ISTANBUL BILGI UNIVERSITY

INSTITUTE OF SOCIAL SCIENCES

FINANCIAL ECONOMICS MASTER'S DEGREE PROGRAM

Income Estimation Model for Individual Customers

ÇAĞRI SARGAN
117626001

PROF. DR. CENKTAN ÖZYILDIRIM

ISTANBUL
2021

Income Estimation Model for Individual Customers

Bireysel Müşteriler için Gelir Tahmin Modeli

Çağrı Sargan
117626001

Tez Danışmanı: PROF. DR. CENKTAN ÖZYILDIRIM
İstanbul Bilgi Üniversitesi
Jüri Üyesi: DOÇ. DR. ENDER DEMİR
İstanbul Medeniyet Üniversitesi
Jüri Üyesi: DR. DENİZ İKİZLERLİ
İstanbul Bilgi Üniversitesi

Tezin Onaylandığı Tarih: 27.01.2021

Toplam Sayfa Sayısı: 60

Key Words

- Credit Risk
- Income Estimation Model
- Machine Learning

Anahtar Kelimeler

- Kredi Riski
- Gelir Tahmin Modeli
- Makine Öğrenmesi

ACKNOWLEDGEMENT

I would like to thank Prof. Cenktan Özyıldırım for all his guidance, insightfull comments and encouragement. As a result of great collaboration, I believe we have prepared a study that will make a great contribution to the literature.

Furthermore, I owe a deep sense of gratitude to my friend Müge Kartal, who always shared her experiences with me throughout my study period. Her devotion and helpfulness gave me enormous strength in this process.

Last but not least, I would also like to thank to my family and my fiancee for their, encouragement, endless support and understanding in completion of this paper. I dedicate this thesis to them.

Contents

Tables

Charts

Figures

ABSTRACT

The aim of this thesis is to show the increase in predictive power of the income estimation model used during individual product allocation in financial institutions by using machine learning modeling techniques. There are hundreds of factors that affect a customer's income. Although most of these factors are the customer's own information, macroeconomic indicators can cause an impact on individuals' income. In the literature, generally traditional modeling techniques have been used to estimate the income of the customers, and in this study, a modeling study has been carried out by using boosting and bagging algorithms. Compared to regression-based modeling performances, it has been observed that the performance of boosting-based models has more explanatory power. With this study, it is aimed to create a more accurate revenue estimation mechanism for customers. In this way, credit limits will be defined to customers in direct proportion to their ability to pay, and default rates in the portfolio will be minimized with correct product allocation. Within the scope of the study, model validation tests were performed and it was determined that the model performance for the validation sample provided the most descriptive results with the XGBoost algorithm.

Key Words: Income Estimation Model, Machine Learning, Credit Risk, Boosting, Bagging, Non Performing Loans

ÖZET

Bu tez çalışmasının amacı finansal piyasalarda bireysel ürün tahsisi sırasında kullanılan gelir tahmin modelinin machine learning modelleme teknikleri kullanılarak tahmin gücünün yükselişini göstermektir. Bir müşterinin gelirini etkileyen yüzlerce faktör bulunmaktadır. Bu faktörlerin büyük bir bölümü müşterinin kendine ait bilgileri olmakla birlikte makroekonomik göstergelerde kişilerin gelirlerinde etkiye sebep olabilmektedir. Literatürde genel olarak geleneksel modelleme teknikleri kullanılarak müşterilerin gelirleri tahmin edilmeye çalışılmış olup bu çalışmada ise boosting ve bagging algoritmaları kullanılarak modelleme çalışması gerçekleştirilmiştir. Regresyon bazlı modelleme performanslarına kıyasla boosting tabanlı modellerin performanslarının daha fazla açıklayıcılığa sahip olduğu görülmüştür. Bu çalışma ile müşteriler için daha doğru bir gelir tahmin mekanizmasının oluşturulması amaçlanmıştır. Bu sayede müşterilere ödeyebilme güçleriyle doğru orantıda kredi limitleri tanımlanabilecek ve doğru ürün tahsisi ile portföydeki batık oranları minimum seviyeye gelebilecektir. Çalışma kapsamında model validasyon testleri gerçekleştirilmiş ve fazladan örneklem kitlesi için model performansının XGBoost algoritması ile en açıklayıcı sonuçları sağladığı tespit edilmiştir.

Key Words: Gelir Tahmin Modeli, Makine Öğrenmesi, Kredi Riski, Boosting, Bagging, Batık Krediler

1 INTRODUCTION

The automatic lending process in financial markets has been increasing recently. Banks aim to make a customer-based risk analysis, and then complete a systematic assessment and complete the lending process, especially when providing products to individual customers. This risk analysis is an approach that shows the ability of the customer to pay for the relevant product and behind it, using statistical methodologies and information indicating the customer's behavior in financial markets. After the customer's risk analysis is done, application-based evaluation mechanisms are triggered for the relevant customer, and the disbursement approval/rejection process is completed. The most important indicator data here are the customers' past credit product payment performance and the limit risk ratios in the sector. However, when a product of the customer is approved, the income of the relevant customer is the most important variable used in determining the amount of the relevant product and how much risk the customer may take.

With the legislative amendment by the BRSA, which explains the way of assigning the Credit Card Limit assignment structure, started to be implemented in February 2014, banks are able to define a maximum limit of 4 times their income for their customers. This limit is the customer's roof limit, and the credit card limit amount in the entire banking system of the customers cannot exceed this amount. Also, banks are obliged to verify customers' declared income. With this law, banks now have an obligation to estimate the income of their customers. However, banks actively use these income estimates for the credit card product and other individual products. A comparison is made between the income declared by the customer and the income estimated by the bank, and a limitation process is made for the products according to the final accepted income amount.

Financial institutions always want to sell products to portfolios that they find right according to their own risk appetite. It markets products for certain periods of time to customers of their choice, without the customer's application. The most important indicator in the marketing of such products is the calculated income amount for the

customer. The relevant bank makes an income estimation for its customer, then looks at the limits in the sector and calculates the limit gap. Then it can offer the customer a credit card limit increase or a new loan product. Banks should be able to calculate this calculation with high accuracy in order to maintain or increase their market share.

One of the most important indicators showing the portfolio structure of a company is Non Performing Exposure ratio (NPL) in financial institutions. It shows what percentage of the portfolio by volume is a bad portfolio, and the higher the ratio, the banks have to allocate provisions and capital at this rate. In this case, it may cause banks to not reach their profit targets. The default amount of a loan contains a parallel relationship with the risk amount of that loan. Banks determine the loan amounts according to the income of the customers and then make a limit process. While making this calculation, the income of the customer is calculated beforehand, the monthly loan installments and the living amount of the customer are deducted from this income, and the amount of the relevant product is calculated according to this income gap.

This study aims to show the calculation of the income prediction modeling of customers, which are of such importance in financial institutions, by using machine learning methods. Regression-based income estimation calculations have been made in the literature, and tree-based machine learning algorithms have been used in this study.

2 LITERATURE SURVEY

2.1 Income Estimation Survey

There are few academic studies in the literature on income estimation. Studies generally focus on the analysis of demographic and macroeconomic factors affecting income.

Francisco, Whigham, Filho, and Zambaldi (2008), in their studies, examined the relationship between the Brazilian household electricity consumption amount and household income using the Geographic weighted regression method. They concluded that there is a significant relationship between them.

Kibekbaev and Duman (2016), tried to estimate the income of customers by using bank data by using regression algorithms. Traditional linear regression results were found to perform comparably to more sophisticated nonlinear and two-stage models.

As stated in the BRSA's policy, banks have to verify the income declared by the customers by their internal approaches or documentation. In parallel with this directive, Chakraborty, Hui, and Bader (2007) created an income verification methodology using the credit bureau information and loan application information of the customers in their study. They develop the model using the Mars (Multivariate adaptive regression splines) method. They explain that Treenet (gradient boosting technique) techniques will also be used in the important variable elimination step in complex data. They use this technique in variable elimination steps while developing MARS.

Saavedra and Twinam (2020) tried to estimate the occupational income score with machine learning methodologies using demographic and geographic input variables. Occupational income scores systematically underestimate income gaps due to race and gender.

Koskinen (2005) conducted a study to estimate different salary groups using age, working year, and GDP variables. According to the analysis results, salary bands behave differently from each other according to the relevant variables. According to the study, the margin of error in the model was lower in the middle salary group than in the high and low salary groups. This study shows that wage grouping is meaningful in income estimation.

2.2 Methodology Overview

Machine learning algorithms have been popularly used in risk management models recently. The fact that it contains a much higher level of model performance compared to traditional algorithms such as regression has increased the interest and studies in these methodologies. In this context, especially the methodologies using boosting and bagging algorithms are preferred more because of their high performance and fast processing. With the use of ML algorithms, non-linear relationships are included in the calculation. The uncertainty about how much the explanatory variables affect the model score, which is one of these models' biggest problems, has been solved and used with algorithms written on open source platforms. Researchers observed that the tree-based models are more stable than ones based on multilayer artificial neural networks. (Provenzano, Trifiro, Datteo, Giada, Jean, Riciputi, Pera, Spadaccino, Massaron, & Nordio, 2020)

Petropoulos, Siakoulis, Stavroulakis andKlamargias (2018), Basel committee carried out a PD model development study for corporate customers using Deep learning Neural Network, Xgboost, and traditional algorithms by using 10 years of Greek data. In the study, 3 years of out of time data were used. It showed that Deep learning Neural Network and Xgboost algorithms have higher performance than traditional methodologies, logistic regression, and discriminant analysis methods.

Munkhdalai, Namsrai, Lee and Ryu (2019) compared the machine learning model algorithm and the expert-based credit risk model on a performance basis using 3 years of US data between 1998-2001. The results of the study showed that if the financial institutions used the machine learning model algorithm instead of their

internal methodologies in their lending processes in 2001, the expected loan loss provisions would be at a much lower level. In the relevant process, machine learning algorithms, especially the deep neural network and Xgboost algorithms, were seen as the algorithms with the highest performance.

Fenerich, Steiner, Neto, Tochetto, Tsutsumi, Assef, and Santos (2020) analyzed the performance levels among models using Bayesian Network, Decision Tree and Support Vector Algorithm, which are 3 different machine learning algorithms in Unbalanced Brazilian financial instutations data, and 95.2% accuracy was achieved during the development period with the Support Vector Machine algorithm.

Yu (2017) used random forest and xgboost algorithms, one of the ensemble machine learning models, within the scope of online leading credit risk prediction and estimated the risk score of the customers. In this study, the 10-variable Xgboost model showed a higher K-S score performance than the random forest.

Wang nad Ni (2019) implemented a risk model with Xgboost and tried 5 different feature selection algorithms while setting up this model: Weight by Gini, weight by Chi-Square, hierarchical variable clustering, weight by correlation, weight by information. Also, random search and bayesian tree-structured Parzen (TPE) as estimators have tried two hyperparameter optimization approaches and compared the performance of each model. As a result of the analysis, the highest performance Xgboost model was obtained using the hierarchical clustering optimal feature selection and the bayesian TPE hyperparameter methods. In the related study, it was also mentioned that the effects of the feature selection process differ according to the modeling technique. Although the feature selection process for logistic regression does not show much change based on method, the feature selection process designed based on Gini is preferred. In the Xgboost algorithm, the feature selection algorithms change the model performances significantly.

Belkhayat (2018) compared the gradient boosting algorithms; XGB and LGBM models with each other, and the Multi-Layer Perceptron methodology in his modeling study. As a result of this study, both two gradient boosting algorithms

perform better than MLP. In the performance comparison between XGB and LGBM, XGB showed higher accuracy.

The ultimate goal of machine learning algorithms is to make adequate predictions and minimize the error. The power of the model both on model development data and model test data should be sufficient to use it in the bank's strategies. To make accurate predictions, besides using correct and complete data, the choice of methodology is also crucial. For this purpose, instead of using one learner, there is an adopted method that combines many weak learners to obtain more robust learner. This method is known as the ensemble method, and it includes different types of combination methodologies of the learners. The most common ensemble methods are bagging and boosting algorithms.

The cross-validation technique is used during the selection of hyperparameters in order to test the overfitting problem of boosting and bagging methodologies. Cross-validation basically tests the performance consistency in the samples that the model did not see during the development phase. It divides the whole data set into a specified number of parts, develops a model for some of the divided parts, and tests the developed model in the other part. It does this iteratively. Instead of determining and developing the model for a single model development audience and test audience, the model will be developed and tested separately in different train and test groups by creating a cross-validation structure with a specified number of floors (k), and the overfitting problem will be minimized.

Kohavi (1995), analyzed model performance by using k-fold cross-validation and bootstrapping techniques and stated that the most appropriate model selection condition was the cross-validation technique using k coefficient 10

Ramezan, Warner, and Maxwell (2019) applied 3 different cross-validation techniques in their study for tuning classifiers parameters. In the study, k fold, leave one out, and Monte Carlo methodologies are used. Although they were stated that the methodologies had minimal effect on the model's performance, they were

decided to choose the k-fold method because the working time of the k-fold methodology is shorter than the other methodologies.

3 DATASET DESCRIPTION

The conditions for creating the final model development sample, time horizon, and feature selection process are described in the following sections. Besides, target adjustments and target consistency control in model development and validation samples are detailed in the following section.

3.1 Dataset Creation

Within the scope of the study, Bank data was used between June 2019 and March 2020. June, September, and December 2019 data sets were used for training and January, March 2020 data sets were used for validation sample. Data sources such as CRB information, credit account information and balances, demographic information, credit card transactions, salary and income-related information have been used. Training dataset separated into %80 train and %20 test set. Data sources such as CRB information, credit account information, and balances, demographic information, credit card transactions, salary and income-related information have been used. A total of 216 variables were created, and then these variables went through the elimination steps, and the final variables were selected.

In the data set, firstly, during the determination of the data set in which the model was trained, using the determined representation variables, it was checked whether the train set represented the validation set or not. Later, for each of the variables in the data sources, the missing value, stability, trend analysis, anomaly control, and significance of the variables were examined. In Boosting and Bagging methodologies, since there is no condition for non-correlation between variables, correlated variables were used during model development.

Inflation adjustment has been made to the most up-to-date data period for amount variables such as limit and risk. The inflation adjustment is based on the inflation indices published by the Turkish Statistical Institute.

One of the most challenging steps in income models is determining the amount of income, which is the target variable. Although the additional income of customers

with documented information in the bank such as salary amount, rental income and interest income on term deposits is calculated and the target variable is calculated, for customers whose documented revenues are not in the bank, it is necessary to ensure the reliability of the declared income by the customer before being used directly. For customers who do not have documented income, the income declared by the customers was used by applying certain control points. These control points are basically based on comparing the income declared by the customer with their debts in the sector for which the customer is currently able to pay. If the income declared by the customer contains information that is very different from the debt installment amount or credit card limit in the sector, the data with this income has been removed from the modeling sample.

The most frequently used and meaningful variables in income estimation models are the variables that show the monthly solvency of the customer. In particular, information on products (credit cards) restricted by law and have limits proportional to income form an important part of the income model. In addition, information such as the customer's average spending amount for different items (basic and luxury needs separately), the average monthly entry into the deposit account, the customer's work type/occupation, and age have a significant place in the income forecast. In addition to these, while considering the expenses of the customers, it should also be taken into account whether they experience payment difficulties.

3.2 Feature Selection Process

A total of 216 variables were created as a result of the information obtained from data sources. While determining the variables to be included in the final models, first of all, categorical variables were removed from the population in accordance with the algorithms to be used to increase the speed of model working process. Later, variables with a missing rate of more than 95% were eliminated according to the significance of the variables. For each modeling technique, univariate variables with low significance and instability problems were finally removed from the modeling population. The variables that will eventually enter the modeling process

are applied by applying a recursive process. The variables that constitute the most appropriate level of complexity and model performance are selected.

