Wiley
International Journal of Intelligent Systems
Volume 2024, Article ID 5585575, 21 pages
https://doi.org/10.1155/2024/5585575

WILEY

Research Article

Product Recommendation System With Machine Learning Algorithms for SME Banking

Ilker Met,[1] **Ayfer Erkoc**,[2] **Sadi Evren Seker**,[3] **Mehmet Ali Erturk**,[3] **and Baha Ulug**[4]

[1]*Analitical Banking, Ziraat Bank, İstanbul, Turkey*
[2]*Operational Center, Ziraat Bank, Ankara, Turkey*
[3]*Istanbul University, İstanbul, Turkey*
[4]*Bahcesehir University, İstanbul, Turkey*

Correspondence should be addressed to Ayfer Erkoc; ayfcelik@ziraatbank.com.tr

Received 21 May 2023; Revised 12 February 2024; Accepted 13 August 2024

Academic Editor: Riccardo Ortale

In the present era, where competition pervades across all domains, profitability holds crucial economic importance for numerous companies, including the banking industry. Offering the right products to customers is a fundamental problem that directly affects banks' net revenue. Machine learning (ML) approaches can address this issue using customer behavior analysis from historical customer data. This study addresses the issue by processing customer transactions using a bank's current account debt (CAD) product with state-of-the-art ML approaches. In the first step, exploratory data analysis (EDA) is performed to examine the data and detect patterns and anomalies. Then, different regression methods (tree-based methods) are tested to analyze the model's performance. The obtained results show that the light gradient boosting machine (LGBM) algorithm outperforms other methods with an 84% accuracy rate in the light gradient boosting algorithm, which is the most accurate of the three methods used.

Keywords: data analysis; machine learning; product recommendation system

1. Introduction

In the early 2000s, data analytics played a crucial role in determining branching and service points, which propelled the banking and finance sector to develop customer-specific and digital services [1]. With the advent of digitalization and changes in customer behavior, this sector felt the need to keep up with the times. Today, given the enormous volume of daily transactions processed by banks, classical processing methods fall short in handling the complex data analysis tasks at hand and the complexity of the transaction processing process that is required to process the transactions. In today's business trends, leveraging data for insights has become a crucial production factor for banks to gain a competitive advantage. The significance of data analytics has heightened due to the fact that data serve as the primary raw material for information, which, in turn, impacts decision-making processes. To obtain economic value from data, the corrected data must be processed using advanced analytical methods [2]. As a result, machine learning (ML) has become a key technology in recent years and has been adapted to clean, process, analyze, and reveal hidden patterns in the data. Like other industries that harness the power of data, ML techniques have also made their way into the banking sector's business processes.

Some of the significant data analytic studies can be listed below:

1. Customer Churn Analysis and Churn Prediction [3, 4]
2. Customer Retention [3–5]
3. Bankruptcy Prediction, Fraud Detection, Price Prediction, and Credit Scoring [5, 6]
4. Customer Profiling and Document Classifying [7–9]
5. Banking Product/Service Recommendation [9–12]

6. Performance Management and Target Setting [13]

Customer churn rate or customer churn score can be defined as the probability that customers have stopped using a product or service or have switched service providers in the financial service sector due to poor customer service. Identifying customers' needs and suggesting appropriate products is one of the most effective strategies for preventing customer churn in the financial industry. In today's competitive landscape, losing customers is easy and gaining new ones is difficult. With an estimated 1.5 million bank customers churning annually, it is crucial to analyze customer data to increase loyalty and predict potential churn [3].

Another issue in which customer data are analyzed and used as decision support systems is customer retention. Customer retention in banking can be expressed as the percentage of a bank's current customers who remain loyal during a period of interest. By understanding customer needs and offering an individualized service, the bank can increase customer satisfaction and retention, creating mutual value for the customer and the bank [4].

The churn rate prediction can be followed up by the product recommendation system for customer retention. Besides these analyses, the financial and banking markets benefit from ML approaches to solving nonlinear tasks such as bankruptcy prediction and credit card risk evaluation, stock price and portfolio recommendations, fraud detection, behavioral finance, and data analytics. From 2011 to 2021, a bibliometric analysis of 348 articles indexed by Scopus in Q1 and Q2 reveals a significant increase of 34% in AI and ML studies from 2017 to the present [5]. Credit scoring schemes are another hot topic [6], investigated for small- and medium-sized enterprises (SMEs) and corporate accounts. Also, analyses on unstructured data analysis problems (such as corporate documentation) are performed with the ML approach. The purpose of the study was to attempt to automate the analysis of unstructured data for commercial bank customers and to identify the source of the data. Term frequency–inverse document frequency (TF–IDF) extracts essential keywords after documents are cleansed. Then, specific ML algorithms are used for classification or for analysis. First, it is critical to determine the target customer group at the beginning of the main analysis of the banking system. Profiling studies can overcome this problem. The second difficulty arises from the fact that the behavior of newly acquired customers has yet to be discovered. This problem can be solved by using demographic-based profiling [8]. In such cases, hybrid algorithms and unsupervised ML techniques are used [9]. To prepare an appropriate information filtering and distribution system based on customer preferences that can inform them, the new services and opportunities offered have become an essential issue for banks. With this information, the filtering and distribution system can be an automated "product recommendation system" that offers a personalized solution for each client in short form, and the proposal will be unique for each client or specific for each customer segment [10–12]. In the literature, there are studies on the use of automatic ML methods to distribute periodic targets to bank branches also [13].

In the financial sector, where competition is intense, speed is one of the most important phenomena. The main motivation of using ML methods in the study is born from the need for a fast solution that can respond to customer needs faster and instantaneously. The increase in daily transaction volumes brings along the problem of speed in analyzing these big data. Recommending the product suitable for the customers' needs instantly and in the shortest possible time, even without the customer's request, allows financial actors to save costs, increase their efficiency, and improve their management skills to be more competitive. This study has significant implications for two critical areas: firstly, the growing need for banking analytics, and secondly, the applicability of recommender systems across various industries. The first is the development of a new type of recommender system. The research problem in the paper is focused on the challenge of offering the right banking products to customers to increase banks' net revenue, which is fundamental in the highly competitive banking industry. The paper addresses this issue by utilizing state-of-the-art ML approaches to process customer transaction data, specifically using a bank's current account debt (CAD) product. The study aims to identify patterns and insights from historical customer data to recommend banking products more accurately, leveraging ML algorithms to improve the efficiency and effectiveness of product recommendations to customers. The main contribution of this study is as follows: Ziraat Bank is the first financial institution to use the presented product/service recommender system on the SME scale. In this study, we address the issue of recommendation engines for SMEs. We propose a model for the banking sector that utilizes ML methods found in the literature. However, this model and its algorithms can be adapted and used in other industries and sectors beyond banking. Section 2 presents a most recent literature review about recommendation engines and data analysis methods used in the banking sector. In Section 3, the data analysis methodology is presented. In Section 4, the model and different data analysis algorithms are discussed. Section 5 illustrates the hyperparameter tuning, Section 6 evaluates the model, Section 7 discusses customer-specific cases, and Section 8 concludes the paper with the cross-industry standard processing for data mining (CRISP-DM) approach.

2. Literature Review

Product recommendations to customers are based on profiling studies. Recommenders or recommendation systems are tools designed to help users decide on specific tasks. Usually, these systems filter available options through prior usage information to the user behavioral, or product segmentation is a prerequisite for research in determining the target customer for the product. Customers can be profiled based on size and behavior and programmed according to their product usage habits, such as the characteristics of customers using ATMs or digital channels or the characteristics of customers using credit. After the target customer group is determined through profiling studies, product

recommendation systems can be activated. For example, by profiling customers who use loans and analyzing their product usage habits, financial institutions can make targeted product recommendations or offers to both new loan applicants and existing customers. ML methods are used intensively in the profiling and product recommendation stages. Among these methods, which appear as complementary methods in the literature, the use of artificial neural networks (ANNs) [14], classification methods, and support vector machine (SVM) [15] algorithms is remarkable. These methods can be used with data within the institution as well as by analyzing economic news and user comments from social media channels, and general product recommendations can be created.

Product recommendation systems can often be combined with supervised, unsupervised, or ensemble methods. These systems are typically grouped into two main categories. The first is collaborative filtering (CF), which recommends items based on the similarities between users. The recommended items are those preferred by the users. The second involves content-based systems, matching user and product profiles. Users are recommended items that have the greatest overlap with their profile [16]. The general models used in the banking sector include matching the product and the customer, identifying similar groups, and suggesting similar products to the target audience. In both methods, the preference for ML methods over classical methods has been increasing in recent years. Different ML approaches are investigated by the researchers, such as deep learning on credit card scoring by SME customers and deep learning on credit card scores. At Ziraat Bank, the SME customer segment refers to companies with an active operating size of less than 500 million and fewer than 250 employees. Research findings illustrate that using deep learning on credit scoring is not promising, and XGBoost outperforms deep learning in most cases [17]. Feature selection [18] and ML approaches are presented for credit scoring. Chi-square, information gain, and gain ration are used to test for feature selection, and Bayesian, naïve Bayes, SVM, and decision tree (C5.0) are evaluated for ML. As a result, random forest (RF) classifier performs better with 93% accuracy in classification, and chi-square performs well in feature selection. The authors in [19] address the problem of stock market recommendation through sentiment analysis. According to the study, the investor's buying or selling behavior may depend on emotional feelings. China's stock market was selected as a study base to measure investors' emotions and to build a recommendation system based on Guba sentiment. Also, researchers highlight the need for a specific recommendation system for monetary and banking markets [20].

Recommendation methods are categorized as content-based, knowledge-based, and CF [21]. According to the study, content-based methods are more suitable for text content such as news. However, the second category, CF methods, took the attention of researchers and applied to different fields in recommender systems. Unfortunately, the CF technique depends on the user's historical data, and a cold-start problem is one of the disadvantages where initial data are necessary to operate. The third category in recommender systems is knowledge-based systems that come with having the algorithm is developed using predefined rule sets.

A review of financial service recommendations examines the system structure, application areas, and various categories of strategies utilized in the finance sector. The study also discusses different application areas and use cases. However, the authors limit their focus to users with active entities and their interactions. Thus, the review focuses on the individual use cases and the data mining algorithms (CF, content filtering, hybrid methods, knowledge-based methods, and case-based reasoning [CBR]). Cited papers in the study offer investors the best actions on products or services such as load and market. Also, finance-related banking recommendation analyses briefly discussed that tools help bank managers to take the best action regarding loan decision processes [22].

The authors in [8] argue that CF does not apply to banking since user rating about products and services is unavailable as it is in movies or music. Also, the cold-start problem is a significant issue in the banking applications of CF. The study addresses this issue with a five-stage hybrid system.

Fog computing aims to use edge resources instead of moving every data processing operation to the cloud. This approach is used in baking product recommendations to bring customers' cost-effective, secure, agile, and transparent services with fog-oriented banking architecture (FOBA) [23]. Fog architecture requires processing close enough to where the data are generated. In the current architecture, the fog layer is devised between the user and the cloud, where the fog layer is responsible for product recommendations using ML algorithms within bank offices. And the cloud layer is responsible for CBR to validate and evaluate the recommendation's success. Authors use a generated dataset (credit and debit cards, saving accounts, mortgages, loans, and other investment products) with domain knowledge experts to test and validate the model. The system's performance is evaluated with a 0.719 mean value of Spearman's rank correlation coefficient.

The authors in [10] design a Bon card product recommendation system that targets customers with no Bon card purchase history. A hybrid technique is presented to optimize CF in Bon card transactions. The K-means algorithm clusters customers into market segments, and generalized singular value decomposition (SVD) is used for dimension reduction. The dataset is obtained from a private bank in Iran which contains 46 different customers' transaction information from the point of sale (POS) located in 50 different locations. The results show that the hybrid system has a lower error rate (RMSE) than classical techniques. An alternative recommendation system for Bon card users is presented by the authors in [24]. SVD is used in the study to recommend POS services to customers. The study uses a dataset from a private bank in the Iranian study [25] addressing the problem of offering stock recommendations with quantitative algorithms to connect different stocks to find necessary relationships. The study's primary aim is to provide stock investment strategies to users using CF. The

algorithm offers five stock options with the highest rating to buy and recommends an optimum period to sell to start a new investment process. The method presented here is based on data from the Chinese stock market between 2014 and 2018. The results illustrate an investor can profit annually 11.42% more using the algorithm.

