Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 26, Sayı 3, 420-426, 2022

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 26, Issue 3, 420-426, 2022

DOI: 10.19113/sdufenbed.1073504

Mining Housing Features to Classify Housing Unit Price

Betül KAN KILINÇ*[1], Simay MİRGEN[2],

[1] Eskişehir Technical University, Science Faculty, Department of Statistics, 26470, Eskişehir, Turkey
[2] Vakıfbank Yunusemre Branch, Yunusemre Ave. Odunpazarı Eskişehir, 26120, Turkey

(Alınış / Received: 15.02.2022, Kabul / Accepted: 28.10.2022, Online Yayınlanma / Published Online: 20.12.2022)

Keywords
Ordinal,
Logistic,
Classification,
Cross-validation

Abstract: In data mining, classification builds an interdisciplinary field upon from statistics, computer science, mathematics and many other disciplines. There are numerous statistical applications where parametric and non-parametric methods are frequently used to train data to estimate mapping function. In this study, two of the most widely used techniques are applied to a real dataset. The goal of the study is to compare the classification success of ordinal logistic regression and the classification trees and to predict a categorical response. For this purpose, the potential factors affecting the housing unit price for sale as being the dependent variable with three classes in Eskişehir were examined. The real data set was split into three as train, validation and test groups. The classification performance of the techniques was demonstrated with 5-fold cross validation technique. According to the results, a more successful classification was made with the classification trees algorithm.

Konut Özellikleri Madenciliğiyle Konut Birim Fiyatlarını Sınıflandırması

Anahtar Kelimeler
Sıralı,
Lojistik,
Sınıflandırma,
Çapraz Geçerlilik

Öz: Sınıflandırma, istatistik, bilgisayar bilimi, matematik ve diğer birçok disiplin arasında veri madenciliği ile ortak bir alan yaratır. Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi sınıflandırmak için sıklıkla kullanılan parametrik ve parametrik olmayan pek çok istatistiksel uygulamalar bulunmaktadır. Bu çalışmada yaygın olarak kullanılan iki sınıflandırma tekniği kullanılmıştır. Çalışmanın amacı, sıralı lojistik regresyon ve sınıflandırma ağaçları tekniklerinin sınıflandırma başarısını karşılaştırmaktır. Bu amaçla, Eskişehir' de üç sınıflı bağımlı değişken olarak ele alınan konut birim fiyatlarını konut birim fiyatlarını etkileyen potansiyel faktörler incelenmiştir. Gerçek veri seti, eğitim, doğrulama ve test olmak üzere üç gruba bölünmüştür. Bu tekniklerin sınıflandırma başarısı, 5 katlı çapraz geçerlilik ile gösterilmiştir. Elde edilen sonuçlara göre, daha başarılı bir sınıflandırma, sınıflandırma ağaçları algoritmasıyla elde edilmiştir.

1. Introduction

Classification is one of the main issues in data mining studied by scientists from numerous disciplines. Major classification techniques can be found in a literature search. Among the classification methods, logistic regression and classification trees can be used for classification purpose. The aim of both methods is to determine the relationship between predictors and a particular outcome with the qualitative characteristic and find the best fitting model.

Initial studies for the use of logistic regression model were developed by Berkson (1944) and this model was used by Finney (1971) as an alternative to probit analysis for biological experiments [1,2]. There are many statistical literatures on the dichotomous outcome variable [3,4]. Also, Aranda-Ordaz (1981) and Johnson (1985) studied the goodness of fit for a logistic model [5,6]. The most well-known models produced by several scientists for estimating an ordinal outcome variable are the proportional odds (PO) models [7,8]. Most of these models have their own assumptions and pre-defined underlying

* Corresponding author: bkan@eskisehir.edu.tr

relationships between dependent and predictors. Although this method is popular, it can appropriately not handle with the potential nonlinear relationship between variables [9]. Also, performance of these regression models is affected by outliers and multicollinearity.

Pakgohar et al. (2010) studied Classification and Regression Trees (CART) and Multinomial Logistic Regression to investigate drivers' characteristics in the resulting crash severity [10]. They found that the CART method provided more simpler, precise and easier results to interpret.