Friedman (2001), defines the relative importance of a single feature to be obtained by averaging over all classes as follows:

$$\hat{I}^2_{jk} = \frac{1}{M}\sum_{m=1}^{M} \hat{I}^2_j(T_{km})$$

Where T_{km} is the tree induced for the kth class at iteration m. The quantity $\hat{I}_{jk}$ can be interpreted as the relevance of predictor variable x_j in separating class k from the other classes. The overall relevance of x_j can be obtained by averaging over all classes.

$$\hat{I}_j = \frac{1}{K}\sum_{k=1}^{K} \hat{I}_{jk}$$

Although there are many variable elimination techniques in the literature, the process of eliminating variables according to their relative importance level is a straightforward and applicable method.

PSI test was used to calculate variable stability. It is conducted to examine the differences in the distribution of raw numeric factors among data sets.

$$PSI = \sum_{i=1}^{T}\left(s_i^1 - s_i^2\right)\cdot ln\left(\frac{s_i^1}{s_i^2}\right)$$

$$s_i^j = \frac{n_i^j}{N^j}$$

where n_i^j is the number of counterparts in bucket i in jth sample and N^j is the total number of counterparts in jth sample

There are two different types of variable importance calculation for the output of Random Forest algorithm. Both techniques aim to measure the predictive power of the variables. While more powerful variables lead to more adequate outcomes,

variables that have lower predictive power make the algorithm complicated and slower with a lower impact on the outcomes. The first technique to measure variable importance is defined as a mean decrease in accuracy. In order to calculate variable importance with a mean decrease in accuracy technique, the out-of-bag sample is split during the growth of each tree in a random forest. The out-of-bag data is used for shuffling all the values of the chosen variables to calculate the importance of it. The remaining variables are kept the same while the values of the chosen variables is shuffled. Then, the decrease in accuracy of the outcome on the out-of-bag data is measured. After the application of this process on each tree in the random forest, the mean decrease of the accuracy is obtained for all variables. The second technique is called mean decrease in node impurity. To specify the decrease in node impurity, Gini index is used. Each time of chosen of the variable to split a new node, the sum of the decrease in accuracy is measured for all trees. In order to find the mean of the decrease, the sum of the decrease is divided by the number of the trees in the random forest.

Since the second technique is biased in splitting node with variables which have many classes, the result may be biased. For this reason, the first technique is chosen to define to calculate variables' importance.

Feature selection steps are done separately for each model segment and are explained in the following sections.

4 METHODOLOGY

This study aims to estimate the income of individual customers by using boosting and bagging algorithms, which are ensemble algorithms. It is aimed to create a strong estimator by combining the related algorithms with weak learners, and while doing this, a low bias and a low variance are aimed. Ensemble methodologies can basically be called combining different models. The iterations in the boosting algorithms are affected by the previous errors, and the bagging algorithms work independently from the previous errors. Alfaro, Gamez, & Garcia, 2013). Gradient Boosting algorithms have become a very important and frequently used methodology in machine learning models due to the performance they produce, especially in recent years.

The name of bagging is the short version of bootstrap aggregating. Bagging was proposed by Leo Breiman in 1996. For each iteration of training, samples are chosen by bootstrapping and used for training and validation purposes. The samples are selected at random and replaced into a training set again. After the selection of the samples, the key point is to determine the base model to be aggregated. Models are trained with base methodology such as decision tree independently from each other concurrently and combined with the averaging process. This approach helps to reduce variance and concern of over fitting with using bootstrapped samples and cross validation process. Random Forest is the most popular bagging algorithm, and it contains many decision trees.

Alfaro, Gamez, and Garcia (2013) define the bagging algorithm as follows:

Repeat for $b = 1, 2, \ldots\ldots, B$

a. Take a bootstrap replicate T_b of the training set T_n
b. Construct a single classifier $C_b(x_i) = \{1, 2, \ldots\ldots, k\}$ in T_b

ii. Combine the basic classifiers $C_b(x_i)$, $b = 1, 2, \ldots\ldots, B$ by the majority vote (the most often predicted class) to the final decision rule $C_f(x_i) = \text{argma}x_{j \in Y} \sum_{b=1}^{B} \mathrm{I}(C_b(x_i) = \mathrm{j})$

The other main algorithms of ensemble methods are boosting methods. Many boosting algorithms are available, but Freund and Schapire (1999) first proposed the boosting algorithm. Unlike bagging algorithms, weak learners are trained sequentially in boosting algorithms. In addition, while the learners have the same weight in bagging algorithms, different weighting approach is applicable in boosting method. This means that the misclassified observations have more weight in the next training step. Because in the boosting algorithm, the aim is the minimize error that comes from the previous learner. The most important difference between the boosting algorithms is the weighting training data sets. The most popular boosting algorithms; AdaBoost, Gradient Boosting, Extreme Gradient Boosting (XGB) and LightGradient Boosting Machine (LGBM).

Schapire (2013) defines the boosting algorithm as follows that:

Given(x_1, y_1), ……. , (x_m, y_m) where $x_i \in$K, $y_i \in \{-1,1\}$.

Initialize: $B_1(i) = {}^1/_m$ *for* $i = 1$, ……. ,m.

For $t = 1$, …… ,T:

- Train weak learner using distribution B_t
- Get weak hypothesis l_t: K $\rightarrow$ {-1,1}.
- Aim: select l_t with low weighted error:

  $\varepsilon_t = Pr_{i \sim l_t}[s_t(x_i) \neq y_i]$

- Choose $\sigma_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
- Update, for i = 1, ……. ,m:

$$B_t(i) = \frac{B_t(i)\exp(-\sigma_t y_t s_t(x_i))}{Z_t}$$

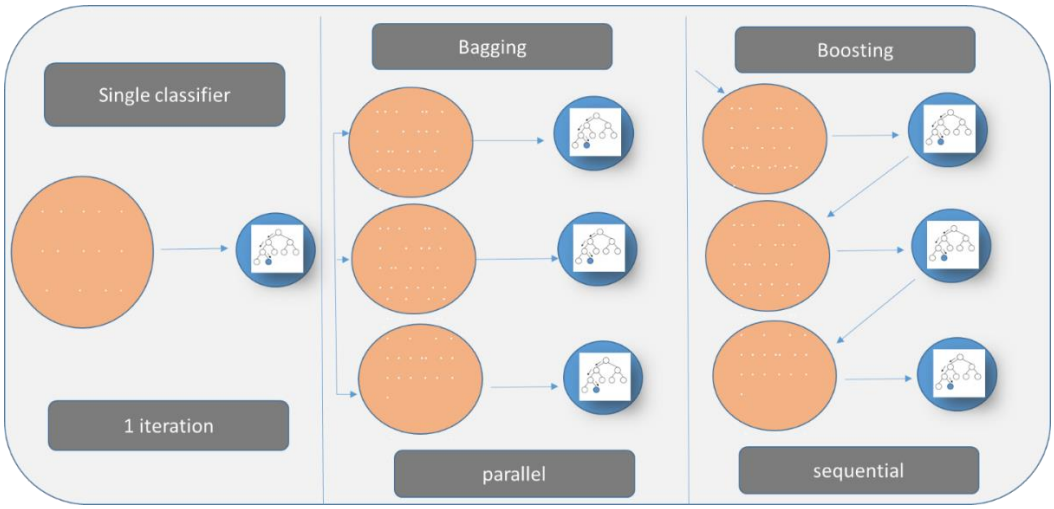
Where Z_t is a normalization factor (chosen so that $B_{t+1}(i)$ will be a distribution).

Output the final hypothesis:

$$S(\mathrm{x}) = \text{sign}\left(\sum_{t=1}^{T} \sigma_t \, s_t(x)\right)$$

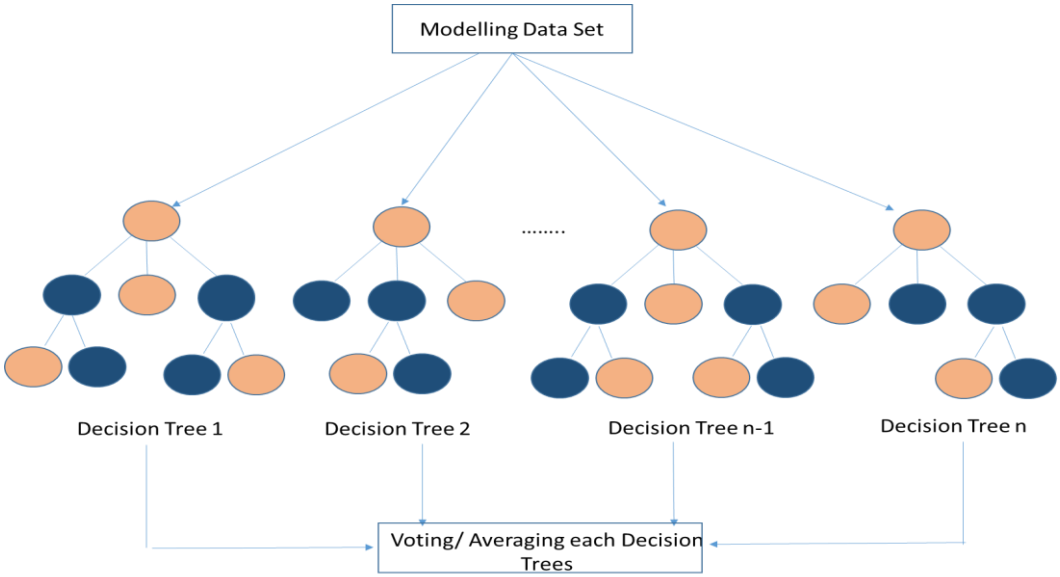
Kotsiantis and Pintelas (2004) states that Boosting algorithms are considered stronger than bagging on noise-free data. However, bagging is much more robust than boosting in noisy settings.

Figure 1: Boosting&Bagging Model Processes



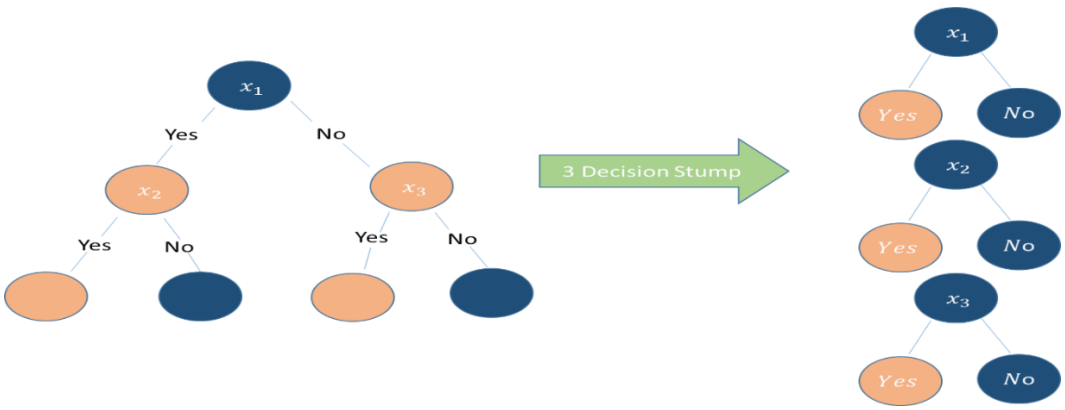
In this study, individual customers' income estimation was estimated using Random Forest, XGBoost, and LightGBM algorithms instead of classical regression algorithms. The main reason for this is that regression-based models do not have high accuracy in income models. Although its implementation is more difficult and more complex than regression, it is aimed to make an income model with higher accuracies with boosting and bagging algorithms.

Figure 2: Random Forest Model Process



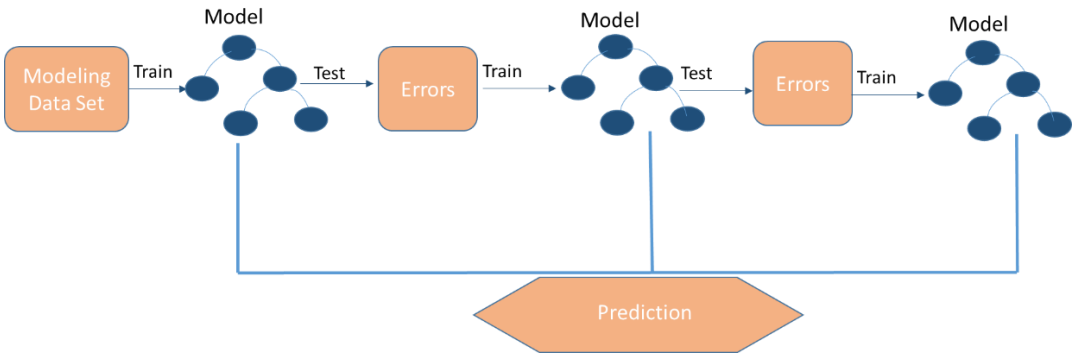
The random forest algorithm basically defines each tree as an explanatory variable and applies the final estimation process by taking the average of the predictions of each of these trees. Instead of applying the decision from a single decision tree, this type of application reduces the overfitting problem by averaging the explanatory decision trees.

Figure 3: AdaBoost Model Process



AdaBoost is a methodology that aims to convert weak classifiers into a single strong classifiers and split decision trees with a single split called stumps. It is one of the first boosting algorithms to appear. Flayeh and Davami (2013, p.58) states that "*AdaBoost is a classification algorithm which calls a given weak learner algorithm repeatedly in a series of rounds. A weak learner is a learning algorithm that performs just slightly better than random guessing and finds a separation boundary between two classes (positive and negative). AdaBoost combines a number of weak learners to form a strong learner in order to achieve better separation between classes. The strong learner is a weighted majority vote of the weak learners*".

Figure 4: XGBoost Model Process

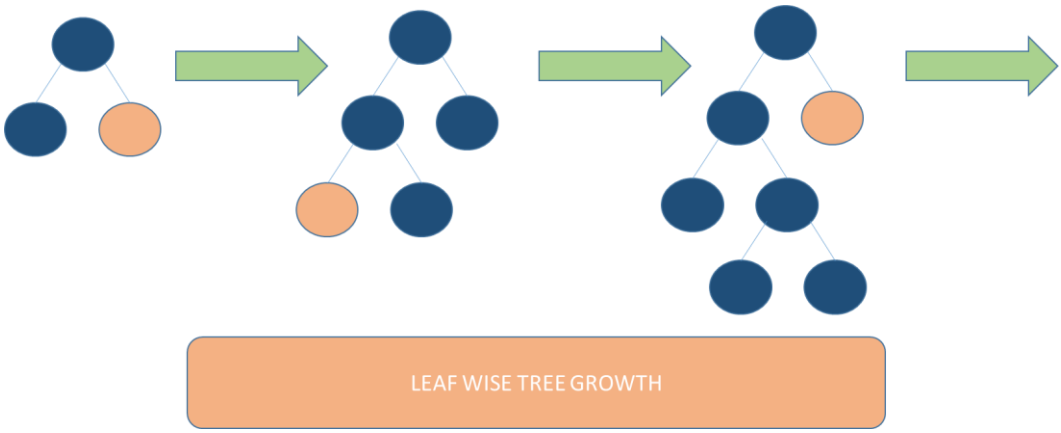


XGBoost is the algorithm of gradient boosting, which is optimized efficiently and flexibly. It is an algorithm that has been used quite popular in Kaggle competitions recently. It works very fast compared to other gradient boosting algorithms. The main difference between XGBoost and AdaBoost is that XGB determines learners based upon the selected loss function.

Dias, Forti, and Witarsa (2018) stated that the boosting algorithms have better performance in the customer risk models he made comparing boosting algorithms and logistic regression, and there are two reasons for this. i) Boosting algorithm provides more interaction between variables and can predict possible scenarios better than regression and ii) Due to the methodology of using variables in boosting

algorithms, it is not common to eliminate any variable because it presents inverse interpretation of the business sense. In logistic regression, variables can be eliminated more easily if their effects on the model result are low.

