

Credit risk analysis using boosting methods

S. B. COŞKUN AND M. TURANLI

Abstract

The use of credit for various occasions has become a routine in our lives. In return, banking and financial institutions require to determine whether the loan demands from them contain any risk. Accordingly, these institutions have been increased their activities in determining whether credit rating models from past credit records of the person applying for the loan works properly. Machine learning-based technologies have opened a new era in this field. AI and machine learning based methods for credit scoring are currently implemented by banking or non-banking financial institutions. Employed models are to extract meaningful features from the required data in which wide variety of information available. In this study, credit risk assessment is conducted using boosting methods such as CatBoost, XGBoost and Light GBM. To this aim, Kaggle Home Credit Default Risk dataset is used and the effect of crediting tendency on the results is also considered. The results have shown that gradient boosting methods provide results that are close to each other, and crediting tendency produces better AUC score in CatBoost while it causes a small decrement in AUC score of XGBoost and LightGBM.

Mathematics Subject Classification 2000: 62P05

Keywords: Credit risk, Boosting, Catboost, XGBoost, LightGBM

1. INTRODUCTION

As an important part of debt based monetary system, banking system has the major activity of lending money to people in need of money that makes credit risk evaluation an important issue in financial risk management. A possible bankruptcy of a financial institution would possibly cause a serious butterfly effect. Regarding the problem addressed above, the credit rating system for the loan applicants is crucial that makes it an important task in which loan applicant's past credit information and other potential credit-related data are used to evaluate the repayment ability.

Machine learning-based technologies have made groundbreaking effect in credit risk analysis. In the literature, many studies exist considering the issue. Machine learning techniques including logistic regression, kNN, k-means clustering, SVM, neural networks, decision trees, naïve Bayes, XGBoost, catboost, adaboost, LightGBM had been used recently in credit risk analysis applied to various dataset

(Pandey *et al.*, 2016, Qiu *et al.*, 2019, Li *et al.*, 2019, Menendez, 2019, Pillai *et al.*, 2019, Tian *et al.*, 2020, Wang *et al.*, 2020a, Wang *et al.*, 2020b, Turjo *et al.*, 2021, Ponsam, *et al.*, 2021, Guo and Zhou, 2022).

In this study we investigate the performance of some of the boosting methods applied to a real-life dataset, *i.e.*, Kaggle Home Credit Default Risk dataset (www.kaggle.com/c/home-credit-default-risk). Qiu *et al.* (2019) used same data set and analyzed dataset using logistic regression, random forest and LightGBM and concluded that LightGBM achieved the best result. This study employs XGBoost, Catboost and Light GBM to propose effective credit scoring models based on high dimensional, weakly correlated sparse datasets. To this aim, first the models are developed without performing any parameter tuning technique which are called as base models. Then, parameter tuning is employed to improve the model performances. The results are discussed in detail.

2. MATERIAL AND METHODS

In this section, the data used and the information about the methods employed in the study are described. In this section, the data used and the information about the methods employed in the study are described.

2.1. Dataset

The dataset used in the study is home credit default risk data set from Kaggle (www.kaggle.com/c/home-credit-default-risk). This dataset was the subject of a competition in Kaggle with a prize of seventy thousand usd for which the competitors were asked to predict if loan applicants could repay a loan. The description of the dataset is given in Fig. 1. The dataset includes the information of 307511 customers and is an unbalanced one such that 92% of the data constitutes noncrediting class information while the remaining 8% is about the crediting decision. Each line has 122 columns in which 106 of them have numerical values while 16 of them include categorical information. Five-fold validation is used where 80% of the data is processed for training and validation while 20% percent is used for testing.

The application_train.csv file shown in Fig.1 has 122 columns and 307511 lines and the file also includes TARGET column which provides decision information for all credit applications including 1s and 0s. Hence, the problem becomes a binary classification problem. SK_ID_CURR columns shows client ID. The 1s in the TARGET column describes the client who repaid their loan on time and the 0s in the TARGET column describes the clients who delayed or were unable to pay their loan payments. Most of the data are continuous and discrete, however, string data does also exist. Besides, there are missing data in the dataset. While some of these missing data are really missing data, some of them were entered incompletely due to the absence of that feature, that is, 0.

