.7395
Logistic Regression		0.6371	0.5999	0.6179	102340	60917 40634	34702 67638	0.6305	0.6411

6. Conclusion

The primary purpose of this study is to estimate the probability of transition to employment with high accuracy. Therefore, the possibility of a two-way transition was not estimated, and only the employment of unemployed individuals into employment or staying of the current situation was examined so that it can be considered as a one-way transition. This study aims to feature the power and the accuracy of forecasting. In doing so, it aims to use relatively big data. Firstly, prediction power can be increased with greater accuracy through big data. Secondly, this is an excellent way for Turkey's labour market data that have not been created before. In order to standardize HLFS surveys, this big data have been tried to be created.

On the other hand, it is attempted to look at as wide a range as possible, thus trying to present a broader perspective. For this purpose, data were collected from the beginning of the 2000s. Therefore, HLFS data published by TURKSTAT between 2004-2016 were compiled and aggregated.

All variables regarding the employment status of individuals in the current period are excluded from the dataset. The main thing that is predicted is the employment situation in the current period. Therefore, only personal, experiential, and past information about individuals are reserved.

The current period status of the individuals was considered as unknown. If unemployed individuals could find a job in the next period, they were coded as (1), and if they could not, they were coded as (0). Therefore, the transition possibilities of individuals fit the dichotomous model. The logit model in econometric methodology solves these problems. Similarly, in machine learning methodology, it is accepted as a classification problem. For this purpose, effectual classifiers, one dimensional and deep artificial neural networks are used in machine learning methodology. In ANN models, the activation function of the last layer is taken as the sigmoid function, which can estimate the probability.

When the performance levels of all probabilities were compared, the performance of 5-layer deep artificial nerve areas was found to be quite high. Accordingly, considering the 43 characteristics of an individual, the probability of transition to employment in the next period can be estimated at 74%. This prediction power outperforms all other models. The most important reason for this is that deep artificial neural networks include non-linear combinations of all variables used in the model. In terms of machine learning classifiers, XgBoost, and Random Forest classifiers could reach 67% accuracy rate to estimate the current state of unemployed individuals in the Turkish labour market. The next followers are Extra Trees Classifier and Bagging 11 from machine learning methodology have nearly 63% accuracy rate, which is also shared and achieved by logistic regression from the econometric methodology. One layered that is shallow, the neural network model could get only 50%, which the percentage is also predicted by chance from any individual that has no knowledge about the labour market. So its power could be regarded as a complete failure.

According to the relatively higher predictive power of estimates for the probability of individuals' employment transition, this model based on deep ANN could be considered very important for the Turkish labour market. Policymakers, micro simulators, researchers who prepared projections for the Turkish labour market and employment agencies can perform more successful estimates by using predictions based on this kind of big data and neural networks based on deep learning.

References

Abadi, M., Agarwal, A., Barham, P., and Zheng, X., 2015, TensorFlow: Large-scale machine learning on heterogeneous systems, Software available from tensorflow.org.

Abdulkadiroglu, A. and Sönmez, T., 2013, Matching markets: Theory and practice, Advances in Economics and Econometrics, 1, 3–47.

Acemoglu, D. and Autor, D., 2011, Lectures in labour economics, Manuscript. http://economics. mit. edu/files/4689.

Becker, G., 1964, Human capital: a theoretical and empirical analysis, with special reference to education. National bureau of economic research publications: General series. National Bureau of Economic Research; distributed by Columbia University Press.

Bulutay, T., 1992, A general framework for employment in Turkey, State Institute of Statistics, Turkey.

Bulutay, T., 1995, Employment, unemployment and wages in Turkey, International Labour Organization.

Chollet, F., 2015, Keras, https://keras.io.

Cohn, E., Kiker, B. F. and De Oliveira, M.M., 1987, Further evidence on the screening hypothesis, Economics Letters, 25(3), 289–294.

Fountain, C., 2005, Finding a job in the internet age, Social Forces, 83(3), 1235–1262.