Figure 5: LightGBM Model Process



Unlike other boosting algorithms, LightGBM grows vertically relative to the tree leaf wise. Other Boosting algorithms work with growth level-wise tree growth. While level based growth maintains a balanced tree structure, leaf wise growth focuses on dividing the leaf that will reduce the loss the most and aims to minimize errors. Although this situation is more prone to overfitting, the possibility of overfitting is further reduced if the number of data used is high. LGBM works faster than other algorithms because it supports GPU learning, and it uses less memory. It is called light because of its speed. Also, when using the LGBM algorithm, it can work in categorical variables.

Taha and Malebary (2020) studied to make predictions using different algorithms within the scope of Credit Card fraud detection and then compared the performance of each model used. They stated that LGBM methodology has the highest explanatory power when comparing the performances of random forest, logistic regression, the radial support vector machine, the linear support vector machine, k nearast neighbors, decision tree, and naive bayes.

4.1 Hyperparameter

Hyperparameters are coefficients determined before starting the learning process in any modeling process. Although these coefficients vary according to modeling techniques, they can affect the model's performance and overfitting in machine learning modeling algorithms. In the machine learning model, the process that tries to find out which combinations of the parameters belonging to the model will create the best performance for the model is called hyperparameter optimization and these parameters will bring the loss function to the lowest level. In their study on hyperparameters, Yang and Shami (2020) studied different automatic hyperparameter optimization methods. They showed that the tunings made had a direct effect on the model performances. Also, Yang and Shami stated that "two types of parameters exist in machine learning models: one that can be initialized and updated through the data learning process, named model parameters, while the other, named hyperparameters, can not be directly estimated from data learning and must be set before training an ML model because they define the architecture of an ML model". Hyperparameters have a high effect on model results, especially in tree-based modeling techniques. Although there are many hyperparameter tuning algorithms, the most popular are; Manual search, Grid search, Random search, and automated hyperparameter tuning methods. Bergstra and Bengio (2012) stated that the hyperparameter values determined by random selection are more efficient than other methods. According to Bergstra and Bengio (2012), not every hyperparameter is of the same importance, and more complex methods, therefore, require a lot of time. They state that each hyperparameter can easily be found with a random search. Due to the large number of data samples and variables used within the scope of this study, it has been tried to find the best hyperparameter values with the grid search method.

Details of the important hyperparameters used for Random Forest, XGBoost, LightGBM, and AdaBoost are explained below.

Most important hyperparameters for Random Forest:

n-estimator: shows the number of trees in the forest model. The default value is 10 different trees

max-depth: the maximum number of depths allowed to build a tree.

Max-features: maximum number of features considered when splitting a node.

min_samples_split: Minimum number of samples required to split an internal leaf node. The default value is 2.

min_samples_leaf: Minimum number of samples required to be at a leaf node. The default value for this parameter is 1

bootstrap = method for sampling data points. Determines the bootstrap samples are used when building trees.

Most important hyperparameters for AdaBoost:

n-estimator: The maximum number of estimators at which boosting is terminated. The default value is 50.

learning-rate: Shows how each tree contributes to the overall results. It shrinks the contribution of each classifier. There is a trade-off between the number of estimators and the learning rate. The default value is 1. Friedman (1999) states that small values ($v <= 0.1$) lead to much better generalization error.

Most important hyperparameters for Extreme Gradient Boost(XGB):

Booster: which booster to use. It can be gbtree or dart for tree-based models.

Nththread: The number of parallel threads used to run XGBoost.

Eta: Step size shrinkage used in the update to prevents overfitting. Shrinks the feature weights to make the boosting process more conservative.

Gamma, alpha and lambda: parameters are used for tree pruning in order to apply a different learning method in case the homogeneity of the samples classified

by decision trees in the decision tree-based XGBoost algorithm is below a certain threshold value. The larger gamma is, the more conservative the algorithm will be.

Max_depth: It is the parameter that determines how deep the decision tree will branch out. Increasing this value will make the model more complex and more likely to overfit.

Min_child_weight: Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning. The larger min_child_weight is, the more conservative the algorithm will be.

Subsample: Subsample ratio of the training instances. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees, and this will prevent overfitting.

Colsample_bytree: It is the parameter that determines what percentage of model variables will be used randomly while creating decision trees. Subsampling occurs once for every tree constructed.

Scale_pos_weight: It is the parameter that provides a granularity of model predictive values in data sets where positive and negative target variable percentages are not evenly distributed.

Most important hyperparameters for Extreme Gradient Boost (XGB):

N-leaves: This parameter is used to set the number of leaves to be formed in a tree.

Min_data_in_leaf: minimal number of data in one leaf. It can be used to deal with over-fitting

max_depth: limit the max depth for the tree model. This is used to deal with over-fitting when #data is small.

min_child_samples: Minimum number of data needed in a child (leaf)

Learning_rate: Boosting learning rate. You can use the callbacks parameter of the fit method to shrink/adapt learning rate in training using reset_parameter callback

colsample_bytree: It is the parameter that determines what percentage of model variables will be used randomly while creating decision trees. Subsampling occurs once for every tree constructed.

In this study, using grid search, certain experimental values for the relevant parameters for each model were determined separately, and the parameter values that maximize the model performance in the model development sample and test data were determined as the final values.

4.2 Model Segmentation

One of the most important processes in the model development phase is segmentation. Model segmentation can be made according to the usage area of the model, the change in the data sources used, or the variables that can determine the different behaviors in the model development sample. Within the scope of this study, it has been determined that the average income of individual customers varies according to their education level and age. For this reason, the model development sample is divided into 9 groups according to the age and educational status of the customer. Relevant segments and data quantities of each segment are shown in the table below.

Table 1 Number of Counts based on Segment and Year Month

Segment	Education Level	06'2019	09'2019	12'2019	01'2020	03'2020	Total
1	Unknown& Elementary	67,603	128,482	145,937	101,452	120,526	564,000
2	High School& Associate Degree	194,100	301,134	300,492	218,486	252,552	1,266,764
3	Bachelor& Master	209,318	230,411	232,210	216,648	244,072	1,132,659
	Total						2,963,423

As shown in Table 1, a data set of 2.9 million, including the model development sample and the validation sample, has been created. It is seen that the bachelor & master segment has the largest share in this population. The model development sample is randomly divided into 80% train and 20% test for each segment.

Table 2 Monthly Average Income based on Education Level and Year Month

Segment	Education Level	06'2019	09'2019	12'2019	01'2020	03'2020
1	Unknown& Elementary	4,128	3,703	3,806	4,384	4,418
2	High School& Associate Degree	4,311	3,911	3,901	4,371	4,430
3	Bachelor& Master	7,252	7,046	7,183	7,667	7,734

It is seen in table 2 that as the level of education increases, the average income increases especially in the bachelor & master segments. In the table below, the distribution of the monthly average income amount according to education levels according to the data of TURKSTAT is given. Consistent with Table 2, it was observed that the higher the education level, the higher the average income level.

Table 3 Monthly Average Income based on Education Level by TURKSTAT

Segment	Education Level	2018	2019
1	Below High School Level	1,953	2,236
2	High School Level	2,540	2,843
3	Bachelor& Master	3,896	4,324

5 APPLICATION OF MODELS

Models were developed separately for 3 education segments decided within the scope of model segmentation. The feature selection steps of the relevant segments and each model study developed were evaluated separately.

5.1 Unknown & Elementary Education Level Model Development Scope

Model development work was executed for customers with unknown or elementary education levels. In the modeling phase, the dataset is divided into train, test, and validation samples. Train and test samples have 80% and 20% weight, respectively. The model validation sample includes the period January 2020 and March 2020, which are not used in the model development phase.

Table 4 Datasets Distribution

Datasets	Count
Model Development Sample (%80)	273,618
Model Test Sample (%20)	68,404
Model Validation Sample	221,978

5.1.1 Feature Selection Process XGBoost

After the datamarts were formed, a total of 219 variables were created. These variables were then evaluated in the stability and correlation elimination. In the stability analysis, the stability of the variables between the model development population and the validation population for all variables was evaluated with the PSI test. Variables above 0.25 according to the PSI value were excluded from the model development sample. In total, 7 variables were eliminated due to instability problems. PSI values of 7 eliminated variables are shown in the table below.

Table 5 Variable PSI Information

#	Variable Name	PSI
1	LAST_KKB_PRODUCT_TENURE	1.03
2	TOT_EDU_PMNT_TL_L12M	0.49
3	KKB_APP_CNT_L6M	0.43

4	OTH_BNK_OTH_ACC_EFT_CNT _L12M	0.27
5	OTH_BNK_OTH_ACC_EFTAMT_ TL_L12M	0.26
6	AVG_OTH_FN_SRV_PUR_AMT_ TL_L12M	0.25

Although model performance in boosting and bagging algorithms is not affected by the correlation of variables with each other, an elimination process has been applied for the variables with more than 70% correlation in order to reduce the margin of error. Explanatory variables of the variables correlated with each other were examined with the target, and then the variable that was less explanatory with the target was eliminated. In total, 63 variables were eliminated in the correlation step. The target explanatory (Gain) and correlations of the variables and the details of the eliminated variables are given in the tables below.

Table 6 Top 10 Variable Gains

#	Variable Name	Gain
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	63.2%
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	12.2%
3	AVG_RETLOAN_PMNT_AMT_TL_L12M	4.0%
4	AVG_MAIL_ORD_PUR_TRX_CNT_L12M	2.8%
5	OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	2.0%
6	MILAT_FLAG	1.0%
7	INDV_CC_AVG_PUR_FX_AMT_TL_L12M	0.8%
8	KKB_MINUMUM_CC_OPEN_DATE_TERM	0.8%
9	AVG_PERS_FIN_ASSET_L12M	0.8%
10	CUSTOMER_AGE	0.6%

As shown in Table 6, the variables with the first 10 gains constitute approximately 90% of the total weight. Descriptions of 10 variables are shown in the table below.

Table 7 Description of Top 10 Variables

#	Variable Name	Variable Description
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	The customer's total credit card limit in the banking sector
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	The average outflow amount from the demand deposit balance in the last 12 months

3	AVG_RETLOAN_PMNT_AMT_TL_L12M	Average amount of installments paid to YKB installment loans in the last 12 months
4	AVG_MAIL_ORD_PUR_TRX_CNT_L12M	Average number of mail orders in the last 12 months
5	OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	EFT amount made to other banks in the last 12 months
6	MILAT_FLAG	Having a credit card limit before/after the credit card limit determination law
7	INDV_CC_AVG_PUR_FX_AMT_TL_L12M	The average monthly amount of foreign currency expenditures made by personal credit cards in the last 12 months.
8	KKB_MINUMUM_CC_OPEN_DATE_TERM	The number of months from the oldest open credit cards to the relevant date
9	AVG_PERS_FIN_ASSET_L12M	Average TL amount of personal financial assets in the last 12 months
10	CUSTOMER_AGE	Customer Age

63 variables were eliminated from the total correlation elimination, and the relationship strength of these variables to the target is low, and their total contribution to the model seems to be approximately 4%.

After the variable elimination, 74 variables were eliminated during the modeling phase because the level of significance was not sufficient according to the XGBoost methodology. It means that these variables don't add value to the XGB model. 75 variables remained in the model development step. Summary statistics details of these variables are given in the table below. Summary statistics details of these variables are given in the table below.

Table 8 Final Variables Summary Statistics

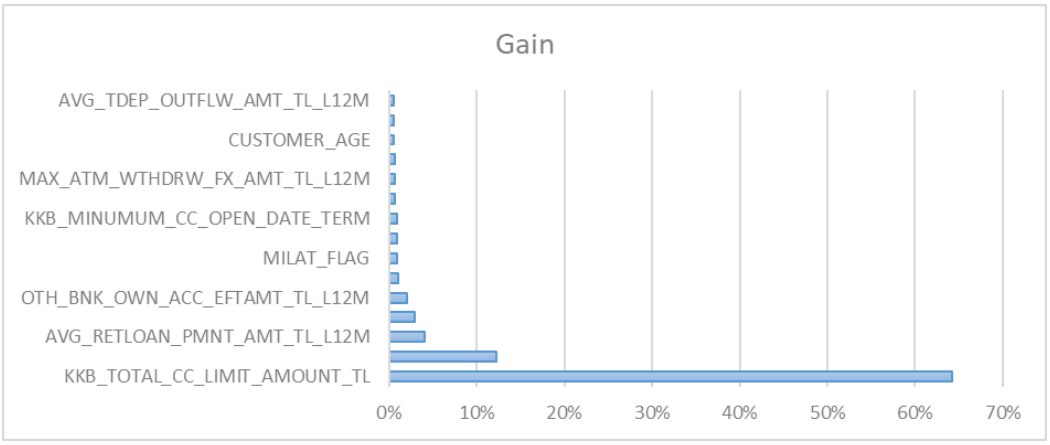
Variable Names	Min	1st Qu.	Median	3rd Qu.	Max
KKB_TOTAL_CC_LIMIT_AMOUNT_TL	0	2600	7600	11200	221900
AVG_DDEP_OUTFLW_AMT_TL_L12M	-2090776	-5640	-3641	-2694	0
AVG_RETLOAN_PMNT_AMT_TL_L12M	0	98	612	1168	34831
AVG_MAIL_ORD_PUR_TRX_CNT_L12M	0	0	0	0	31
OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	0	0	0	1350	1621340
INDV_CC_AVG_PUR_FX_AMT_TL_L12M	-178	0	0	0	23110
MILAT_FLAG	0	0	0	1	1
AVG_PERS_FIN_ASSET_L12M	0	46	382	1469	7084747
KKB_MINUMUM_CC_OPEN_DATE_TERM	0	49	90	137	367
NOF_BH_INTRNT_BANK_LOGIN_L12M	0	0	0	2	1445
MAX_ATM_WTHDRW_FX_AMT_TL_L12M	0	0	0	0	13564

INDV_CC_AVG_PUR_AMT_TL_L6M	-892	0	0	628	53901
CUSTOMER_AGE	18	32	38	45	86
AVG_ATM_WTHDRW_AMT_TL_L12M	0	1172	1804	2419	25115
AVG_TDEP_OUTFLW_AMT_TL_L12M	-1243180	0	0	0	0
MZC_AVG_CSH_LIM_AMT_TL_L12M	0	5569	14357	30796	1165656
AVG_TPORT_PUR_AMT_TL_L12M	-107	0	0	0	10975
AVG_HOTEL_PUR_AMT_TL_L12M	-1134	0	0	0	30045
AVG_DDEP_EUR_PSTV_BAL_TL_L12M	0	0	0	0	896649
AVG_TDEP_USD_BAL_TL_L12M	0	0	0	0	7003350
AREA_HOUSEHOLD_INCOME_AMT_TL	0	2390	3346	4528	65609
TOT_SWIFT_AMT_TL_L12M	0	0	0	0	1910763
AVG_CR_INST_AMT_TL_L12M	0	0	0	584	24270
NOF_BH_MOB_BANK_LOGIN_L12M	0	19	77	152	5065
AVG_HGS_PMNT_TL_WITH_DDEP_L12M	0	0	0	0	311
KKB_MAX_OPEN_OD_LIM_AMT_TL	0	500	1500	2750	145250
KKB_TOTAL_OPEN_CREDIT_COUNT	0	3	4	6	37
AVG_DDEP_USD_PSTV_BAL_TL_L12M	0	0	0	0	1096048
AVG_CAFE_PUR_AMT_TL_L12M	0	0	0	0	3827
AVG_BILL_PMNT_CNT_WITH_DD_L12M	0	0	3	19	439
KKB_MAX_2ND_CC_LIM_AMT_TL_L3Y	0	0	1150	3170	71900
KKB_MAX_CLS_CR_INST_AMT_TL_L3Y	0	0	311	688	221485
MAX_ATM_WTHDRW_AMT_TL_L12M	0	1500	1700	2200	5500
AVG_INSUR_PUR_AMT_TL_L12M	-827	0	0	0	20603
AVG_CLOTHING_INST_CNT_L12M	0	0	0	0	176
KKB_TOTAL_CC_RISK_AMOUNT_TL	0	0	1113	3555	185732
KKB_TOTAL_OPEN_CC_COUNT	0	1	1	2	11
AVG_TBC_LIQR_PUR_AMT_TL_L12M	-6461	0	0	0	5872
KKB_OPEN_TOT_PMT_AMT_LM	0	0	0	700	321198
CUSTOMER_TENURE	3	31	79	146	442
YKB_CC_TENURE	0	0	8	27	356
AVG_GLASSWARE_PUR_AMT_TL_L12M	-144	0	0	0	5600
AVG_OTH_BILLPMNTTL_WTH_DD_L12M	0	0	0	0	98983
NOF_ATM_WTHDRW_TL_TRX_L6M	0	12	19	30	964
LAST_STMT_INDV_CC_INST_BAL_PCT	0	0	0	0	266
AVG_SH_BG_LTHR_PUR_AMT_TL_L12M	-28	0	0	0	3000
AVG_TDEP_EUR_BAL_TL_L12M	0	0	0	0	2088689
AVG_ASSN_CLUB_PUR_AMT_TL_L12M	0	0	0	0	3766
NOF_BH_ATM_TRX_L12M	0	40	83	144	7948
AVG_TELECOM_PMNTTL_WTH_DD_L12M	0	0	0	0	8040
AVG_PATISSERIE_PUR_AMT_TL_L12M	0	0	0	0	6757
KKB_DIFF	0	0	1	1	7
AVG_CC_CADV_BAL_TL_L12M	0	0	0	0	21592

