İktisat Araştırmaları Dergisi • Journal of Research in Economics • Cilt: 3 / Issue • Sayı: 1 Mart • March 2019 • ss/pp. 58-75
ISSN: 2636-8307 • DOI: 10.24954/JORE.2019.29

PREDICTION OF TRANSITION PROBABILITIES FROM UNEMPLOYMENT TO EMPLOYMENT FOR TURKEY VIA MACHINE LEARNING AND ECONOMETRICS: A COMPARATIVE STUDY

TÜRKİYE İÇİN MAKİNE ÖĞRENMESİ VE EKONOMETRİ YOLUYLA İŞSİZLİKTEN İSTİHDAMA GEÇİŞ OLASILIKLARININ TAHMİNİ: KARŞILAŞTIRMALI BİR ÇALIŞMA

Yasin KÜTÜK *
Bülent GÜLOĞLU **

Abstract

In this study, it is mainly aimed to predict transition probabilities of individuals who are previously unemployed and get employment or stay unemployed. In order to do that, Household Labor Force Surveys conducted in Turkey are merged and matched from 2004 to 2016. Information about individuals only consists of individual characteristics and qualifications since there should not be any informative clue about the present situation. To predict those probabilities, logistic regression analysis as econometric approach, a shallow neural network and machine learning classification algorithms are run in order to compare them. The results indicate that classification in machine learning is slightly better than logistic regression and shallow neural network. While XGBoost classifier and Random Forest get 67% accuracy, logistic regression can predict only 63% of an individual's transition and shallow neural network forecasts 51%.

Keywords: Employment, Transition Probability, Machine Learning, Classification

JEL Classification: J21, C38, C45

Öz

Bu çalışma içerisine, esasen daha önce işsiz olan bireylerin istihdama geçiş veya işsizlikte kalma olasılıklarını tahmin etmek amaçlanmaktadır. Bu amaçla, Türkiye'de yapılan Hanehalkı İşgücü Anketleri

* Corresponding Author, Ph.D. Candidate at Istanbul Technical University and R.A at Altınbaş University, Department of Economics, SEASS, Social Sciences Campus, Büyükdere Caddesi, No: 147, PK.34349, Esentepe, Şişli / Istanbul, Turkey, Tel:(+90) 212.604.0100 (2071). yasinkutuk@itu.edu.tr | yasin.kutuk@altinbas.edu.tr, Orcid: 0000-0003-2304-8309

** Istanbul Technical University, Department of Economics, Management Faculty, Harbiye Mahallesi, 34367 Maçka, Beşiktaş / Istanbul, Turkey, Tel: (+90) 212.293.1300 (2029). guloglub@itu.edu.tr

Makale Gönderim Tarihi: 15.02.2019
Yayına Kabul Tarihi: 01.03.2019

2004 ile 2016 arasında birleştirilmiş ve eşleştirilmiştir. Veriler, bireylerle ilgili mevcut istihdam/işsizlik durumu hakkında herhangi ipucu içermeyecek şekilde, bireysel özellik ve niteliklerle oluşturulmuştur. Bu geçiş olasılıklarını tahmin etmek ve bunları karşılaştırmak amacıyla, ekonometrik yaklaşım olarak lojistik regresyon analizi, tek katmanlı yapay sinir ağı ve yapay öğrenme sınıflandırma algoritmaları uygulanmıştır. Sonuç olarak, yapay öğrenme algoritmalarının, lojistik regresyon ve tek katmanlı sinir ağından görece daha iyi olduğunu göstermektedir. XGBoost sınıflandırıcısı ve Rassal Orman Karar Ağaçları algoritmaları %67 doğruluk ile, lojistik regresyon bir bireyin geçiş olasılığını yalnızca %63 düzeyinde ve tek katmanlı yapay sinir ağları ise % 51'ini tahmin edebilmektedir.

