IJCEAS

International Journal of Contemporary Economics and Administrative Sciences
ISSN: 1925 – 4423
Volume: XIII, Issue: 1, Year: 2023, pp. 084-100

Determining Consumer Default Risk with Data Mining Techniques: An Empirical Analysis on Turkey

Begüm ÇIĞŞAR [1]
Semra BOĞA [2]
Deniz ÜNAL [3]

Received: 22.04.2022, Accepted: 13.4.2023
DOI Number: 10.5281/zenodo.8332194

Abstract

The aim of this study, which deals with consumer default risk, is to reveal the financial, socioeconomic, and demographic determinants of default risk at household level. Credit risk was investigated with various variables by applying data mining methods to the data set obtained from the Turkish Statistical Institute, Household Income and Living Conditions Survey covering the years 2016, 2017, 2018. Analyses were carried out using the WEKA data mining program. The findings of the study revealed that variables such as gender, age, marital status, education level, health status, employment status, region of residence and income status are important determinants of default. The findings of the study are thought to be an important reference for lenders in terms of risk assessment. In addition, the findings are expected to shed light on policy makers in terms of regulations to be applied to financial markets.

Key words: Big Data, Data Mining, Default Risk

JEL Code: C81, D12, D14, G51

1. Introduction

Globalization trends that accelerated all over the world in the post 1980 period, together with technological developments, have been a factor that reshaped Turkey's financial system. In Turkey, which has adopted neoliberal economic policies with the 24 January 1980 Decisions, credit markets have also been an area where these policies have had significant reflections. In fact, while the share of the private sector and households in total loans was 3.6% in 1980, this ratio increased to

[1] PhD, Independent Researcher, Turkey, begumcigsar@gmail.com, http://orcid.org/0000-0002-8453-985X
[2] Assoc. Prof., PhD, Final International University, Turkey, semra.boga@final.edu.tr, http://orcid.org/0000-0003-2799-9080
[3] Prof., PhD, Çukurova University, Turkey, dunal@cu.edu.tr, http://orcid.org/0000-0002-4095-3039

6.1% in 1988 (Kjellström, 1990). In the following years, with the deepening experienced in the banking and financial sectors, the number of individual loan customers, including credit card customers, exceeded 34 million in 2021 (BDDK, 2022). Although these developments have increased the financial development in Turkey, the default risk of individuals has become an important factor of fragility, especially in economic downturns. In this con- text, it is important to reveal which factors determine the individual default risk, both for lenders and for the policies to be applied for financial markets. Studies dealing with debt default at the household level in the literature are mostly empirical and the main purpose of these studies is credit scoring, which develops rates according to whether lenders make their payments on time (Devaney and Lytton, 1995). When the studies that deal with consumer default with empirical analysis are examined, it is seen that some of the studies consider mortgage defaults, while others focus on credit card and other non-securitized debt defaults.

The first studies in the field focused mainly on the mortgage loan default risk, and it was revealed that the ratio between the mortgage loan taken and the current value of the real estate and the income ratio used to pay the debt were important determinants of the default risk (Campbell and Dietrich, 1984; Lawrence and Arshadi, 1995; Wong et al., 2004). Studies conducted especially in the post-1980 period have shown that personal and demographic characteristics such as education, income, gender, age, as well as financial variables, are important in explaining consumer default (Alfaro and Gallardo, 2012). In terms of non-securitized loans, Avery et al. (2004) found that individuals who have been married for a long time have lower default rates than individuals who are newly married or divorced. This is explained by the fact that married couples are less sensitive to income shocks, in other words they tend to have two incomes. On the other hand, considering gender, it has been observed that males tend to have higher default probabilities. Sharma and Zeller (1997) argue that women are less likely to default because they choose less risky projects. These findings are confirmed by Stavins (2000), who tested the predictors of inability to pay credit card on time and default, and also found that married couples, older individuals, and better educated and higher-income individuals all had a lower probability of de- fault. Using the 1999-2001 data, Özdemir and Boran (2004), who tested the loan payment performance in Turkey with demographic and financial variables, proved that financial variables are more effective in terms of debt payment performance. On the other hand, Karan and Arslan (2008), who discussed the socioeconomic and demographic determinants of household credit risk with binomial logit estimation for Turkey, reached significant relationships between household assets, savings and business characteristics and credit risk. In their study conducted for China, Lin et al. (2017) revealed that in addition to demographic characteristics of borrowers such as age, marital status, education level, financial indicators such as monthly income, income-payment ratio are also important determinants of individual default. Dendramis et al. (2018)'s study on Greece revealed that especially the economic recession, high financial pressure and high political stability are factors that increase the default risk of consumers. In a recent study on Portugal by Silva