AREA_UNIVERSITY_EDUCATION_PCT	0	0	0	0	1
AVG_COSMETIC_PUR_AMT_TL_L12M	-6	0	0	0	2980
AVG_AUTO_PMNT_CNT_WITH_DD_L12M	0	0	0	4	179
AVG_COMNC_PUR_AMT_TL_L12M	-55	0	0	6	9494
AVG_FURNTR_PUR_AMT_TL_L12M	-1079	0	0	0	24445
NOF_NBH_ATM_TRX_L12M	0	21	50	97	7525
AVG_ACSORY_PUR_AMT_TL_L12M	-40	0	0	0	3300
AVG_MTV_PMNT_TL_WITH_DDEP_L12M	0	0	0	0	4360
AVG_OTHER_SRV_INST_CNT_L12M	0	0	0	0	53
INDV_CC_VDAY_PUR_AMT_TL_L3Y	-11623	0	0	0	16286
NOF_BH_BRANCH_VISIT_L12M	0	0	1	3	117
AVG_DON_PMNT_TL_WITH_DDEP_L12M	0	0	0	0	3257
MTV_AMT_TL_WITH_CC_L12M	0	0	0	0	4360
AVG_SPRT_ENT_PUR_AMT_TL_L12M	-108	0	0	0	3688
AVG_TOY_SHOP_PUR_AMT_TL_L12M	-4	0	0	0	1346
AVG_OFFC_SPLY_PUR_AMT_TL_L12M	0	0	0	0	22500
AVG_SOUVENIR_PUR_AMT_TL_L12M	-3	0	0	0	5290
AVG_DDEP_CHF_PSTV_BAL_TL_L12M	0	0	0	0	8497
AVG_BOOK_MSC_PUR_TRX_CNT_L12M	0	0	0	0	35
INDV_CC_AVG_CSH_PNT_EXP_L12M	-9	0	0	0	584
MAX_OVERDUE_DAY_CNT_L12M	0	0	0	5	4581
AVG_JEWELLERY_PUR_AMT_TL_L12M	0	0	0	0	28980

Gains of the most important variables are shown in the chart below.

Chart 1 Most Important Variables Gain



5.1.2 Feature Selection Process LightGBM

As mentioned in the XGBoost step, first of all, variables with instability problems between the model development sample and the validation sample were removed from the modeling phase. Since modeling and validation samples did not change at this stage, the variables that were eliminated are the same variables as XGB. According to the PSI analysis results, 7 variables were eliminated due to instability problems.

The remaining variables were included in the correlation analysis. Among the variables correlated according to the binary correlation results, the variable with the highest explanatory value with the target was selected as the final variable. As a result of the performance analysis at the LGBM variable level, 43 variables were excluded from the modeling stage because they were not significant. Explanation results with a target on a variable basis by applying the LGBM methodology are given in the table below.

Table 9 Top 10 Variable Gains

#	Variable Name	Gain
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	57%
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	12%
3	KKB_MAX_CC_LIM_AMT_TL_L3Y	4%
4	AVG_RETLOAN_PMNT_AMT_TL_L12M	3%
5	INDV_CC_AVG_PUR_AMT_TL_L6M	2%
6	AVG_MAIL_ORD_PUR_TRX_CNT_L12M	2%
7	INDV_CC_AVG_PUR_FX_AMT_TL_L12M	2%
8	OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	1%
9	KKB_MAX_2ND_CC_LIM_AMT_TL_L3Y	1%
10	AVG_RESTAURANT_PUR_AMT_TL_L12M	1%

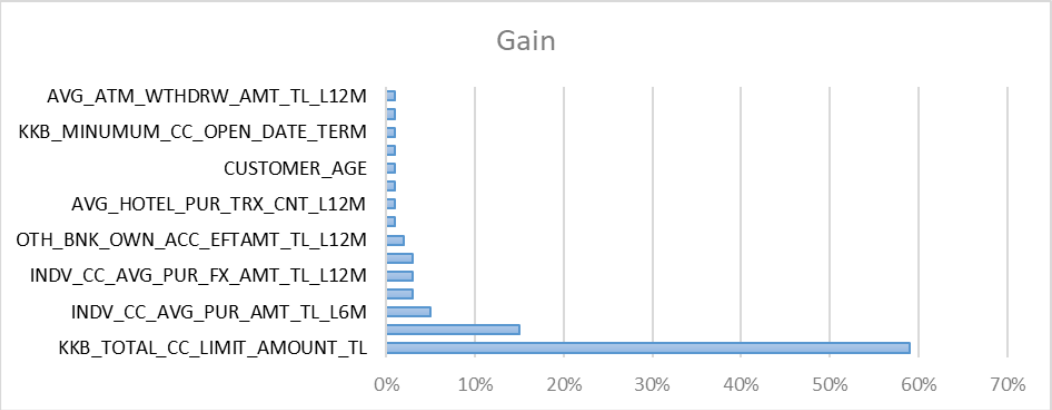
The first 10 variables with the highest Gain are specified in Table 8. These variables make up 85% of the total explanatory. In LGBM modeling, KKB_TOTAL_CC_LIMIT_AMOUNT_TL and AVG_DDEP_OUTFLW_AMT_TL_L12M variables appear to be the 2 variables with the highest explanatory power. Descriptions of 10 variables are shown in the table below.

Table 10 Description of Top 10 Variable

#	Variable Name	Variable Description
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	The customer's total credit card limit in the banking sector
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	The average outflow amount from the demand deposit balance in the last 12 months
3	KKB_MAX_CC_LIM_AMT_TL_L3Y	The highest credit card limit amount of the customer in the last 3 years
4	AVG_RETLOAN_PMNT_AMT_TL_L12M	Average amount of installments paid to YKB installment loans in the last 12 months
5	INDV_CC_AVG_PUR_AMT_TL_L6M	The average amount of all purchases made from personal credit cards in the last 6 months.
6	AVG_MAIL_ORD_PUR_TRX_CNT_L12M	Average number of mail orders in the last 12 months
7	INDV_CC_AVG_PUR_FX_AMT_TL_L12M	The average monthly amount of foreign currency expenditures made by personal credit cards in the last 12 months.
8	OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	EFT amount made to other banks in the last 12 months
9	KKB_MAX_2ND_CC_LIM_AMT_TL_L3Y	The second highest credit card limit amount of the customer in the last 3 years
10	AVG_RESTAURANT_PUR_AMT_TL_L12M	The average amount of expenses made in the "Restaurant" sector in the last 12 months.

As a result of the correlation analysis, 77 variables were eliminated, and 73 variables were used in the modeling stage. Unlike the XGB model, 29 different variables are included in the modeling phase. Gains of the most important variables are shown in the chart below.

Chart 2 Most Important Variables Gain



5.1.3 Feature Selection Process Random Forest

As mentioned in the XGBoost and LGBM steps, first of all, variables with instability problems between the model development sample and the validation sample were removed from the modeling phase. Since modeling and validation samples did not change at this stage, the variables that were eliminated are the same variables as XGB and LGBM. According to the PSI analysis results, 7 variables were eliminated due to instability problems.

The remaining variables were included in the correlation analysis, and among the variables that correlated according to the binary correlation results, the variable with the highest explanatory value with the target was selected as the final variable. As a result of the performance analysis at the Random Forest variable level, 44 variables were excluded from the modeling stage because they were not significant. Explanatory results with the target on a variable basis by applying Random Forest methodology are given in the table below.

Table 11 Top 10 Variable Importance Level

#	Variable Name	Mean Decrease in Accuracy
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	12.607
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	10.818
3	KKB_MAX_CC_LIM_AMT_TL_LY	7.736

4	KKB_MAX_CC_LIM_AMT_TL_L3Y	7.696
5	MZC_ALL_CSH_LOAN_RISK_AMT_TL	5.509
6	AVG_RETLOAN_PMNT_AMT_TL_L12M	5.065
7	KKB_MAX_OPEN_CR_RISK_AMT_TL	4.516
8	KKB_MINUMUM_CC_OPEN_DATE_TERM	4.48
9	AVG_ATM_WTHDRW_AMT_TL_L12M	4.388
10	KKB_TOTAL_OPEN_CR_RISK_AMT_TL	4.022

As in XGB and LGBM modeling, KKB_TOTAL_CC_LIMIT_AMOUNT_TL and AVG_DDEP_OUTFLW_AMT_TL_L12M variables appear to be the 2 variables with the highest explanatory power.

Table 12 Description of Top 10 Variables

#	Variable Name	Variable Description
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	The customer's total credit card limit in the banking sector
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	The average outflow amount from the demand deposit balance in the last 12 months
3	KKB_MAX_CC_LIM_AMT_TL_LY	The highest credit card limit amount of the customer in the last 1 year
4	KKB_MAX_CC_LIM_AMT_TL_L3Y	The highest credit card limit amount of the customer in the last 3 years
5	MZC_ALL_CSH_LOAN_RISK_AMT_TL	The customer's most up-to-date cash risk at all banks in Memzuç system
6	AVG_RETLOAN_PMNT_AMT_TL_L12M	Average amount of installments paid to YKB installment loans in the last 12 months
7	KKB_MAX_OPEN_CR_RISK_AMT_TL	The highest risk amount among risk amounts of all retail loans obtained from CRB query
8	KKB_MINUMUM_CC_OPEN_DATE_TERM	The age in months of the cards whose opening date is the oldest among the open / closed cards on the query date obtained from CRB query
9	AVG_ATM_WTHDRW_AMT_TL_L12M	It is the average of the amount of transactions withdrawn from the ATM in TL currency type within the last 12 months
10	KKB_TOTAL_OPEN_CR_RISK_AMT_TL	The total open risk amount in all open retail loan products of the customer obtained from CRB query

As a result of the correlation analysis, 88 variables were eliminated and the modeling phase was continued with 30 variables with the highest explanatory value. Summary statistics and gains of the relevant variables are shown in the tables below.

Table 13 Final Variables' Importance Level

#	Variable Name	Mean Decrease in Accuracy
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	15.935
2	AVG_ATM_WTHDRW_AMT_TL_L12M	11.195
3	MILAT_FLAG	11.149
4	AVG_DDEP_OUTFLW_AMT_TL_L12M	10.860
5	AVG_RETLOAN_PMNT_AMT_TL_L12M	8.376
6	AVG_TOT_CR_BAL_TL_L12M	6.812
7	TOT_CR_LIM_AMT_TL_L12M	6.425
8	KKB_MAX_CC_LIM_AMT_TL_LY	6.302
9	CUSTOMER_AGE	5.612
10	KKB_TOTAL_OPEN_CR_INST_AMT_TL	5.591
11	CUSTOMER_TENURE	5.526
12	NOF_ATM_WTHDRW_TL_TRX_L6M	5.181
13	NOF_ATM_WTHDRW_TL_TRX_L12M	5.091
14	MZC_AVG_CSH_LIM_AMT_TL_L12M	5.049
15	AVG_ASSN_CLUB_PUR_TRX_CNT_L12M	4.808
16	KKB_MAX_CC_LIM_AMT_TL_L3Y	4.703
17	OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	4.678
18	AVG_GPL_PMNT_AMT_TL_L12M	4.631
19	MAX_ATM_WTHDRW_AMT_TL_L12M	4.580
20	KKB_MINUMUM_CC_OPEN_DATE_TERM	4.547
21	NOF_BH_ATM_TRX_L12M	4.503
22	AVG_PERS_FIN_ASSET_L12M	4.402
23	INDV_CC_AVG_CSH_PNT_BAL_L12M	4.299
24	KKB_MAX_OPEN_CR_RISK_AMT_TL	4.254
25	KKB_MAX_CLS_CR_USG_AMT_TL_L3Y	4.174
26	WPS_L12M	4.092
27	MAX_CC_PMNT_AMT_TL_L12M	4.006
28	NOF_NBH_ATM_TRX_L12M	3.984
29	KKB_OPN_CR_RSK_AMT_TL_WHOUT_HL	3.954
30	KKB_MAX_2ND_CC_LIM_AMT_TL_L3Y	3.778

5.1.4 Feature Selection Process Linear Regression

The most widely used income estimation models in the banking sector are developed using linear regression. The main reasons for this are that the interpretability and explanation of the model are simpler, and the working speed of the model is faster than other modeling methods, which increases the desire to use this method. Linear regression methodology was run using the same data set and variables in order to compare the boosting and bagging algorithms mentioned above. After testing the basic assumptions of the regression, as mentioned in the next steps, the variables with instability problems were first eliminated. The 7 variables that were eliminated are variables that were also eliminated in XGBoost, LGBM, and Random Forest models.

Each variable was modeled as a single variable with the target variable. Then variables with a p-value greater than 0.05 were excluded from the modeling sample by looking at the explanation of the variables. At this stage, 24 variables were eliminated. Afterward, binary correlation analysis was performed between variables, and variables with high R ^ 2 values were selected as final variables. In the correlation phase, 100 variables were eliminated. The final remaining 88 variables were included in the regression stage, and at this stage, backward variable elimination was performed according to Akaike Information Criteria (AIC) values. The variables that will not affect the model performance and cause less than 0.01% change of the model AIC that is developed using 88 variables were removed in the modeling mass, and modeling study was carried out using 20 variables with the highest effect on the model. Summary statistical indicators of the final variables are given in the table below.

Table 14 Final Variable Summary Statistics

Coefficients	Estimate	Std. Error	t value
(Intercept)	1934.78	12.04	160.65
AVG_EDU_PUR_AMT_TL_L12M	0.73	0.03	27.52
AVG_RNT_PMNTAMTTL_WITH_DD_L12M	2.03	0.08	25.55
MAX_ATM_WTHDRW_FX_AMT_TL_L12M	0.43	0.01	28.78

TOTAL_SWIFT_CNT_L12M	1354.35	38.76	34.94
KKB_MAX_CLS_CR_USG_AMT_TL_L3Y	0.01	0	28.26
AVG_MAIL_ORD_PUR_TRX_CNT_L12M	563.82	22.08	25.54
AVG_DDEP_USD_PSTV_BAL_TL_L12M	0.03	0	30.93
AVG_INSUR_PUR_AMT_TL_L12M	1.12	0.03	34.42
INDV_CC_AVG_PUR_FX_AMT_TL_L12M	1.61	0.03	50.3
AVG_TDEP_USD_BAL_TL_L12M	-0.01	0	-32.96
MILAT_FLAG	-320.96	8.47	-37.87
KKB_TOTAL_OPEN_CC_COUNT	-302.71	4.89	-61.9
AVG_HOTEL_PUR_TRX_CNT_L12M	1711.87	34.61	49.46
AVG_DDEP_OUTFLW_AMT_TL_L12M	-0.02	0	-42.57
AVG_ATM_WTHDRW_AMT_TL_L12M	0.22	0	53.54
AVG_ASSN_CLUB_PUR_TRX_CNT_L12M	3444.8	60.84	56.62
AVG_PERS_FIN_ASSET_L12M	0.02	0	70.53
AVG_RETLOAN_PMNT_AMT_TL_L12M	0.46	0.01	77.87
OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	0.02	0	100.09
KKB_TOTAL_CC_LIMIT_AMOUNT_TL	0.15	0	207.28

5.1.5 Model Selection Process

In the modeling phase, Random Forest, XGBoost, Linear Regression, and LightGBM methodologies were used. The performance of the relevant models in model development and validation samples was examined. Then the final methodology to be used was decided. Within the scope of the performances of the relevant models, mean absolute error (MAE), mean percentage error (MAPE), and root mean square error (RMSE) values were examined. Detailed information on the performance of the model developed using the XGBoost, Linear Regression, LGBM, and Random Forest methodologies are shown below.