There are several efforts performed on CF with different ML approaches to financial ecosystems, such as dynamic embedding with neural networks to build history-augmented collaborative filtering (HCF) [26]. Also, hybrid strategies with multiple methods are used in CF. For example, in [27], the authors first apply swarm particle optimization and K-means clustering to determine loyal and churning customers. The CF algorithm then proposes alternative services to potential churning customers to increase customer loyalty. Classical CF algorithm brings issues such as cold start, and broadly user-related content may lead to data sparsity problems. Additionally, the growing number of users will increase the amount of data available, resulting in performance problems. To address these issues, the authors use predefined dynamic factors of clients' mutual funds and a graphical system to lower graph computation costs. Furthermore, the recommender system focuses on recommending mutual funds to the customers [28].

The study presents a data processing framework that can be used as an input for the financial recommender system [29]. The proposed method converts demographic data with purchase history to matrix form in CFs [30]. The primary approach is to cluster customers into segments, and each segment's demographic attributes are evaluated [31]. Knowledge-based recommender systems are used in the banking sector to address rink-aware funding recommendations using fuzzy linguistic methods [32]. Based on the context, financial stock recommendation systems can be grouped as personalized and nonpersonalized. Most of the literature studies focus on individual customers, and small and medium enterprises and corporate bank customers are neglected or not considered. This research's main contribution is to develop a recommendation framework for banking products that target SMEs and the broader banking community.

3. Methodology and Model

This section outlines the methodology and model employed during the study. As demonstrated in Figure 1, the research methodology adheres to the CRISP-DM [33] methodology.

- Understanding the business processes: Initial steps to define the actual problem.
- Data preprocessing stage: The data prepreprocessing operations are decided in this stage.
- Model phase: A ML or statistical model is developed on the persistent problem and data sources.
- Evaluation phase: A general review and evaluation of the previous steps so far is made in this step. Overall performance is tested to determine whether it meets the success criteria set in the first step (problem definition phase).
- Product phase: The output of all steps in the study can be packed as a production-ready solution in such scenarios.

In this study, the business understanding and data analysis steps involved analyzing the business flow, consulting with banking experts, and conducting a literature review. Data format standards were established in line with the banking database structure, and each data collected from the banking database and data warehouse was analyzed by banking experts. The opinions of these experts regarding the data were also gathered during the analysis phase. Finally, the data were integrated and versioned while validation and unification issues were addressed. The result was a denormalized big data format version. The system is designed to stay up to date with an interval of 6 months. Within this period, the most recent algorithm changes are updated on the system. In addition to updating algorithms, the trained models are retrained to ensure the system works on the most recent historical data.

3.1. Exploratory Data Analysis (EDA). EDA is a method that is used to investigate data to discover patterns that are not clear. In addition, EDA is used to better understand the data to ensure that hypotheses and assumptions are valid within statistical frameworks with graphical representations [34].

3.2. Libraries, Frameworks, and Tools. In this study, Python programming language was used as the primary platform in this study, along with additional libraries, given below:

1. Numpy: Numpy is an essential library used for scientific calculations in Python. Multidimensional arrays (arrays), masked arrays, and matrices are a library that contains mathematical, logical, shape manipulation, sorting, and selection operations for fast operations on arrays [35].
2. Pandas: Basic data manipulation, extract, transform, and loading (ETL) operations, and data preparation steps are handled by the Pandas library. Pandas is also essential for the ML libraries in Scikit learn [36].
3. Scikit-learn: Scikit-learn is one of the most used libraries in the field of artificial ML. It contains linear regression, logistic regression, decision trees, RF, and many basic artificial learning methods [37].
4. Matplotlib: Matplotlib is a Python library for visualizing data. It works integrated with almost all visualization methods [38].
5. Shap: Explainable artificial intelligence (XAI) library for bridging the input features and the outcomes of the models [39].

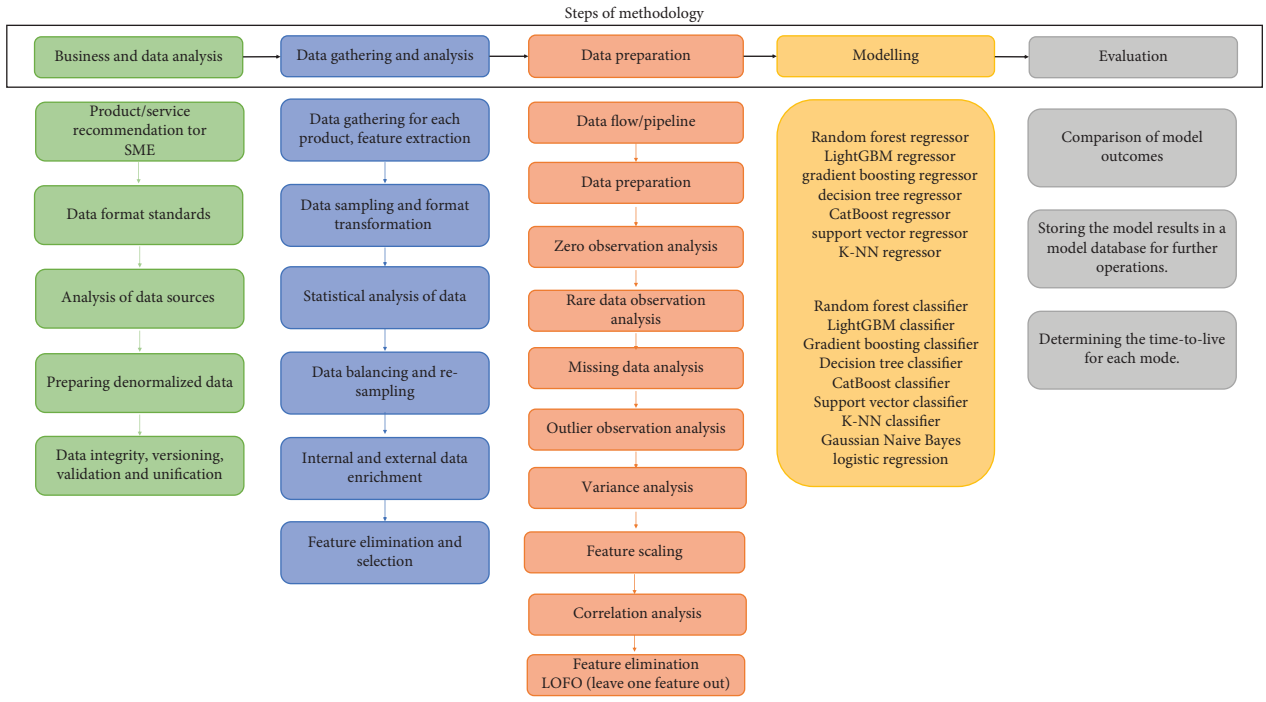

Figure 1: Steps of the methodology.

3.3. Definition and Characteristics of Dataset. Customer data were used in the study. In accordance with GDPR rules, clarification text was obtained from each of our customers, and no data that could identify individuals and cause data breaches were used in the information used. The clarification text was prepared by our bank's lawyers in compliance with the GDPR, and written consent was obtained from our customers by informing them that the information would be used in shaping marketing policies.

The traces left by the customers working with our bank as a data source can be monitored by the persons who have access rights within the framework of the information security policies of our bank and the authorization rules determined by the Banking Regulation and Supervision Agency. As for the product recommendation resulting from the model, the customer is asked whether he/she wants to purchase the product by stating the benefit that the product will provide to the customer through the relevant customer representative and bank channels (ATM and Internet banking).

The dataset consists of information about customers using a bank's CAD product, namely, CAD. The data consist of 240.963 rows and 677 columns. Each row represents the bank customers. The columns are the features expressed as behavioral data that emerge from the relationship of these customers with the bank. Dataset is summarized in Appendix 1. Table 1 provides a detailed overview of the selected features:

3.4. Working on the Data. The data import process is relatively easy with third-party Python libraries. The data can be imported through the comma separated value (CSV) or Excel (Microsoft Excel) file, or even the data source can be a SQL database within a private or cloud environment.

In the first stage, the objective is to gain an understanding of what needs to be done without delving into specifics. The subsequent aim is to identify the most pertinent variables for the given issue. Although this stage may be slow and cumbersome, insights gleaned from the dataset will make it easier to tackle potential issues in the subsequent stages.

At the EDA stage, the main goal is to clean the dataset to continue the analysis. During this cleaning process, unnecessary or useless variables are identified and reorganized. Duplicate rows or duplicate columns should be removed. In addition, variable naming standards should be established, and illogical or unsightly column names should be changed.

3.5. Understanding the Variable. As described earlier, variable selection is one of the essential steps when performing ML operations on the data. The description of the categorical and numerical variables is given in Table 1.

3.6. Data Preprocessing Steps. It consists of 676 independent variables and 1 dependent (target) variable. The dataset is generated by selecting 10,000 samples from 732.000 observation samples, and the EDA steps were completed. During the preparation of the dataset, we followed the standards of the CRISP-DM method given below.

1. Zero analysis: Two columns with all zero values were deleted without any analysis in the dataset.
2. Rare data observation analysis: The threshold value for rare observation was set at 0.01. One column containing values to be included in the rare class by

Table 1: Feature set.

Feature group	Variables	Variable definition	Variable number	Data type
Firmographic	Age, duration of customer, address, number of subcompany...	Represents the age of the customer (covers the period from the year of establishment of the company) and duration of being a bank customer, number of address changes notified to the trade registry gazette, subcompanies and sectors and sector types worked with ...	30	Numerical
Financial information	Deposits, loans, and credit cards...	Represents the average bank loan balance, current deposit balance, term deposit balance, investment account balance, overdraft account balance, and credit card spending, average foreign currency deposit balances, product types and numbers in other banks...	132	Numerical
Financial calculations	Credit and credit card usage rates, customer profitability...	Represents the loan debt, total credit card debt, total overdraft account debt, and customer profitability and relating supply information	152	Numerical
Payment behavior	Payment behavior...	Represents overdue payment number, prepayment number, transfer number, number of EFTs, clear debt number, duty of restitutions, relating supply information for payment ...	204	Numerical
Banking products	Banking product processes Numbers and use of amounts...	Represents the process number and use of bank cards, credit cards, insurance, loans, and deposit products, debtor current account, check, bill, and letter of guarantee	68	Numerical
Channel usage	Login and process Number of channel usages...	Represents the login and process numbers via Internet, mobile, branch, ATM, telephone banking, and IVR options (daily, monthly, weekly, and yearly)	48	Numerical
Campaigns/ opportunity	Number of campaigns/opportunities...	Represents the monthly number of campaigns defined for customers and the number of answers given to campaigns positively and negatively, with asks later (daily, monthly, weekly, and yearly)	16	Numerical
Complaints/requests	Number of complaints and requests and resolution times...	Represents the numbers of complaints and requests with minimum, maximum, and average resolution times, respectively (daily, monthly, weekly, and yearly)	16	Numerical
Salary information	Salary...	The number of institutions (retirement fund, employee) to which the salary is paid and the salary amounts breakdowns	11	Categorical/numerical

remaining below this threshold value was detected, and rare variables in the column were aggregated under rare.

3. Missing data observation analysis: It was set as the 25% threshold value. Columns with more than 25% blank rows proportionally may cause noise if included in the modeling. As a result of the analysis, 15 variables were removed from the dataset.
4. Outlier analysis: Multiple outlier data analysis was performed using the local outlier factor [40] method. Six observations were removed from the dataset.
5. Variance analysis: The threshold value for the analysis of variance, which is expressed as the sum of the squares of the deviations of the values in a column from the arithmetic mean, was defined as 0.05; 414 columns that did not pass this threshold were removed from the dataset.
6. Feature scaling: One Hot Encoding [41] was applied for categorical variables and Robust Scaler [42] was applied for numeric variables.
7. Correlation analysis: A threshold value of 0.95 was defined. Ninety-nine variables were removed from the dataset.
8. Feature elimination analysis: A total of 126 variables, which had a negative effect on the accuracy of the model results established with light gradient boosting machine (LGBM) regressor [43], were removed from the dataset. R^2 scores are shown in Table 2.