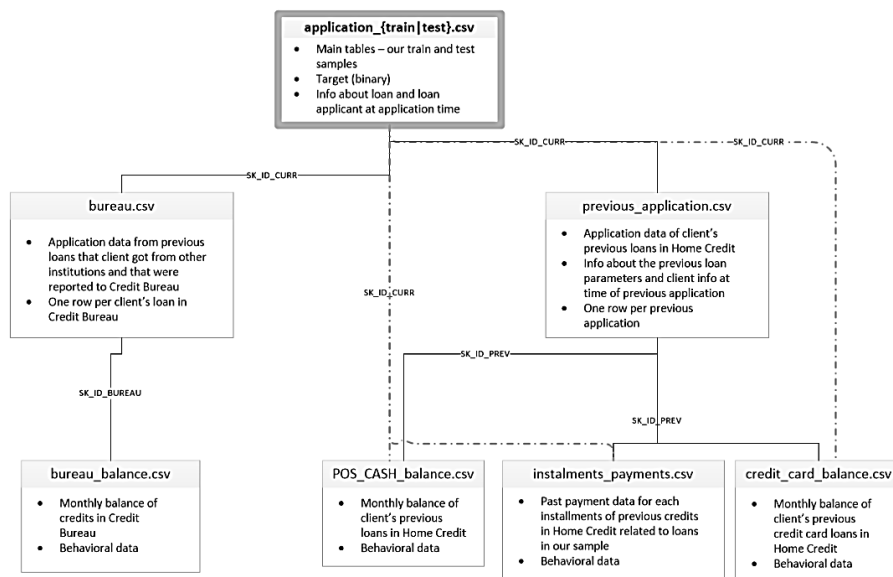


Fig. 1. Descriptions for the columns in the various data files of home credit default risk dataset (www.kaggle.com/c/home-credit-default-risk).

2.2. The Methods

This work includes the application of some of boosting methods which are tree-based classification techniques such that CatBoost, XGBoost and LightGBM. These methods are composed of simple building block models which are sometimes called

as weak learners (James *et al.* 2021, Yalcin and Bagdatli Kalkan, 2022). We start with the general information about boosting and then move to the methods used in the study.

2.2.1. Boosting

Boosting is an ensemble technique for improving the predictions resulting from a decision tree. An ensemble method uses many base models that are sometimes known as weak learners due to their mediocrity. It can be applied to both regression and classification problems.

Although the history of decision trees goes back two decades ago (www.explorium.ai/blog/the-complete-guide-to-decision-trees), decision tree algorithms which are called as CARTs (Classification and Regression Trees) proposed by Breiman *et al.* (1984) can be applied to both regression and classification problems.

In boosting the trees are grown sequentially such that each tree is grown using information from previously grown trees. New trees are created to reduce the errors from the predictions of existing sequence of trees (Giussani, 2019, Jiang, 2021). The boosting methods used in this study are, XGBoost, CatBoost and LightGBM.

2.2.2. XGBoost

XGBoost (eXtreme Gradient Boosting) was proposed by Chen and Guestrin (2016) based on the weak-learners, providing an improvement by training new trees sequentially. XGBoost became so popular in machine learning applications due to its speed and performance. The method applies a shrinkage factor that controls the weighting of new trees added to overcome overfitting. XGBoost adds regular terms and directly uses the first derivative and second derivative value of the loss function to reduce the risk of overfitting (Zhang and Gong, 2020). The XGBoost algorithm can automatically use the CPU for multithreaded parallel computation and apply Taylor's second-order expansion on the loss function. Meanwhile, over-learning is avoided by taking tree model complexity as a regular term in the target function (Liang *et al.*, 2019).

2.2.3. CatBoost

CatBoost (categorical boosting) is a gradient boosting decision tree-based algorithm introduced by the engineers of Yandex (Prokhorenkova *et al.*, 2018). CatBoost is a gradient boosting library which is capable of handling categorical data. Instead of using binary substitution of categorical values, the method performs a random permutation of the dataset and calculates the average label value for the example with the same category value placed before the given one in the permutation (Jhaveri *et al.*, 2019). This is a major advance called ordered boosting that were introduced in CatBoost which is a new boosting schema. This method can overcome prediction shift caused by gradient bias, and further enhance the generalization ability of the model (Zhang *et al.*, 2020). The other major advantage is its new algorithm for categorical features.