Table 15 Model Performance Comparison

		MAE	MAPE	RMSE
XGB	DEV	728	17%	1727
	TEST	756	17%	2039
	VAL	859	16%	2216
LGBM	DEV	769	18%	1913
	TEST	792	18%	2112
	VAL	911	18%	2346
Random Forest	DEV	597	14%	1608
	TEST	768	18%	2035

	VAL	858	17%	2207
Linear Regression	DEV	979	24%	2168
	TEST	984	24%	2246
	VAL	1156	24%	2569

While using each methodology and performing the modeling study, hyperparameters were determined by manual search. While proceeding to the final modeling stage, a grid search technique was applied for the selected champion model and final parameters were decided. As shown in Table 12, among the 4 methodologies, the model with the highest performance among the model development sample appears to be the Random Forest. However, a high level of deterioration was observed in the performance of the relevant methodology in test and validation samples. This situation is an indication that the model is doing overfitting. Due to this learning, the performance of the model decreases in the data not used in the model development phase. Also, as can be seen in the relevant table, the XGBoost model is the second model with the highest performance, and the model performance produced results consistent with the development and validation samples. The linear regression methodology, which is the traditional methodology, has shown the lowest performance compared to other methodologies.

After the selected XGB methodology, hyperparameters were tested in different combinations by applying grid search technique instead of manual hyperparameters determined by using expert decision to determine the final hypermaparameters. As a result of the study, the parameters that have lower MAE values and that will minimize the performance difference between the model development sample and the test sample have been determined. The performance of the model performance results developed using the final hyperparameters found using Grid search in model development, testing and validation samples are shown in the table below.

Table 16 Model Perfomance with Grid Search

	DEV	TEST	OOT
MAE	694.89	757.82	855.18
MAPE	17.2%	17.8%	17.0%
RMSE	1,513.92	1,943.48	2,075.44

As can be seen from the table above, grid search used in determining the hyperparameter contributed to the improvement of the model performance.

5.2 High School Education Level Model Development Scope

Model development work was executed for customers with a high school education level. In the modeling phase, the dataset is divided into train, test, and validation samples. Train and test samples have 80% and 20% weight, respectively. The model validation sample includes the period January 2020 and March 2020, which are not used in the model development phase.

Table 17 Datasets Distribution

Datasets	Count
Model Development Sample (%80)	636,581
Model Test Sample (%20)	159,145
Model Validation Sample	471,038

5.2.1 Feature Selection Process XGBoost

After the datamarts were formed, a total of 219 variables were created. These variables were then evaluated in the stability and correlation elimination. In the stability analysis, the stability of the variables between the model development population and the validation population for all variables was evaluated with the PSI test. Variables above 0.25 according to the PSI value were excluded from the model development sample. In total, 4 variables were eliminated due to instability problems. PSI values of 4 eliminated variables are shown in the table below.

Table 18 Variable PSI Information

#	Variable Name	PSI
1	LAST_KKB_PRODUCT_TENURE	0.941
2	AVG_OTH_FN_SRV_PUR_AMT_TL_L12M	0.664
3	MTV_AMT_TL_WITH_CC_L12M	0.538
4	KKB_APP_CNT_L6M	0.269

Explanatory variables of the variables that are correlated with each other were examined with the target, and then the variable that was less explanatory with the target was eliminated. In total, 41 variables were eliminated in the correlation step. The target explanatory (Gain) and correlations of the variables and the details of the eliminated variables are given in the tables below.

Table 19 Top 10 Variable Gains

#	Variable Name	Gain
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	66.60%
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	7.70%
3	AVG_RETLOAN_PMNT_AMT_TL_L12M	6.60%
4	MILAT_FLAG	1.90%
5	KKB_MINUMUM_CC_OPEN_DATE_TERM	1.20%
6	TOT_SWIFT_AMT_TL_L12M	1.20%
7	MAX_ATM_WTHDRW_FX_AMT_TL_L12M	1.00%
8	INDV_CC_AVG_PUR_AMT_TL_L12M	0.70%
9	AVG_CR_INST_AMT_TL_L12M	0.70%
10	OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	0.60%

As can be seen in Table 15, the variables with the first 10 gains constitute approximately 90% of the total weight.

41 variables were eliminated from the total correlation elimination, and the relationship strength of these variables to the target is low. Their total contribution to the model seems to be approximately 4%.

After the variable elimination, 85 variables were eliminated during the modeling phase because the level of significance was not sufficient according to the XGBoost methodology, and 77 variables remained to the model development step. Summary statistics details of these variables are given in the table below. Summary statistics details of these variables are given in the table below.

Table 20 Final Variables Summary Statistics

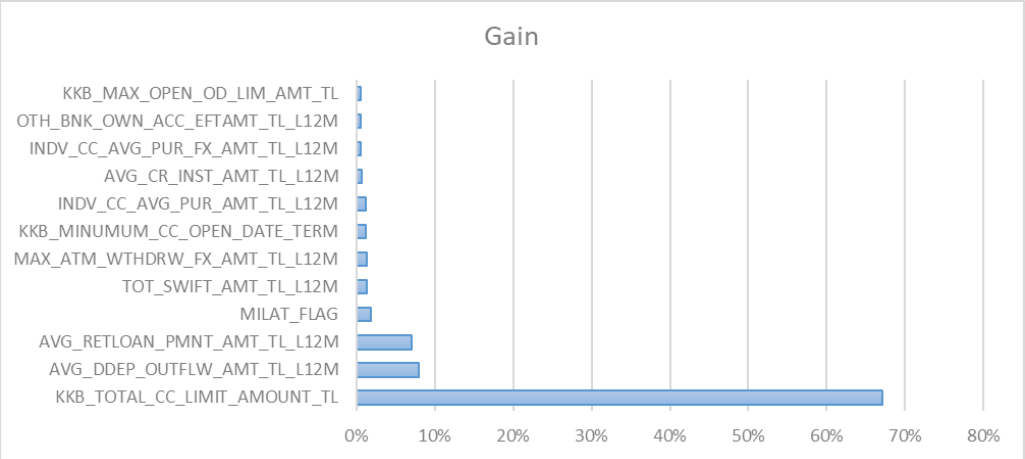
Variable	Mean	Std Dev	Minimum	Maximum	1st Pctl	99th Pctl
KKB_TOTAL_CC_LIMIT_AMOUNT_TL	8,683	7,416	0	245,250	0	33,040
AVG_DDEP_OUTFLW_AMT_TL_L12M	-6,190	9,082	-1,415,916	0	-34,279	-972

AVG_RETLOAN_PMNT_AMT_TL_L12M	813	856	0	45,027	0	3,498
MILAT_FLAG	0	0	0	1	0	1
TOT_SWIFT_AMT_TL_L12M	48	6,193	0	4,785,416	0	0
MAX_ATM_WTHDRW_FX_AMT_TL_L12M	21	350	0	13,760	0	0
KKB_MINUMUM_CC_OPEN_DATE_TERM	97	64	0	373	0	265
INDV_CC_AVG_PUR_AMT_TL_L12M	647	1,149	-28,734	56,658	0	4,990
AVG_CR_INST_AMT_TL_L12M	363	582	0	62,628	0	2,313
INDV_CC_AVG_PUR_FX_AMT_TL_L12M	11	146	-30,264	33,045	0	209
OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	7,657	22,457	0	2,910,255	0	82,770
KKB_MAX_OPEN_OD_LIM_AMT_TL	3,057	5,934	0	170,000	0	34,400
KKB_MAX_CLS_CR_USG_AMT_TL_L3Y	10,831	17,172	0	2,100,000	0	73,000
AVG_ATM_WTHDRW_AMT_TL_L12M	1,685	1,089	0	39,210	0	5,082
AREA_HOUSEHOLD_INCOME_AMT_TL	4,224	3,103	0	65,609	0	15,222
OTH_BNK_OTH_ACC_EFT_CNT_L12M	7	13	0	566	0	56
AVG_TPORT_PUR_AMT_TL_L12M	10	62	-252	8,263	0	194
AVG_PERS_FIN_ASSET_L12M	3,487	16,434	0	2,524,521	0	52,456
YKB_CC_TENURE	25	43	0	373	0	213
AVG_HOTEL_PUR_TRX_CNT_L12M	0	0	0	11	0	1
CUSTOMER_TENURE	95	77	3	465	4	288
MAX_OVERDUE_DAY_CNT_L12M	5	32	0	3,705	0	41
KKB_MAX_2ND_CC_LIM_AMT_TL_L3Y	2,177	2,699	0	85,000	0	11,200
AVG_INSUR_PMNT_TL_WITH_DD_L12M	30	367	0	76,000	0	480
AVG_TOT_CR_BAL_TL_L12M	9,550	21,241	0	1,925,825	0	94,121
KKB_MAX_OPEN_CR_INST_AMT_TL	597	773	0	77,920	0	2,804
AVG_EDU_PUR_AMT_TL_L12M	12	148	-1,251	22,000	0	269
MAX_ATM_WTHDRW_AMT_TL_L12M	1,758	787	0	7,000	0	4,000
KKB_TOTAL_OPEN_CREDIT_COUNT	4	3	0	38	0	13
KKB_MAX_OPEN_CREDIT_USG_AMT_TL	19,193	31,772	0	3,380,420	0	143,000
AVG_ASSN_CLUB_PUR_TRX_CNT_L12M	0	0	0	50	0	0
AVG_CR_PMNT_AMT_TL_L12M	835	2,786	0	350,444	0	10,685
AVG_ELC_BILLPMNTTL_WTH_DD_L12M	46	90	0	26,140	0	251
INDV_CC_VDAY_PUR_AMT_TL_L3Y	36	242	-5,010	25,000	0	658
NOF_BH_MOB_BANK_LOGIN_L12M	134	148	0	8,439	0	643
AVG_BOOK_MSC_PUR_TRX_CNT_L12M	0	0	0	107	0	1
KKB_TOTAL_CC_RISK_AMOUNT_TL	2,901	3,987	0	203,665	0	17,141
AVG_FAST_FOOD_PUR_AMT_TL_L12M	6	21	0	3,280	0	85
AVG_BILL_PMNT_TL_WITH_DD_L12M	79	249	0	104,310	0	480
KKB_TOTAL_OPEN_OD_LIM_AMT_TL	3,921	6,840	0	170,000	0	38,100
NOF_ATM_WTHDRW_TL_TRX_L6M	26	26	0	1,002	0	110
AVG_TBC_LIQR_PUR_AMT_TL_L12M	5	60	-78	15,477	0	93
AVG_TLCOM_PMNT_TL_WITH_CC_L12M	7	28	0	3,407	0	102
AVG_TDEP_EUR_BAL_TL_L12M	80	4,856	0	1,870,861	0	0

AVG_FURNTR_PUR_AMT_TL_L12M	29	203	-993	29,254	0	604
AVG_CC_CADV_BAL_TL_L12M	72	339	0	25,680	0	1,400
NOF_NBH_ATM_TRX_L12M	72	80	0	5,687	0	361
AVG_PHARMACY_PUR_AMT_TL_L12M	7	25	0	4,669	0	95
KKB_TOTAL_OPEN_PERS_LOAN_CNT	1	2	0	37	0	8
AVG_DDEP_USD_PSTV_BAL_TL_L12M	427	5,131	0	1,120,038	0	11,257
KKB_MAX_CLS_CR_INST_AMT_TL_L3Y	818	2,763	0	257,385	0	12,073
NOF_BH_BRANCH_VISIT_L12M	2	4	0	227	0	16
TOT_EDU_PMNT_TL_L12M	2,107	5,874	0	1,755,144	0	17,690
AVG_MAIL_ORD_PUR_TRX_CNT_L12M	0	0	0	22	0	1
NOF_BH_ATM_TRX_L12M	109	100	0	5,858	0	460
AVG_UTIL_BILL_PUR_AMT_TL_L12M	21	123	-4,400	13,868	0	333
AVG_TDEP_BAL_TL_L12M	2,044	13,900	0	1,326,750	0	52,816
CUSTOMER_AGE	34	9	18	92	20	57
AVG_ACSORY_PUR_AMT_TL_L12M	3	30	-92	10,000	0	54
AVG_WHT_APPLNC_PUR_AMT_TL_L12M	68	269	-1,598	14,662	0	1,026
MZC_ALL_NONCSH_LOAN_RSK_AMT_TL	6	537	0	150,000	0	0
AVG_MTV_PMNT_TL_WITH_DDEP_L12M	34	107	0	6,086	0	562
AVG_INSUR_INST_CNT_L12M	0	1	0	147	0	3
OTH_BNK_OTH_ACC_EFTAMT_TL_L12M	9,156	33,480	0	5,658,492	0	116,001
AVG_BILL_PMNT_CNT_WITH_CC_L12M	4	11	0	289	0	60
WPS_L12M	-42	6,537	-999,999	2	0	2
MZC_AVG_CSH_LIM_AMT_TL_L12M	25,676	32,661	0	1,155,643	0	153,950
AVG_SHPNG_CNTR_PUR_AMT_TL_L12M	10	56	-615	12,600	0	167
AVG_BILL_PMNT_CNT_WITH_DD_L12M	18	23	0	622	0	93
AVG_SHOE_BG_LTHR_INST_CNT_L12M	0	1	0	25	0	3
KKB_TOTAL_OPEN_CC_COUNT	2	1	0	10	0	4
NOF_BH_INTRNT_BANK_LOGIN_L12M	5	24	0	4,188	0	99
KKB_OPEN_TOT_PMT_AMT_LM	2,300	4,264	0	622,896	2	15,080
AVG_HEALTH_SRV_PUR_AMT_TL_L12M	6	45	-40	9,500	0	125
AVG_HGS_PMNT_TL_WITH_CC_L12M	1	6	0	500	0	26
AVG_DON_PMNT_TL_WITH_DDEP_L12M	1	24	0	5,000	0	4
AVG_SOUVENIR_PUR_AMT_TL_L12M	2	14	-9	4,380	0	33

The most important variables with the highest gain among 77 variables used in the final modeling phase are given in the table below. These variables make up 91% of the total gain.

Chart 3 Most Important Variables' Gain



5.2.2 Feature Selection Process LightGBM

As mentioned in the XGBoost step, first of all, variables with instability problems between the model development sample and the validation sample were removed from the modeling phase. Since modeling and validation samples did not change at this stage, the variables that were eliminated are the same variables as XGB. According to the PSI analysis results, 4 variables were eliminated due to instability problems.

The remaining variables were included in the correlation analysis and among the variables that correlated according to the binary correlation results, the variable with the highest explanatory value with the target was selected as the final variable. As a result of the performance analysis at the LGBM variable level, 30 variables were excluded from the modeling stage because they were not significant. Explanation results with a target on a variable basis by applying the LGBM methodology are given in the table below.