et al. (2020), the default risk of consumer loans was analyzed with a logistic regression model. The results of the analysis showed that the risk of default increases with the increase in the credit margin of the customers, the loan term and the age of the customer, on the other hand, the fact that the customers have more credit cards reduces this risk. Dendramis et al. (2021), on the other hand, in which they investigated the individual default risk in Greece, showed that loan-specific variables have a bigger impact than macroeconomic variables on determining the default risk. The findings of the study showed that the share of loan amounts in individuals' personal income significantly affects their ability to pay debt.

The aim of this study, which deals with consumer default risk, is to reveal the financial, socioeconomic, and demographic determinants of default risk at household level. Credit risk was investigated with various variables by applying data mining method to the data set obtained from the Turkish Statistical Institute Household Income and Living Conditions Survey for the years 2016, 2017 and 2018. Analyses were carried out using the WEKA data mining program. The study has two important contributions to the literature. First of all, it will contribute to the expansion of the limited literature investigating debt default at the household level in Turkey. The other is to show the use of data mining method, which is used in different disciplines in the literature, in debt default analysis at household level.

2. Material Method

The datasets used in this study consist of Turkish Statistical Institute (TUIK) surveys for the years 2016, 2017 and 2018. The data for each year is approximately 56 thousand units. From these surveys, the household head over the age of 15 was selected and the analyzes were applied. In the study, the TUIK database, which includes the socio-demographic, demographic and economic data of the individuals, was used to determine the individual credit risk. The reason why the years obtained from TUIK were 2016, 2017 and 2018 was the desire to examine the credit risk of households in Turkey comparatively for three years before the pandemic. In addition, since there are variables such as age, gender, marital status, employment status, etc. in the household data, it is thought that it will be effective in determining the credit risk. As a matter of fact, these variables included in the study are the variables that are frequently used in the literature to determine credit risk (Papouskova and Hajek, 2019; Arora and Kaur, 2020).

The socioeconomic and demographic variables selected from the questionnaires are listed in Table 1. In the obtained dataset, whether or not to pay the loan debts in the last 12 months (class variable) was determined as the dependent variable. All data mining analyses was conducted using WEKA Program.

Table 1. Attributes List

Attribute Name	Description
Age	Age
Gender	Gender (1 = Male, 2 = Female)
Marital Status	Marital Status (1= Married, 2=Single)
Education	Education Level (1 = Illiterate, 2 = Primary School, 3 = Secondary School, 4 = High School, 5 = Higher Degree)
Work	Working Status (1 = Working, 2 = Looking for a Job, 3 = Retired, 4 = Other (Non-Active: Seasonal Worker, Part Time Worker etc.))
Health	Health (1 = Good, 2 = Medium, 3 = Poor)
Region	Region (1 = Mediterranean, 2 = Aegean, 3 = Marmara, 4 = Black Sea, 5 = Central Anatolia, 6 = Eastern Anatolia, 7 = Southeastern Anatolia)
Housing	Housing Status (1 = Paying Rent, 2 = Not Paying Rent)
Revenue	Individual Revenue (1 = Low Income, 2 = Medium Income, 3 = Higher Income)
Home loan	Non-payment of house rent, interest-bearing debt repayment, or home loan payment within the last 12 months (1 = No, 2 = Yes)
Bills	Non-payment of electricity, water, and gas bills within the last 12 months (1 = No, 2 = Yes)
Class	Non-payment of credit card installments and other debt payments within the last 12 months (1 = No, 2 = Yes)

Source: Authors' calculations

WEKA Data Mining Software and Algorithms

Waikato Environment for Knowledge Analysis (WEKA) is an open source data mining software used in the analysis of big data. In this study, WEKA software was used to compare the classification algorithms and to interpret the analyzes. The reason for choosing WEKA is that it has a big data-oriented working system, it is an open source software, the variety of classification algorithms and there are more than one algorithm suitable for our model among these algorithms. Moreover, while WEKA provides a variety of comparison criteria and ease of calculation, the data preparation process is also easy. In addition to these features, the WEKA program was used while determining the credit risk in many similar studies, and the WEKA program was preferred in the study. (Examples of studies using WEKA: Ince and Aktan, 2009; Yu et al.,2010; Hamid and Ahmed, 2016; Aksu and Dogan, 2019; Torvekar and Game, 2019).