2.2.4. LightGBM

LightGBM (light gradient boosting machine) is a lightweight gradient booster that was developed by Microsoft Research (Ke *et al.* 2021) to increase the training time performance of the XGBoost algorithm. The prediction results showed that XGBoost and LightGBM are superior to the traditional GBDT (Xie *et al.* 2019). LightGBM is an improved version of XGBoost with four aspects improved (Zhang and Gong, 2020). First, LightGBM's algorithm incorporates gradient-based one-side sampling algorithm. Second, LightGBM uses a histogram to identify the optimal segmentation point. Thirdly, LightGBM reduces the feature dimension to a certain extent by means of exclusive feature bundling. Fourthly, LightGBM uses a leaf-wise algorithm with depth limitation instead of the traditional level-wise, which improves accuracy and prevents overfitting.

3. ANALYSIS AND DISCUSSION

The study aims to develop boosting models that predict whether the clients will be able to pay off the home loan they applied for. Based on the information provided by the developed models, banking institutions can create a portfolio of risky customers.

At first, the data must be prepared for the analysis. To this aim, data pre-processing steps were conducted using Numpy (numpy.org) and Pandas (pandas.pydata.org) in Jupyter (jupyter.org) environments. In addition, scikit-learn (scikit-learn.org) and LightGBM (github.com/microsoft/LightGBM) libraries were employed in developing machine learning models.

Many of the missing data in the data set are left blank even though they mean 0 (zero). However, this type of data is not missing, it means that it does not show the related property. For example, the credit card number of a client should be written as 3 if he has 3 cards, if he has 1 card, 1 should be written, however, if he does not have any credit card, 0 should be written, it should not be left incomplete. Hence, missing data in the dataset that actually declares a number, or a class is filled with 0 (zero).

Before starting the exploratory data analysis, it is useful to first look at the distribution of the TARGET dependent variable. While 91.9% of clients have difficulty in paying their home loans, 8.07% of them can pay their home loans on time. That means the data set is an unbalanced data set such that the TARGET column includes 0s, 91.9% of column data, 1s, 8.07% of column data. The first problem that may be encountered because of the model working in this way will be that the model learns the 0 class very well, but cannot identify the customers in the 1 class, that is, it cannot learn this class. To prevent this issue, hyper parameters that are effective for imbalanced problems of ensemble models are used.

In developing models, the data is divided into three parts such that 60% is for training, 20% is for validation and 20% for testing. Models are trained in the training set, optimum hyperparameters are found in the validation set, and the final model is decided in the testing set. Models are developed via LightGBM, XGBoost and Catboost, “5-fold cross validation” for the best score and “grid search” methods for optimum parameters. Models are fitted according to the AUC value. In addition, parameter tuning is performed to optimize the AUC value.

Since the Catboost processes categorical variables itself, no additional encoding process is required. Hence, categorical variables are applied to the model as strings without data preprocessing. It is necessary to apply data encoding to categorical

variables for the LightGBM model. Hence, label encoding is applied using scikit-learn. Label encoding is also employed in the XGBoost model.

Some evaluation metrics with the confusion matrix shown in Fig.2 are used to evaluate the performance of the models.

Confusion Matrix		Predicted Class	
		0	1
Actual Class	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

Fig. 2. Confusion Matrix

In the confusing matrix, TP means that the outcome is predicted positive and it's true, TN means that the outcome is predicted negative and it's true, FP means that the outcome is predicted positive and it's false and finally FN means that the outcome is predicted negative and it's false. As per the elements of confusion matrix, following variables are defined as metrics to evaluate and compare the performances of different models (towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

Recall, given in Eq.(1), represents the ratio such that from all positive classes, how many we predicted correctly.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

Precision, given in Eq.(2), represents the ratio such that from all classes we have predicted as positive, how many are actually positive.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

Accuracy, given in Eq.(3), represents the ratio such that from all classes (positive and negative), how many of them we have predicted correctly. With the information provided by Eqs.(1) and (2) following metric can be defined.

$$\text{F1-score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

It is difficult to compare two models with low precision and high recall. Hence, F1-score is used to measure Recall and Precision at the same time. F1-score, given in Eq.(4), employs Harmonic Mean instead of Arithmetic Mean by punishing the extreme values more.