Anahtar Kelimeler: İstihdam, Geçiş Olasılığı, Yapay Öğrenme, Sınıflama

JEL sınıflaması: J21, C38, C45

1.Introduction

Not only in strong, rich and wealthier economies, but in all economies, more or less, unemployment is a phenomenon that exists not only in theoretical textbooks and articles, but also in everyday life or society. As ILO determines, "*the definition of unemployment covers people who are: out of work, want a job, have actively sought work in the previous four weeks and are available to start work within the next fortnight; or out of work and have accepted a job that they are waiting to start in the next fortnight*" [1]. This definition includes only those who are actively seeking jobs, but the duration of active job search is not determined. For this reason, this active job search may be short or long. Important factors determining this period vary from the economic environment involved to the characteristic and experiential characteristics of the person. Economic environments vary from country to country. While some countries that may be export-oriented and can provide current account surpluses can create job opportunities in line with growth rates due to rich economic environment, some countries based on labor-intensive sectors can provide less employment opportunities due not to chronically reaching high-value-added technology intensive sectors. From the perspective of labor economics, the significant point is that how an economy can deal with unemployment with successful policies within its economic environment. History of unemployment says most of countries have not been passed in that issue. Historically, there has been a marked increase in unemployment rates, throughout the countries. This increase in unemployment rate is quite obvious in Table 1:

Table 1. Unemployment Rates all over the World

Country	1950-73	1974-83	1984-93	1994-98
Belgium	3.0	8.2	8.8	9.7
Finland	1.7	4.7	6.9	14.2
France	2.0	5.7	10.0	12.1
Germany	2.5	4.1	6.2	9.0
Italy	5.5	7.2	9.3	11.9
Netherlands	2.2	7.3	7.3	5.9

1 http://www.ilo.org/ilostat-files/Documents/description UR EN.pdf Accessed: 10.11.2017.

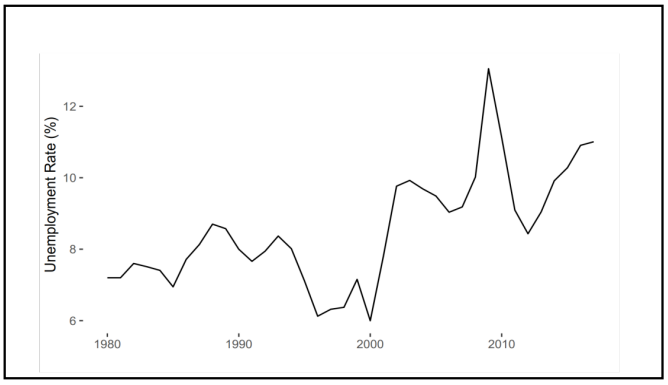
Norway	1.9	2.1	4.1	4.6
Sweden	1.8	2.3	3.4	9.2
United Kingdom	2.8	7.0	9.7	8.0
Ireland	n.a.	8.8	15.6	11.2
Spain	2.9	9.1	19.4	21.8
Average	2.6	6.0	9.2	10.7
Australia	2.1	5.9	8.5	8.6
Canada	4.7	8.1	9.7	9.4
United States	4.6	7.4	6.7	5.3
Average	3.8	7.1	8.3	7.8
Japan	1.6	2.1	2.3	3.4

Source: 1950-83 from Maddison et al. (1995) and updates from OECD, Labour Force Statistics accessed: 15.11.2017

There are many arguments for these fluctuations in history, but the reasons for all of them depend on several factors. Historically, despite the fact that the population has nearly doubled in all of the above countries, in all economies, except Japan, Italy and Netherlands, the unemployment ratio has tripled. When the average of each period is taken into account, the apparent increase is more pronounced. Moreover, all these countries, starting from the 1950s, have advanced to developed economies. Unemployment in developing countries presents worse scenarios than economically developed countries.

Turkey, the world's 17th economy, at the moment this study was carried out [2], is one of these developing economies. Having achieved significant growth rates for the last 15 years, Turkey did not present the same success in creating employment opportunities when the trend of unemployment rates is compared with others. After 2000, having experienced two crises, one in 2001 is national, the latter in 2008-2009 was international, unemployment rates has doubled regardless of getting back prior to 2000s' level. These increases have not been reverted back as these jumps can be easily seen in Figure 1:

Figure 1. Unemployment Rates in Turkey



2 https://www.worldatlas.com/articles/the-world-s-25-biggest-economies.html

The main causes of unemployment in Turkey result from the inability to achieve the desired level of industrialization, the quality remained limited due to the excess weight of the share of rural communities, inability to access the level of investment required. These reasons can quickly increase the unemployment of young people and carry structural problems. One of the major differences between developed countries and Turkey; while the level of education is high in the unemployed, the level of education in the underemployed is low in Turkey. Another major difference, in contrast to developed countries, is the unemployment rate in Turkey is higher among educated people; the cause of this is that the jobs are far from being satisfactory and unsuitable for individuals looking for work for the first time.