Table 21 Top 10 Variable Gains

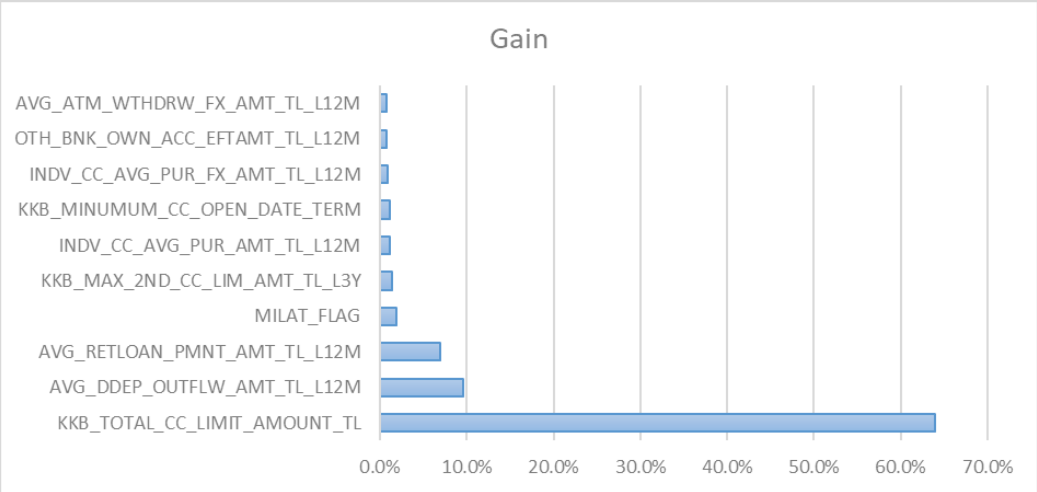
#	Variable Name	Gain
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	63.90%
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	9.50%
3	AVG_RETLOAN_PMNT_AMT_TL_L12M	6.90%
4	MILAT_FLAG	1.80%

5	KKB_MAX_2ND_CC_LIM_AMT_TL_L3Y	1.40%
6	INDV_CC_AVG_PUR_AMT_TL_L12M	1.20%
7	KKB_MINUMUM_CC_OPEN_DATE_TERM	1.20%
8	INDV_CC_AVG_PUR_FX_AMT_TL_L12M	0.90%
9	OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	0.80%
10	AVG_ATM_WTHDRW_FX_AMT_TL_L12M	0.80%

The first 10 variables with the highest Gain are specified in Table 17. These variables make up 88% of the total explanatory. In LGBM modeling, KKB_TOTAL_CC_LIMIT_AMOUNT_TL and AVG_DDEP_OUTFLW_AMT_TL_L12M variables appear to be the 2 variables with the highest explanatory power.

As a result of the correlation analysis, 48 variables were eliminated, and 116 variables were used in the modeling stage. Unlike the XGB model, 52 different variables are included in the modeling phase. Gains of the most important variables are shown in the chart below.

Chart 4 Final Variables Gain



5.2.3 Feature Selection Process Random Forest

As mentioned in the XGBoost and LGBM steps, first of all, variables with instability problems between the model development sample and the validation

sample were removed from the modeling phase. Since modeling and validation samples did not change at this stage, the variables that were eliminated are the same variables as XGB and LGBM. According to the PSI analysis results, 4 variables were eliminated due to instability problems.

As a result of the performance analysis at the Random Forest variable level, 51 variables were excluded from the modeling stage because they were not significant. Explanatory results with a target on the variable basis by applying Random Forest methodology are given in the table below.

Table 22 Top 10 Variable Importance Level

#	Variable Name	Importance
1	MILAT_FLAG	13.254
2	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	11.062
3	AVG_DDEP_INFLW_AMT_TL_L12M	9.633
4	AVG_TOT_CR_BAL_TL_L12M	9.29
5	KKB_MINUMUM_CC_OPEN_DATE_TERM	8.509
6	AVG_ATM_WTHDRW_AMT_TL_L12M	7.474
7	AVG_RETLOAN_PMNT_AMT_TL_L12M	7.24
8	AVG_DDEP_OUTFLW_AMT_TL_L12M	7.222
9	CUSTOMER_TENURE	6.411
10	NOF_ATM_WTHDRW_TL_TRX_L6M	6.312

KKB_TOTAL_CC_LIMIT_AMOUNT_TL and MILAT_FLAG variables appear to be the 2 variables with the highest explanatory power.

As a result of the correlation analysis, 81 variables were eliminated, and the modeling phase was continued with 30 variables with the highest explanatory value. The Mean Decrease Accuracy of the relevant variables is shown in the tables below.

Table 23 Final Variables' Importance Level

#	Variable Name	Importance
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	33.23
2	AVG_RETLOAN_PMNT_AMT_TL_L12M	20.04
3	MILAT_FLAG	19.72
4	AVG_DDEP_INFLW_AMT_TL_L12M	18.43

5	NOF_BH_ATM_TRX_L12M	15.25
6	AVG_ATM_WTHDRW_AMT_TL_L12M	15.18
7	CUSTOMER_AGE	14.96
8	YKB_CC_TENURE	13.26
9	NOF_NBH_ATM_TRX_L12M	12.18
10	CUSTOMER_TENURE	11.95
11	MAX_CC_PMNT_AMT_TL_L12M	11.69
12	NOF_ATM_WTHDRW_TL_TRX_L6M	11.59
13	AVG_DDEP_OUTFLW_AMT_TL_L12M	11.02
14	TOT_CR_LIM_AMT_TL_L12M	10.95
15	KKB_MAX_OPEN_CREDIT_USG_AMT_TL	10.69
16	MAX_OVERDUE_DAY_CNT_L12M	10.51
17	AVG_TOT_CR_BAL_TL_L12M	10.26
18	KKB_MAX_OPEN_CR_INST_AMT_TL	10
19	KKB_MINUMUM_CC_OPEN_DATE_TERM	9.59
20	AVG_CR_PMNT_AMT_TL_L12M	8.04
21	INDV_CC_VDAY_PUR_AMT_TL_L3Y	7.79
22	OTH_BNK_OTH_ACC_EFTAMT_TL_L12M	7.73
23	KKB_MAX_OPEN_OD_LIM_AMT_TL	6.71
24	MAX_MOB_PHN_PMNTTL_WTH_DD_L12M	6.59
25	KKB_TOT_OPEN_CR_CNT_L6M	6.29
26	NOF_NBH_MOB_BANK_LOGIN_L12M	6.21
27	AVG_MTV_PMNT_TL_WITH_DDEP_L12M	5.59
28	KKB_TOTAL_OPEN_OD_LIM_AMT_TL	5.16
29	NOF_ATM_WTHDRW_FX_TRX_L12M	4.01
30	AVG_SPRT_ENT_PUR_AMT_TL_L12M	3.13

5.2.4 Feature Selection Process Linear Regression

Linear regression methodology was run using the same data set and variables to compare the boosting and bagging algorithms mentioned above. After testing the basic assumptions of the regression, as mentioned in the next steps, the variables with instability problem were first eliminated. The 4 variables that were eliminated are variables that were also eliminated in XGBoost, LGBM, and Random Forest models.

Each variable was modeled as a single variable with the target variable. Then variables with a p-value greater than 0.05 were excluded from the modeling sample

by looking at the explanation of the variables. At this stage, 15 variables were eliminated. Afterward, binary correlation analysis was performed between variables, and variables with high R ^ 2 values were selected as final variables. In the correlation phase, 84 variables were eliminated. The final remaining 116 variables were included in the regression stage, and at this stage, backward variable elimination was performed according to Akaike Information Criteria (AIC) values. The variables that will not affect the model performance and cause less than 0.01% change of the model AIC that is developed using 116 variables, were removed in the modeling mass, and modeling study was carried out using 25 variables with the highest effect on the model. Summary statistical indicators of the final variables are given in the table below.

Table 24 Final Variable Summary Statistics

Coefficients:	Estimate	Std. Error	t value
(Intercept)	2296.46	7.36	311.89
NOF_NBH_ATM_TRX_L12M	-0.47	0.03	-13.7
KKB_TOTAL_OPEN_OD_LIM_AMT_TL	0.05	0	27.44
KKB_MAX_OPEN_OD_LIM_AMT_TL	-0.06	0	-27.71
AVG_HOTEL_PUR_TRX_CNT_L12M	596.42	16.34	36.5
KKB_MAX_OPEN_CREDIT_USG_AMT_TL	-0.01	0	-27.01
AVG_EDU_PUR_AMT_TL_L12M	0.6	0.02	35.67
NOF_BH_ATM_TRX_L12M	-0.77	0.03	-27.69
KKB_TOTAL_CC_RISK_AMOUNT_TL	0.03	0	36.31
KKB_MAX_2ND_CC_LIM_AMT_TL_L3Y	0.04	0	34.55
KKB_TOTAL_OPEN_CREDIT_COUNT	-55.62	1.49	-37.31
KKB_TOT_OPEN_HLOAN_USG_AMT_TL	0.01	0	31.58
INDV_CC_AVG_PUR_FX_AMT_TL_L12M	0.91	0.02	52.28
INDV_CC_AVG_PUR_AMT_TL_L12M	0.12	0	43.3
AVG_PERS_FIN_ASSET_L12M	0.01	0	72.09
AVG_DDEP_INFLW_AMT_TL_L12M	0.02	0	55.81
KKB_MAX_CLS_CR_USG_AMT_TL_L3Y	0.01	0	46.61
KKB_TOTAL_OPEN_CC_COUNT	-189.37	3.52	-53.87
MILAT_FLAG	-342.26	4.91	-69.76
NOF_ATM_WTHDRW_FX_TRX_L12M	216.56	3.78	57.26
OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	0.01	0	55.21
OTH_BNK_OTH_ACC_EFT_CNT_L12M	14.89	0.21	72.29
TOT_SWIFT_AMT_TL_L12M	0.07	0	88.22

AVG_ATM_WTHDRW_AMT_TL_L12M	0.22	0	90.58
AVG_RETLOAN_PMNT_AMT_TL_L12M	0.55	0	119.98
KKB_TOTAL_CC_LIMIT_AMOUNT_TL	0.12	0	247.52

5.2.5 Model Selection Process

Within the scope of the performances of the relevant models, mean absolute error (MAE), mean percentage error (MAPE), and root mean square error (RMSE) values were examined. Detailed information on the performance of the model developed using the XGBoost, Linear Regression, LGBM, and Random Forest methodologies are shown below.

Table 25 Model Performance Comparison

		MAE	MAPE	RMSE
XGB	DEV	749	17%	1631
	TEST	759	17%	1699
	VAL	825	16%	1849
LGBM	DEV	743	17%	1666
	TEST	753	17%	1677
	VAL	821	17%	1843
Random Forest	DEV	313	7%	749
	TEST	748	17%	1700
	VAL	800	16%	1782
Linear Regression	DEV	939	22%	1891
	TEST	940	22%	2045
	VAL	1033	22%	2170

While using each methodology and performing the modeling study, hyperparameters were determined by manual search. While proceeding to the final modeling stage, a grid search technique was applied for the selected champion model, and final parameters were decided. As in the Unknown & Elementary segment, the model with the highest performance in the model development sample appears to be the Random Forest. However, the performance of the Random Forest model in validation and test samples significantly decreases. Although the cross-validation methodology was used when determining the relevant methodology parameters, it was observed that the overfitting problem could not be solved. XGB

and LGBM models performed very close to each other. Performance stability of both models continues in validation and test samples. The linear regression method shows the lowest performance in parallel with the previous segment.

After the selected XGB methodology, hyperparameters were tested in different combinations by applying grid search technique instead of manual hyperparameters determined by using expert decision to determine the final hyperparameters. As a result of the study, parameters that have lower MAE values as in the previous segment and that will minimize the performance difference between the model development sample and the test sample have been determined. The performance of the model performance results developed using the final hyperparameters found using Grid search in model development, testing, and validation samples are shown in the table below.

Table 26 Model Performance with Grid Search

	DEV	TEST	VAL
MAE	717.38	744.07	811.39
MAPE	17%	17%	17%
RMSE	1,483.73	1,643.35	1,780.19

It has been observed that the performance of the model results developed by using Grid search has a better performance than other model results.

5.3 Bachelor & Master Education Level Model Development Scope

Model development work was executed for customers with bachelor&master education levels. In the modeling phase, the dataset is divided into train, test, and validation samples. Train and test samples have 80% and 20% weight, respectively. The model validation sample includes the period January 2020 and March 2020, which are not used in the model development phase.

Table 27 Datasets Distribution

Datasets	Count
Model Development Sample (%80)	537,551
Model Test Sample (%20)	134,388
Model Validation Sample	460,720

5.3.1 Feature Selection Process XGBoost

After the datamarts were formed, a total of 219 variables were created. These variables were then evaluated in the stability and correlation elimination. In the stability analysis, the stability of the variables between the model development population and the validation population for all variables was evaluated with the PSI test. Variables above 0.25 according to the PSI value were excluded from the model development sample. In total, 2 variables were eliminated due to instability problems. PSI values of 4 eliminated variables are shown in the table below.

Table 28 Variable PSI Information

#	Variable Name	PSI
1	LAST_KKB_PRODUCT_TENURE	0.941
2	AVG_OTH_FN_SRV_PUR_AMT_TL_L12M	0.664

Explanatory variables of the variables that are correlated with each other were examined with the target, and then the variable that was less explanatory with the target was eliminated. In total, 89 variables were eliminated in the correlation step. The target explanatory (Gain) and correlations of the variables and the details of the eliminated variables are given in the tables below.

Table 29 Top 10 Variable Gains

#	Variable Name	Gain
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	66.60%
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	7.70%
3	AVG_RETLOAN_PMNT_AMT_TL_L12M	6.60%
4	MILAT_FLAG	1.90%

5	KKB_MINUMUM_CC_OPEN_DATE_TERM	1.20%
6	TOT_SWIFT_AMT_TL_L12M	1.20%
7	MAX_ATM_WTHDRW_FX_AMT_TL_L12M	1.00%
8	INDV_CC_AVG_PUR_AMT_TL_L12M	0.70%
9	AVG_CR_INST_AMT_TL_L12M	0.70%
10	OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	0.60%

As shown in Table 24, the variables with the first 10 gains constitute approximately 90% of the total weight.

41 variables were eliminated from the total correlation elimination, and the relationship strength of these variables to the target is low. Their total contribution to the model seems to be approximately 4%. 124 variables remained in the model development step. Summary statistics details of these variables are given in the table below. Summary statistics details of these variables are given in the table below.