Although there are numerous feature elimination or dimension reduction techniques, we have had some limitations such as explainability and hardware limitations. Since the overall framework will be implemented in the banking, the bank managers requested an explainable system in each step. So explainability is one of the main motivations for not using dimension reduction techniques such as principle component analysis (PCA) or linear discriminant analysis (LDA). Also, this study is replacing a manual system in the baking recommendation, so we did not need to dive into the null hypothesis or proving the system validity, such as chi-square. Leave-one-feature-out (LOFO) method also covers many other feature selection methodologies inclusively [44]. In the method, the target variable and the entire dataset are first combined to create a base model. The features are all present in this model. In order to create a new model, remove each feature from the dataset one at a time. After each feature is removed, the model's performance is evaluated. The effectiveness of the model that was produced after each feature was eliminated is contrasted with the effectiveness of the base model. A feature is deemed important for the model if model performance suffers noticeably when it is removed. For each feature in the dataset, this procedure is repeated. As a result, each feature's effect on the performance of the model is ranked or sorted. This ranking enables the listing of features in terms of priority.

Table 2: Feature selection and LGBM regression R^2 scores.

Feature selection method	LGBM regression R^2 score
LOFO	83.23
Forward selection	79.010
Backward selection	79.6500
Stepwise selection	81.800

4. Modeling

After the data preprocessing step is completed, the data are divided into 2 as training data and test data with a certain ratio. The training set is the part of the dataset that is presented for the purpose of training the model and will enable it to reveal the pattern in the model with various algorithms. Generally, 75%–80% of all data are used for model training. The test set usually covers 20%–25% of the main dataset. The trained model is presented without the target variable, and the model's estimates are used to compare the model's predictions with the actual target values. This process helps to determine how accurately the established model makes predictions. Using two datasets, different models are trained, and their success is measured with the test set. The best fit model for the data is determined based on the performance of the models, each with different approaches to the data. The selection method is often done by looking at the scores, which are evaluated using various metrics separately for classification and regression models. The graphical summary of the 3-phase analytical study is shown in Figure 2.

The proposed system works in real time with pretrained customer data. In this architecture, real-time requests are received by the model, and evaluated results are returned as real-time responses. The model is updated with the 6-month periods mentioned in previous sections. The designed recommendation model includes a robust security architecture. In this architecture, the roles and role groups defined in the system, which data the users can access, and for how long are defined. With this approach, the customer representative authorized in the system can only access the required data of the customer defined for him. Also, this system's users are bank customer representatives, and end users (or customers) cannot access them. In addition, the system compiles with the Personal Data Protection Law in Turkey [45, 46].

4.1. Logistic Regression. Logistic regression [47] is a frequently preferred ML method in classification problems where the dependent variable is a categorical variable. Although it is called regression, it is widely used in linear classification problems. With logistic regression, not only the predicted class but also the probability of belonging to that class is specified as the output in the classification methods. The probability presented here also indicates how strong that prediction actually is. The higher these probability values, which are between 1 and 0, produced as a result of model estimation, the stronger the predictions are made by the model.

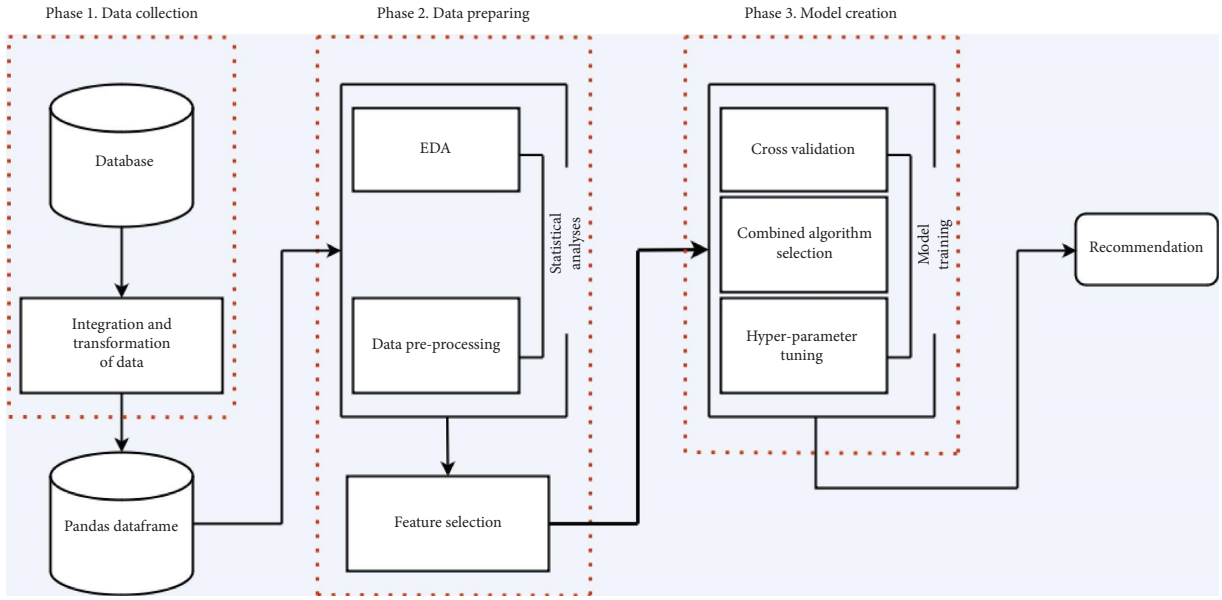

Figure 2: Modeling overview.

4.2. Tree-Based Methods. Tree-based learning algorithms are among the most used supervised learning algorithms. In general, they can be adapted to the solution of all the problems (classification and regression) considered. These methods have evolved into decision trees [48], RF [49], bagging [49], and boosting methods, respectively, by increasing their performance over the years. They are widely used in data science problems due to their adaptability for solving a wide variety of problems and their high success performance.

4.3. Decision Trees. Decision trees, which are widely used due to their ability to work with multiple outputs in classification and regression problems, can also be easily applied to complex datasets. Data preprocessing steps are relatively easy and simple. In the decision tree, while choosing the root node, it is tried to choose the column that can decompose the dataset as meaningful as possible. The decision tree stops when it reaches the predefined max_depth hyperparameter or fails to obtain a purer subset. Concepts such as Gini and Entropy are used to explain this situation. Entropy tends to produce a more balanced tree, while Gini tends to decompose the class with higher frequency. Classification and Regression Trees (CART) is a special version of decision trees with the Gini coefficient on the tree nodes [50]. If the model has been overfitted, the max_depth hyperparameter is usually decremented first. However, the overfit situation is the weakest point of decision trees and different methods have been developed for this. Training scores, cross-validation scores, and scalability of the model are shown in Figure 3.

4.4. Bagging. The biggest problem of the simple but very high performance of decision trees in estimation is that they can easily overfit the model. These models, which are very successful in train data, cannot show the same performance when the dataset is changed. Bagging methods have been developed to overcome this problem [49]. Accordingly, it can be said that it emerged from the idea of training independent trees together instead of a tree [51]. The most popular ML technique used in the bagging method is RF. Multiple decision trees are used in the RF. Bagging is when multiple weak data can work better than a single strong data. It improves accuracy in classification and regression problems; it reduces variance. The biggest disadvantage is that it is prone and open to bias.

4.5. RFs. The RF method enables generating various models and creating classifications by training each decision tree on a different observation sample over multiple decision trees. Due to its ease of use and flexibility, it has been adopted by the industry and its use has become widespread as it handles both classification and regression problems. RF also offers a variety of recommendation strategies and widely deployed in the recommender systems. For example, recommender systems for IoT sensors [52] or fertilizer recommender for the leaf diseases [53] or content recommendation for the natural language processing studies [54] are some of the implementations of RF in recommender systems. It offers the opportunity to explore the dataset again and deeper by creating various models on the dataset [55]. Training score, cross-validation score, and scalability of the model are shown in Figure 4.

4.6. Boosting. Boosting is one of the best methods for classification problems. It is good at handling missing data. The biggest disadvantage is the performance problem that occurs in practice due to the increasing complexity of the algorithm at each step. Boosting creates new data sequentially. In boosting algorithms, trees are connected to

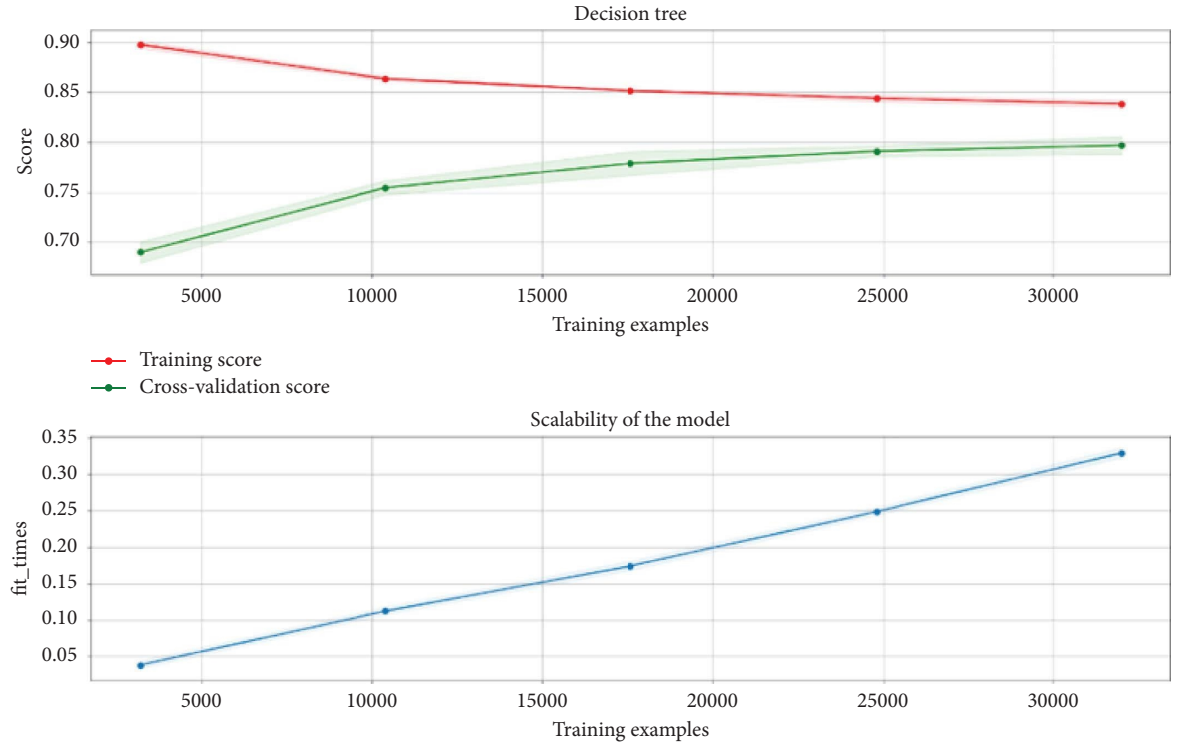

Figure 3: Decision trees.

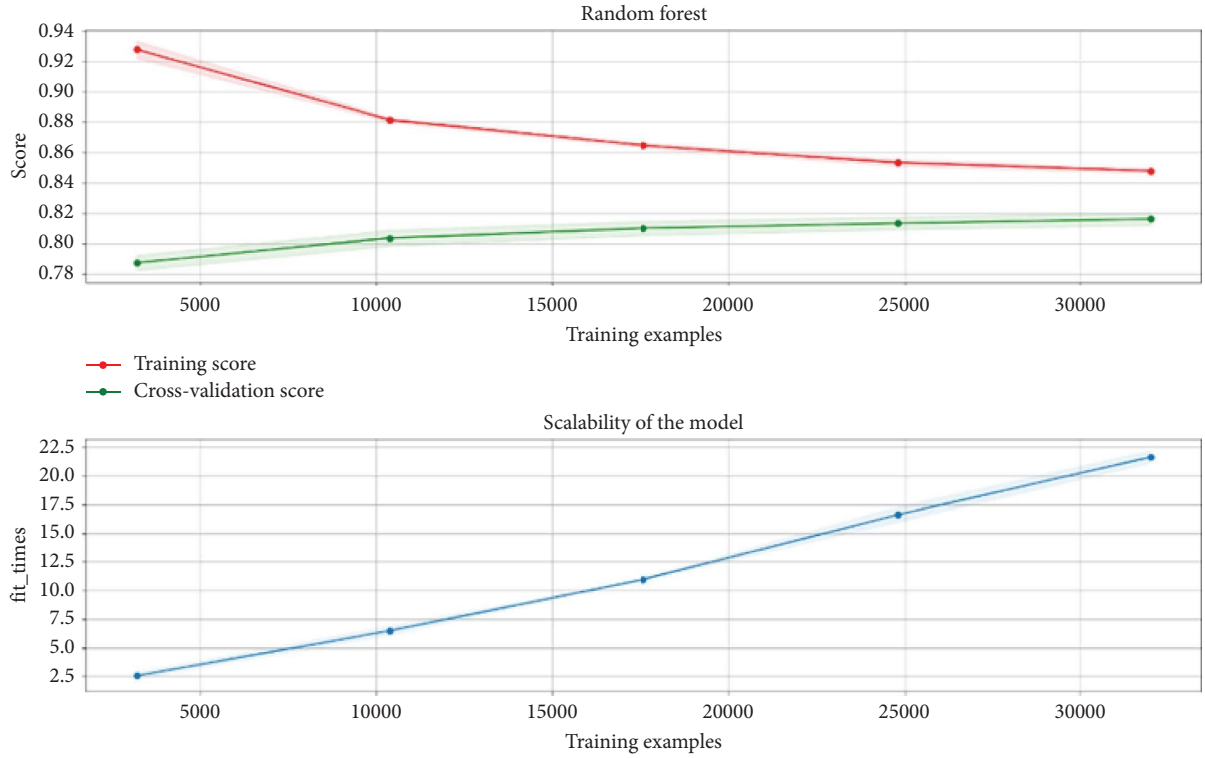

Figure 4: Random forest.

each other and each classifier is trained considering the success of previous classifiers. After each training, the weights are redistributed. It increases the weight of misclassified data. The best and common boosting techniques are gradient boosting and XGBoost [56], LGBM [43], and CatBoost [57].