In order to deal with unemployment, the private sector in employment, which was initiated in 2017, became the forefront of employing the public sector to support the social security premium that should be paid to the employer. In addition to this, the tax that will be deducted from the salary of the employee will be paid by the state. According to the explanation given at the beginning, the employer who wants to benefit from the support of "one sibling" will pay the first month's tax, premium and salary of the new worker, according to the new incentives to be introduced in the coming period. Every two months from the second month, the expenses will be covered by the Unemployment Fund. Measures put into effect from the beginning of 2017 seem to be to provide temporary fixes by transferring private sector funds from the public sector rather than solving the problem permanently. The payment of the bill to the unemployment fund, which the public sector will undertake, brings many new debates.

The state of the economy is very important in the course of an individual's employment or transition to it. As mentioned before, the other most important factors are all about personal abilities. Moreover, in economies where unemployment is high, individual developments gain more importance because of the increased competition. In this sense, individuals attach importance to their personal development, receive more training, establish and expand links with their business networks, and acquire new skills. Turkey's labor market which is competitive in this perspective is a market where individual talents come to the fore. In that sense, numerous universities have opened to increase human capital, the number of individual vocational courses supported and subsidized by the government has been increased and the quality of pre-university education has been raised.

This study, at this point, examines individual skills, ingenuities and characteristics for the transition of individuals into employment. However, here it is necessary for individuals to be unemployed in accordance with the above ILO definition in the previous period, to move employment in the next period or to keep their situation. This is called as transition which can be calculated with probability. Since the calculation of this transitivity is important for policy makers or institutions, previous studies are not enough because both the estimation and prediction power are relatively low and the datasets are limited. As the second decade of the millennium brings renovated, strengthened and enriched tools to humankind in many areas, the most powerful one is data science. This study uses new estimation methods to increase predictability and forecast ability by

using machine learning techniques which use all the blessings of data science. Later, these results are compared to econometric ones.

In the following sections, first, a review of literature that covers job finding in the view of experimental studies in labor market, then some key figures are shown for Turkey. Later, the methods used in this study are presented with their data as well, and the results follow. Finally, the conclusion will summarize the whole study and provide some basic insight to the policy makers. Limitations and further studies conclude this study.

2.Literature Review

In the job market, all workers compete with each other. In a complete information and perfect competitive labor market, labor earns only marginal product according to classical theory. Again, a rational worker can invest as much as the cost of investment in his/her human capital (Becker, 1964) when the marginal investment of his/her human capital is positive. While it is positive, individual invests. Investment in human capital may be split into two main categories as offered: general and specific (Becker, 1964). General human capital arises if there is no firm-specific differentiation among workers and firms are not responsible for and do not tend to pay it to worker (Acemoglu and Shimer, 1999). Specific training is under the control of firms in order to increase workers' productivity while it helps to boost economic growth in a country by reducing quit rates and turnovers, increasing labor efficiency (Higashi, 2002; Parsons, 1972; Donaldson and Eaton, 1976). In the nature of specific investment, it is not certain that it accumulates during employment (Bougheas and Georgellis, 2004) leading to an increase in productivity and uncompetitive specializations. Wasmer (2006) and Cingano (2003) conclude specific investment in human capital and propose that firm-specific skills provide higher returns rather than others.

Empirical results show that the return on education is always positive, but researchers do not reach a common result on the subject of its amount. Dearden (1998), Ashenfelter et al. (1999) and Psacharopoulos (1994) find the effect of education around 5%, Mincer (1974), Krueger and Ashenfelter (1992) and Becker (1964) reach the effect of schooling on earnings around 11% level. In macroeconomics, Acemoglu (2002) conducts a study in less-developed regions and finds evidence that return on education has a multiplier effect on less developed countries. Blundell et al. (1999), Mincer and Ofek (1982) draw attention to the fact that gender differentiation on rate of return of education is in favor of male workers.

Tansel and Tasci (2004) uses 2001-2001 HLFS on the Turkey data to estimate unemployment transition probabilities, which may be the first application querying labor market transitions in Turkey. Proportional risk models, log-logistic and log-normal models were used in the study. The results indicate that urban or married individuals are more likely to be unemployed than other groups, but women are less likely to be unemployed than men. At the educational level, the probability of unemployment of college or undergraduate graduates is higher than other education groups, and the probability of becoming unemployed decreases with age. After that,

Tasci and Tansel (2005) estimated long-term unemployment durations with two-stage probit models using TURKSTAT's 2000-2001 Household Labor Force Surveys. Living in cities, for both types of unemployment, is reducing unemployment while women are increasing the possibility of long-term unemployment. Geographically, long-term unemployment decreases as we move from west to east. At the educational level, the probability of long-term unemployment decreases as the number of primary school graduates increases, but it still has positive effect, while the effect of long-term unemployment on graduates is negative. The appearance of gender discrimination in the labor market in Turkey, due to undergo more difficult than expected employment of women transition from unemployment is examined and this employment are high influence under finding a job in their field (İlkkaracan, 2012). Tunali and Ercan (2003) also indicates that although vocational school education is expected to help the graduates ease their way into the labor market, statistics reveal that a significant majority continues with higher education. Tasci and Ozdemir (2006) find that the transition from long-term unemployment to employment may be damaged by living in rural areas, being female, older and having lower education.