As the first step, the models are developed without any hyperparameter tuning operation. These models are defined as base models. Below, the results for the base models are compared in Table I. The number of crediting decision (1) in the TARGET column of the testing set is 4965 and the number of non-crediting decision (0) in the TARGET column of the testing set is 56538.

As the first observation from Table I, the 0-class has much better precision, recall and f1-score values when compared to the 1-class. This behavior is due to the unbalanced nature of the data. Almost 92% of the data belong to 0-class that lead to high precision and recall values and hence high f1-score. Since the 0-class represents non-crediting, with such a model banking institution will decide not to give the home credit to the client. All base models have the same precision, recall and f1-score values. The precision 0.92 for the class-0 means, 92% of the non-crediting (class-0) decision is correct non-crediting decision in all non-crediting decisions. Hence, 8% of non-crediting decision is given for the clients who can repay the loan. Recall for the class-0 is 100% that lead to a full-success in non-crediting decision. While this model is very safe for the institution, most clients who can repay the loan cannot get a loan from the bank because the model tends to give non-crediting decision, *i.e.* tends to predict the target as 0, due to its unbalanced nature in favor of non-crediting. Such a result can be characterized as an issue caused by the base models and if it becomes widespread in the country, it may become a social issue in the society.

Table I. Performances of the base models

Model	Class	Precision	Recall	F1-score	Accuracy	AUC
Catboost	0	0.92	1.00	0.96	0.92	0.51356
	1	0.51	0.03	0.06		
XGBoost	0	0.92	1.00	0.96	0.92	0.75550
	1	0.48	0.04	0.07		
LightGBM	0	0.92	1.00	0.96	0.92	0.75171
	1	0.48	0.03	0.06		

Crediting decision which is class-1, has a precision of 0.48 and 0.51 according to the base models. That means, only approximately half of the crediting decision by the base models are correct that lead to the half the clients credited by the institution cannot repay the loan. Recall for the class-1 is 0.03 and 0.04 according to base models. Hence, only 3-4% of the clients who can repay the loan, is credited. In other words, approximately 150-200 clients can reach the credit only over the whole true class-1 data that is 4965 clients who can repay the loan. Accuracies for all the base models are the same and also are high due to the unbalanced data. The XGBoost base model has the greatest AUC score. Although it is smaller, the LightGBM base model has a very close AUC score when compared to the XGBoost model. The smallest AUC score belongs to the Catboost base model. Since the higher the AUC the better the classifier (Burkov, 2019), one can conclude that the most successful model out of three base model is XGBoost model. However, the difference between the XGBoost and the LightGBM base models' AUC score is about 0.5%. Hence, we can qualify the LigthGBM model as successful in comparison to the XGBoost base model.

To improve the performance of the model we perform parameter tuning since the model does not predict the class-1 well. Grid search algorithm is used for the tuning. Important parameters after tuning are obtained as shown in Table II. Besides, the performance of the models after applying parameter tuning is shown in Table III.

Table II. Parameters obtained after grid search algorithm

Parameter	Catboost	XGBoost	LightGBM
learning_rate	0.1	0.1	0.1
scale_pos_weight	13	15	13
metric	AUC	AUC	AUC
max_depth	7	10	5
num_leaves	-	64	64

From Table III, it is observed that precision of the class-0 is increased to 0.95 and 0.96 that means 95-96% of the non-crediting (class-0) decision is correct non-crediting decision in all non-crediting decisions. However, recall values are decreased to 0.73, 0.75 and 0.77 for the Catboost, the LightGBM and The XGBoost models respectively. Hence, from all non-crediting decisions we only predict 73-77% of them correctly. That behavior of the models is due to the crediting tendency that exists after parameter tuning.