4.6.1. GBM. It improves the results of decision trees using the gradient descent algorithm. Splitting the dataset into multiple subdatasets as in a RF is not implemented in this approach [43]. GBM and its variation such as lightGBM are widely used in recommendation systems. For example, one research is about creating a multipurpose recommender system [58], another research is about crop recommendation [59], or another research is about diet recommendation [60]. A similar study on crop recommendation with deep learning shows the high demand on hardware and a relatively bigger dataset for the study [61]. A decision tree is built using the dataset as it is, and a new decision tree is created based on its errors. Thus, hundreds of sequential decision trees are obtained. Training score, cross-validation score, and scalability of the model are shown in Figure 5.

4.6.2. XGBoost. XGBoost has brought some improvements over GBM, such as the use of regularization, pruning, and parallelization to prevent overlearning [56]. Similar to the previous algorithms, XGBoost also has a wide range of recommender system implementation. For example, a general-purpose recommender engine [62], or a mobile package recommender [63] or a movie recommender system [64] can be built on the XGBoost algorithm. Training score, cross-validation score, and scalability of the model are shown in Figure 6.

4.6.3. LGBM. Although its success performance is slightly lower than XGBoost, LGBM is preferred in the early stages of projects due to its speed. Thanks to its faster modeling capability, it provides increased productivity for feature engineering steps. Training score, cross-validation score, and scalability of the model are shown in Figure 7.

4.7. CatBoost. CatBoost is a fast, scalable, high-performance gradient boosting library for decision tree used for classification and regression, and it supports CPU and GPU computing. CatBoost is also one of the widely implemented algorithms for the recommender systems, such as stroke risk, online recommender system [65], agriculture crop cultivation recommender [66], or medical patient monitoring and alarm threshold recommendation [67]. Training score, cross-validation score, and scalability of the model are shown in Figure 8.

5. Hyperparameter Tuning

ML models are simply based on mathematical functions that represent the relationship between dependent and independent variables (target-features). Apart from the parameters learned during modeling, there are also external parameters that increase the success of the model and are determined manually. These may differ for each ML model. The process of finding the most appropriate hyperparameter combination is based on a success metric determined by a ML algorithm [68].

5.1. Grid Search. A separate model is created with all combinations for the hyperparameters and their values that are desired to be tested in the model, and the most successful hyperparameter set is determined according to the specified metric. Guaranteed to identify the best-performing hyperparameter set as all combinations have been tried [69]. It is quite suitable for small data. However, when working with a large dataset or when the number and value of hyperparameters to be tried are increased, the number of combinations will increase exponentially. Considering that each established model is tested with cross-validation, the cost will increase tremendously. Therefore, RandomSearchCV method can be preferred as an alternative. In this study, grid search was used for parameter tuning.

5.1.1. Train Error

5.1.1.1. Best Parameters.

```
RF: {'max_depth': 8, 'max_features': 25, 'min_samples_split': 2, 'n_estimators': 500}
LGBM: {'colsample_bytree': 0.8, 'learning_rate': 0.01, 'max_depth': 8, 'n_estimators': 1000}
GBM: {'max_depth': 5, 'max_features': 25, 'min_samples_split': 10, 'n_estimators': 200}
CART: {'max_depth': 8, 'min_samples_split': 10}
CatBoost: {'depth': 10, 'iterations': 1000, 'learning_rate': 0.01}
XGBoost: {'colsample_bytree': 0.8, 'learning_rate': 0.01, 'max_depth': 8, 'n_estimators': 500}
```

Parameters obtained as a result of the grid search were determined for decision tree regressor. Accordingly, a value of 8 for the max_depth parameter and a value of 10 for the min_samples_split parameter were determined in order to establish the best model.

Parameters obtained as a result of the grid search were determined for RF regressor. Accordingly, a value of 8 for the max_depth parameter, a value of 25 for the max_features parameter, a value of 2 for the min_samples_split parameter, and a value of 1000 for the n_estimators parameter were determined in order to build the best model.

Parameters obtained as a result of the grid search were determined for gradient boosting regressor. Accordingly, a value of 8 for the max_depth parameter, a value of 25 for

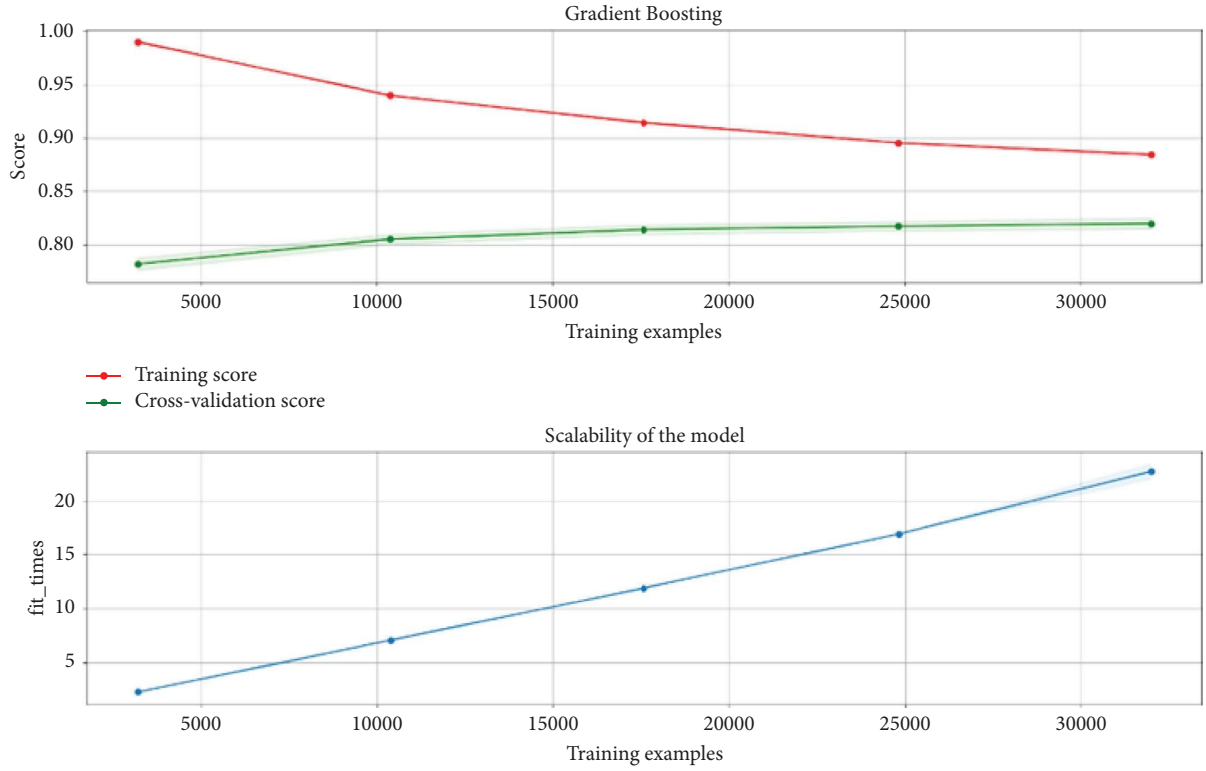

Figure 5: GBM.

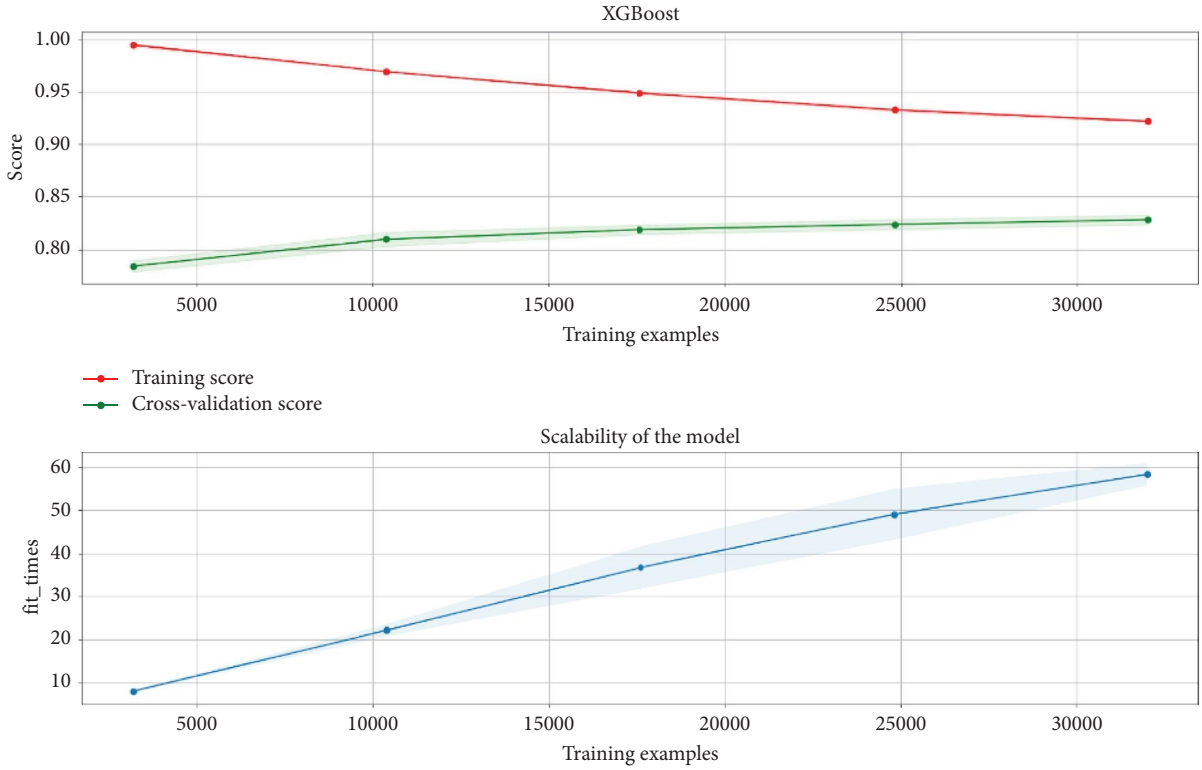

Figure 6: XGBoost.