3.The Data

The data are collected from Household Labor Force Surveys (HLFS) conducted in between 2000 and 2016 by the Turkish Statistical Institute (TURKSTAT). In total, 7,323,449 individuals were surveyed by means of a set of questionnaires consisting of extensive information that were classified under seven headings such as household characteristics, individual's characteristics, employment status, income, unemployment and inactivity, past work experience, labor status one year before the survey. Total number of variables is 110 with three of them derived from other ones. The data as a csv file costs 3.18 GB to an Ext4 [3] file-system, which is an example of a big-data that it is hard to deal with classical computational methods. Rather, the data is compressed under Ext4 system by using "gzip" compression with a level of 6 by means of R that saves the data as an ".rdata' in order to process it with ease. The size of the big data after compression is reduced by up to 15%.

However, since this research is restricted only with both previously unemployed individuals and surveys done in 2003-2000 are not compatible to catch individuals who are previously unemployed and who are employed or unemployed at the survey date, the data merged are filtered against them [4] to focus only on those who are interested in this study. So, given that information, the data are downsized to 203,891 individuals who will be split as train set (188,208 individuals) and test set (15,683). Due to the inconsistency of the surveys, a map for variables is generated to match them. However, even if the same question is asked to individual in the next year, the answers may be changed by TURKSTAT. So, the matching process for questions is not enough, the answers must be compared in terms of both content encodings and means. Variables are categorized into four groups: characteristics, employment, unemployment and past information. Later, some of

3 Fourth extended file-system, generally used by Linux OSs.

4 Individuals who are not in labor force, institutional population and are not actively searching for a job with the definition of ILO are dropped.

them are processed to derive certain variables. For instance, it is asked when the individual had a job in his/her past. [5] The answer may not give a clear opinion, so the difference in between survey year and the answer is used to get the period of inactivity. [6] The tabulation matrix, Table 2, is given below by year:

Table 2. Tabulation of HLFS (2004-2016)

Year	t0	t1_Unemp	t1_Emp
2004	Unemp	13426	7722
2005	Unemp	15136	8139
2006	Unemp	14110	7559
2007	Unemp	13685	7287
2008	Unemp	14024	7404
2009	Unemp	18625	10244
2010	Unemp	20097	9833
2011	Unemp	18223	7977
2012	Unemp	15941	6841
2013	Unemp	14997	6756
2014	Unemp	15064	7067
2015	Unemp	15497	7339
2016	Unemp	15066	7383
		SUM	203,891

Note: t0 means previous term while t1 means current period of survey.

The data have so many NAs in both qualitative and quantitative features. To be able to run a ML classifier that is so sensitive to NAs, 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 median's of that feature for the current survey. Results do not show significant changes.

4.Methodology

As it was previously stated, since the dependent variable (feature) is a dichotomous one that indicates current employment or unemployment, it should be analyzed with discrete (or qualitative) choice models in econometrics or supervised-classification problem in machine

5 JOBLEFT YEAR

6 INACTIVE=SURVEY YEAR – JOBLEFT YEAR

learning literature. Greene (2003) splits discrete choice models as binary, multinomial or ordered. [7] So, in econometric terms, the problem can be regarded as a binary choice problem.

