Kepes
http://scholopress.com/kepes-journal

2021 Volume 19 Issue 3 pp. 70-81

Predicting Long-Term Deposit Openings of Bank Customers Using Decision Tree and Random Forest Classification

Mohammadreza Shahriari[1], Mohammad Hesam Asoodeh[2]

[1]Associate Professor & Faculty of Industrial Management, South Tehran Branch, Islamic Azad University, Tehran, Iran
[2]Ph.D Candidate, Department of Technology Management, Islamic Azad University, Dubai, UAE

Abstract

In recent years, banks have faced challenges in providing credit facilities due to customers' credit risk, prompting the implementation of customer validation systems to mitigate risk. Credit risk directly affects bank profitability, making it a significant concern for banks. Classification and clustering can both be valuable tools for analyzing customer behavior. This study focuses specifically on the classification process for bank customers' data, with the goal of identifying those who open long-term deposits and those who do not. Independent variables describing customer performance within the banking system are used in the classification process, with the decision tree and random forest methods being employed.

Keywords: Classification, Decision tree, Random Forest, Long-term deposits, Forecast

1. Introduction

Economic growth and development are crucial goals for policymakers and decision-makers around the world. The pursuit of these goals has become one of the most important topics in economics and has garnered significant attention from economists. One of the primary objectives of national planning is to create opportunities for increasing national capital and promoting economic growth, which is directly linked to investment. Banks play a critical role in this process by granting credit facilities to producers and merchants through various financial instruments, such as civil partnerships, Mudarabah, and other Islamic contracts. The ability of banks to attract a diverse range of deposits is a key factor in their ability to contribute to the economic development of a country and generate profits.

The banking sector has become increasingly competitive in recent years, with the entry of new institutions motivated by the sector's profitability. This has led to a rapid pace of change in banking services, driven by advancements in communication facilities, globalization, deregulation, technological advances, and the development of information technology. The profitability of banks is influenced by a range of internal and external factors. Internal factors are within the bank's control and include activities such as managing deposits, credit risk, and liquidity. External factors, on the other hand, are environmental factors that govern

[1] Corresponding author

DOI 10.5281/zenodo.7936583#44

the non-banking market and are beyond the control of management.

Bank deposits are among the most important internal factors that affect a bank's profitability. A decrease in bank deposits can increase a bank's liquidity risk, put it at risk of bankruptcy, and make it difficult to grant credit facilities. In countries with high inflation rates, depositors may be less motivated to deposit their capital in banks, which can further exacerbate this issue. However, long-term deposits can play a significant role in generating profits and enabling effective planning for bank managers.

The use of data-based analysis has become increasingly important in the banking industry, with many institutions leveraging data to inform their decision-making processes. In this context, we will discuss research that has been conducted on using data-based analysis to improve banking systems.

F. Rahnama et al compared the predictive power of artificial neural networks with other forecasting methods used in bank liquidity forecast in an article entitled "Compare power of classic time series models and artificial neural network in predicting the liquidity needs of banks" and used classic time series models such as moving averages, Holtz model and ARIMA model as well as multi-layer perceptron artificial neural network model in forecasting liquidity items. They concluded that since neural network model has the least prediction error for the last months of 2006, it is introduced as the best model (Rahnamaye Roudpashti & Mousavi, 2012).

M. Qavamzadeh forecasted the Tehran Stock Exchange Index in Tehran University in 1976 in his master thesis entitled "Forecast in Organized Trading Markets". He used price of two shares in Tehran Stock Exchange, world oil price and world gold price. In simulations carried out in this research, first an ARIMA model was fitted to relevant series and its performance was examined (Qavamzadeh, 1997). Klara Stowisk forecasted inflation in 2007 using ARMA model, in an article entitled "Forecast with ARMA models", a case study of Slovenia inflation (Stoviček, 2007).