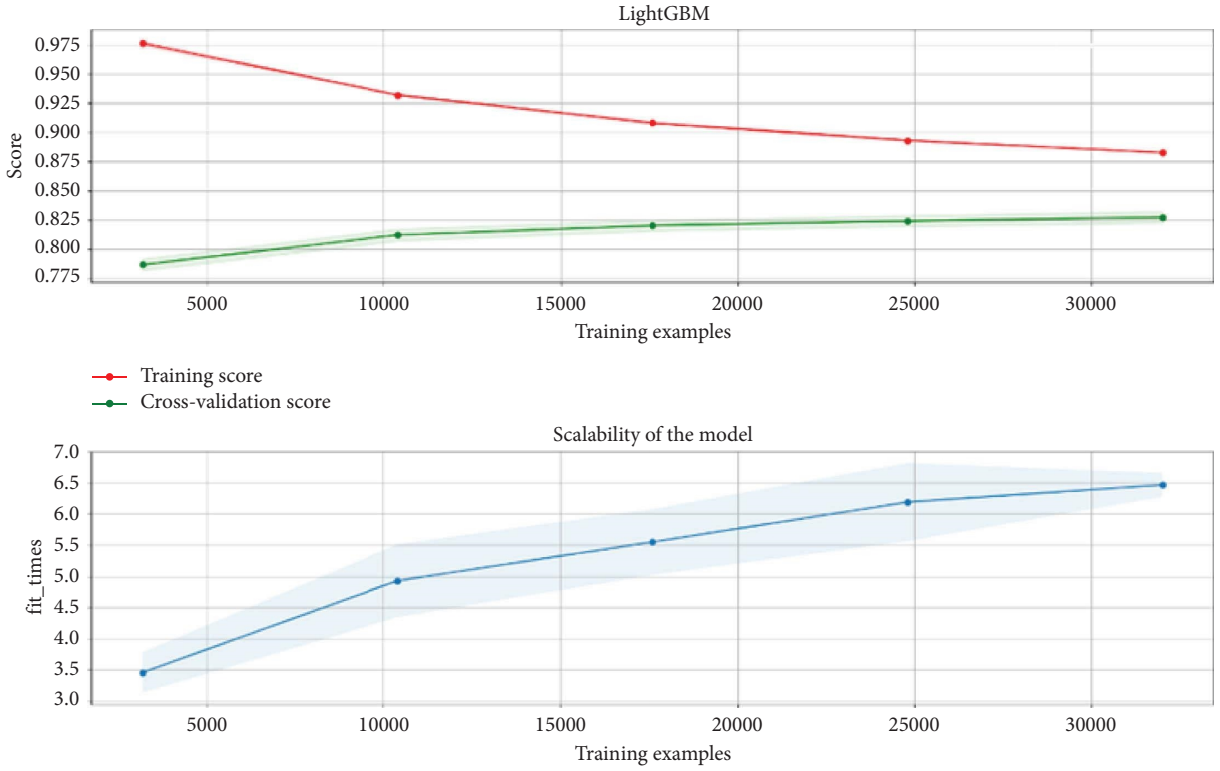

Figure 7: LGBM.

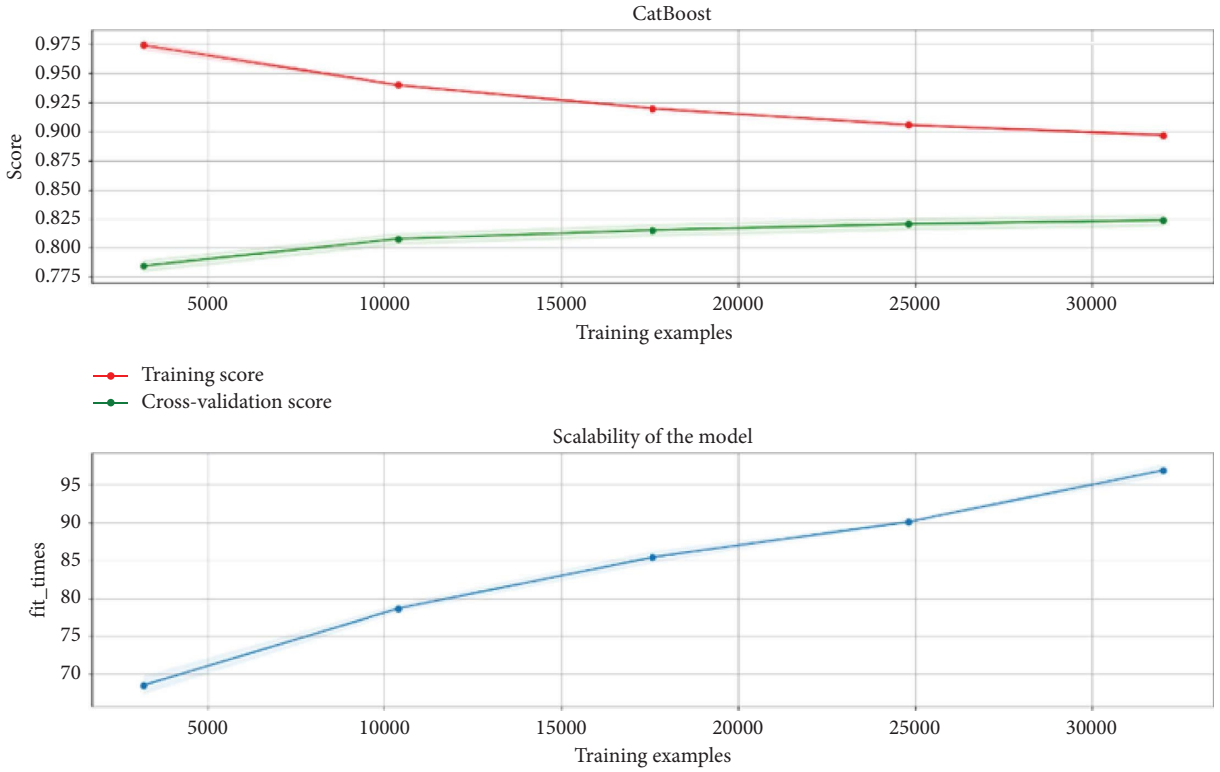

Figure 8: Cat boost.

Table 3: Model scores and metrics.

Model	R^2 score	MAE	MSE	MedAE
CART	0.8101	0.0791	0.0430	0.0000
Random forests	0.8244	0.0866	0.0398	0.0004
GBM	0.8293	0.0882	0.0387	0.0100
LGBM	0.8364	0.0809	0.0370	0.0026
XGBoost	0.8347	0.0810	0.0375	0.038
CatBoost	0.8328	0.0841	0.0379	0.0037

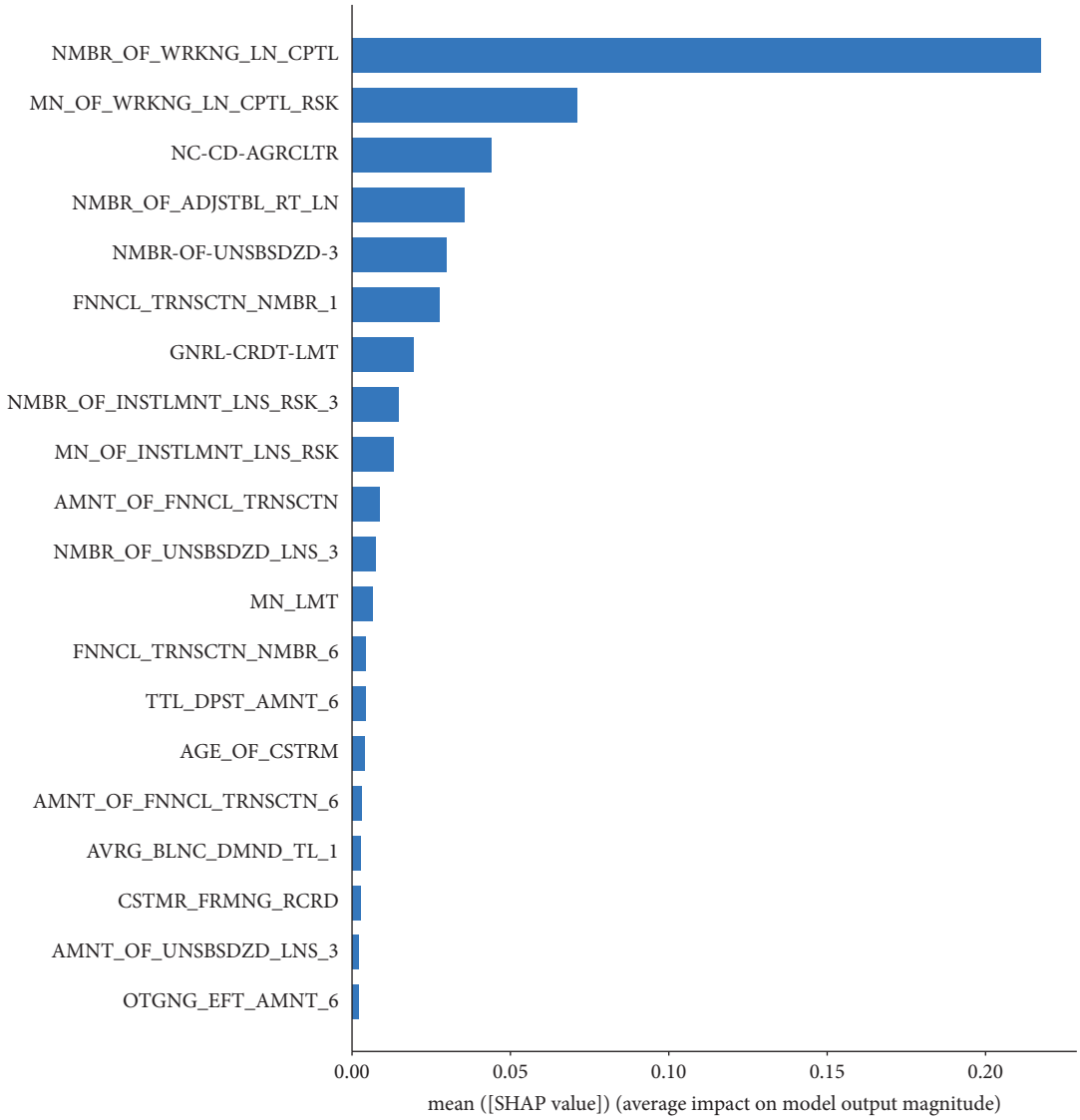

Figure 9: Variable importance.

Table 4: Test predictions.

Customer_id	y_test	rfr_y_pred	lgbmr_y_pred	gbr_y_pred	dtcr_y_pred	catbr_y_pred	xgbr_y_pred
334	0.000	0.151	0.178	0.219	0.104	0.150	0.142
1854	0.000	0.009	0.062	0.152	0.000	0.058	0.045
9479	1.000	0.774	0.901	0.702	0.636	0.817	0.920
3476	0.000	0.000	0.070	0.131	0.000	0.066	0.053
2366	0.000	0.247	0.148	0.229	0.104	0.270	0.115

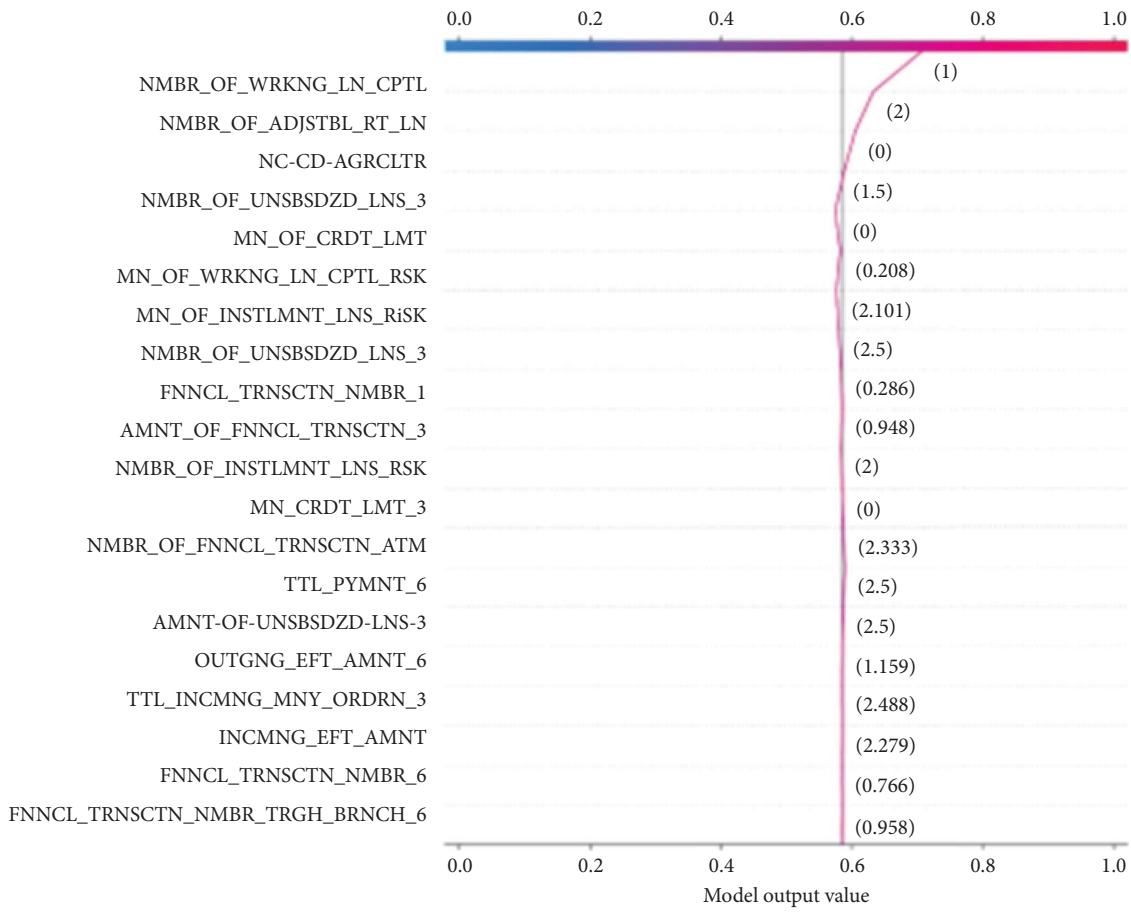

Figure 10: Predictor importance ranking for customer ID: 38091.

the max_features parameter, and a value of 200 for the n_estimators parameter were determined in order to build the best model.