Gale, D. and Shapley, L. S., 1962, College admissions and the stability of marriage, The American Mathematical Monthly, 69(1), 9–15.

Groot, W. and Oosterbeek, H., 1994, Earnings effects of different components of schooling; human capital versus screening. The review of Economics and Statistics, 76, 2, 317–321.

Hall, R. E., 1979, A theory of the natural unemployment rate and the duration of employment, Journal of monetary economics, 5(2), 153–169.

İlkkaracan, İ., 2012, Why so few women in the labour market in turkey? Feminist Economics, 18(1), 1–37.

James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013, An introduction to statistical learning, Springer.

Kaitz, H. B., 1970, Analyzing the length of spells of unemployment, Monthly Labour Review, 93, 11-20.

Kroch, E. A. and Sjoblom, K., 1994, Schooling as human capital or a signal: some evidence, Journal of Human Resources, 29, 156–180.

Kütük, Y. and Güloğlu, B., 2019, Prediction of transition probabilities from unemployment to employment for turkey via machine learning and econometrics: A comparative study, Journal of Research in Economics, 3(1), 58–75.

Lynch, L. M., 1989, The youth labour market in the eighties: Determinants of re-employment probabilities for young men and women, The Review of Economics and Statistics, 71, 1, 37–45.

Mincer, J., 1974, Schooling, experience, and earnings, Human Behavior & Social Institutions No. 2, National Bureau of Economic Research, Inc., New York, NY.

Mincer, J. and Ofek, H., 1982, Interrupted work careers: Depreciation and restoration of human capital, Journal of human resources, 17, 3–24.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E., 2011, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12, 2825–2830.

Reynolds, L.G., 1951, The structure of labour markets, New York, Harper.

Roth, A.E., 1984, The evolution of the labour market for medical interns and residents: a case study in game theory, Journal of political Economy, 92(6), 991–1016.

Schultz, T. W., 1961, Investment in human capital, The American Economic Review, 51(1), 1–17.

Şenses, F., 1996, Structural adjustment policies and employment in Turkey, New Perspectives on Turkey, 15, 65–93.

Spence, M., 1978, Job market signalling, In Uncertainty in Economics, 281–306, Elsevier.

Stigler, G.J., 1961, The economics of information, Journal of political economy, 69(3), 213–225.

Tansel, A. and Tasci, H. M., 2004, Determinants of unemployment duration for men and women in Turkey, IZA Discussion Paper No. 1258.

Tunali, I. and Ercan, H., 2003, Background study on labour market and employment in Turkey, European Training Foundation, Torino.

Van Rossum, G. and Drake Jr, F. L., 1995, Python tutorial, Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Appendix 1: Descriptive Statistics of Quantitative Features

Statistic	N	Mean	St. Dev.	Min	Max
age	203,891	32.150	11.160	15	99
tr coming year	3,569	15.360	11.580	0	71
living place	3,569	15.360	11.580	0	71
educ year	203,891	7.545	3.947	0	15
experience	203,891	14.970	11.650	0	83
deflator	203,891	1.114	0.291	0.681	1.655
weight hlfs	203,891	160.800	79.950	5.620	754.200
hh population	203,891	4.616	2.234	1	31
hh population estimated	203,891	4.485	2.183	1	31
inactive	52,774	7.486	9.646	0	73