Table 30 Final Variables Summary Statistics

Variable Name	Mean	Std Dev	Minimum	Maximum	1st Pctl	99th Pctl
KKB_TOTAL_CC_LIMIT_AMOUNT_TL	17,186	16,488	0	311,443	0	83,700
AVG_DDEP_OUTFLW_AMT_TL_L12M	-10,954	19,674	-1,810,229	0	-84,011	-854
AVG_MAIL_ORD_PUR_TRX_CNT_L12M	0	0	0	86	0	1
AVG_RETLOAN_PMNT_AMT_TL_L12M	1,226	1,727	0	57,327	0	7,492
OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	23,819	56,244	0	3,361,799	0	245,491
CUSTOMER_AGE	35	8	18	94	22	59
INDV_CC_AVG_PUR_AMT_TL_L12M	1,581	2,791	-143,748	94,470	0	13,237
AVG_PERS_FIN_ASSET_L12M	13,295	64,596	0	9,262,543	0	222,790
NOF_ATM_WTHDRW_FX_TRX_L12M	0	1	0	127	0	6
AVG_ASSN_CLUB_PUR_TRX_CNT_L12M	0	1	0	380	0	1
INDV_CC_AVG_PUR_FX_AMT_TL_L12M	100	645	-150,014	68,985	0	2,176
AVG_TBC_LIQR_PUR_AMT_TL_L12M	11	76	-5,716	16,371	0	228
YKB_CC_TENURE	32	51	0	369	0	251
NOF_BH_INTRNT_BANK_LOGIN_L12M	13	35	0	2,187	0	156

OTH_BNK_OTH_ACC_EFTAMT_TL_L12M	24,173	65,159	0	########	0	247,850
AREA_HOUSEHOLD_INCOME_AMT_TL	5,630	4,777	0	65,609	0	23,063
TOT_SWIFT_AMT_TL_L12M	611	28,127	0	9,843,338	0	0
KKB_MAX_2ND_CC_LIM_AMT_TL_L3Y	4,179	5,113	0	150,000	0	24,150
AVG_ATM_WTHDRW_AMT_TL_L12M	1,530	1,307	0	36,778	0	5,875
KKB_MINUMUM_CC_OPEN_DATE_TERM	120	67	0	755	3	290
AVG_TDEP_BAL_TL_L12M	6,097	35,670	0	4,375,500	0	134,450
NOF_BH_MOB_BANK_LOGIN_L12M	140	158	0	10,040	0	687
MILAT_FLAG	0	0	0	1	0	1
AVG_MTV_PMNT_TL_WITH_DDEP_L12M	48	145	0	5,578	0	660
TOT_EDU_PMNT_TL_L12M	2,942	11,748	0	2,035,253	0	37,417
AVG_DDEP_EUR_PSTV_BAL_TL_L12M	923	8,293	0	1,236,720	0	20,739
AVG_CR_INST_AMT_TL_L12M	369	921	0	51,934	0	3,722
KKB_MAX_OPEN_OD_LIM_AMT_TL	3,670	6,209	0	147,525	0	35,800
KKB_MAX_CLS_CR_USG_AMT_TL_L3Y	17,217	34,711	0	3,270,050	0	142,500
AVG_INSUR_PUR_AMT_TL_L12M	60	263	-2,987	43,371	0	850
MAX_ATM_WTHDRW_AMT_TL_L12M	1,623	866	0	9,000	0	4,000
MZC_AVG_CSH_LIM_AMT_TL_L12M	43,134	64,413	0	7,700,060	0	283,079
AVG_CLOTHING_INST_CNT_L12M	2	3	0	168	0	14
AVG_TDEP_OUTFLW_AMT_TL_L12M	-3,646	22,121	-3,405,267	0	-80,712	0
AVG_TPORT_PUR_AMT_TL_L12M	52	187	-499	21,652	0	759
AVG_AUTO_INDST_PUR_AMT_TL_L12M	30	166	-2,964	26,000	0	481
AVG_DDEP_USD_PSTV_BAL_TL_L12M	2,091	14,558	0	2,240,334	0	45,510
CUSTOMER_TENURE	102	85	3	473	4	303
INDV_CC_VDAY_PUR_AMT_TL_L3Y	85	483	-20,884	47,362	0	1,427
INDV_CC_NEWYEAR_PUR_AMT_TL_L3Y	130	625	-20,000	76,577	0	2,015
NOF_NBH_ATM_TRX_L12M	50	59	0	2,271	0	262
AVG_ELC_BILLPMNTTL_WTH_DD_L12M	49	136	0	47,874	0	271
AREA_UNIVERSITY_EDUCATION_PCT	0	0	0	1	0	1
INDV_CC_AVG_CSH_PNT_EXP_L12M	4	15	-82	2,273	0	60
AVG_TDEP_USD_BAL_TL_L12M	3,022	41,398	0	9,064,019	0	78,060
AREA_HIGH_SCHOOL_EDUCATION_PCT	0	0	0	2	0	1
KKB_TOTAL_CC_RISK_AMOUNT_TL	5,278	7,657	0	241,910	0	36,335

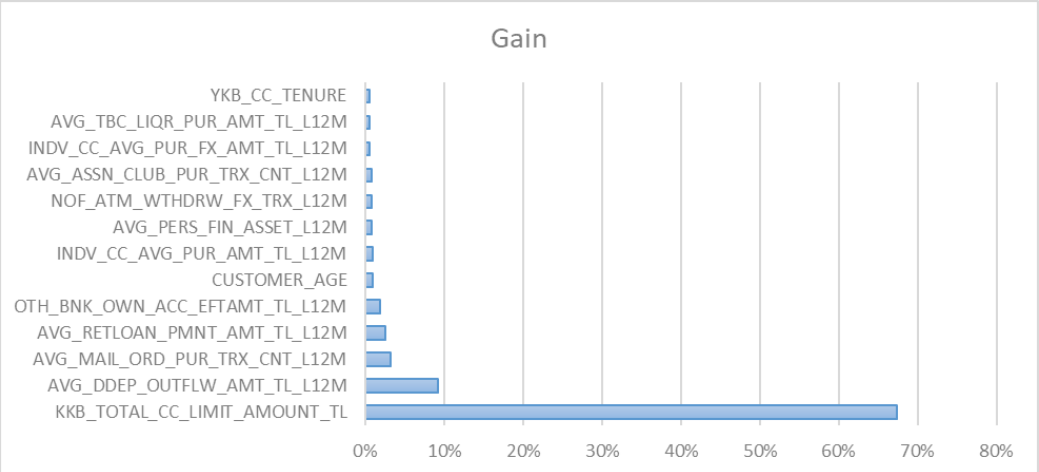
KKB_APP_CNT_L6M	2	4	0	252	0	19
AVG_HOTEL_PUR_AMT_TL_L12M	54	322	-750	62,311	0	1,042
AVG_FURNTR_PUR_AMT_TL_L12M	72	339	-4,920	45,500	0	1,256
AVG_HGS_PMNT_TL_WITH_DDEP_L12M	1	9	0	500	0	40
AVG_HEALTH_SRV_PUR_AMT_TL_L12M	20	108	-343	18,383	0	333
NOF_BH_BRANCH_VISIT_L12M	2	3	0	243	0	14
AVG_MOB_PHN_PMNTTL_WTH_DD_L12M	46	76	0	7,341	0	330
AVG_INSUR_PMNT_TL_WITH_DD_L12M	37	1,010	0	250,000	0	480
MAX_OVERDUE_AMT_TL_L12M	306	1,143	0	373,137	0	4,345
NOF_BH_ATM_TRX_L12M	92	87	0	4,029	0	394
AVG_CC_CADV_BAL_TL_L12M	48	430	0	77,484	0	1,079
AVG_RNT_PMNTAMTTL_WITH_DD_L12M	10	160	0	8,800	0	0
AVG_ACSORY_PUR_AMT_TL_L12M	8	34	-48	9,300	0	107
KKB_TOT_OPEN_HLOAN_USG_AMT_TL	20,433	64,611	0	2,630,000	0	280,000
AVG_TDEP_EUR_BAL_TL_L12M	824	15,854	0	3,492,042	0	11,806
AVG_AUTO_PMNT_CNT_WITH_DD_L12M	13	22	0	584	0	90
AVG_UTIL_BILL_PUR_AMT_TL_L12M	28	191	-980	25,979	0	467
AVG_BILL_PMNT_TL_WITH_CC_L12M	29	125	0	45,157	0	285
MTV_AMT_TL_WITH_CC_L12M	53	162	0	6,547	0	750
AVG_TELECOM_PMNTTL_WTH_DD_L12M	23	44	0	9,185	0	128
KKB_MAX_OPEN_OD_RISK_AMT_TL	733	1,877	0	144,600	0	6,728
AVG_SPRT_ENT_PUR_AMT_TL_L12M	16	95	-2,295	19,053	0	281
AVG_EDU_PUR_AMT_TL_L12M	64	470	-4,063	85,275	0	1,697
NOF_ATM_WTHDRW_TL_TRX_L6M	23	20	0	893	0	85
AVG_SHPNG_CNTR_PUR_AMT_TL_L12M	28	107	-578	19,000	0	373
AVG_GAS_PMNT_TL_WITH_DDEP_L12M	39	101	0	19,174	0	325
AVG_TRAVEL_PUR_AMT_TL_L12M	52	248	-3,283	58,280	0	929
AVG_BOOK_MSC_PUR_TRX_CNT_L12M	0	1	0	110	0	3
AVG_CAR_RNTAL_PUR_AMT_TL_L12M	9	82	-608	12,223	0	218
KKB_OPEN_TOT_PMT_AMT_LM	3,377	7,094	0	905,375	3	24,522
KKB_TOTAL_OPEN_CREDIT_COUNT	5	3	0	43	1	13
AVG_CAFE_PUR_AMT_TL_L12M	4	37	-20	9,620	0	80
AVG_OTHER_SRV_INST_CNT_L12M	0	1	0	58	0	3
AVG_WATER_PMNT_TL_WITH_DD_L12M	18	52	0	28,000	0	131

AVG_OFFC_SPLY_PUR_AMT_TL_L12M	5	40	-64	9,632	0	98
AVG_OTH_BILLPMNTTL_WTH_DD_L12M	53	998	0	550,000	0	1,011
AVG_TOY_SHOP_PUR_AMT_TL_L12M	6	33	-35	5,765	0	113
AVG_TPORT_INST_CNT_L12M	0	1	0	125	0	4
AVG_COSMETIC_PUR_AMT_TL_L12M	11	48	-275	8,383	0	171
AVG_COMNC_PUR_AMT_TL_L12M	39	127	-114	15,160	0	534
AVG_DDEP_GBP_PSTV_BAL_TL_L12M	64	2,171	0	433,392	0	0
AVG_HSPTL_HLTH_PUR_AMT_TL_L12M	32	156	-644	17,890	0	577
KKB_MAX_CLSD_OD_LIM_AMT_TL_L3Y	280	1,701	0	363,000	0	4,650
KKB_TOT_OPEN_CR_CNT_L6M	1	1	0	44	0	6
AVG_SOUVENIR_PUR_AMT_TL_L12M	4	22	-17	6,180	0	58
NOF_WKD_INTRNT_BANK_LOGIN_L12M	2	6	0	397	0	27
AVG_HGS_PMNT_TL_WITH_CC_L12M	2	11	0	500	0	50
AVG_JEWELLERY_PUR_AMT_TL_L12M	25	257	-444	74,041	0	458
AVG_TECH_PUR_AMT_TL_L12M	8	96	-720	31,500	0	156
AVG_DON_PMNT_TL_WITH_DDEP_L12M	5	73	0	10,000	0	100
AVG_GAS_PMNT_TL_WITH_CC_L12M	14	88	0	45,157	0	270
KKB_TOTAL_OPEN_CC_COUNT	2	1	0	15	0	5
AVG_CAR_WASH_PUR_AMT_TL_L12M	0	7	0	1,400	0	10
AVG_GLASSWARE_PUR_AMT_TL_L12M	7	61	-391	20,000	0	123
MEMZUC_ALL_NON_CASH_LIM_AMT_TL	110	11,903	0	3,900,000	0	0
AVG_SUPERMKT_INST_CNT_L12M	0	1	0	71	0	2
AVG_TLCOM_PMNT_TL_WITH_CC_L12M	8	27	0	2,754	0	105
AVG_CAR_TIRE_PUR_AMT_TL_L12M	7	78	-100	30,625	0	183
AVG_CONST_INST_CNT_L12M	0	0	0	25	0	1
AVG_OTH_FN_SRV_INST_CNT_L12M	0	0	0	27	0	1
OWN_FUND_FLG_L12M	0	0	0	1	0	1
AVG_AUTO_INDST_INST_CNT_L12M	0	0	0	24	0	1
AVG_JEWELLERY_INST_CNT_L12M	0	0	0	16	0	1
AVG_HOUSE_APPLNC_INST_CNT_L12M	0	0	0	20	0	0
OWN_STOCK_FLG_L12M	0	0	0	1	0	1
OWN_BES_INSUR_FLG_L12M	0	0	0	1	0	1
WPS_L12M	-14	3,858	-999,999	2	0	2

AVG_DDEP_CHF_PSTV_BAL_TL_L12M	13	1,134	0	285,981	0	0
MZC_ALL_NONCSH_LOAN_RSK_AMT_TL	33	5,406	0	2,035,000	0	0
AVG_CAR_RNTAL_INST_CNT_L12M	0	0	0	19	0	0
AVG_TDEP_GBP_BAL_TL_L12M	61	13,080	0	7,956,988	0	0
KKB_DIFF	1	1	0	11	0	2
OWN_DASK_INSUR_FLG_L12M	0	0	0	1	0	1
AVG_COPY_PRINT_INST_CNT_L12M	0	0	0	56	0	0
OWN_BOND_FLG_L12M	0	0	0	1	0	0
AVG_TDEP_CHF_BAL_TL_L12M	6	1,014	0	326,171	0	0
AVG_TBC_LIQR_SHP_INST_CNT_L12M	0	0	0	2	0	0

The most important variables with the highest gain among 124 variables used in the final modeling phase are given in the table below. These variables make up 90% of the total gain.

Chart 5 Most Important Variables' Gain



5.3.2 Feature Selection Process LightGBM

As mentioned in the XGBoost step, first of all, variables with instability problems between the model development sample and the validation sample were removed from the modeling phase. Since modeling and validation samples did not change at this stage, the variables that were eliminated are the same variables as XGB.

According to the PSI analysis results, 2 variables were eliminated due to instability problems.

The remaining variables were included in the correlation analysis. Among the variables that correlated according to the binary correlation results, the variable with the highest explanatory value with the target was selected as the final variable. As a result of the performance analysis at the LGBM variable level, 22 variables were excluded from the modeling stage because they were not significant. Explanation results with the target on a variable basis by applying the LGBM methodology are given in the table below.

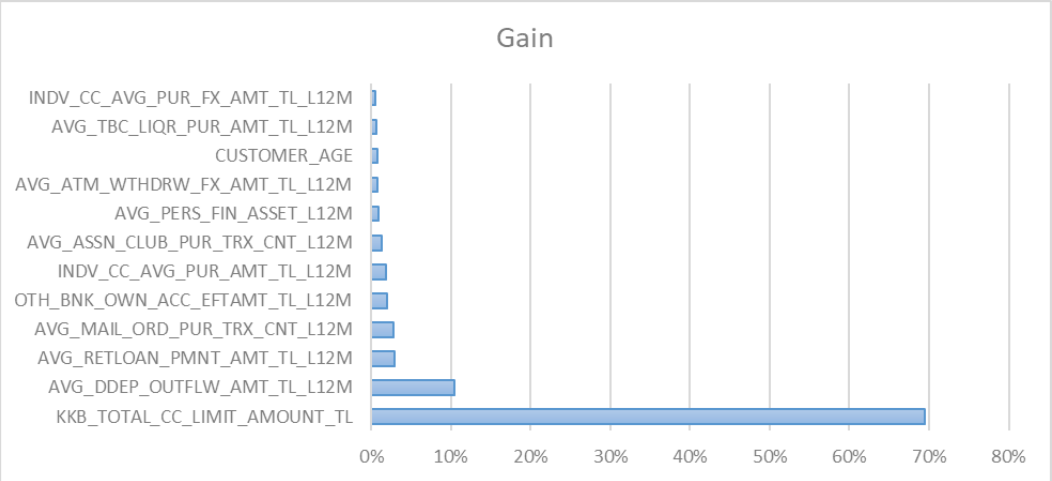
Table 31 Top 10 Variable Gains

#	Variable Name	Gain
1	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	68%
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	11%
3	AVG_MAIL_ORD_PUR_TRX_CNT_L12M	3%
4	AVG_RETLOAN_PMNT_AMT_TL_L12M	3%
5	OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	2%
6	INDV_CC_AVG_PUR_AMT_TL_L12M	1%
7	AVG_ASSN_CLUB_PUR_TRX_CNT_L12M	1%
8	CUSTOMER_AGE	1%
9	AVG_PERS_FIN_ASSET_L12M	1%
10	AVG_ATM_WTHDRW_FX_AMT_TL_L12M	1%

The first 10 variables with the highest Gain are specified in Table 26. These variables make up 92% of the total explanatory. In LGBM modeling, KKB_TOTAL_CC_LIMIT_AMOUNT_TL and AVG_DDEP_OUTFLW_AMT_TL_L12M variables appear to be the 2 variables with the highest explanatory power.

As a result of the correlation analysis, 91 variables were eliminated, and 112 variables were used in the modeling stage. Unlike the XGB model, 11 different variables are included in the modeling phase. Gains of the most important variables are shown in the chart below.