Since the aim of the research is to determine the default/non-default status of individuals and this is a classification problem, classification algorithms were used. There are several classification algorithms in different groups in the WEKA program. In this study, it is aimed to estimate the default status of individuals by using six different classification algorithms under Bayes, Function and Tree groups

(under Bayes Group: BayesNet, Naive Bayes; Under Function Group: Logistic, MLP; Under Tree Group: J48, Random Forest). The reason for using these algorithms is that they have high estimation performances in determining the default risk and are frequently used in the literature (Yu et al., 2010; Aksu and Dogan, 2019; Ince and Aktan, 2009; Torvekar and Game, 2019). What makes a classification algorithm powerful is not only its high performance, but also its speed, durability, and easy interpretation (Han et al., 2012). Therefore, when creating a model, it is necessary to measure such features and determine which algorithm is most suitable for the data set. For this reason, statistical criteria were used to compare the performances of algorithms in this study (Examples of studies using similar criteria: Ince and Aktan, 2009; Torvekar and Game, 2019; Papouskova and Hajek, 2019).

The concept of classification and the classification algorithms used in this study are briefly mentioned below and comparison criterion are explained next sub-section.

Classification: The most commonly used method in data mining is to classify the features in the data according to their similarity levels (Gorunescu, 2011). The purpose of classification is to assign unknown datasets to known classes (Han et al., 2012). In order to make classifications, Naive Bayes, Bayesian networks, J48, Random Forest, Multi-Layer Perceptron (MLP), Logistic regression algorithms that give the best results among the algorithms in Weka 3.9 data mining software are discussed in this section (Arora and Suman, 2012). These algorithms, the basic features of which are given below, were compared using six different criteria in order to determine which one is the most suitable for the data.

J48: This algorithm, when the data is separated according to the variables, determines the class with the highest information gain and enables the branching (Quinlan, 2014).

Random Forest (RF): In this algorithm (Breiman, 2001), classification is done by using more than one tree instead of using one tree during the classification process. Each tree generates a classifier. These produced classifiers vote among themselves and the algorithm determines the classifier with the most votes. This selected classifier is used to classify the data when new data is given (Kantardzic, 2003).

Bayes Group: Bayesian classifiers are statistical classifiers. They are algorithms that determine the probability of data groups belonging to a particular class. Naive Bayes and BayesNet algorithms were used in this group (Gorunescu, 2011; Alpaydın, 2004).

Logistic Regression (LR): Like all regression models, logistic regression is based on estimating the relationship between a dependent variable and independent

variables. It gets its name from the logarithmic function used to convert it to a linear regression model. It works with categorical data (Sharma et al., 2015).

Multi-Layer Perceptron (MLP): MLP is one of the artificial neural network algorithms (Fausett, 1984). Artificial neural networks make generalizations by collecting information from the dataset, and then apply this information to the new dataset, to enable decision-making (Alexander and Mortton, 1990).

Comparison Criterion

Comparisons were made according to Root Mean Squared Error (RMSE), Receiver Operating Characteristic (ROC) area, Accuracy (AC), Recall (R), Precision (P), F measure criteria. After the mentioned comparison criteria are briefly introduced in this section, the values obtained by using the data for the years are given in Tables 3, 4 and 5.

RMSE: The first criterion used in comparisons is the root mean squared error criterion, known as the square root of the mean squared error (MSE). As known MSE is defined as the average of the squares of the errors. Then the formula for the RMSE is given as follows,

$$\text{RMSE}=\sqrt{\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{n}} \qquad (1)$$

As it is a measurement of errors, it is better to take small values (Han et al., 2012). For defining the ROC area, AC, R, P, F measure criteria, we must first give the confusion matrix.

Confusion Matrix: A confusion matrix is a table often used to describe the performance of a classification model when the actual values are known. The structure of the confusion matrix is given Table 2 (Witten and Frank, 2011; Han et al., 2012). This table compares the performance of known/actual versus predicted values.

Table 2. Confusion Matrix

		Prediction Class	
	Population= P+N	Positive (PP)	Negative (PN)
Actual Class	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

Source: Han et al., 2012

To understand the confusion matrix, one must first understand its content.

TP- True Positive: Being able to say "true" to a "true" result. *TN- True Negative:* Being able to say "false" to a "false" result. *FP- False Positive:* Saying "false" to a "true" result.

FN- False Negative: Saying "true" to a "false" result. With the expressions defined here, the relationship between actual and estimated values is summarized in a crossly (Arora and Suman, 2012). Also, by using these table values, the values of the following comparison criteria can be calculated.