Ince and Trafalis in an article entitled "Integrated model for exchange rate prediction" showed how he had used ARIMA models and neural networks in predicting exchange rates; the results show that neural networks provide better and more accurate answers (Ince & Trafalis, 2006). Hossein Hadipour predicted stock price in Food Industry Group with guidance of Reza Tehrani at Higher Institute of Education and Research, Management and Planning in 2003, in his master's thesis entitled "Determine the best model for predicting stock prices in food and beverage industry group of Tehran Stock Exchange". Firstly, he gathered weekly stock prices of years 1998, 1999 and first 9 months of 2000 for these companies and predicted the last 14 weeks of 2000 using Excel, Eviews and Statgraph software and Exponential Smoothing and Moving Average models and Box Jenkins; then, compared MSE index of these methods with each other. Finally, the research insists on the fact that there is no specific model for predicting stock prices of food and beverage industry groups, and each series of stock price data has its own special trends, features, and limitations. One must first use the trends and features of data time series to predict the stock price of each company, then use the appropriate model by methodology of option forecasting models, and then forecast company stock price using that model (Hadipour, 2003).

Ali Rajabzadeh Qatrami examined forecasting methods and their combination with each other aimed to reduce prediction error with guidance of Adel Azar, Tarbiat Modares University in his master's thesis entitled "Combined evaluation of forecasting methods and presentation of an optimal model for forecasting stock prices in Tehran Stock Exchange". The forecasting methods are time series (univariate) forecasting methods that rely on historical data in order to estimate future values. Several researches have been conducted on combination of forecasting methods, the results of which have shown a great reduction in forecasting errors. Each of time series forecasting methods are called individual methods and the methods combination is called combined methods. Multivariate regression model was used in order to combine individual methods which are able to perform various statistical tests in verifying model; also, the result values of individual methods, independent variables and combined prediction, the dependent variable were defined. The collected data is related to stock price of Pars Electric Company for three years in Tehran Stock Exchange, and forecasting stock price was done with different methods for fourteen periods; six methods (moving average, linear smoothing, Holt, first-order autoregressive, power trend, second-order trend) of these methods were more consistent and robust with data and had less error and were selected and used in combination (Rajabzadeh Qatarmi, 1998).

Mohammad Botshekan wrote his thesis entitled "stock price forecast using neural-fuzzy networks and its comparison with linear forecasting models" with guidance of Reza Raei, Mohammad Reza Mehrgan's consultant at Tehran University's Faculty of Management in 2000. One of artificial intelligence techniques called neural-fuzzy networks (ANFIS) was used in this research in order to forecast stock price and finally, the model ability in forecasting stock price was compared to linear ARIMA models. The survey results show the superiority and priority of ANFIS network in stock price forecasting compared to ARIMA models (Botshekan, 2000).

Also, Sinaii, Mortazavi and Teimouri Asal predicted stock prices in Tehran stock exchange using neural networks and provided evidence of chaotic behavior of index in an article entitled "forecast Tehran stock exchange index using artificial neural networks in 2005". Neural networks used in this study are multilayer perceptron (MLP) types which are trained with algorithm method after error propagation. The results showed that neural networks perform better than ARIMA linear model in forecast (Sinaii et al., 2005).

Sarafraz and Afsar in an article entitled "Examine the factors affecting gold price and present a forecast model based on fuzzy neural networks in 2005" in Economic Research Magazine used Fuzzy neural networks method based on Takagi-Sugno model in order to forecast gold price after examining the historical importance of gold and factors affecting its price fluctuation; then compared the results of this prediction with results obtained from regression method and showed the superiority of fuzzy neural networks in forecasting price of gold compared to regression method (Sarafaraz & Afsar, 2005).

Fatemeh Al-Sadat Mirfeizi wrote her thesis entitled "forecast five-year deposits of Tejarat Bank based on ARMA model" with guidance of Maryam Khalili Eraqi and Kambizpeikarjou consulting at Tehran University of Science and Research University in 2008-2009. Researcher used ARMA and ARIMA model presented by Box and Jenkins in 1970 in this research in order to predict the time series of Tejarat Bank's five-year deposits and concluded that bank trend of five-year deposits is downward (Mirfeizi, 2009).