Chart 6 Final Variables Gain



5.3.3 Feature Selection Process Random Forest

As mentioned in the XGBoost and LGBM steps, first of all, variables with instability problems between the model development sample and the validation sample were removed from the modeling phase. Since modeling and validation samples did not change at this stage, the variables that were eliminated are the same variables as XGB and LGBM. According to the PSI analysis results, 2 variables were eliminated due to instability problems.

As a result of the performance analysis at the Random Forest variable level, 38 variables were excluded from the modeling stage because they were not significant. Explanatory results with the target on variable basis by applying Random Forest methodology are given in the table below.

Table 32 Top 10 Variable Importance Level

#	Variable Name	Importance
1	MILAT_FLAG	18.055
2	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	15.344
3	NOF_BH_MOB_BANK_LOGIN_L12M	14.47
4	AVG_DDEP_INFLW_AMT_TL_L12M	13.806
5	AVG_MAIL_ORD_PUR_TRX_CNT_L12M	13.384
6	CUSTOMER_AGE	12.104

7	KKB_MINUMUM_CC_OPEN_DATE_TERM	11.233
8	CUSTOMER_TENURE	11.034
9	AVG_DDEP_OUTFLW_AMT_TL_L12M	10.288
10	AVG_RETLOAN_PMNT_AMT_TL_L12M	9.919

KKB_TOTAL_CC_LIMIT_AMOUNT_TL and MILAT_FLAG variables appear to be the 2 variables with the highest explanatory power.

As a result of the correlation analysis, 80 variables were eliminated, and the modeling phase was continued with 30 variables with the highest explanatory value. The Mean Decrease Accuracy of the relevant variables is shown in the tables below.

Table 33 Final Variables' Importance Level

#	Variable Name	Importance
1	AVG_MAIL_ORD_PUR_TRX_CNT_L12M	50.04
2	AVG_DDEP_OUTFLW_AMT_TL_L12M	43.754
3	CUSTOMER_TENURE	39.048
4	MILAT_FLAG	37.727
5	KKB_MINUMUM_CC_OPEN_DATE_TERM	35.229
6	NOF_BH_INTRNT_BANK_LOGIN_L12M	33.593
7	AVG_ATM_WTHDRW_AMT_TL_L12M	33.546
8	AVG_PERS_FIN_ASSET_L12M	33.544
9	KKB_MAX_OPEN_OD_LIM_AMT_TL	32.553
10	CUSTOMER_AGE	31.471
11	MZC_AVG_CSH_LIM_AMT_TL_L12M	29.203
12	NOF_BH_MOB_BANK_LOGIN_L12M	27.898
13	NOF_BH_ATM_TRX_L12M	26.007
14	INDV_CC_AVG_PUR_FX_AMT_TL_L12M	25.712
15	INDV_CC_AVG_CSH_PNT_BAL_L12M	24.78
16	NOF_NBH_ATM_TRX_L12M	23.686
17	AVG_TOT_CR_BAL_TL_L12M	23.194
18	KKB_TOTAL_CC_LIMIT_AMOUNT_TL	22.462
19	AVG_RETLOAN_PMNT_AMT_TL_L12M	20.16
20	AREA_UNIVERSITY_EDUCATION_PCT	19.768
21	AVG_CR_INST_AMT_TL_L12M	17.553
22	KKB_MAX_CC_LIM_AMT_TL_L3Y	16.08
23	NOF_NBH_MOB_BANK_LOGIN_L12M	14.168
24	OWN_LIFE_INSUR_FLG_L12M	14.146
25	AVG_ATM_WTHDRW_FX_AMT_TL_L12M	9.478
26	KKB_TOTAL_OPEN_CR_INST_AMT_TL	9.311

27	MAX_ATM_WTHDRW_FX_AMT_TL_L12M	8.151
28	KKB_MAX_2ND_CC_LIM_AMT_TL_L3Y	8.004
29	NOF_ATM_WTHDRW_FX_TRX_L12M	6.034
30	KKB_MAX_OPEN_CR_INST_AMT_TL	4.4

5.3.4 Feature Selection Process Linear Regression

Linear regression methodology was run using the same data set and variables in order to compare the boosting and bagging algorithms mentioned above. After testing the basic assumptions of the regression, as mentioned in the next steps, the variables with instability problems were first eliminated. The two variables that were eliminated are variables that were also eliminated in XGBoost, LGBM, and Random Forest models.

Each variable was modeled as a single variable with the target variable. Then variables with a p value greater than 0.05 were excluded from the modeling sample by looking at the explanation of the variables. At this stage, 19 variables were eliminated. Afterward, binary correlation analysis was performed between variables, and variables with high R ^ 2 values were selected as final variables. In the correlation phase, 94 variables were eliminated. The final remaining 103 variables were included in the regression stage, and at this stage, backward variable elimination was performed according to Akaike Information Criteria (AIC) values. The variables that will not affect the model performance and cause less than 0.01% change of the model AIC developed using 103 variables were removed in the modeling mass, and modeling study was carried out using 26 variables with the highest effect on the model. Summary statistical indicators of the final variables are given in the table below.

Table 34 Final Variables Summary Statistic

Variable Name	Estimate	Std. Error	t value
(Intercept)	2136.4	8.3	257.14
AVG_TDEP_EUR_BAL_TL_L12M	0	0	24.47
KKB_TOT_OPEN_HLOAN_USG_AMT_TL	0	0	18.32
AREA_UNIVERSITY_EDUCATION_PCT	804.4	21.1	38.13
KKB_TOTAL_CC_RISK_AMOUNT_TL	0	0	27.8

KKB_TOT_OPEN_CR_CNT_L6M	-53.8	2	-27.38
NOF_BH_MOB_BANK_LOGIN_L12M	-0.5	0	-27.7
AVG_HOTEL_PUR_TRX_CNT_L12M	572.4	16.3	35.06
NOF_BH_ATM_TRX_L12M	-0.9	0	-32.87
AVG_EDU_PUR_AMT_TL_L12M	0.6	0	34.8
KKB_TOTAL_OPEN_CREDIT_COUNT	-48.5	1.5	-32.98
KKB_MAX_2ND_CC_LIM_AMT_TL_L3Y	0	0	37.54
INDV_CC_AVG_PUR_AMT_TL_L12M	0.1	0	47.05
INDV_CC_AVG_PUR_FX_AMT_TL_L12M	0.9	0	50.87
AVG_DDEP_OUTFLW_AMT_TL_L12M	0	0	-46.05
KKB_APP_CNT_L6M	21.7	0.4	48.99
KKB_MAX_CLS_CR_USG_AMT_TL_L3Y	0	0	48.87
AVG_PERS_FIN_ASSET_L12M	0	0	70.01
MILAT_FLAG	-332.6	5	-67.16
KKB_TOTAL_OPEN_CC_COUNT	-200.6	3.5	-57.19
NOF_ATM_WTHDRW_FX_TRX_L12M	217.4	3.8	57.59
OTH_BNK_OWN_ACC_EFTAMT_TL_L12M	0	0	62.26
OTH_BNK_OTH_ACC_EFT_CNT_L12M	16.1	0.2	76
AVG_ATM_WTHDRW_AMT_TL_L12M	0.2	0	91.2
TOT_SWIFT_AMT_TL_L12M	0.1	0	87.61
AVG_RETLOAN_PMNT_AMT_TL_L12M	0.5	0	122.51
KKB_TOTAL_CC_LIMIT_AMOUNT_TL	0.1	0	254.44

5.3.5 Model Selection Process

Within the scope of the performances of the relevant models, mean absolute error (MAE), mean percentage error (MAPE), and root mean square error (RMSE) values were examined. Detailed information on the performance of the model developed using the XGBoost, Linear Regression, LGBM, and Random Forest methodologies are shown below.

Table 35 Model Performance Comparison

		MAE	MAPE	RMSE
XGB	DEV	1203	18%	2076
	TEST	1400	19%	2858
	VAL	1519	19%	2925
LGBM	DEV	1421	19%	2887
	TEST	1449	20%	2948
	VAL	1593	20%	3175
Random Forest	DEV	589	8%	1319

	TEST	1410	19%	3031
	VAL	1459	18%	3025
Linear Regression	DEV	940	22%	1888
	TEST	941	22%	2040
	VAL	1034	22%	2166

While using each methodology and performing the modeling study, hyperparameters were determined by manual search. While proceeding to the final modeling stage, a grid search technique was applied for the selected champion model and final parameters were decided. As in the other segments, the model with the highest performance in the model development sample appears to be the Random Forest. However, the performance of the Random Forest model in validation and test samples significantly decreases. While the performance of the XGB, Regression LGBM models is close, the XGB model provides higher model performance.

Table 36 Model Performance with Grid Search

	DEV	TEST	VAL
MAE	1,182.91	1,350.86	1,473.60
MAPE	18%	18%	18%
RMSE	2,075.16	2,767.86	2,856.50

Hyperparameters were found by applying grid search methodology to the selected XGB model. It has been observed that the performance of the final model increases when using grid search. However, as shown in Table 35, MAE and RMSE results were found to have lower values and, therefore higher performance when linear regression was used.

6 CONCLUSION

Customer income is one of the most important factors that financial institutions take into account in their loan granting processes. The ability of customers to pay according to their income is calculated and then the limit for the requested product is defined. Therefore, it is a process that has a direct impact on the risk that financial institutions will face from the customer. If this process is not managed properly, deterioration in the loan portfolio, which is the most important asset item of financial institutions, non performing loans and loan loss provisions and losses may arise. Especially in countries where income stability is low, estimating income for customers becomes more important and becomes difficult due to uncertainties. Although there are studies on the factors affecting income in the literature, there are a limited number of articles in terms of income estimation. In this study, using machine learning algorithms evaluated within the scope of developing analytical solution methodologies, modeling techniques that can make high-performance revenue estimation have been tried.

In the study, alternative methodologies that banks can use when estimating income for customers were used, and the performances of these methodologies were compared. Within the scope of the study, besides traditional regression modeling, boosting and bagging based XGBoost, LightGBM, and Random Forest methodologies were used. Before the model development phase, the data set was divided into 3 segments according to the customer's educational level. A total of 12 different models have been developed. In the modeling studies conducted in these 3 segments, the XGBoost model in the High School and Unknown & Elementary education level segment displayed higher and stable performance compared to other methodologies. In contrast, in the Bachelor & Master segment, XGBoost, LGBM and Linear Regression methodologies performed similar to each other. This shows that the boosting algorithm performs better than other methodologies.

In the study, the validation sample, which has a different time interval in the model development sample, was also used to measure performance with the model development sample. The stable performance of the model in the validation sample

compared to the model development sample will prolong the life span of the model. Also, cross-validation was used while determining hyperparameters in boosting and bagging algorithms. Since related methodologies cause overfitting problems, using cross-validation can be said as the most important approach.

While modeling on a data set with a large number of data and features will increase the predictive power, it may cause the algorithms to slow down or even stop working. Especially random forest methodology can work much slower than other methodologies due to its structure. For this reason, before starting the modeling study, it is recommended to decrease the number of variables significantly with technical tests such as correlation, stability, and explanatory power and to establish the model development sample by sampling to increase the working speed of the model. Also, before starting modeling, conducting a segmentation study that makes a difference in terms of business management will contribute to the reduction of the data set, the more meaningful evaluation of the variables and the prevention of bias.

During the determination of the final model, hyperparameter estimates were made by applying manual and grid search techniques separately. Detection of hyperparameters has a direct impact on the final model performance. In the study, an increase in model performance was observed when grid search was used. However, it is recommended to use different hyperparameter techniques in future studies.

The addition of macroeconomic variables and the application of different hyperparameter techniques in future income model estimation studies will have an improving effect on the modeling performances.

REFERENCES

1. Alfaro, E., Gámez, M., & Garcia, N. (2013). adabag: An R package for classification with boosting and bagging. *Journal of Statistical Software*,*54*(2), 1-35.

2. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281-305.

3. Chakraborty, A., Hui, K. & Bader, F. (2007). Method and system for income estimation. *U.S. Patent Application No. 11/288,073*. Washington, DC: Kilpatrick Stockton LLP

4. Di Cellio Dias, P. C., Forti, M. & Witarsa, M. (2018). A comparison of gradient boosting with logistic regression in practical cases. *SAS Global Forum Proceedings, Paper 1857*, 1-25. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1857-2018.pdf

5. Fenerich, A., Steiner, M. T. A., Neto, P. J. S., Tochetto, E., Tsutsumi, D., Assef, F. M., & dos Santos, B. S. (2020). Use of machine learning techniques in bank credit risk analysis. *Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería*, *36*(3), 1-15.

6. Fleyeh, H., & Davami, E. (2013). Multiclass adaboost based on an ensemble of binary adaboosts. *American Journal of Intelligent Systems*, *3*(2), 57-70.

7. Francisco, E., Whigham, P., Filho, F. & Zambaldi,F. (2008). A consumer income predicting model based on survey data: An analysis using geographically weighted regression (GWR). *Latin American Advances in Consumer Research*, *2*, 77-83.

8. Freund, Y., & Schapire, R. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*,*14*(5), 771-780.

9. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29* (5) 1189-1232.

10. Kahavi R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Appears in the International Joint Conference on*

Articial Intelligence (IJCAI), 14 (2), 1137-1145. https://ai.stanford.edu/~ronnyk/accEst.pdf

11. Kibekbaev, A., & Duman, E. (2016). Benchmarking regression algorithms for income prediction modeling. *Information Systems, 61*, 40-52.

12. Koskinen, L., Nummi, T., & Salonen, J. (2005). *Modeling and predicting individual salaries: A study of Finland's unique dataset* (Finnish Centre for Pensions Working Papers No.2). https://www.julkari.fi/bitstream/handle/10024/129098/ModellingandpredictingindividualsalariesastudyofFinlandsuniquedataset.pdf?sequence=1

13. Kotsiantis, S., & Pintelas, P. (2004). Combining bagging and boosting. *International Journal of Computational Intelligence,1*(4), 324-333.

14. Munkhdalai, L., Munkhdalai, T., Namsrai, O. E., Lee, J. Y., & Ryu, K. H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability*, *11*(3), 699-722.

15. Omar, K.B. (2018*). XGBoost and LGBM for Porto Seguro ' s Kaggle challenge : A comparison* [Semester Project]. Swiss Federal Institute of Technology Zurich. https://pub.tik.ee.ethz.ch/students/2017-HS/SA-2017-98.pdf

16. Petropoulos, A., Siakoulis, V., Stavroulakis, E. & Klamargias A. (2018). *A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting*. IFC Bulletins Chapters in: Bank for International Settlements (ed.): Vol.49. Paper presented at Ninth IFC Conference on "Are post-crises statistical initiatives completed?", Basel. https://www.bis.org/ifc/publ/ifcb49_49.pdf

17. Provenzano, A. R. , Trifiro, D. , Datteo, A. , Giada, L. , Jean, N. , Riciputi, A., Le Pera, G. , Spadaccino, M. , Massaron, L., & Nordio, C. (2020). *Machine learning approach for credit scoring*. arXiv preprint arXiv:2008.01687v1. https://arxiv.org/pdf/2008.01687.pdf

18. Ramezan, C. A., Warner, T.A., & Maxwell, A.E. (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification.*Remote Sensing, 11*(2), 1-21.

19. Saavedra, M., & Twinam, T. (2020). A machine learning approach to improving occupational income scores. *Explorations in Economic History*,*75* (101304), 1-10.

20. Schapire, R.E. (2013) Explaining AdaBoost. In B. Schölkopf, Z. Luo & V. Vovk (Eds), *Empirical inference: Festschrift in honor of Validimir N. Vapnik* (pp. 37–52). Springer-Verlag.

21. Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access*, *8*, 25579-25587.

22. Wang, Y., & Ni, X. S. (2019). *A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization*. arXiv preprint arXiv:1901.08433. https://arxiv.org/ftp/arxiv/papers/1901/1901.08433.pdf

23. Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316.

24. Yu, X. (2017). *Machine learning application in online lending risk prediction*. arXiv preprint arXiv:1707.04831. https://arxiv.org/ftp/arxiv/papers/1707/1707.04831.pdf