Classification and regression trees (CART) are flexible, fast and accurate and often preferred in statistical applications. They can be easily applied and interpreted as they are more robust to the presence of outliers [11,12]. CART is a nonparametric model with no predefined definition of underlying relationships between exploratory variables and the dependent variable. Due to the increasing computational resources after 90' s, the more flexible methods of data mining became available.

Friedman was the first who applied recursive partitioning method [13]. Several applications for classification purpose can be found in a literature search [14,15].

Nagalla et al. (2017) examined the driver's gap acceptance behaviour by using nonparametric data-mining techniques namely, support vector machines, random forests and decision trees [16]. They modeled the gap acceptance behaviour of the driver to predict whether the gap would be accepted or rejected. Similarly, decision trees are used to identify the main factors affecting the severity of road accidents [17].

It is well-known that the tree actually simplifies the classification process. The tree grows by the responses to the questions asked to the independent variables. Hence, they are seen as tree branches so that the variables affecting the dependent variable and the importance of these variables in the model can be examined visually without data complexity.

In this study, it is aimed to compare the modelling success of these two methods by using k-fold cross validation. This paper is organized as follows. Ordinal logistic regression and classification and regression trees methods are described first. Next section introduces how the data are collected from a website. After then, application section includes the results from ordinal logistic regression and classification tree algorithm that were performed to classify a response variable with three classes. Last section presents the conclusion.

2. Methodology

2.1. Ordinal logistic regression

In the logistic regression model, the ordered logistic regression model is used if the dependent variable has at least three categories and the categories are ordered from small to large in a natural order [18-24]. For example, the severity of the disease (from the least severe to the most severe) can be measured on scales such as customer satisfaction (not satisfied, slightly satisfied, very satisfied) [24,25].

Ordinal logistic regression (OLR) model is one of appropriate technique to determine the relationship between dependent variables with unequal ordered categories and independent variables. In order to obtain the ordinal logistic regression model, a link function is used. Logit, probit and cloglog are the link functions often used [26]. For an ordinal logistic regression, the regression coefficients do not depend on the categories of the ordered response, hence the regression coefficients estimated using the link function are the same at each cut-off point (threshold value) [27,28].

An important approach in the formation of the model is that it is assumed to be rearranged from an unobserved continuous latent dependent variable under the influence of the ordered categorical dependent variable. Second, there is the assumption of parallel lines, which is considered the most prominent feature of the ordinal logistic regression model. The parallel regression (lines) assumption is that the regression coefficients (except the intercept) are assumed to be equal in all categories of the ordered dependent variable. That means that the relationship between the predictors and the dependent variable does not vary according to the categories of the dependent variable. In the ordinal logistic regression, when this assumption is hold with j-1 logistic comparison for the dependent variable having J categories, there are α_{j-1} cut off points and j-1β parameters [29,30].

In an OLR model, the number of categories of the dependent variable are more than two levels. The model estimates the probability being at or below a specific category of the dependent variable given a group of independent variables. The ordinal logistic regression model can be expressed in the logit form as:

$$\begin{aligned} ln(Y_j) &= logit(\pi(\boldsymbol{x})) \\ &= ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \\ &= \alpha_j + (-\beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p) \end{aligned} \quad (1)$$

where $\pi_j(\boldsymbol{x}) = \pi(Y \leq j|x_1, x_2, \dots, x_p)$ is the probability of being at or below category j, given a set of predictors, j=1,2,...,J-1, α_j are the cut points and $\beta_1, \beta_2, \dots, \beta_p$ are the logit coefficients [29-32].

The equal logit slope (propor proportional odds assumption) can be expressed by Brant test [33].

2.2. Classification and regression trees (CART)

The classification and regression tree algorithm (CART) is a non-parametric statistical technique that is used to analyse and estimate the values of both categorical and continuous dependent variables. It is very common in applications since it does not require any assumptions for the data set. Decision tree models can be used to determine the relationship between variables and to make predictions from the data. In these models, the leaves arise with the responses from the questions of the independent variable. This forms the tree branches visually and show the independent variables that affect the dependent variable.