Chen, Peng and Abraham in an article entitled "Model Stock index using HiRBF in 2007" used a HiRBF (HiRBF) network model which is an RBF neural network type in order to predict three international rates in current conventional transactions (Chen et al., 2006). Lin Yu, Kin Kyung Lai and Wang in an article entitled "General learning of multi-stage RBF neural network in forecasting exchange rates in 2008" used a multi-stage RBF model in forecasting exchange rates (Yu et al., 2008).

The classifier builds a model and uses it in forecasting classes of unknown objects in order to distinguish between objects belonging to different classes. These classes are defined in advance but not differentiated and sorted (Han J. & Kamber, 2006). Zhang and Zhu stated that classification and forecast is process of identifying a set of common features and models that describe and distinguish data classes or concepts (Zhang & Zhou, 2004). Common classification methods include Neural Networks, The Naïve Bayes Networks, Decision Trees, and Support Vector Machine. Such classification tasks are used in detecting credit card, health insurance and car insurance frauds, corporate fraud and other types of frauds. Classification is one of the most common learning models used in data mining in order to detect financial fraud (Le Khac & Kechadi, 2010). Classification is a two-step process. In first step, the model is trained using a training sample. The sample is organized in a number of rows (Tuples) and columns (Features). One of features, the category title feature contains values that indicate the predefined category to which each row belongs. This step is known as supervised learning. In second step, the model tries to classify the objects that do not belong to training sample and form a test (validation) sample (Kerkaus E. et al., 2007).

Decision tree is one of the most important classification tools in different fields. Decision trees are predictive decision support tools that create a picture of observations for possible outcomes (Han J. & Kamber, 2006). The decision tree classifies subjects based on feature values. Leaves of this model represent forecast, nodes represent a feature in a subject to be classified, and branches represent the value that a node can take and shows features sharing (Phua et al., 2010). A decision tree can be planted using algorithms based on machine learning such as: CART, Iterative Dichotomizer3 (ID3) and C5/4 algorithm. Decision trees are commonly used in detecting credit card fraud, car insurance fraud and corporate fraud (Phua et al., 2010).

Study by Dipali et al. in 2019 is among the researches that can be mentioned in recent years who investigated data mining techniques use in banking systems aimed to detect fraud and manage customer relationships (Kamath et al., 2019). Also, Kanchana et al. in 2020 investigated phishing attacks on banking sites using classification techniques in data mining (Kanchana et al., 2020). In another article, Shahbazi presented a credit rating model for bank customers using decision tree-based classification algorithms (Shahbazi, 2020). Dinser et al. presented a method based on data mining in order to evaluate customer satisfaction from mobile software in the banking industry (Dinçer et al., 2020). Kaur et al. used data mining in forecasting credit risk of bank customers (Kaur et al., 2019). Another study by Hosni et al. was published and they studied digitization and data mining in banking systems (Hassani et al., 2018).

1. Problem Statement

Importance of attracting financial resources is so important and vital for banks of today's world that it has created a very intense competition in this field. Attract more deposits and provide better and faster services to customers are among the main reasons of providing various banking services such as bank telephones, smart cards, establishing branches outside the mainland or using advanced technology in new areas; some of them include banking through television and mobile phones, extensive use of internet and increased prevalence of electronic commerce in economic activities field. People's deposits in banks are important for two reasons: First, deposited money by people in banks increases the lending power of banks, therefore banks can provide more new loans using collected savings in production and investment affairs. The second importance of people's increased deposits in banks is when people prefer to keep their money in banks and spend less money, the money amount in circulation decreases and as result inflation rate decreases and purchase power increases. These factors encourage people in different ways to refer banks and make deposits and help economic prosperity of country.