Accuracy Rate (AC): Accuracy is a criterion that is widely used to measure the success of a model. It calculates the ratio of correct predictions to all data. In other words, accuracy rate for the correct predictions can be calculated by following formula (Han et al., 2012).

$$AC = (TN + TP) / (TP + FP + FN + TN) \quad (2)$$

Precision (P): Precision is the rate of correctly predicted positive observations over the total predicted positive observations (Witten and Frank, 2011; Han et al., 2012).

$$P=TP / (TP + FP) \quad (3)$$

Recall (R): The fraction of correctly predicted positive observations among all the observations in the class (Witten and Frank, 2011; Han et al., 2012).

$$P=TP / (TP + FN) \quad (4)$$

F-Measure: F-Measure is the harmonic mean of the Precision and Recall values (Witten and Frank, 2011; Han et al., 2012).

$$F\text{-Measure}= (2.P.R) / (P+R) \quad (5)$$

where P: Precision and R: Recall. The reason of using the harmonic mean is that we should not ignore the extreme cases.

ROC Area: The area under the ROC curve which gives the predictive performance of the algorithms. Indicates the goodness of the accuracy of the algorithm as its value approaches to 1 (Gorunescu, 2011; Han et al., 2012).

While applying each algorithm one by one to the model, our expectation was to determine the algorithm that gives the best accuracy rate, RMSE, precision, recall, ROC area, and F-measure values to ensure that the most accurate algorithm is used for each data set. The disadvantage of the selection process here is that it is time consuming to execute some algorithms.

3. Findings

In this section, the performances of the algorithms are listed according to the comparison criteria for each year. The results are as in Tables 3, 4 and 5.

As can be seen from the Tables 3, 4 and 5 for the data of 2016, 2017 and 2018, the LR algorithm has the smallest RMSE, the highest ROC area and Recall values for all years compared to other algorithms. In addition, it is seen that F measure and Accuracy criteria are provided in favor of the LR algorithm in 2017, F-Measure in 2018 and Accuracy in 2016. In other words, the LR algorithm is the best algorithm with all five criteria in 2017 and four criteria in the other two years.

Table 3. Performance of Algorithms for the 2016 Year's Data

Algorithms/Criterions 2016	Accuracy	RMSE	ROC Area	Precision	Recall	F-Measure
BayesNet	83.597	0.3413	0.840	0.835	0.836	0.835
Naive Bayes	83.695	0.342	0.839	0.836	0.837	0.836
MLP	83.704	0.336	0.829	0.833	0.837	0.835
Logistic	84.212	0.327	0.845	0.832	0.842	0.835
J48	84.109	0.335	0.805	0.836	0.841	0.838
Random Forest	82.229	0.359	0.808	0.813	0.822	0.817

Source: Authors' calculations

Table 4. Performance of Algorithms for the 2017 Year's Data

Algorithms/Criterions 2017	Accuracy	RMSE	ROC Area	Precision	Recall	F-Measure
BayesNet	85.324	0.330	0.848	0.851	0.853	0.852
Naive Bayes	85.298	0.331	0.846	0.850	0.853	0.852
MLP	85.241	0.325	0.836	0.845	0.852	0.848
Logistic	85.893	0.315	0.854	0.852	0.859	0.855
J48	85.867	0.324	0.799	0.853	0.859	0.855
Random Forest	84.843	0.329	0.835	0.839	0.848	0.842

Source: Authors' calculations

Table 5. Performance of Algorithms for the 2018 Year's Data

Algorithms/Criterions 2018	Accuracy	RMSE	ROC Area	Precision	Recall	F-Measure
BayesNet	85.691	0.322	0.844	0.908	0.923	0.916
Naive Bayes	85.641	0.323	0.842	0.909	0.921	0.915
MLP	86.161	0.314	0.834	0.902	0.937	0.919
Logistic	86.323	0.307	0.850	0.898	0.945	0.921
J48	86.568	0.317	0.773	0.906	0.938	0.921
Random Forest	84.698	0.339	0.805	0.897	0.924	0.910

Source: Authors' calculations

Similarly, in the study conducted with 2015 TUIK data (Çığşar and Ünal, 2019).

), it was seen that the LR algorithm had the smallest RMSE, the highest ROC area and Recall values, and the F-Measure and Accuracy values also gave results in favor of LR. Accordingly, it can be stated that the LR is a suitable algorithm for the analysis of such data (Çığşar and Ünal, 2019; Husejinovic et al., 2018; Vangaveeti et al., 2020). For these reasons, it was decided to investigate the default risks of individuals according to the results of the LR algorithm.