Table III. Performances of the models with parameter tuning

Model	Class	Precision	Recall	F1-score	Accuracy	AUC
Catboost	0	0.96	0.73	0.83	0.72	0.69171
	1	0.18	0.65	0.28		
XGBoost	0	0.95	0.77	0.85	0.76	0.73912
	1	0.18	0.58	0.28		
LightGBM	0	0.96	0.75	0.84	0.74	0.74952
	1	0.18	0.62	0.27		

Now, the new models tend to give more crediting decision. Decreases in the recall values are directly reflected in the accuracy values as well. In the base models, precision for the class-1 was about 0.50. In the new models, the precision value is 0.18. Only 18% of the crediting decision (1) by the new models are correct that lead to 82% the clients credited by the institution cannot repay the loan. Recall for the new class-1 is increased to 0.58, 0.62 and 0.65 for the XGBoost, the LightGBM and the Catboost models respectively. Hence, 58-65% of the clients who can repay the loan, is now credited. In other words, approximately 2880-3227 clients can reach the

credit now over the whole true class-1 data that is 4965 clients who can repay the loan. This amount was 150-200 for the base models. There is a large increase in crediting decision between the base models and the models after parameter tuning. However, the new models also decide to give credit to the clients who cannot repay the loan that is the reason for the decrease in precision. There is a slight decrease in AUC scores by 0.29% and 2.2% for the LightGBM model and the XGBoost model respectively, while the AUC score of the Catboost model is increased by approximately 34.7%. After parameter tuning, the AUC score of the Catboost model increased significantly while the AUC scores of the LightGBM and the XGBoost models remains almost the same. After the evaluation of both the base models and the models after parameter tuning, one can conclude that the LightGBM model and the XGBoost model performed well in this study.

4. CONCLUSIONS

In this study, credit risk assessment is conducted on Kaggle Home Credit Default Risk dataset using some of the popular boosting techniques such as CatBoost, XGBoost and Light GBM. As the first step, the models are developed without any hyperparameter tuning operation which are called as base models. The results for the base models have shown that all methods have the same or very close precision, recall, f1-score and accuracy values. However, the XGBoost and the LightGBM base models have greater AUC score compared to the CatBoost base model. Besides, the XGBoost base model has the highest AUC score albeit by a small margin. The most important thing that has been observed in the base models is that they learn the 0s very well because of the unbalanced data set. However, the basic models do not predict the 1s very well due to the reasons related to the data. This has resulted in the base models tending to not credit. As a solution to this issue, we performed parameter tuning so that the models can learn 1s better. Hence, hyperparameter tuning is performed via grid search technique. The new models tend to credit more when compared base models. However, recall values and accuracy values are reduced for these new models. Much more clients can reach the credit now, but the new models also decide to give credit to the clients who cannot repay

the loan. After parameter tuning, the AUC score of the Catboost model increased significantly while the AUC scores of the LightGBM and the XGBoost models do not change much. When all the work performed in the study are evaluated, it can be concluded that the LightGBM model and the XGBoost model perform well even in the base models due to their successful score metrics. However, it should be taken into account that this success is based on the overfitting of the dominant class in the unbalanced data set used in the study.

REFERENCES

- BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R.A., STONE, C.J., 1984. *Classification and Regression Trees*, Chapman & Hall/CRC.
- BURKOV, A. 2019. *The Hundred Page Machine Learning Book*, Andriy Burkov.
- CHEN, T., GUESTRIN, C., 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM*
- GIUSSANI, A., 2019. Applied Machine Learning with Python, Egea S.p.A. *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
- GUO, W., ZHOU, Z.Z., 2022. A comparative study of combining tree-based feature selection methods and classifiers in personal loan default prediction. *Journal of Forecasting*, 41, 1248-1313.
- JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R., 2021. *An Introduction to Statistical Learning with Applications in R*, 2nd ed., Springer, New York.
- JHAVERI, S., KHEDKAR, I., KANHARIA, Y., JASWAL, S. 2019. Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns, *Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)*, 27-29 March, Erode, India, 1170-1173.
- JIANG, H., 2021. *Machine Learning Fundamentals*, Cambridge University Press.
- KE, G., MENG, Q., FINLEY, T., WANG, T., CHEN, W., MA, W., YE, Q., LIU, T.Y., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.
- LI, Y., 2019. Credit Risk Prediction Based on Machine Learning Methods. *14th International Conference on Computer Science & Education (ICCSE 2019)*, Toronto, Canada, 19-21 August 2019, 1011-1013.
- LIANG, Y., WU, J., WANG, W., CAO, Y., ZHONG, B., CHEN, Z., LI, Z. 2019. Product Marketing Prediction based on XGboost and LightGBM Algorithm, *Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition AIPR 2019*, August 16–18, Beijing, China, 150-153.
- MELENDEZ, R., 2019. Credit Risk Analysis Applying Machine Learning Classification Models. *Intelligent Computing - Proceedings of the Computing Conference CompCom 2019, Advances in Intelligent Systems and Computing*, vol.997, 804-814.
- PANDEY, T.N., MOHAPATRA, S.K., JAGADEV, A.K., DEHURI, S., 2017. Credit Risk Analysis using Machine Learning Classifiers. *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)*, Chennai, India, 1-2 August 2017, 1850-1854.