Bank deposits can be predicted based on importance of resources (bank deposits) in achieving bank's aims in terms of attracting suitable resources, granting long-term facilities to production sector, and need for proper liquidity management and reducing liquidity risks.

As you know, banks use two main sources in allocating resources and lending:

- ✓ Equity
- ✓ Types of attracted deposits.

Considering the balance sheet of a bank, we understand that equity has a very small share in allocating resources and credit to real and legal entities, and the main source of granting various banking facilities is available deposits (Mirfeizi, 2009).

Therefore, financial planners and decision makers of bank must look at available deposits with a scientific perspective in order to predict the trend of this necessary resource and take necessary and effective measures in preventing the resources reduction and its increase.

This research aimed to forecast opening of long-term deposits by bank customers using random forest approach. In this regard, random forests performance is compared based on basic classification methods such as decision trees and results are reported.

2. Decision Tree

The decision tree is one of powerful and common tools of classification and forecast. Unlike neural networks, decision trees produce rules. That is, the decision-making tree explains its forecast as series of rules, while neural networks express only final prediction, and its form remains hidden in network itself. The decision tree algorithm starts by selecting the test that performs the best separation for categories. The most important aim of classification is to obtain a model for forecast. Therefore, we use a set of data called "training data" which is a set of variables and records. In the next steps, the same is done for lower nodes with less data to obtain the best rules. Finally, the tree becomes so big that no better separation can be done for node data. At this step, the created tree's effectiveness must be measured. For this purpose, a set of records or experimental data is used which are different from primary data created by tree. The measured criterion is: the percentage of data that is correctly classified and predicted category is the same as categories.

Decision trees are those classify samples by their categories based on their feature values. Each node of decision tree represents feature of sample to be classified, and each branch represents the value that node can take. Clustered samples start from parent node and are clustered based on their feature values. Decision trees are a way of presenting a series of rules that lead to a category or value.

Disadvantages of decision tree (Wang & Lee, 2006)

- They are not suitable in cases where the learning aim is to estimate a function with continuous values.
- The error probability is high in cases with a large number of categories and a small training sample.
- Generate a decision tree has a high computational cost.
- There is a possibility of false relationships.

Advantages of the decision tree (Wang & Lee, 2006)

- Decision tree requires complex calculations to categorize data.
- Decision tree presents us which fields with variables have important effect on our prediction and classification.
- The applied production rules can be extracted and understood.
- Unnecessary comparisons are omitted in this structure.
- There is no need to estimate distribution function.
- Data preparation for decision tree is simple or unnecessary.
- It is possible to confirm a model in decision tree using statistical tests.

The basic algorithm for decision tree induction is a greedy algorithm that creates decision trees in a top-down recursive divide-and-conquer method. The algorithm steps are as below:

Create a node N.

1) If samples are all of the same category C then
2) Return node N as leaf node named with category C.
3) If feature list is empty, then.
4) Return node N as leaf node named with majority of sample set.
5) Select test feature, a feature from features list with the highest information gain
6) Name node N with test feature
7) For each known value of a_i the test feature
8) Grow a branch from node N for special feature condition a_i=
9) Suppose S_i as a set of samples with test feature of a_i

10) If S_i is empty, then:

11) Attach a sheet labeled Majority Class Samples

12) Otherwise, attach the node that returns following function.

There is a common approach that decision tree induction algorithms can use in order to avoid overgrowing of training data which is pruning deployed decision tree. If two trees use the same type of test and have similar forecast accuracy, the tree with fewer leaves will be preferred. In summary, one of the most useful characteristics of decision trees is their comprehensibility.