In machine learning literature, classification problems can be solved with numerous methodologies. But here, the models which are shallow are preferred. The reason why shallow models are chosen is that since the econometric methodology can be regarded a one-layered neural network with some inputs that are weighted with their coefficients, an activation function that is a sigmoid and one output that is probabilities which are estimated with marginal coefficients is produced consequently. The whole econometric process, therefore, can be called as a shallow neural network which consists of one perceptron. It also allows comparing easily both estimation methodologies. Hence, the algorithms used in this study are given below:

Table 3. Machine Learning Algorithms

Function F	**Methods**
Naïve Bayesian Classifier	Gaussian
	Bernoulli
Support Vector Machines	Support Vector Classification (SVC)
	Linear SVC
Decision Trees	Extra Trees Classifier
	CART
	Logistic Model Trees
	Gradient Boosted Trees (XGBoost)
Discriminant Analysis	Linear Discriminant Analysis
	Quadratic Discriminant Analysis
Stochastic Gradient Descent	SGD Classification
Neighbors Classifier	KNeighbors
Ensemble Methods	Random Forest
	Bootstrapped Aggregation (Bagging)
	Adaboost
Neural Networks	Shallow/Wide Artifical Neural Network

All modules except the neural network were run under Python thanks to the Scikit-Learn (Pedregosa et al., 2011), the marvelous package, while shallow neural network was developed under Keras (Chollet et al., 2015). All were developed with and written in Python.

In these machine learning algorithms, the data must be split into two parts [8]: Train and test data (Breiman and Spector, 1992). An incisive separation of the data makes it vulnerable in prediction. In order to prevent that problem, Stratified K-Fold with randomized but not duplicated is a well suited algorithm to split the data (James et al., 2013). Strata are determined by the years, in the

7 and event counts, but it is not appropriate the case of this study.

8 Separation of cross validation/development data is skipped.

data, it is 13. So at first, Stratified K-Fold algorithm splits the data into 13 strata. Then each stratum is shuffled. Later, shuffled observations are attained into a test set at 8% approximately without duplication. The shape of the train data is [188208,43], while test's is [15683,43].

The final data reached, after being filtered by several conditions stated above, total observations which survived are 203,891 who were unemployed in the previous year of the survey date. To deal with the computationally costly process of ML algorithms, all quantitative variables/features are normalized in 0 and 1 according to the survey year. The variables that are used are classified according to their types within a broad classification: quantitative and qualitative variables. NAs in quantitative variables are accepted as another category since all the level of that kind of variables has levels. [9] Hence for the quantitative variables, this process is simplified. All NAs are filled with the median value for the current year. Since median is more robust to outliers and large datasets, it is preferred (Rousseeuw and Bassett Jr, 1990).

Afterwards, a well-known econometric approach, logistic regression, for discrete/binary choice observations is applied to estimate transition probabilities for individuals who were unemployed in the previous term and moves to employment or stayed unemployed again for the current survey period.

5.Results

Descriptive statistics for the variables that are used in this study are given in Appendices. Quantitative features are summarized in Table 5 while qualitative features are in Table 6. Even the variable that was used to estimate population projections, WEIGHT_HLFS is also utilized to estimate transition probabilities, as it conveys regional information to estimation. Again, the deflator is also used since it has a small but significant role to convey information about the state of the economic environment for Turkey to the estimation processes. Therefore, no other variables were derived and used in conjunction with the variables specified in the descriptive statistics.

Table 4. Classifier Results

Classifier	Classes	Precision	Recall	F1-score	Support	Confusion Matrix		Accuracy	AUC
Random Forest	0	0.76	0.75	0.75	93740	70549	23191		
Train	1	0.75	0.76	0.76	94468	22852	71616		
	Avg / Total	0.76	0.76	0.76	188208			0.7554	0.8486
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
XGBoost	0	0.73	0.73	0.73	93740	68501	25239		

9 or factors, categories etc.

Train	1	0.73	0.74	0.73	94468	24955	69513		
	Avg / Total	0.73	0.73	0.73	188208			0.7333	0.8243
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
Ada Boost	0	0.65	0.6	0.62	93740	56347	37393		
Train	1	0.63	0.67	0.65	94468	30916	63552		
	Avg / Total	0.63	0.63	0.63	188208			0.6371	0.6948
Ada Boost	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
Extra Tree	0	1	1	1	93740	93740	0		
Train	1	1	1	1	94468	26	94442		
	Avg / Total	1	1	1	188208			OVERFIT	OVERFIT
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
						60917	34702		
Logistic Regression		0.6371	0.5999	0.6179	102340	40634	67638	0.6305	0.6300

Later, the algorithms stated in Table 3 are run to get transition probabilities for the same sample again. For the Support Vector Classifier (SVC), radial basis function is used as kernel with 1.0 penalty parameter. Linear Support Vector uses squared hinge loss function. All tree based classifiers, Extra Tree Classifier and CART, since the data uses two classes which contain impurities, criterion contains only Gini for splitting the classes. Learning rate α for the XGBoost classifier is determined as 0.1 while the shallow network has 0.01. Logistic Regression Trees, Stochastic Gradient Descent Classifier and Linear Support Vector have their regularization terms prepared with L2, the ridge regression is utilized for them not to emerge over-fitting. However, Extra Tree Classifier gives over-fitted results due to catching information about next term employment for individuals.