In classification procedure, the objective is to develop a tree-based model that classifies observations into one of k pre-determined categories. The final tree model can be obtained as a series of conditional probabilities (posterior probabilities) of category membership given a set of covariate values. For each terminal node, a probability distribution for category membership is obtained as in the following where the probabilities are of the form:

$$\hat{P}\left(C_j \middle| \boldsymbol{x} \in T_A\right) \text{ where } \sum_{j=1}^{k} \hat{P}\left(C_j \middle| \boldsymbol{x} \in T_A\right) = 1 \qquad (2)$$

Here, T_A is a terminal node defined by the set of predictors $\boldsymbol{x}$.

The terminal nodes are found by a number of binary splits chosen to minimize the overall 'loss' of the resulting tree. One method is to construct classification trees so that the overall misclassification rate is minimized. In classification problems, the prior knowledge is represented by prior probabilities of an observation being from category j, denoted by π_j. Note that, $\sum_{j=1}^{k} \pi_j = 1$. The modeling process is the cost or loss incurred by classifying an object from category j as being from category i and vice versa. The aim of the classification process regarding to the growth of the tree is to avoid making the most costly misclassifications on our training data set.

There are different criteria to split the tree. In this paper, Gini index is used and can be defined for node c as:

$$\text{Gini(c)} = 1 - \sum_{\text{j}} \text{p}^2(\text{j}|\text{c}) \qquad (3)$$

where $p(j|c) = p(j,c)/p(c)$, $p(j,c) = \pi_j N_j(c)/N_j$ and $p(c) = \sum_j p(j,c)$, j; number of target classes, π_j; prior probability for class j, $p(j|c)$; conditional probability of a case being in class j provided that is in node m, $N_j(c)$; number of cases of class j of node m, N_j; number of cases of class j in the root node.
Gini index measures the degree of purity of the node. The tree procedure achieves the maximum purity in the node, hence the best split is the one that minimizes Gini index. This achieves the maximal tree that overfits the data. The complexity of the tree is decreased by pruning the tree [12].

2.3 Confusion matrix

Confusion matrix is a nxn table that represents the true positives, false positives, true negatives, false negatives and misclasifying counts. Table 1 reports the counts in each cell as defined:

Table 1. 3-class Confusion matrix

Actual	Prediction			False Negative	Recall
	Class 1	Class 2	Class3		
Class 1	c1	c2	c3	c2+c3	c1/(c1+c2+c3)
Class 2	c4	c5	c6	c4+c6	c5/(c4+c5+c6)
Class 3	c7	c8	c9	c7+c8	c9/(c7+c8+c9)
False positive	c4+c7	c5+c8	c6+c9		
Precision	c1/(c1+c4+c7)	c5/(c2+c5+c8)	c9/(c3+c6+c9)		

Accuracy: $$(\text{c1} + \text{c5} + \text{c9}) / \sum_{i=1}^{9} ci \qquad (4)$$

The misclassfying counts are all of the counts except c1, c5 and c9. Predictive values (positive and negative) reflect the performance of the prediction. Positive Prediction Value (PPV or precision) represents the proportion of positive samples that are correctly classified to the total number of positive predicted samples. On other hand, Negative Predictive Value (NPV), inverse precision, measures the proportion of negative samples that are correctly classified to the total number of negative predicted samples. Prevalence is all positives over total samples.

3. Application

In this study, it was aimed to classify the housing unit price by both ordinal logistic regression and classification trees. Then, the comparison of the

classification performance was tested by using 5-fold cross validation. The potential variables that may affect housing unit prices in Eskisehir province, Turkey, were considered for comparisons. The data published from October 2018 to May 2019 on a widely website was used in the analysis. The number of 280 houses for sale in Eskisehir were used. The characteristics of the houses used in the anaysis are features are price of houses, age, the type of room, land, bedroom, the number of bathrooms, garage, social environments around (shopping center, hospital, school, sport center, etc.), with/without balcony, the floor of the apartment, with/without elevator and the area where the apartment is located.

The ratio of the house unit price, ($TL/m^{2)}$ was calculated by dividing house prices by their square meters. It was classified into three categories in an ordered structure. To determine the category borders, the housing unit prices which were published online by Central Bank of the Republic of Turkey (TL/m^2) for 2018 on March 11, 2019 were used (EVDS, Electronic Data Delivery System). The minimum price 2.118,52 (TL/m^2) and the maximum 2.314,83 (TL/m^2) were used to create the boundary of the categories such as low, expensive, and moderate elsewhere.