Since TUIK is an official statistical institution, the reliability and validity studies carried out by TUIK were accepted and analyzes were made. Before building the model, the data preparation process was carried out and missing data was checked. In data mining, the data is divided into train and test sets in order to prevent over fitting of the data. While the model is trained in the train sets, it is tested in the test set. In this study, this process was done by selecting 10-fold cross validation on WEKA platform. With this operation, the dataset is divided into 10 parts, one fold is reserved for testing, and the train operation is performed on the remaining 9 folds.

Determining Credit Risks by Logistic Algorithm

In the previous section, the best algorithm was determined as Logistic Regression. Before applying this algorithm to the data, the WEKA attribute selection panel was used to decide on the variables to be included in this analysis. Chi-Square analysis was chosen in this panel, and as a result of the analysis, it was decided that all variables should be in the model. The assumptions necessary to perform all analyzes were checked and provided on the WEKA platform, both with graphics and under the data selection platform. In the created logistic model, the interpretations were made according to the non-default status, which is the sub-category of the dependent variable class. Then, ODDs ratios related to which variables are important for credit risk and default situation are calculated and given in Table 6.

ODDs Ratio is a measure of effect size, particularly important in Bayesian statistics and logistic regression. When calculating ODDs ratio, two groups are considered. Firstly, ODDs values are calculated by taking the ratio of the probability of having or not having a special situation. For example, let $P_1$and P_2 represent the probability of an event occurring in the first and second groups, respectively. Then ODDs values of the groups can be calculated P_1 / (1- P_1) and P_2 / (1- P_2), respectively. The rate of the ODDs values of these two groups gives the ODDs ratio value as follows: P_1 / (1- P_1) / P_2 / (1- P_2).

If the ODDs ratio value is greater than 1, the event/situation considered is more likely for the first group, and if it is less than 1, the opposite situation is true (Sharma et al., 2015).

Table 6. Odds Ratios for Years

Variables	2016		2017		2018	
	Subgroup	**ODDs Ratio**	**Subgroup**	**ODDs Ratio**	**Subgroup**	**ODDs Ratio**
Gender	Female	0.7906	Female	0.9323	Female	0.8577
Marital Status	Single	1.2077	Single	1.2196	Single	1.2022
Education Level	Secondary School	0.8771	Secondary School	0.8999	Secondary School	0.8671
	Higher Degree	1.238	Higher Degree	1.3776	Higher Degree	1.3889
Health	Good	1.2117	Good	1.17	Good	1.1917
	Medium	0.8501	Poor	0.7820	Poor	0.8289
Working Status	Looking for a Job	0.6980	Looking for a Job	0.8057	Looking for a Job	0.6240
	Other (Non- Active)	1.3803	Other (Non- Active)	1.2316	Other (Non- Active)	1.2301
Region	Mediterranean	0.7220	Mediterranean	0.7774	Mediterranean	0.7233
	Black Sea	1.1451	Central Anatolia	1.1875	Marmara	1.1541
Housing	Not Paying Rent	0.9083	Not Paying Rent	1.090	Not Paying Rent	1.1527
Home Loan	Non-Payment Home Loan	0.3486	Non- Payment Home Loan	0.3215	Non- Payment Home Loan	0.3232
Bills	Non-Payment Bills	0.0813	Non-Payment Bills	0.0723	Non-Payment Bills	0.0795
Revenue	Medium Income	0.9179	Low Income	1.2525	Low Income	1.1476
	Higher Income	1.4996	Higher Income	0.9049	Higher Income	0.9147

Age	–	1.0204	–	1.0204	–	1.0207

Source: Authors' calculations

It was observed that the risk of default for men was lower than women at all three years. For example, while women were 0.7906 times less likely to not defaulting in 2016 than men, this rate was 0.9323 times less in 2017 and 0.8577 times less in 2018.

Considering the default status of household heads according to the marital status variable, it was concluded that unmarried individuals were able to pay their debts more regularly than married ones in all three years. Obtained odds ratio values show that the odds of not defaulting in singles in 2016 are 1.2077 times higher than married ones, 1.2196 times in 2017 and 1.2022 in 2018.

According to their educational status, the probability of not defaulting is higher for individuals having higher education degree than in other education groups. These rates are 1.238 times, 1.3776 times and 1.3889 times higher in 2016, 2017 and 2018, respectively.