The results indicate that the ML algorithm is relatively better than an econometric one (see Table 4, the Logistic Regression). In order to keep comparability between econometric and machine learning solutions, the shallow neural network is prepared as one layered since other algorithms including logistic regression has one process. Top two algorithms which get close accuracy rates are XGBoost Gradient Descent Classifier and Random Forest Classifier, their accuracies are 67%, 66%. Both of them use decision trees which is so suitable for the classification problems. As the classes that are employed or unemployed at the survey date but are unemployed in previous

term are not imbalanced, namely the ratio between two classes do not exceed 95% in any year, these classifiers successfully distinguish the classes by using characteristic information about individuals.

AdaBoost Classifiers follows these top two, where accuracy rate is around 63.6%. Extra-Tree classifier comes from later than AdaBoost Classifiers and gets 63.5% which is quite similar to AdaBoost's. But these latter classifiers perform so close to econometric approach since logistic regression which has a determined cut-off probability gets nearly 63% accuracy (see Table 4). In order to predict t1 term's unemployment (0) or employment (1), the cut-off value is taken as 0.5. So, if the probability is greater than the cut-off, the label gets a 1 which indicates predicted t1's situation is employment, otherwise, it is 0, that is individual stays unemployed in survey date. While logistic regression is beaten by four machine learning classifiers, it can successfully beat the rest of classifiers that are listed in Table 3 and resulted in Appendix in Table 7.

When it is elaborated why machine learning estimation methods could beat econometric ones, the following reasons come to mind. As stated in the literature, Mincer's (1974) equation establishes a non-linear equation where experience is taken as squared as independent variable. Since, especially XGBoost and Random Forests are completely superior tools when there is a non-linearity between features (explanatory variables) in a classification problem, they can deal with the problem rather than linear solutions even if the non-linear features do not feed the model. Second, linear models are good at monotonic relationships. Non-linearity breaks the assumption about monotonicity, therefore, machine learning estimation methods work splendidly. Theoretically, tree-based models in principle can approximate functions regardless of shape, whereas linear models can only produce functions with a linear shape with respect to a chosen set of features. Third is related to cut-off point. In logistic regression, it is required to determine cut-off as 0.5 where the z-value turns 0 indicating the class is not determined. However, in machine learning algorithms, this cut-off value does not depend on any value. Owing to determine the classes the path of trees that reach 1 or 0, there is no cut-off value which should be optimal.

The last concerns the size of the data, by virtue of processing relatively big data, computational issues take more time than expected. XGBoost and Random Forest classifiers can run in approximately 0.0007% of the time logistic regression and can beat it with ease, machine learning estimation methods can be chosen if the task needs to be solved within a certain time.

6.Conclusion

In this study, the main aim is to predict how individuals who are unemployed in previous term pass to employment or stay at current state with his/her characteristics and qualifications. In order to do so, Turkish labor market is handled with care by using relatively bigger data. Side task is to look at Turkish labor market with recent estimation methods which provide more accuracy compared to previous ones.

In order to look at transition from unemployment from a broader angle, the time span was chosen as wide as possible. Household labor force surveys which were decided on to calculate labor related issues were utilized from the beginning 2000. However, since it is impossible to derive the state of previous terms, starting year is accepted as 2004, then the data are expanded to 13 years, to 2016. Most of the variables are matched as much as possible. Though, some of them can give a clue to the present state of the individual in labor market, these variables are excluded. Survived variables are only related with characteristics, previous term and present qualifications of individuals.

Previous year state of individuals is known, however, present state is taken as unknown, which can be unemployment (0) or employment (1). Since the dependent variable that should be predicted is a dichotomous one, the problem turns out to be a binary choice model. Binary choice models can be solved with logistic regression in econometrics and classification methods in machine learning literature. For machine learning literature, classifiers and a shallow neural network with a sigmoid activation function that can be used to compute transition probabilities of each individual are used to compare them with logistic regression. The family of classifiers is Naïve Bayesian, Support Vector Machines, Decision Trees, Discriminant, Stochastic Gradient Classifier and Ensemble techniques.