As the ordinal logistic regression has a primary effect on variable selection due to the assumptions it has to provide, the variables that did not provide the parallel regression assumption were ignored. According to the results of parallel lines test, chi-square test statistic was obtained as 30.678 with 10 degrees of freedom (p=0.000). This result stated that the null hypothesis was rejected at 0.05 significance level. Then, it was decided to use with the variables that were provided by the test of parallel lines. Hence, the status of garden, the status of elevator and the variable district were used for further comparisons. In all the analysis, SPSS 24.0 and R-Studio 1.0.153 (Mac OS X 10.12.6) were used.

The number of observations and the percentages for the unit price, type, elevator and district are shown in Figure 1. The 'Low' category consists of 137 observations as the 48.93% of all data, the 'moderate' category consists of 81 observations at 28.93%, and the 'High' category consists of 62 observations at 22.14%. Note that the independent variables that are considered to be effective on the unit price are the type (with- coded as 1/without garden-coded as 0), elevator, and district. These three variables were chosen because they provide parallelism assumption. The variable 'garage' was coded as 1 (with property garage), 0 (without the property garage) elsewhere. As shown in Figure 1, there are 245 observations without garden at 87.50% and 35 observations with garden at 12.50%. The variable 'elevator' consists of two categories coded as 1 (with property elevator) and elsewhere 0 (without the property elevator). As shown in Figure 1, it consists of 127 observations elevator at 45.36% without and 153 observations with elevator at 54.64%. The variable 'district' consists of two locational categories: Odunpazarı or Tepebaşı. It can be also seen from the figure that, 122 apartments located in Odunpazarı at 43.57% and 158 observations in Tepebaşı at 56.43%.

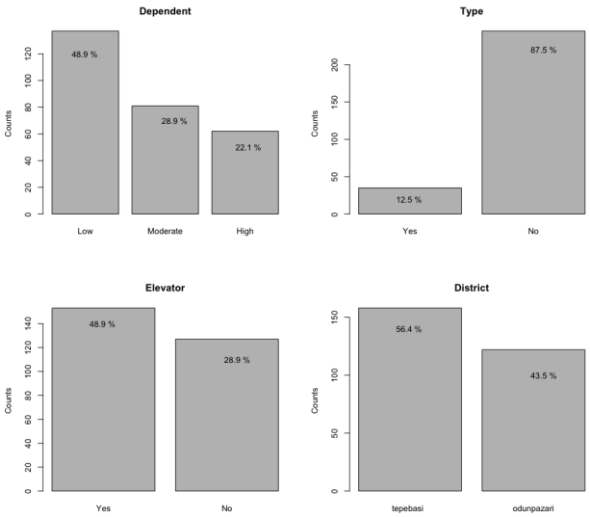

Figure 1. Bar charts for variables: dependent, type (with/without garden), elevator, and district.

In this paper, the comparison of the classification performance of the parametric and nonparametric techniques was discussed and compared. For this purpose, one of the parametric methods, ordinal logistic regression, which is a special form of logistic regression, was used. On the other hand, due to the nature of the dependent variable was categorical, a non-parametric method, the classification trees was performed. There are similar studies that use and compare both techniques for othe real World problems such as churn analysis and disease and evapotranspiration prediction [34-36]. In order to demonstrate the performance success of both techniques, the validity of the models was tested by 5-fold cross validation. Using the results, statistical inferences about the best model was explained.

4. Results

The model fitting for the ordinal logistic regression (OLR) is based on likelihood values. The log likelihood value indicates the probability that the dependent variable is estimated well by the estimators. The log Likelihood value is approximately chi-square distributed under the null hypothesis that indicated model was good without estimators, whereas the alternative hypothesis stated that the model was not good without estimators. According to these results, the -2logL statistic was rejected with a value of p = 0.000 at 0.05 (= alpha) significance level (since p <0.05). Therefore, the OLR model with the estimators provided a statistically significant contribution.