When the health variable is examined, it is seen that the probability of not defaulting is higher for household heads who state that their health is good (1.2117 times higher for 2016, 1.1700 for 2017 and 1.1917 times higher for 2018).

When the employment status is examined, it is seen that the probability of default of the job seekers is higher than the other employment status groups. For example, according to 2018 data, job seekers are 0.624 times less likely to not be in default than other groups. On the other hand, the non-default status of the individuals in the non-working group is higher than the other working groups for the years 2016, 2017, 2018.

Considering the probability of default on the basis of regions, it is seen that the region with the highest probability of default is the Mediterranean region.

According to the housing status, the probability of not defaulting of those who do not pay rent compared to those paying rent, is higher in years other than 2016.

It is seen that those who default on their home loan payments and cannot pay their bills are more likely to default on other loan debts as well. For example, the probability of not defaulting on other loans of people who defaulted on their home loans in 2018 is 0.3449 times less than other groups. Again, the probability of not defaulting on credit debts of individuals who say they could not pay their bills in the same year is 0.0838 times less than those who can pay.

While the probability of not defaulting was only in favor of high-income people in 2016, it is seen that the probability of not defaulting in 2017 and 2018 is 1.2525 and 1.1476 times higher, respectively, compared to the others. In 2015, this rate was 1.0316 times, again in favor of low-income people (Çığşar and Ünal, 2019).

With the Attribute selection process performed in 2015, it was determined that the age variable did not have a significant effect on the model (Çığşar and Ünal, 2019). However, the opposite situation was encountered in this study. The age variable was also included in the model as it contributed significantly to the model in 2016, 2017 and 2018. Looking at these years, it is seen that the age odds ratios are similar. According to odds ratio, it is observed that the probability of not defaulting increases as the age increases.

4. Discussion and Conclusion

In this study, using the data set obtained from the Household Income and Living Conditions Survey conducted by TUIK, the determinants of debt default at the household level in Turkey were analyzed with the WEKA data mining program. In the analysis for 2016, 2017 and 2018, the default risk for men was found to be lower than women in all three years. Again, every three years, unmarried individuals were found to have higher debt repayment ability than married individuals. It was found that the probability of default of individuals with high education level, good health status and working status was lower than the individuals who were in the opposite situation. The analysis presented that the region with the highest probability of individual default in Turkey is the Mediterranean Region. Except for 2016, the risk of default was found to be higher for homeowners than rent payers. On the other hand, it is seen that those who default on their housing loan payments and cannot pay their bills are more likely to default on other loan debts as well. Considering the income level, while the probability of not defaulting in 2016 was only in favor of high income earners, it is seen that the probability of not defaulting in this group in 2017 and 2018 is 1.2525 and 1.1476 times higher than the others, respectively. In the study, it was revealed that age is an important determinant of default; It has been observed that the probability of not defaulting increases with increasing age.

The findings of the study highlights that the risk of default in Turkey differs according to demographic and socioeconomic status. It was observed that the default risk of men was lower than women in all three years. This is a valuable finding since it differs from the general acceptance in literature that the men are riskier in terms of default probability. Our finding was also supported by the research of Durango-Gutiérrez (2021) conducted for Bolivia and Colombia which revealed a lower default risk for men suggesting that the gender effect in default may vary according to the socio-economic conditions of the countries. The situation in Turkey can be explained by the fact that women are employed for lower wages. According to DISK-AR's report published in 2020, men's incomes are 31.4% higher

than women's in Turkey. In fact, the difference in income for self-employed people reaches even up to 80%. In this context, an important tool in preventing women's default will be to prevent gender inequalities in wages.

Another important finding of the study is that single individuals have higher debt repayment ability than married people. Although the risk of default for married couples is seen to be less in the literature, Yap et al. (2011) also revealed that singles have a higher debt-paying capacity. In Turkey, this situation can be explained by the lifestyle of singles in Turkey. While single people in developed countries can live in a separate house and with a single salary due to their socio-economic and cultural environment, many single people in Turkey continue to live with their families. Therefore, this situation becomes a factor that reduces the probability of default. In this context, it will be an important element in determining the default risk that the creditors should review the status of individuals living alone in addition to their marital status while making risk assessments.

In the study, in parallel with the literature, it was observed that the risk of default decreases as the level of education, health, and working status increase. These findings offer important clues for designing a healthy-functioning financial environment, especially for policy makers. Different from the existing literature, regional differences were also examined and the Mediterranean region was found to be the region with the highest default probability in Turkey. This may be due to the fact that the main economic activities of the Mediterranean region are agriculture and tourism where temporary work is intense. Diversifying and developing the economic activities of the region is important in terms of preventing the risk of default.