Parameters obtained as a result of the grid search were determined for LGBM regressor. Accordingly, a value of 0.8 for the colsample_bytree parameter, a value of 0.01 for the learning_rate parameter, a value of 8 for the max_depth parameter, and a value of 1000 for the n_estimators parameter were determined to build the best model.

Parameters obtained as a result of the grid search were determined for XGBoost regressor. Accordingly, a value of 0.8 for the colsample_bytree parameter, a value of 0.01 for the learning_rate parameter, a value of 8 for the max_depth parameter, and a value of 500 for the n_estimators parameter were determined to build the best model. Parameters obtained as a result of the grid search were determined for CatBoost regressor. Accordingly, a value of 10 for the depth parameter, a value of 1000 for the iterations parameter, and a value of 0.01 for the learning_rate parameter were determined in order to establish the best model.

6. Model Evaluation

Parallel to the CRISP-DM methodology, the fifth step is the model evaluation, and the evaluation methodology is implemented with R^2 score, mean absolute error (MAE), mean squared error (MSE), median average error (MedAE), as already discussed and mentioned in many studies in Sections 3 and 4 [70–72].

While the precision and accuracy ratios of the system are decided on a model basis with model scores and merits (R^2, MAE, MSE, and MedAE), another success criterion in SME banking is measured by the sales success of the recommended product in the field. With the customer reports monitored on a monthly basis, the sales and usage rates of the recommended product by customers are monitored at levels parallel to the model success (average 85% sales and target realization success).

One of the most prominent applications of ML is in the field of recommendation systems in the financial sector, as in other sectors. ML algorithms can effectively study customer behavior and use this information to make future predictions, which are faster and more efficient than traditional methods. Algorithms such as CF content–based filtering and SVMs are highly successful in making efficient recommendation models [73]. Interest rate risk, the risk of changing personnel or fast-changing customer behavior in parallel with the sectoral conjuncture, is one of the risks that are always present in the financial field. For this reason, models are frequently updated in order to keep up with changes and risks. Our success rate is 85%, and we find it successful when we examine the average realization rate of

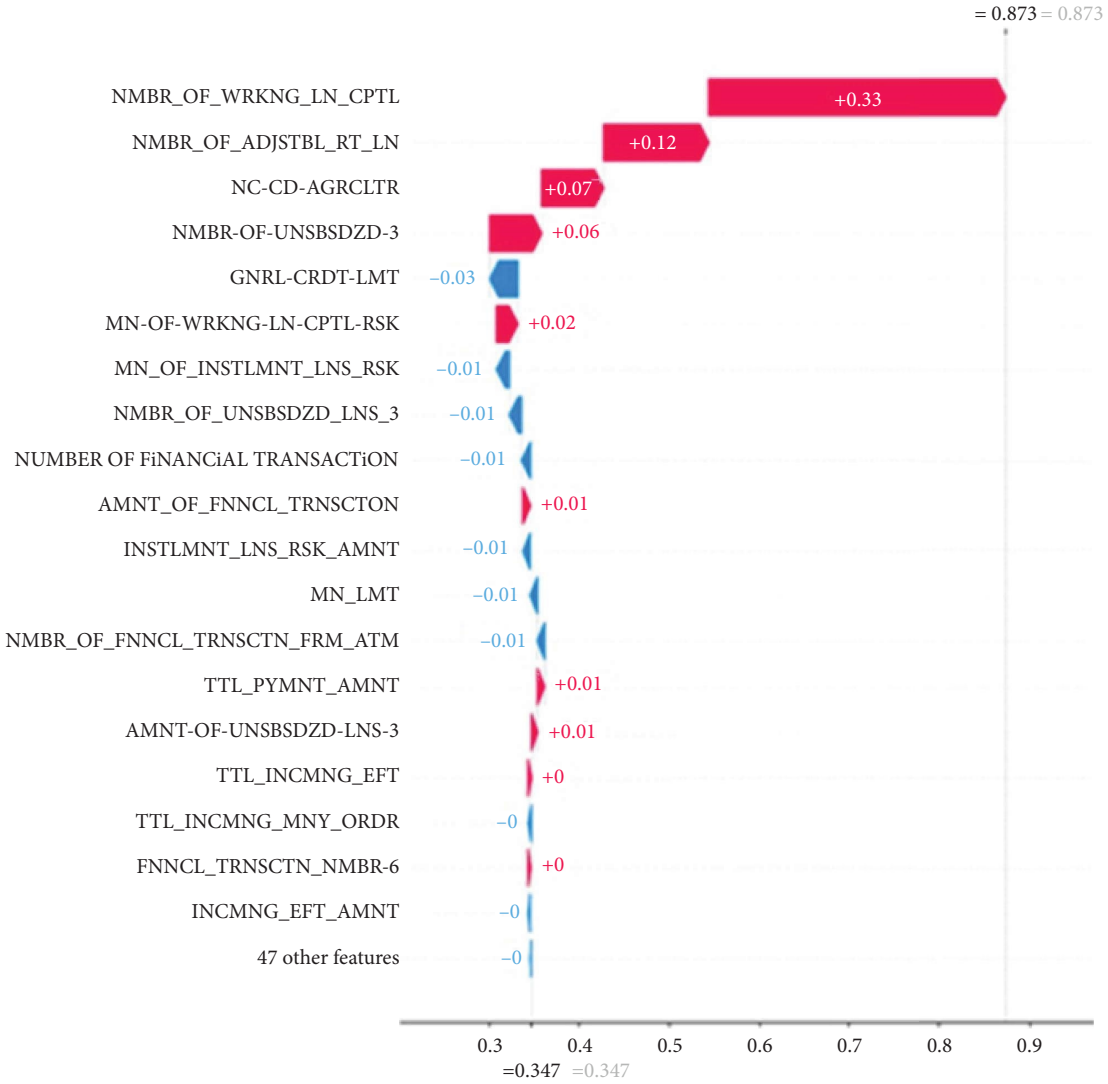

Figure 11: Effect sizes of variables for customer ID: 38091.

the sales targets we give to our branches every 3 months, which we previously prepared using ML methods as a reference [13], considering their contribution to our profitability. When we examine the target realization rates distributed classically before using the recommendation system and ML methods, the average increase of 11%–15% we achieved led us to believe that our models were successful.

In Ziraat Bank's SME banking, customers are segmented into 3 general categories including product usage, assets, and financial size, i.e., SME small, SME medium, and SME large customers. Therefore, the limits of the proposed product are also evaluated within the framework of these segmentation limits. For example, the limit of the product recommended to a small SME customer and the limit of the product recommended to a large SME customer are within the limits of the customer's financial gap and the average financial size of the segment. In this way, the financial product usage limits of the customers, as part of the product recommendation system, both meet customer needs and have a positive impact on profitability, while working in harmony with the bank's other systems.

6.1. Test Errors. CatBoost regressor is also experienced on the dataset and the optimum parameters are set to 1000 epochs and 0.01 learning rate. Optimum models are trained with these parameters. Thus, test results that produce much more accurate results for model evaluation will be observed. Results are shown in Table 3.

6.2. Test Predictions. Figure 9 shows the importance of variables for model estimation. In Table 4, sample customer forecasts are summarized.

7. Discussion

In this study, we proposed a model for a recommendation engine for the banking sector, and the findings were

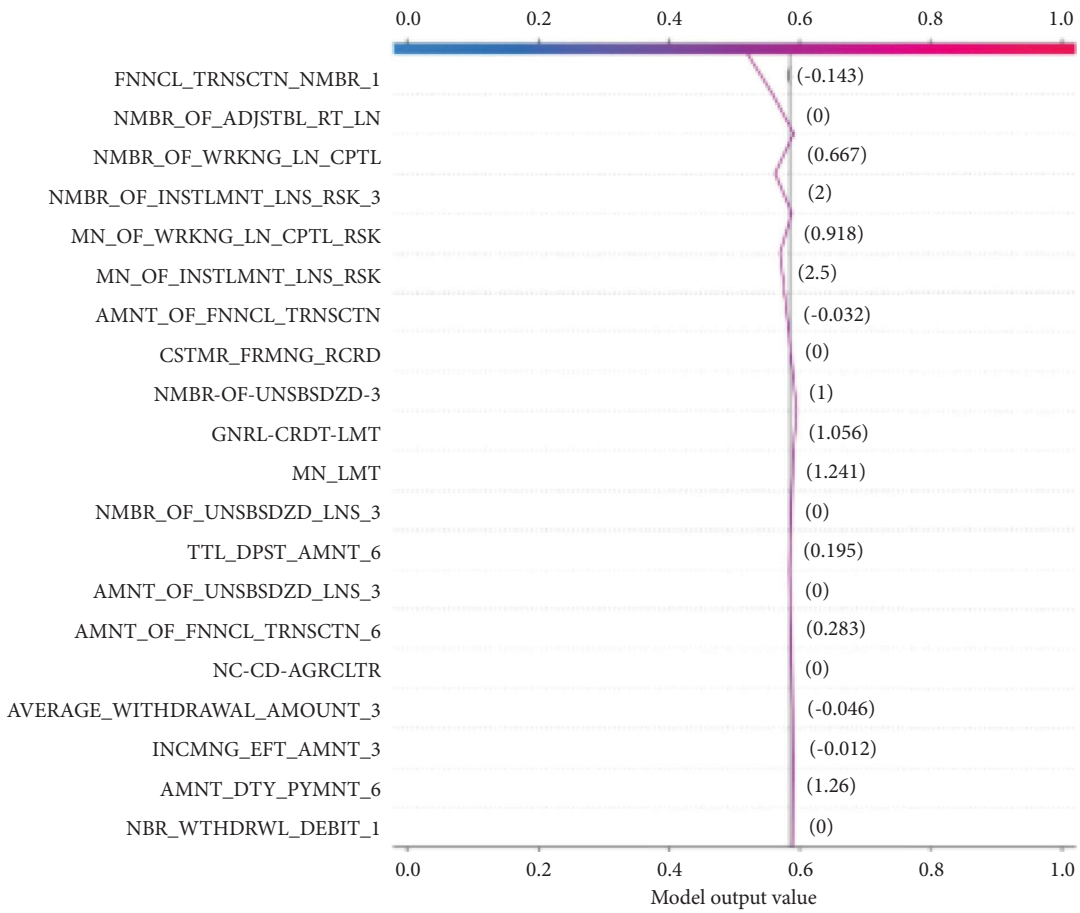

Figure 12: Predictor importance ranking for customer ID: 31238.

evaluated. The study's primary focus is to develop a recommendation engine for a recommendation rather than user interfaces; therefore, we did not include it. In addition, the system obtained in this study is used by bank employees rather than end users. In this section, we will discuss the obtained results with customer-specific cases.

7.1. Customer-Specific Cases. Customer ID: 38091

Customer 38091: LGBM Model's prediction for the recommendation system: 0.873459857538759.

customer_id	y_test	rfr_y_pred	lgbmr_y_pred	gbr_y_pred	dtcr_y_pre	catbr_y_pred	xgbr_y_pred
38091	1.000	0.915	0.873	0.865	0.996	0.795	0.883

According to the model of the customer recommendation system, customer 38091 is one of the customers with a high propensity to buy the product (87.3%) and will buy nmbr_of_wrking_ln_cptl product most likely. Results are shown in Figures 10 and 11.