Appendix 2: Descriptive Statistics of Qualitative Features

gender	**age.group**	**birth loc**	**abroad exp6**
Min. :1.00	Min. : 4.00	Min. :1	Min. :1
1st Qu.:1.00	1st Qu.: 5.00	1st Qu.:1	1st Qu.:2
Median :1.00	Median : 7.00	Median :1	Median :2
3rd Qu.:1.00	3rd Qu.: 9.00	3rd Qu.:1	3rd Qu.:2
Max. :2.00	Max. :14.00	Max. :2	Max. :2
		NA's :70381	NA's :100874
relative type	**recent school grad k**	**foet99 k**	**nuts1**
Min. : 1.00	Min. :0.00	Min. : 1	Min. : 1.00
1st Qu.: 1.00	1st Qu.:2.00	1st Qu.: 6	1st Qu.: 3.00
Median : 3.00	Median :3.00	Median : 6	Median : 6.00
3rd Qu.: 3.00	3rd Qu.:4.00	3rd Qu.:12	3rd Qu.: 9.00
Max. :11.00	Max. :6.00	Max. :21	Max. :12.00
		NA's :171336	
nuts2	**prev residence**	**prev living**	**spouse**
Min. : 1.0	Min. :1	Min. :1	Min. : 1
1st Qu.: 5.0	1st Qu.:1	1st Qu.:1	1st Qu.: 2
Median :11.0	Median :1	Median :2	Median :99
3rd Qu.:19.0	3rd Qu.:1	3rd Qu.:3	3rd Qu.:99
Max. :26.0	Max. :2	Max. :3	Max. :99
	NA's :155343	NA's :160119	NA's :13426

Appendix 2: Descriptive Statistics of Qualitative Features (cont'd)

mother	**father**	**literacy level**	**recent school grad**
Min. : 1	Min. : 1	Min. :1	Min. :1.00
1st Qu.: 2	1st Qu.: 1	1st Qu.:1	1st Qu.:2.00
Median : 3	Median :99	Median :1	Median :2.00
3rd Qu.:99	3rd Qu.:99	3rd Qu.:1	3rd Qu.:2.00
Max. :99	Max. :99	Max. :2	Max. :2.00
NA's :13426	NA's :13426	NA's :122214	
contd school	**contd school year**	**course attandence**	**course aim**
Min. : 1	Min. :1	Min. :1.00	Min. :1
1st Qu.: 4	1st Qu.:1	1st Qu.:2.00	1st Qu.:1
Median : 4	Median :2	Median :2.00	Median :1
3rd Qu.: 5	3rd Qu.:3	3rd Qu.:2.00	3rd Qu.:2
Max. :32	Max. :9	Max. :2.00	Max. :3
NA's :189243	NA's :191939		NA's :202449
maritial status	**job prev**	**jobleft reason**	**nace2 job prev**
Min. :1	Min. :1	Min. : 1	Min. : 1
1st Qu.:1	1st Qu.:1	1st Qu.: 1	1st Qu.: 2
Median :2	Median :1	Median : 3	Median : 3
3rd Qu.:2	3rd Qu.:2	3rd Qu.: 7	3rd Qu.: 4
Max. :9	Max. :2	Max. :13	Max. :11
NA's :17532	NA's :112294	NA's :164341	NA's :164341
isco08 job prev	**jobprev status**	**jobprevyear nace**	**jobprevyear status**
Min. :1	Min. :1	Min. : 1	Min. : 1
1st Qu.:5	1st Qu.:1	1st Qu.: 1	1st Qu.: 1
Median :7	Median :1	Median : 3	Median : 2
3rd Qu.:9	3rd Qu.:2	3rd Qu.: 6	3rd Qu.: 3
Max. :9	Max. :5	Max. :11	Max. :10
NA's :164341	NA's :164341	NA's :151684	NA's :151812

Appendix 3: Impurity Measurement Metrics

ML Agorithm	Set	Gini Impurity	Cohen's (κ)	Entropy	Class Error
ExtraTrees Classifier	Train	0.0000	0.0000	0.0000	0.9997
ExtraTrees Classifier	Test	0.2573	0.0000	0.0000	0.0000
XGB Classifier	Train	0.0030	0.0002	0.3820	0.1744
XGB Classifier.	Test	0.3452	0.0019	0.0001	0.4261
Random Forest Classifier.	Train	0.0000	0.0000	0.0000	0.9696
Random Forest Classifier	Test	0.0594	0.0068	0.0000	0.2651
Bagging Classifier	Train	0.0000	0.0000	0.4999	0.2604
Bagging Classifier	Test	0.0000	0.0000	0.4999	0.2513

Note: These statistics are not available for SVC, LinearSVC and Shallow NN.