The risk of default is an issue that concerns both lenders, borrowers, and policy makers responsible for the management of the economy. In this study, demographic, socioeconomic and regional determinants of the risk of default for Turkey before the Covid-19 pandemic were examined. Consistent with the literature, significant relationships were found between the variables and the risk of default. The findings of the study are expected to shed light on the individual credit evaluations of financial institution, also prediction of the financial risks of banks in terms of household debt default, and thus the development of mechanisms that will provide the establishment of financial stability in Turkey. On the other hand, the fact that variables such as gender, income differences, education, health, and employment status are also associated with the risk of default shows that this risk cannot be avoided only with the efforts of financial institutions and individuals, and that it is important for governments to develop policies to support the healthy functioning of credit markets since it is possible for a large-scale default situation in the economy to create undesirable results such as financial pressure and even crisis. Adding macro-level variables such as employment and income level in addition to financial and demographic variables in future studies will also help develop macro prudential measures that can eliminate financial risks. Considering the transformation created by the Covid-19 crisis in almost every field, the

determinants of individual default should also be analyzed by the researchers for the post-Covid period as well. This study makes an important contribution to the literature as it is one of the few studies that predicts the default risk in Turkey at the household level using data mining.

REFERENCES

Aksu, G., & Dogan, N. (2019). An Analysis Program Used in Data Mining: WEKA.

Alexander, I., & Morton, H. (1990). An Introduction to Neural Computing. London: Chapman and Hall.

Alfaro, R., & Gallardo, N. (2012). The Determinants of Household Debt Default. Revista de Analisis´ Economico, 27 (1), 55-70.´

Alpaydın, E. (2004). Introduction to Machine Learning (pp. 433). The MIT Press Cambridge, Massachusetts London, England.

Arora, N., & Kaur, P. D. (2020). A Bolasso Based Consistent Feature Selection Enabled Random Forest Classification Algorithm: An Application to Credit Risk Assessment. Applied Soft Computing, 86, 105936.

Arora, R., & Suman, S. (2012). Comparative Analysis of Classification Algorithms on Different Datasets Using WEKA. International Journal of Computer Applications, 54(13), 21–25.

Avery, R., Calem, P., & Canner, G. (2004). Consumer Credit Scoring: Do Situational Circumstances Matter? Journal of Banking and Finance, 28(4), 835-856.

BDDK (Banking Regulation and Supervision Agency) (2022). Main Banking Indicators. Ankara.

Breiman, L. (2001). Random Forests. University of California Berkeley.

Campbell, S., & Dietrich, J. (1984). The Determinants of Default on Insured Conventional Residential Mortgage Loans. The Journal of Finance, 38(5), 1569-1581.

Çığşar, B., & Ünal, D. (2019). Comparison of Data Mining Classification Algorithms Determining the Default Risk. Scientific Programming.

Costa e Silva, E., Lopes, I. C., Correia, A., & Faria, S. (2020). A Logistic Regression Model for Consumer Default Risk. Journal of Applied Statistics, 47(13-15), 2879-2894.

Dendramis, Y., Tzavalis, E., & Adraktas, G. (2018). Credit Risk Modelling Under Recessionary and Financially Distressed Conditions. Journal of Banking & Finance, 91, 160-175.

Dendramis, Y., Tzavalis, E., Varthalitis, P., & Athanasiou, E. (2021). What Drives the Default Risk of Restructured Loans. In Money, Trade and Finance (pp. 143-167). Palgrave Macmillan, Cham.

Devaney, S., & Lytton, R. (1995). Household Insolvency: A Review of Household Debt Repayment, Delinquency, and Bankruptcy. Financial Services Review, 4(2), 137-156.

DISK-AR (2020). Gender Pay Gap in Turkey: Women Earn a Third Less Than Men. https://disk.org.tr/2020/09/gender-pay-gap-in-turkey-women-earn-a-third-less-than-men/

Durango-Gutiérrez, M. P., Lara-Rubio, J., & Navarro-Galera, A. (2021). Analysis of Default Risk in Microfinance Institutions under the Basel III Framework. International Journal of Finance & Economics, 8, 1-18.

Dzelihodzic, A., & Donko, D. (2013). Data Mining Techniques for Credit Risk Assessment Task Recent Advances in Computer Science and Applications. Conference, Valencia, Spain, 2013, August.