Customer ID: 31238

Customer 31238: LGBM model's prediction for the recommendation system: 0.082526456782.

customer_id	y_test	rfr_y_pred	lgbmr_y_pred	gbr_y_pred	dtcr_y_pre	catbr_y_pred	xgbr_y_pred
31238	0.000	0.261	0.083	0.077	0.000	0.176	0.187

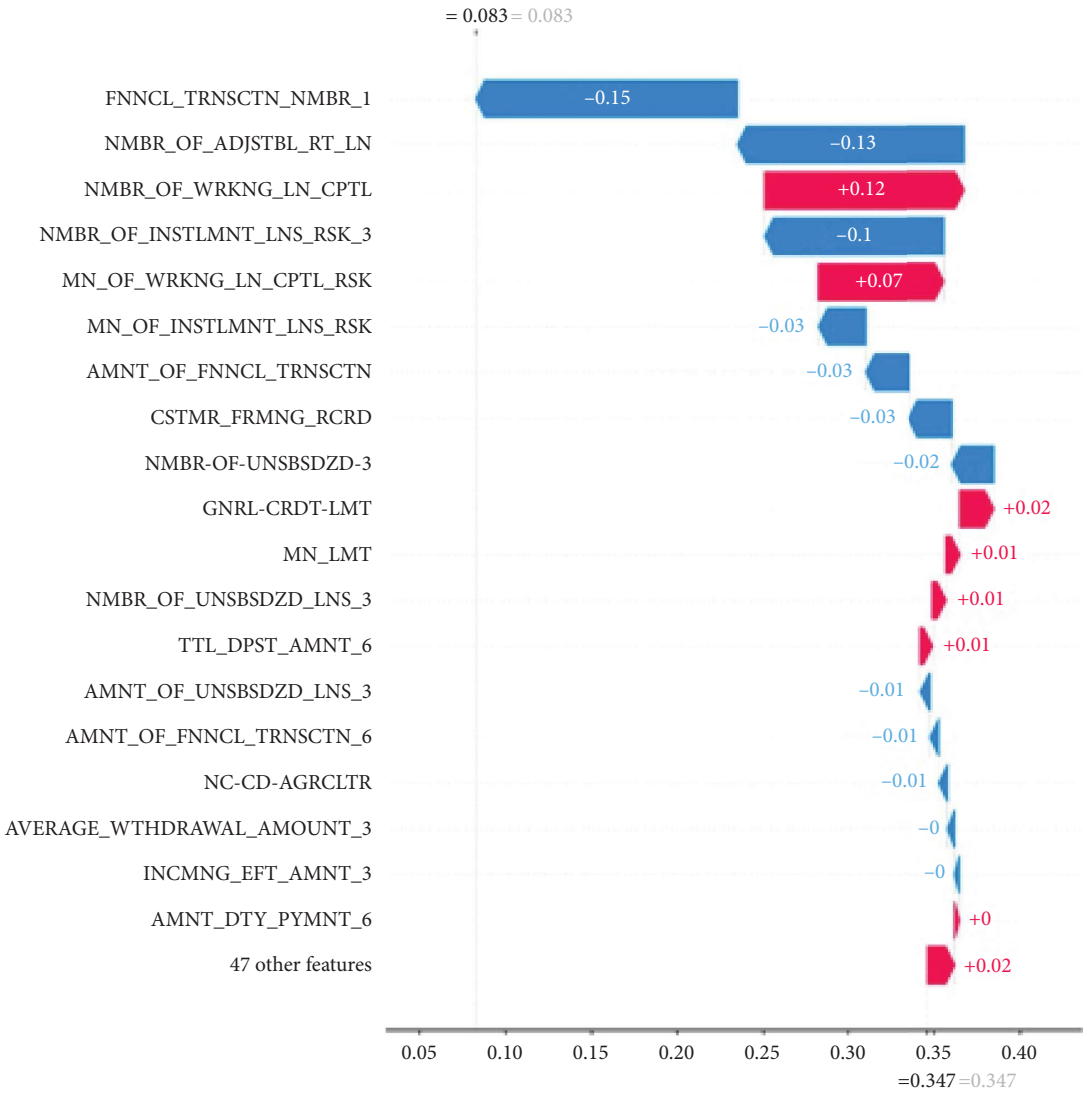

Figure 13: Effect sizes of variables for customer ID: 31238.

According to the model of the customer recommendation system, customer 31238 is one of the customers with a low propensity to buy the product (8.25%). Results are shown in Figures 12 and 13.

8. Conclusions

While the primary internal stakeholders for customer engagement in the banking and finance sector, as in any other business organization, are marketing, sales, and customer service teams, CRM strategies encourage company-wide collaboration as they collect (often dispersed) customer data to improve customer interactions. From this perspective, one of the biggest components of CRM strategies is product recommendation systems. Effective CRM strategies equip the entire organization to leverage customer data in feedback loops to inform product and service offerings, including data from sales, customer service interactions, and marketing campaigns. These data insights can also inform the methods, content, and timing of customer communications. All this is made possible by CRM strategies and the software tools that help to achieve them. The models developed with the product recommendation system are constantly updated every 6 months to include changing customer requests and economic conditions and new algorithms included in the methodology, and the algorithm with the highest success rate is evaluated, and the success of the model commissioned in the field is constantly tested with real data. In this way, both effective and efficient CRM strategies can be invested in and progress is made by continuously increasing the success of the model.

In this study, we aim to review, evaluate, and compare various prevalent ML techniques used in banking recommendation systems, including the use of automated financial analytics. We also offer informative visualizations and general recommendations for selecting a processing pipeline for a recommendation system based on an ensemble approach. While the recommendation system has garnered attention from both researchers and business leaders, modeling customer data for each business case requires

significant time and effort to provide product recommendations to customers. Therefore, this study proposes an automated recommendation system estimation service that uses the LGBM algorithm. This facilitates the creation of a deep learning recommendation prediction model for each business case based on customer behavior. A case study is presented to demonstrate that the GridSearch hyperparameters are automatically adapted until the most suitable model is chosen. This case study showed that it reached a reliability of 0.84. This study can contribute to the automated prediction and evaluation of customer product recommendations in both banking and retail business applications. Moreover, the research is automatically repeatable over time to keep the algorithm performances high. Instead of updating the ML algorithms or addressing the combined algorithm selection and hyperparameter optimization, banks, and financial institutes can benefit from the research outcomes by deploying automated ML in the study that is being conducted in the United States, the researchers wrote in a blog post on Monday. The models are updated periodically and used in CRM activities of Turkey's largest retail bank, Ziraat Bank. Model results with an 84% accuracy rate recommend the product that customers need.

According to the conducted analysis, customers' behavior has a significant impact with recommendation product. This finding has a considerable impact on designing effective client relationship management strategies in the future because the higher the rate of accuracy in detecting and identifying potential customer need for products, the higher the rates of success for more effective and efficient retention strategies.

9. Future of Work

In the specific case of our study, customers who can use the CAD product are identified, and CAD is recommended to these target customers in the recommendation system. While product diversity in SME banking benefits customers in terms of finding appropriate and fast solutions to their financial problems, from a banking perspective, it also increases customer profitability and loyalty. Our model can be adapted to any of the existing product groups in SME banking. In future studies, the study can be expanded with new algorithms, and recommendation systems for the sale of new products in line with banking strategies and goals are thought to provide positive effects on the system in determining the target audience for all product groups and increasing sales.

Data Availability Statement

In accordance with the Banking Law in Turkey, the GDPR, and the decisions and laws of the Banking Regulatory Board, customer data cannot be publicly disclosed as it falls within the scope of banking secrecy.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. *(Supporting Information)*

Appendix1_variable graph: It contains distribution graphs of the variables used in the modeling study.

References

[1] İ. Met, G. Tunali, A. Erkoç, S. Tanrikulu, and M. Ö. Dolgun, "Branch Efficiency and Location Forecasting: Application of Ziraat Bank," *Journal of Applied Finance and Banking* 7, no. 4 (2017): 1–13.

[2] E. Aktan, "Big Data: Application Areas, Analytics and Security Dimension," *Information Management* 1, no. 1 (2018): 1–22.

[3] N. Yahyapour, "Determining Factors Affecting Intention to Adopt Banking Recommender System: Case of Iran," (2008): Luleå University of Technology, Luleå, Sweden, Master thesis.

[4] S. C. Tékouabou, Ş. C. Gherghina, H. Toulni, P. N. Mata, and J. M. Martins, "Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods," *Mathematics* 10, no. 14 (2022): 2379, https://doi.org/10.3390/math10142379.

[5] S. Ahmed, M. M. Alshater, A. E. Ammari, and H. Hammami, "Artificial Intelligence and Machine Learning in Finance: A Bibliometric Review," *Research in International Business and Finance* 61 (2022): 101646, https://doi.org/10.1016/j.ribaf.2022.101646.

[6] N. Gulsoy and S. Kulluk, "A Data Mining Application in Credit Scoring Processes of Small and Medium Enterprises Commercial Corporate Customers," *WIREs Data Mining and Knowledge Discovery* 9, no. 3 (2019): e1299, https://doi.org/10.1002/widm.1299.

[7] K. S. Law and F. L. Chung, "Knowledge-Driven Decision Analytics for Commercial Banking," *Journal of Management Analytics* 7, no. 2 (2020): 209–230, https://doi.org/10.1080/23270012.2020.1734879.

[8] O. Oyebode and R. Orji, "A Hybrid Recommender System for Product Sales in a Banking Environment," *Journal of Banking and Financial Technology* 4, no. 1 (2020): 15–25, https://doi.org/10.1007/s42786-019-00014-w.

[9] D. Zibriczky, "Recommender Systems Meet Finance: A Literature Review," in *2nd International Workshop on Personalization and Recommender Systems in Financial Services* (Bari, Italy, August 2016), 1–10.

[10] A. Sharifihosseini, "A Case Study for Presenting Bank Recommender Systems Based on Bon Card Transaction Data," in *2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*, 72–77, IEEE, Mashhad, Iran, October 2019.

[11] K. Singh, "Banks Banking on Ai," *International Journal of Advanced Research in Management and Social Sciences* 9, no. 9 (2020): 1–11.

[12] G. Shani, D. Heckerman, R. I. Brafman, and C. Boutilier, "An MDP-Based Recommender System," *Journal of Machine Learning Research* 6, no. 9 (2005).

[13] I. Met, A. Erkoç, and S. E. Seker, "Performance, Efficiency, and Target Setting for Bank Branches: Time Series With Automated Machine Learning," *IEEE Access* 11 (2023): 1000–1010, https://doi.org/10.1109/access.2022.3233529.

[14] S. C. T. Chou, C. C. Yang, C. H. Chan, and F. Lai, "A Rule-Based Neural Stock Trading Decision Support System," in *IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering (CIFEr)*, 148–154, IEEE, New York, NY, March 1996.

[15] C. Ma and X. Liang, "Online Mining in Unstructured Financial Information: An Empirical Study in Bulletin News," in *2015 12th International Conference on Service Systems and Service Management (ICSSSM)*, 1–6, IEEE, Hong Kong, China, June 2015.

[16] C. Chakraborty and A. Joseph, "Machine Learning at Central Banks," in *IFC-Bank of Italy Workshop on Data Science in Central Banking*, Rome, October 2021.

[17] B. R. Gunnarsson, S. Vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, "Deep Learning for Credit Scoring: Do or Don't?" *European Journal of Operational Research* 295, no. 1 (2021): 292–305, https://doi.org/10.1016/j.ejor.2021.03.006.

[18] S. K. Trivedi, "A Study on Credit Scoring Modeling With Different Feature Selection and Machine Learning Approaches," *Technology in Society* 63 (2020): 101413, https://doi.org/10.1016/j.techsoc.2020.101413.

[19] Y. Sun, M. Fang, and X. Wang, "A Novel Stock Recommendation System Using Guba Sentiment Analysis," *Personal and Ubiquitous Computing* 22, no. 3 (2018): 575–587, https://doi.org/10.1007/s00779-018-1121-x.

[20] M. Sharaf, E. E. D. Hemdan, A. El-Sayed, and N. A. El-Bahnasawy, "A Survey on Recommendation Systems for Financial Services," *Multimedia Tools and Applications* 81, no. 12 (2022): 16761–16781, https://doi.org/10.1007/s11042-022-12564-1.

[21] Q. Zhang, J. Lu, and Y. Jin, "Artificial Intelligence in Recommender Systems," *Complex & Intelligent Systems* 7, no. 1 (2021): 439–457, https://doi.org/10.1007/s40747-020-00212-w.