By comparing accuracies, specifically the XGBoost classifier which is in Decision Trees, and Random Forest that in Ensemble have nearly the same, 67% accuracy rate to predict present state of individuals in the Turkish labor market. However, logistic regression with the same variables has slightly lower accuracy, 63%; this exceeds half of ML algorithms in general. Since the results obtained belong to the test data, the variables and models can be used to estimate new data produced in the future to predict an individual's transition probability.

It is important to estimate the likelihood of individuals going to employment in the future. Policy-makers, for example, may find individuals who will be employed in the future by using data from the past. They can predict which areas these individuals are educated. Thus, they can develop incentives, subsidiaries for those areas. Moreover, they can invest in education to cover the gap in the human capital that is missing in these areas. Identify common problems in individuals who cannot be employed. For these, they can develop policies aimed at eliminating the problems of re-education, re-training and common problems aimed at increasing their human capital. Based on these estimations, economists can realize their adaptations to develop social policies. Anyone who wants to use human resources effectively may also benefit from these estimates.

References

Acemoglu, D. (2002). Technical change, inequality, and the labor market. *Journal of economic literature*, *40*(1):7 72.

Acemoglu, D. and Shimer, R. (1999). Holdups and efficiency with search frictions. *International Economic Review*, *40*(4):827–849.

Ashenfelter, O., Harmon, C. and Oosterbeek, H. (1999). A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour economics*, *6*(4):453–470.

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.

Blundell, R., Dearden, L., Meghir, C. and Sianesi, B. (1999). Human capital investment: the returns from education and training to the individual, the firm and the economy. *Fiscal studies*, *20*(1):1–23.

Bougheas, S. and Georgellis, Y. (2004). Early career mobility and earnings profiles of german apprentices: Theory and empirical evidence. *Labour*, *18*(2):233–263.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. The x-random case. *International statistical review/revue internationale de Statistique*, 291–319.

Chollet, F. et al. (2015). Keras. https://keras.io.

Cingano, F. (2003). Returns to specific skills in industrial districts. *Labour Economics*, *10*(2):149–164.

Dearden, L. (1998). Ability, families, education and earnings in Britain. Technical report, *IFS Working Papers*.

Donaldson, D. and Eaton, B. C. (1976). Firm-specific human capital: a shared investment or optimal entrapment? *Canadian Journal of Economics*, 462–472.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education, India.

Higashi, Y. (2002). Firm specific human capital and unemployment in a growing economy. *Japan and the world economy*, *14*(1):35–44.

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

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*, *112*. Springer.

Krueger, A. and Ashenfelter, O. (1992). Estimates of the economic return to schooling from a new sample of twins. Technical report, *National Bureau of Economic Research*.

Mincer, J. (1974). Schooling, experience, and earnings. *Human Behavior & Social Institutions*, 2.

Mincer, J. and Ofek, H. (1982). Interrupted work careers: Depreciation and restoration of human capital. *Journal of Human Resources*, 3–24.

Parsons, D. O. (1972). Specific human capital: An application to quit rates and layoff rates. *Journal of political economy*, *80*(6):1120–1143.

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.

Psacharopoulos, G. (1994). Returns to investment in education: A global update. *World Development*, *22*(9):1325–1343.

Rousseeuw, P. J. and Bassett Jr, G. W. (1990). The remedian: A robust averaging method for large data sets. Journal of the American Statistical Association, 85(409):97–104.

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

Tasci, H. M. and Ozdemir, A. R. (2006). Trends in long-term unemployment and determinants of incidence of long-term unemployment in Turkey, *Journal of Economic and Social Research*, 7(2): 1-33.

Tasci, H. M. and Tansel, A. (2005). Unemployment and transitions in the turkish labor market: evidence from individual level data, *IZA Discussion Paper* no. 1663.

Tunali, İ. and Ercan, H. (2003). Background study on labour market and employment in Turkey. European Training Foundation, Torino (June 2003).

Wasmer, E. (2006). General versus specific skills in labor markets with search frictions and firing costs. *American Economic Review*, *96*(3):811–831.