Fausett, L. (1984). Fundamentals of Neural Networks, Architectures, Algorithms and Applications. Prentice-Hall International Edition.

Gan, Q., Luo, B., & Lin, Z. (2008). Identifying Potential Default Loan Applicants-A Case Study of Consumer Credit Decision for Chinese Commercial Bank. SAS Global Forum, (pp. 159).

Gorunescu, F. (2011). Data Mining Concepts, Models, Methods and Algorithms. Springer, (pp. 361).

Hamid, A. J., & Ahmed, T. M. (2016). Developing Prediction Model of Loan Risk in Banks Using Data Mining. Machine Learning and Applications: An International Journal (MLAIJ), 3(1), 1-9.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann, (pp. 740).

Huang, C., Chen, M., & Wang, C. (2007). Credit Scoring a Data Mining Approach Based on Support Vector Machines. Elsevier, 33(4), 847-856.

Husejinovic, A., Keco, D., & Masetic, Z. (2018). Application of Machine Learning Algorithms in Credit Card Default Payment Prediction. International Journal of Scientific Research, 7(10), 425-426.

Ince, H., & Aktan, B. (2009). A Comparison of Data Mining Techniques for Credit Scoring in Banking: A Managerial Perspective. Journal of Business Economics and Management, 10(3), 233-240.

Kantardzic, M. (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, Danvers, (pp. 550).

Karan, M.B., & Arslan, O. (2008). Consumer Credit Risk Factors of Turkish Households. Banks and Bank Systems, 3(1).

Kjellstrom, S. (1990). The Financial Markets in Turkey, World Bank, EMENA Internal Discussion Paper.

Lawrence, C., & Arshadi, N. (1995). A Multinomial Logit Analysis of Problem Loan Resolution Choices in Banking. Journal of Money, Credit and Banking 27(1), 202-216.

Lin, X., Li, X., & Zheng, Z. (2017). Evaluating Borrower's Default Risk in Peer-To-Peer Lending: Evidence from a Lending Platform in China. Applied Economics, 49(35), 3538-3545.

Ozdemir, O., & Boran, L. (2004). An Empirical Investigation on Consumer Credit Default Risk. Discussion Paper. 2004/20. Turkish Economic Association.

Papouskova, M., & Hajek, P. (2019). Two-Stage Consumer Credit Risk Modelling Using Heterogeneous Ensemble Learning. Decision Support Systems, 118, 33-45.

Quinlan, J.R. (2014). C4. 5: Programs for Machine Learning. Elsevier.

Sharma, M., & Zeller, M. (1997). Repayment Performance in Group-Based Credit Programs in Bangladesh: An Empirical Analysis. World Development 25(10), 1731-1742.

Sharma, N., Kuar, A., Gaamdotro, S., & Sharma, B. (2015). Evaluation and Comparison of Data Mining Techniques Over Bank Direct Marketing. International Journal of Innovative Research in Science, Engineering and Technology, 4(8), 7141-7147.

Stavins, J. (2000). Credit Card Borrowing, Delinquency, and Personal Bankruptcy. New England Economic Review, (pp. 15-30).

Torvekar, N., & Game, P. S. (2019). Predictive Analysis of Credit Score for Credit Card Defaulters. Int. J. Recent Technol. Eng, 7(1), 4.

TUIK (Turkish Statistical Institute) (2016, 2017, 2018). Income and Living Conditions Survey.

Van Gestel, T., & Baesens, B. (2008). Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital. Oxford University Press.

Vangaveeti, S.A., Venna, N.L., Kidambi, P.N.S.R.Y., Marneni, H., & KumarMaganti, N.S. (2020). Logistic Regression Based Loan Approval Prediction. JAC: A Journal of Composition Theory, 13(5), 319-325.

Venkatesh, A. & Jacob, S.G. (2016). A Comparative Study on Performance of Classifiers. International Journal of Computer Applications, 145(7).

Witten, I.H., & Frank, E. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Wong, J., Fung, L., Fong, T., & Sze, A. (2004). Residential Mortgage Default Risk and the Loan-to Value Ratio. Hong Kong Monetary Authority Quarterly Bulletin, (pp. 35-45).

Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using Data Mining to Improve Assessment of Credit Worthiness via Credit Scoring Models. Expert Systems with Applications, 38(10), 13274-13283.

Yu, H., Huang, X., Hu, X., & Cai, H. (2010, October). A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation. In 2010 International Conference on Management of e-Commerce and e-Government (pp. 35-38). IEEE.