[22] M. Sridevi, R. R. Rao, and M. V. Rao, "A Survey on Recommender System," *International Journal of Computer Science and Information Security* 14, no. 5 (2016): 265.

[23] E. Hernández-Nieves, G. Hernández, A. B. Gil-González, S. Rodríguez-González, and J. M. Corchado, "Fog Computing Architecture for Personalized Recommendation of Banking Products," *Expert Systems with Applications* 140 (2020): 112900, https://doi.org/10.1016/j.eswa.2019.112900.

[24] A. Sharifihosseini and M. Bogdan, "Presenting Bank Service Recommendation for Bon Card Customers: (Case Study: In the Iranian Private Sector Banking Market)," in *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 145–150, IEEE, Tehran, Iran, December 2018, https://doi.org/10.1109/ICSPIS.2018.8700534.

[25] Z. Zheng, Y. Gao, L. Yin, and M. K. Rabarison, "Modeling and Analysis of a Stock-Based Collaborative Filtering Algorithm for the Chinese Stock Market," *Expert Systems with Applications* 162 (2020): 113006, https://doi.org/10.1016/j.eswa.2019.113006.

[26] B. Barreau and L. Carlier, "History-Augmented Collaborative Filtering for Financial Recommendations," in *Proceedings of the 14th ACM Conference on Recommender Systems*, 492–497, Rio de Janeiro, Italy, September 2020, https://doi.org/10.1145/3383313.3412206.

[27] E. Shafiei Gol, A. Ahmadi, and A. Mohebi, "Intelligent Approach for Attracting Churning Customers in Banking Industry Based on Collaborative Filtering," *International Journal of Industrial and Systems Engineering* 9, no. 4 (2016).

[28] Y. C. Chou, C. T. Chen, and S. H. Huang, "Modeling Behavior Sequence for Personalized Fund Recommendation With Graphical Deep Collaborative Filtering," *Expert Systems with Applications* 192 (2022): 116311, https://doi.org/10.1016/j.eswa.2021.116311.

[29] C. Vaquero-Patricio, N. Van Ommeren, and S. Gil-Begue, "Recommenders in Banking: An End-To-End Personalization Pipeline within ING," in *Proceedings of the 15th ACM Conference on Recommender Systems*, 587–589, Amsterdam Netherlands, September 2021, https://doi.org/10.1145/3460231.3474612.

[30] N. Acharya, A. M. Sassenberg, and J. Soar, "Consumers' Behavioural Intentions to Reuse Recommender Systems: Assessing the Effects of Trust Propensity, Trusting Beliefs and Perceived Usefulness," *Journal of Theoretical and Applied Electronic Commerce Research* 18, no. 1 (2022): 55–78, https://doi.org/10.3390/jtaer18010004.

[31] C. L. Yang, S. C. Hsu, K. L. Hua, and W. H. Cheng, "Fuzzy Personalized Scoring Model for Recommendation System," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1577–1581, IEEE, Brighton, UK, May 2019, https://doi.org/10.1109/ICASSP.2019.8682809.

[32] Á. Tejeda-Lorente, J. Bernabé-Moreno, J. Herce-Zelaya, C. Porcel, and E. Herrera-Viedma, "A Risk-Aware Fuzzy Linguistic Knowledge-Based Recommender System for Hedge Funds," *Procedia Computer Science* 162 (2019): 916–923, https://doi.org/10.1016/j.procs.2019.12.068.

[33] Ş. E. Şeker, "CRISP-DM: Endüstriler Arası Standart İşleme–Veri Madenciliği Için (Cross Industry Standard Processing–Data Mining)," *YBS Ansiklopedi* 5, no. 2 (2018).

[34] J. W. Tukey, *Exploratory Data Analysis* (Berlin, Germany, Springer, 1977).

[35] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," *Computing in Science & Engineering* 13, no. 2 (2011): 22–30, https://doi.org/10.1109/mcse.2011.37.

[36] W. McKinney, "Pandas: a Foundational Python Library for Data Analysis and Statistics," *Python for High Performance and Scientific Computing* 14, no. 9 (2011): 1–9.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-Learn: Machine Learning in Python," *The Journal of Machine Learning Research* 12 (2011): 2825–2830.

[38] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering* 9, no. 3 (2007): 90–95, https://doi.org/10.1109/mcse.2007.55.

[39] S. M. Lundberg, G. Erion, H. Chen, et al., "From Local Explanations to Global Understanding With Explainable AI for Trees," *Nature Machine Intelligence* 2, no. 1 (2020): 56–67, https://doi.org/10.1038/s42256-019-0138-9.

[40] H. Ma, Y. Hu, and H. Shi, "Fault Detection and Identification Based on the Neighborhood Standardized Local Outlier Factor Method," *Industrial & Engineering Chemistry Research* 52, no. 6 (2013): 2389–2402, https://doi.org/10.1021/ie302042c.

[41] L. Jie, C. Jiahao, Z. Xueqin, Z. Yue, and L. Jiajun, "One-Hot Encoding and Convolutional Neural Network-Based Anomaly Detection," *Journal of Tsinghua University* 59, no. 7 (2019): 523–529.

[42] V. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the Influence of Normalization/Transformation Process on the Accuracy of Supervised Classification," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 729–735, IEEE, Tirunelveli, India, August 2020.

[43] M. R. Machado, S. Karray, and I. T. de Sousa, "LightGBM: An Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry," in *2019 14th International Conference on Computer Science & Education (ICCSE)*, 1111–1116, IEEE, Toronto, Canada, August 2019.

[44] F. Cürebal, F. Demirkıran, A. Çayır, and H. Dağ, "Correlation Matrix as a Smart Filter for Malware Classification Using Ensemble of Novel Feature Selection Algorithms," *Research Square* (2023): https://doi.org/10.21203/rs.3.rs-3154479/v1.

[45] Kişisel Verileri Koruma Kurumu, "Personal Data Protection Authority," (2023): https://www.kvkk.gov.tr/en/.

[46] Kişisel Verileri Koruma Kurumu, "Personal Data Protection Law," (2023): https://www.kvkk.gov.tr/Icerik/6649/Personal-Data-Protection-Law.

[47] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic Regression* (New York, NY, Springer-Verlag, 2002).

[48] S. E. Seker, "Real Life Machine Learning Case on Mobile Advertisement: A Set of Real-Life Machine Learning Problems and Solutions for Mobile Advertisement," in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, 520–524, IEEE, Las Vegas, NV, December 2016.

[49] S. E. Seker, "A Real Life Web Based Marketing Optimization Framework With External Data," *Acta Infologica* 2, no. 1 (2018): 45–51, https://doi.org/10.30801/acin.356344.

[50] T. Daniya, M. Geetha, and K. Suresh Kumar, "Classification and Regression Trees With Gini Index," *Advances in Mathematics: Scientific Journal* 9, no. 10 (2020): 8237–8247, https://doi.org/10.37418/amsj.9.10.53.

[51] S. E. Seker and I. Ocak, "Performance Prediction of Roadheaders Using Ensemble Machine Learning Techniques," *Neural Computing & Applications* 31, no. 4 (2019): 1103–1116, https://doi.org/10.1007/s00521-017-3141-2.

[52] M. Uppal, D. Gupta, A. Mahmoud, et al., "Fault Prediction Recommender Model for IoT Enabled Sensors Based Workplace," *Sustainability* 15, no. 2 (2023): 1060, https://doi.org/10.3390/su15021060.

[53] K. D. Priya, A. S. Samyogitha, A. V. K. Reddy, and B. D. Sri, "ENSEMBLED CROPIFY–Crop & Fertilizer Recommender System with Leaf Disease Prediction," in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, 600–604, IEEE, Coimbatore, India, March 2023.

[54] P. Paul and R. P. Singh, "A Weighted Hybrid Recommendation Approach for User's Contentment Using Natural Language Processing," *AIP Conference Proceedings* 2705, no. 1 (2023): https://doi.org/10.1063/5.0148413

[55] M. Shah, H. Kantawala, K. Gandhi, R. Patel, K. A. Patel, and A. Kothari, "Theoretical Evaluation of Ensemble Machine Learning Techniques," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 829–837, IEEE, Tirunelveli, India, January 2023.

[56] T. Chen and C. Guestrin, "Xgboost: A Scalable Tree Boosting System," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794, San Francisco, CA, August 2016.

[57] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting With Categorical Features," *Advances in Neural Information Processing Systems* 31 (2018).

[58] H. Han, Y. Liang, G. Bella, F. Giunchiglia, and D. Li, "LFDNN: A Novel Hybrid Recommendation Model Based on DeepFM and LightGBM," *Entropy* 25, no. 4 (2023): 638, https://doi.org/10.3390/e25040638.

[59] N. Bhatt and S. Varma, "An Enhanced Light GBM Model With Data Analytical Approach for Crop Recommendation," in *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, 1538–1544, IEEE, Tuticorin, India, March 2023.

[60] M. Roy, S. Das, and A. T. Protity, "OBESEYE: Interpretable Diet Recommender for Obesity Management Using Machine Learning and Explainable AI," (2023): https://arxiv.org/abs/2308.02796.

[61] D. A. Ssl, R. Praveenkumar, and V. Balaji, "An Intelligent Crop Recommendation System Using Deep Learning," *International Journal of Intelligent Systems and Applications in Engineering* 11, no. 10s (2023): 423–428.

[62] G. Verma, S. Sengupta, S. Simanta, et al., "Empowering Recommender Systems Using Automatically Generated Knowledge Graphs and Reinforcement Learning," (2023): https://arxiv.org/abs/2307.04996.

[63] C. Zhang and S. Zhang, "A Mobile Package Recommendation Method Based on Grid Search Combined With XGBoost Model," in *Sixth International Conference on Advanced Electronic Materials, Computers, and Software Engineering (AEMCSE 2023)*, 12787 242–248, SPIE, Dalian, China, August 2023, https://doi.org/10.1117/12.3004615.

[64] M. Mohith, K. S. Manikanta, R. Preetham, et al., "A Survey on Movie Recommendation System by Using Machine Learning Algorithm," in *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, IEEE, Erode, India, December 2023.

[65] S. Khan, N. Dekhil, E. Mamatjan, S. Hassan, and Y. Mamatjan, "An Automated Online Recommender System for Stroke Risk Assessment," *CMBES Proceedings* 45 (2023).

[66] M. Hasan, M. A. Marjan, M. P. Uddin, et al., "Ensemble Machine Learning-Based Recommendation System for Effective Prediction of Suitable Agricultural Crop Cultivation," *Frontiers in Plant Science* 14 (2023): 1234555, https://doi.org/10.3389/fpls.2023.1234555.

[67] D. Schmidt, J. Chromik, and B. Arnrich, "Recommender System for Alarm Thresholds in Medical Patient Monitors," in *16th International Conference on Health Informatics*, 74–85, Kyoto Japan, May 2023.

[68] B. Bischl, M. Binder, M. Lang, et al., "Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges," *WIREs Data Mining and Knowledge Discovery* 13, no. 2 (2023): e1484, https://doi.org/10.1002/widm.1484.

[69] B. Halstead, Y. S. Koh, P. Riddle, M. Pechenizkiy, and A. Bifet, "Combining Diverse Meta-Features to Accurately Identify Recurring Concept Drift in Data Streams," *ACM Transactions on Knowledge Discovery From Data* 17, no. 8 (2023): 1–36, https://doi.org/10.1145/3587098.

[70] G. Fan, C. Zhang, J. Chen, P. Li, Y. Li, and V. C. Leung, "Improving Rating Prediction in Multi-Criteria Recommender Systems via a Collective Factor Model," *IEEE*

Transactions on Network Science and Engineering (2023): 1–11, https://doi.org/10.1109/tnse.2023.3270910.

[71] D. T. Tran and J. H. Huh, "New Machine Learning Model Based on the Time Factor for E-Commerce Recommendation Systems," *The Journal of Supercomputing* 79, no. 6 (2023): 6756–6801, https://doi.org/10.1007/s11227-022-04909-2.

[72] S. Sridhar, D. Dhanasekaran, and G. Charlyn Pushpa Latha, "Content-Based Movie Recommendation System Using MBO With DBN," *Intelligent Automation & Soft Computing* 35, no. 3 (2023): 3241–3257, https://doi.org/10.32604/iasc.2023.030361.

[73] S. C. Mana and T. Sasipraba, "A Machine Learning Based Implementation of Product and Service Recommendation Models," in *2021 7th International Conference on Electrical Energy Systems (ICEES)*, 543–547, IEEE, Virtual Conference, February 2021.