Appendices

Table 5. Descriptive Statistics of Quantitative Features

Variable/Feature	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

Table 6. 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

Table 6. 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

Table 7. Classifier Results

Classifier	Classes	Precision	Recall	F 1 - score	Support	Confusion Matrix		Accuracy	AUC
Gaussian	0	0.54	0.48	0.51	93740	44687	49053		
Train	1	0.54	0.6	0.57	94468	37512	56956		
	Avg/Total	0.54	0.54	0.54	188208			0.5401	0.5409
Gaussian	0	0.54	0.47	0.5	7811	3668	4143		
Test	1	0.53	0.6	0.57	7872	3139	4733		
	Avg/Total	0.54	0.54	0.53	15683			0.5357	0.5365
Bernoulli	0	0.61	0.15	0.24	93740	14356	79384		
Train	1	0.52	0.9	0.66	94468	9282	85186		
	Avg/Total	0.56	0.53	0.45	188208			0.5289	0.5340
Bernoulli	0	0.6	0.15	0.25	7811	1208	6603		
Test	1	0.52	0.9	0.66	7872	818	7054		
	Avg/Total	0.56	0.53	0.45	15683			0.5268	0.5320
Logistic	0	0	0	0	93740	0	93740		
Train	1	0.5	1	0.67	94468	0	94468		
	Avg/Total	0.25	0.5	0.34	188208			0.5019	0.5395
Logistic	0	0	0	0	7811	0	7811		
Test	1	0.5	1	0.67	7872	0	7872		
	Avg/Total	0.25	0.5	0.34	15683			0.5019	0.5483
KNeighbors	0	0	0.7	0.82	93740	93740	0		
Train	1	1	0.57	0.73	94468	40605	53863		
	Avg/Total	0.85	0.78	0.77	188208			0.7843	0.9082
KNeighbors	0	0.54	0.78	0.64	7811	6061	1750		
Test	1	0.6	0.34	0.43	7872	5201	2671		
	Avg/Total	0.57	0.56	0.53	15683			OVERFIT	OVERFIT

Note: Avg. is the average of classifier.

Table 7. Classifier Results (cont'd)

Classifier	Classes	Precision	Recall	F1-score	Support	Confusion Matrix		Accuracy	AUC
Decision Tree	0	1	1	1	93740	93740	0		
Train	1	1	1	1	94468	26	94442		
	Avg/ Total	1	1	1	188208			0.9999	1.000
Decision Tree	0	0.62	0.62	0.62	7811	4830	2981		
Test	1	0.62	0.62	0.62	7872	2968	4904		
	Avg/ Total	0.62	0.62	0.62	15683			0.6207	0.6207
Linear Discriminant	0	0.5	1	0.66	93740	93740	0		
Train	1	0	0	0	94468	94468	0		
	Avg/ Total	0.25	0.5	0.33	188208			0.4981	0.6429
Linear Discriminant	0	0.5	1	0.66	7811	7811	0		
Test	1	0	0	0	7872	7872	0		
	Avg/ Total	0.25	0.5	0.33	15683			0.4981	0.6474
Quadratic Discriminant	0	0.61	0.65	0.63	93740	60771	32969		
Train	1	0.63	0.59	0.61	94468	38795	55673		
	Avg/ Total	0.62	0.62	0.62	188208			0.6187	0.6771
Quadratic Discriminant	0	0.6	0.64	0.62	7811	5010	2801		
Test	1	0.62	0.58	0.6	7872	3301	4571		
	Avg/ Total	0.61	0.61	0.61	15683			0.6109	0.6665
SGD	0	0	0	0	93740	0	93740		
Train	1	0.5	1	0.67	94468	0	94468		
	Avg/ Total	0.25	0.5	0.34	188208			0.4981	0.5000
SGD	0	0	0	0	7811	0	7811		
Test	1	0.5	1	0.67	7872	0	7872		
	Avg/ Total	0.25	0.5	0.34	15683			0.4981	0.5000

Table 7. Classifier Results (cont'd)

Classifier	Classes	Precision	Recall	F1-score	Support	Confusion Matrix		Accuracy	AUC
SVM	0	1	0.75	0.85	93740	69922	23818		
Train	1	0.8	1	0.89	94468	11	94457		
	Avg/Total	0.9	0.87	0.87	188208			0.8734	0.9704
SVM	0	0.73	0	0	7811	11	7800		
Test	1	0.5	1	0.67	7872	4	78.680.000		
	Avg/Total	0.62	0.5	0.34	15683			0.5024	0.5046
Linear SVC	0	0.54	0.48	0.51	93740	44686	49054		
Train	1	0.54	0.6	0.57	94468	37524	56944		
	Avg/Total	0.54	0.54	0.54	188208			0.5399	0.5401
Linear SVC	0	0.54	0.47	0.5	7811	3669	4142		
Test	1	0.53	0.6	0.57	7872	3127	4745		
	Avg/Total	0.54	0.54	0.53	15683			0.5365	0.5407
Neural Networks									
Train								0.5018	0.5000
Neural Networks									
Test								0.5019	0.5000