-
- PILLAI, S.G., WOODBURY, J., DIKSHIT, N., LEIDER, A., TAPPERT, C.C., 2019. Credit Risk Analysis Applying Machine Learning Classification Models. *Proceedings of the Future Technologies Conference (FTC) 2019, Advances in Intelligent Systems and Computing*, vol.1069, 107-126.
- PONSAM, J.G., BELLA GRACIA, S.V.J., GEETHA, G., KARPASELVI, S., NIMALA, K., 2021. Credit Risk Analysis using LightGBM and a comparative study of popular algorithms. *4th International Conference on Computing and Communications Technologies (IC CCT)*, Chennai, India, 16-17 December, 634-641.
- PROKHORENKOVA, L., GUSEV, G., VOROBEOV, A., DOROGUSH, A.V., GULIN, A., 2018. Catboost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*.
- QIU, Z., LI, Y., NI, P., LI, G., 2019. Credit Risk Scoring Analysis Based on Machine Learning Models. *6th International Conference on Information Science and Control Engineering (ICISCE) Technologies*, Shanghai, China, 20-22 December 2019, 220-224.
- TIAN, Z., XIAO, J., FENG, H., WEI, Y., 2020. Credit Risk Assessment based on Gradient Boosting Decision Tree. *Procedia Computer Science*, 174, 150-160.
- TURJO, A.A., RAHMAN, Y., KARIM, S.M., BISWAS, T.H., DEWAN, I., HOSSAIN, M.I., 2021. CRAM: A Credit Risk Assessment Model by Analyzing Different Machine Learning Algorithms. *4th International Conference on Information and Communications Technology (ICOI ACT)*, Yogyakarta, Indonesia, 30-31 August, 125-130.
- WANG, Y., ZHANG, Y., LU, Y., YU, X., 2020. A Comparative Assessment of Credit Risk Model Based on Machine Learning – a case study of bank loan data. *Procedia Computer Science*, 174, 141–149.
- WANG, Y., LU, J., QIN, J., ZHANG, C., CHEN, Y., 2020. The Application Study of Credit Risk Model In Financial Institution via Machine-learning Algorithms. *7th International Conference on Information Science and Control Engineering (ICISCE)*, Changsha, China, 18-20 December 2020, 1419-1428.
- XIE, Y., XIANG, W., JI, M., PENG, J., HUANG, Y. 2019. Application analysis of predicting monthly house rental based on XGBoost and LightGBM algorithms, *Comput. Appl. Softw.*, 36(9), 151–155.
- YALCIN, M., BAGDATLI KALKAN, S. 2022. Determining the best estimation model with tree-based machine learning methods: implementation on customer spendings for e-commerce websites, *Advances and Applications in Statistics*, 75, 91-109.
- ZHANG, D., GONG, Y. 2020. The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure, *IEEE Access*, 8, 220990-221003.
- ZHANG, Y., ZHAO, Z., ZHENG, J. 2020. CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China, *Journal of Hydrology*, 588, 125087.

Safa Bozkurt Coşkun, Graduate Student,
Department of Statistics,
Istanbul Commerce University, 34445, Türkiye
Email: safab.coskun@gmail.com

Münevver Turanlı, Professor,
Department of Statistics,
Istanbul Commerce University, 34445, Türkiye
Email: mturanli@ticaret.edu.tr