In the chi-square goodness of fit test, the null hypothesis states that the model was compatible with the data, whereas the alternative hypothesis asserts the opposite. The results from the test indicated that the chi-square test statistic was obtained as 11.30 with the corresponding *p* value (=0.419). The model was compatible with the data at 0.05 significance level.

In OLR model, the likelihood ratio test is used to test the parallel lines assumption in order to interpret the coefficients and odds ratios. The null hypothesis is constructed as follows: the regression coefficients in the model are the same in all categories of the dependent variable. Considering the likelihood ratio test results, the regression coefficients in the model are the same in all categories of the dependent variable at 0.05 significance level with the corresponding *p* value (=0.080). That indicated the assumption of parallel lines assumption was hold.

5. Comparison of Modeling Achievements

In order to measure the classification performance of the models obtained by two methods, the data set was split into three sets such as training (50%), validation (25%) and test (25%) [37]. To avoid overfitting 5-fold cross validation was used, after the training and validation data sets were used for modelling. Accordingly, the best performing technique was selected in the validation data set and the performance comparison was reperformed in the test set that was not used before. For the results obtained in the models that were given in Table 2, the one with high accuracy was preferred.

Table 2. Performance criteria for train, validation and test data sets

Model	Train Acc. (%)	Train Misclassification (%)	Validation Acc.	Validation Misclassification(%)	Test Acc.	Test Misclassification(%)
OLR	71.8	28.2	62.7	37.3	64.8	35.2
CART	73.5	26.5	72.9	27.1	74.1	25.9

Comparing the results from training sets in Table 1, CT provided a more successful model than the OLR with an accuracy of 0.735. The validation data set was used to choose between two methods. CT algorithm provides and accuracy of 0.729 while ordinal logistic regression has an accuracy of 0.627. The results show similarity in the test data (results obtained from data that has never been used) and support the results in the validation data.

The confusion matrices for the three categories of the dependent variable obtained from both models are shown in Table 3. The additional criterion such as sensitivity, specificity, PPV (positive predictive value), NPV (negative predictive value) and prevalence are provided in the Table 3 both for ordinal logistic regression model and classification tree model.

In the OLR model, the rate of 'Low' cases that are correctly classified from the total number of real 'Low' cases in the data set is 0.8302. Similarly, the rate of 'High' cases that are correctly classified from the total number of real 'High' cases in the data set is 0.8667. However, the rate of 'Moderate' cases that are correctly classified from the total number of real 'Moderate' cases in the data set is 0.00.

In the classification tree model, the rate of 'Low' cases that are correctly classified from the total number of real 'Low' cases in the data set and the rate of 'High' cases that are correctly classified from the total number of real 'High' cases in the data set are 0.8113 and 0.8333, respectively. While the sensitivity for 'Moderate' cases is obtained as 0.00 in the OLR model, the sensitivity for the same cases in the CT model is as 0.48. From here it can be said that the CT model performed better in terms of sensitivity.

Tablo 3. Confusion matrix for ordinal logistic regression and classification trees model

	OLR			CT		
Criteria	Low	Moderate	High	Low	Moderate	High
Sensitivity (Recall)	0.8302	0.0000	0.8667	0.8113	0.4800	0.8333
Specificity	0.6364	1.000	0.7692	0.7273	0.8916	0.9487
PPV	0.6875	NaN	0.5909	0.7414	0.5714	0.8621
NPV	0.7955	0.7685	0.9375	0.8000	0.8506	0.9367
Prevalence	0.4907	0.2315	0.2778	0.4907	0.2315	0.2778

Specificity means the probability of a negative decision being correct. In the OLR model, the rate of estimating observations that are not in the 'Low' cases but in other cases ('Moderate and High') is 0.6364. The ratio of observations that are not actually in the 'Moderate' cases is estimated as 1,000 but in other cases and the ratio of observations that are not in the 'High' but in other cases is 0.7692. From this it is seen that the decision which is actually negative is highly likely to be selected as negative. In the analysis of classification trees, specificities were obtained higher than the specificity values in OLR model.

In order to determine the overall frequency of the positive class, the prevalence values are examined. The prevalence values were the same for both OLR and CT models. Kappa statistic that is interpreted as moderate compatibility between the categories [38]

is obtained as 0.41 in OLR model while it is 0.58 in CT model.

Compared to accuracy values obtained from both analyses, the correct classification ratio is 0.648 in the OLR model while it is 0.74 in CT model. Hence the classification trees algorithm overperformed OLR in terms of accuracy.

6. Conclusion

In this study, the housing unit price was predicted by using ordinal logistic regression and classification trees algorithm to compare the classification performance of both techniques. Two appropriate models have been succesfully established with the knowledge of the variables such as garden status, elevator status, age, land, the number of rooms, bedroom and bathroom, amenities, balcony status, floor and district are thought to be effective on housing unit prices in Eskisehir. Due to the existence of the independent variables violating the parallel regression assumption, the number of significant variables providing the assumption are considered for further analysis.

For comparisons, the data set was divided into three parts to measure the classification success of the models. Hence, 50% of all data were used as training data to form a model, 25% were used for validation and 25% were used to test the accuracy of classification rules. The best performing method was selected in the validation data set and the performance comparison was re-performed in the 25% data set which was not used before. In the test dataset, the classification success of the model established with the ordinal logistic regression was not better than the classification success of the model established with the classification tree algorithm in terms of accuracy. Accordingly, the more successful model was obtained by classification trees.

Existing data and the results obtained from the successes of these techniques will guide future studies in different areas of application for data with the same characteristics. Also, a future study may include different methods to examine the performance of the methods when missing values are present both in explanatory variables and in the dependent variable.

Author contribution statements

In this study, the contributions of the authors are as follows: The first author in formation of the idea, design and application and reviewing. The second author in collecting data, assessment of the results, reporting the results.

Declaration of Ethical Code

In this study, we undertake that all the rules required to be followed within the scope of the "Higher Education Institutions Scientific Research and Publication Ethics Directive" are complied with, and that none of the actions stated under the heading "Actions Against Scientific Research and Publication Ethics" are not carried out.

References

[1] Berkson, J. 1944. Application of the Logistic Function to Bio-assay. Journal of the American Statistical Association, 39(227), 357-365.

[2] Finney, D.J. 1971. Probit Analysis. 3rd, Cambridge University Press. Cambridge.

[3] Freeman, D.H. 1987. Applied Categorical Data Analysis. Marcel Dekker Inc., New York.

[4] Cox, D.R. 1970. Analysis of Binary Data. 2nd, Chapman and Hall, London.

[5] Aranda-Ordaz, FJ. 1981. On Two Families of Transformations to Additive for Binary Response. Biometrika, 68(2), 357-363.

[6] Johnson, W. 1985. Influence Measures for Logistic Regression: Another Point of View. Biometrika, 72(1), 59–65.

[7] McCullagh, P. 1980. Regression Models for Ordinal Data. Journal of the Royal Statistical Society. Series B, 42(2), 109-127.

[8] Ananth, C.V., Kleinbaum, D.G. 1997. Regression Models for Ordinal Responses: A Review of Methods and Applications. International Journal of Epidemiology, 26(6), 1323-1333.

[9] Blanco, B., Pino-Mejias, R., Lara, J., Rayo, S. 2013. Credit Scoring Models for the Microfinance Industry Using Neural Networks: Evidence from Peru. Expert System Applications, 40(1), 356-364.

[10] Pakgohar, A., Tabrizi, R.S., Khalilli, M., Esmaeili, A. 2010. The Role of Human Factor in Incidence and Severity of Road Crashes Based on the CART and LR Regression: A Data Mining Approach. Procedia Computer Science, 3(8), 764-769.

[11] Twala, B. 2010. Multiple Classifier Application to Credit Risk Assessment. Expert System Applications, 37, 3326-3336.

[12] Breiman, L., Friedman, J., Olsen, R., Stone, C. 1984. Classification and Regression Trees. Chapman & Hall, New York.

[13] Friedman, J.H. 1991. Multivariate Adaptive Regression Splines. The Annals of Statistics, 19(1), 1-67.

[14] Fu, C.Y. 2004. Combining Loglinear Model with Classification and Regression Tree (CART): An Application to Birth Data". Computational Statistics & Data Analysis, 45(4), 865-874.

[15] Timofeev, R. 2004. Classification and regression Trees (CART) theory and applications. Humbolt University, MSc Thesis, Berlin.

[16] Nagalla, R., Pothuganti, P., Pawar, D.S. 2017. Analyzing Gap Acceptance Behavior at Unsignalized Intersections Using Support Vector Machines, Decision Tree and Random Forests. Procedia Computer Science, 109C, 474-481.

[17] Griselda, L., Juan, De O, Joaquin, A. 2012. Using Decision Trees to Extract Decision Rules From Police Reports on Road Accidents. Procedia Social and Behavioral Sciences, 53, 106-114.

[18] Agresti, A. 2007. An Introduction to Categorical Data Analysis, 2nd, Wiley and Sons, New York.

[19] O'Connell, A.A. 2000. Methods for Modeling Ordinal Outcome Variables. Measurement and Evaluation in Counseling and Development, 33(3), 170-193.

[20] O'Connell, A.A. 2006. Logistic Regression Models for Ordinal Response Variables. Thousand Oaks, SAGE, CA USA.

[21] O'Connell, A.A., Liu, X. 2011. Model Diagnostics for Proportional and Partial Proportional Odds Models. Journal of Modern Applied Statistical Methods, 10(1), 139-175.

[22] Powers, D.A., Xie, Y. 2000. Statistical Models for Categorical Data Analysis. Academic Press, San Diego USA.

[23] Hardin, J.W., Hilbe, J.M. 2007. Generalized Linear Models and Extensions, 2nd, Stata Press, Texas USA.

[24] Montgomery, D.C., Peck, E.A. Vining, G.G. 2013. Introduction to Linear Regression Analysis. 5th, Wiley, USA.

[25] Lawson, C., Montgomery, D.C. 2006. Logistic Regression Analysis of Customer Satisfaction Data. Quality and Reliability Engineering International, 22(8), 971-984.

[26] Liao, T.F. 1994. Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models. Quantitative Applications in the Social Sciences, Sage Publications, 101.

[27] Chen, C.K., Hughes, J. 2004. Using ordinal regression model to analyze student satisfaction questionnaires. Association for Institutional Research, 1, 1-13.

[28] Breslaw, J., McIntosh, J. 1998. Simulated Latent Variable Estimation of Models with Ordered Categorical Data. Journal of Econometrics, 87(1), 25-47.

[29] Kleinbaum, D.G., Klein, M. 2010. Logistic Regression: A Self-Learning Text. 3rd, Springer, New York USA.

[30] Hosmer, Jr D.W., Lemeshow, S., Sturdivant, R.X. 2013. Applied Logistic Regression. John Wiley and Sons, New York USA.

[31] Long, J.S. 1997. Regression Models for Categorical and Limited Dependent Variables. Advanced Quantitative Techniques in The Social Sciences. Sage Publications, 7, 1997.

[32] Long, J.S., Freese, J. 2006. Regression Models for Categorical Dependent Variables Using Stata. 2nd, Stata Press, Texas USA.

[33] Brant, R. 1990. Assessing Proportionality in The Proportional Odds Model for Ordinal Logistic Regression. Biometrics, 46(4), 1171-1178.

[34] Rai, S., Khandelwal, N., Boghey, R., 2020. Analysis of Customer Churn Prediction in Telecom Sector Using CART Algorithm. 1st International Conference On Sustainable Technologies For Computational Intelligence Book Series: Advances in Intelligent Systems and Computing, 1045, 457-466.

[35] Liu, Y.F., Ma, B.Y., Wang, Y. 2021. Study on Prediction Model of Stroke Risk Based on Decision Tree and Regression Model. 2021 IEEE International Conference On Big Data (Big Data), December 15-18, Virtual, 4798-4801.

[36] Tareq, W.K., Shukur, O.B. 2021. Using Cart Approach for Classifying Climatic Status of Mosul City. Journal Of Agricultural And Statistical Sciences, 17, 2325-2331.

[37] Brian, R. 1996. Pattern Recognition and Neural Networks, Cambridge University Press, 354s, Cambridge.

[38] Landis, J.R., Koch, G.G. 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159-174.