3. Random Forest

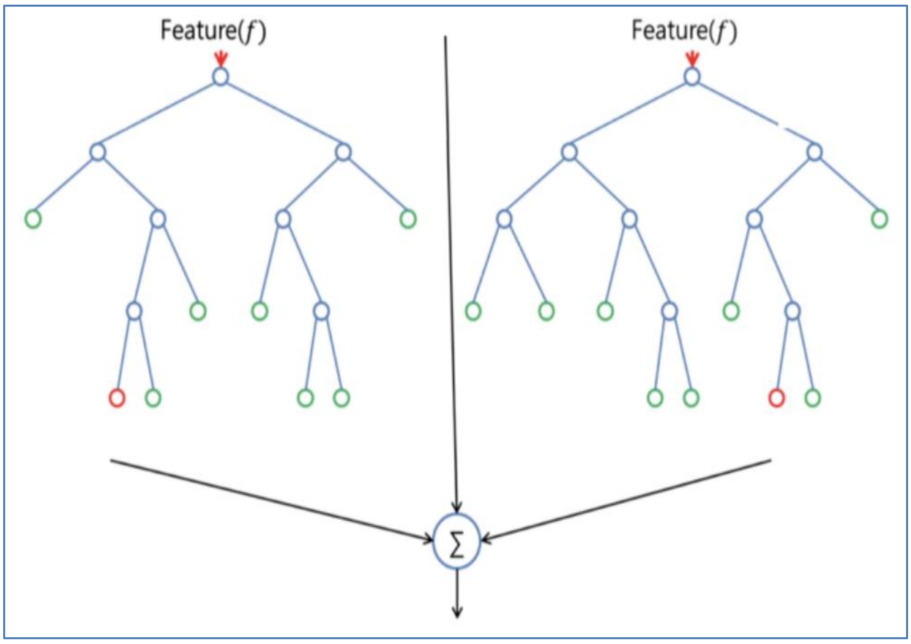
Random Forest is an easy-to-use machine learning algorithm that provides very good results even with no adjust on meta-parameters. Due to simplicity and usability of this algorithm, it is considered as one of the most usable machine learning algorithms for both classification and regression. In this article, we will examine how the random forest works and other important issues around it.

Firstly, one must learn decision tree algorithm which is the building block of a random forest in order to understand how a random forest works. It was studied in detail in previous section. People use decision trees for their decisions and selections every day, even if they don't know what they're using is a machine learning algorithm. However, a decision tree computer model has no prior knowledge and is never able to establish relationships between variables. The model must learn everything about the problem based on the data provided to it. According to their daily experiences, people know how to transform responses from a given procedure into a reasonable forecast; whereas, the model must learn each of these relationships.

Random forest learns to map data to outputs in model's training or fitting phase, as an unsupervised machine learning model. The model is given historical data during training that is relevant to problem domain and the correct amount that model needs to learn in order to make predictions. The model learns the relationships between data (known as features in machine learning) and values that user wants to predict (called targets). The decision tree calculates the best questions to ask in order to get the most accurate estimate. When the model is asked to make a forecast for the next day, it must be given the same data (features) as it was taught during training, so that the model can make an estimate based on learned structure. As humans learn by samples a decision tree learns by experience, the difference is that it has no prior knowledge to apply to problem. Humans are "smarter" than trees in making reasonable estimates before training. However, the decision tree's ability to make forecast surpasses human power after sufficient training with qualitative data. It must be remembered that decision tree has no conceptual understanding of problem and such an understanding is not achieved even after training. From model's perspective, it takes numbers from input and outputs numbers different from what was trained during training. What was stated is the concept of top level of decision tree. In fact, a flowchart of questions that leads to forecast forms a decision tree. Now, there is a big jump from a single decision tree to a random forest.

Random forest is a supervised learning algorithm. As its name suggests, this algorithm creates a random forest. Created forest is actually a group of decision trees. Forest construction by trees is often done by bagging method. The main idea of bagging method is that combination of learning models increases overall results of model. Simply, random forest builds multiple decision trees and merges them together in order to produce more accurate and stable forecasts. One of the random forest advantages is that it can be used for both classification and regression problems, which constitute majority of current machine learning systems. Here, the random forest performance will be described for classification since classification is sometimes considered as the building block of machine learning. Figure 1 presents two random forests made of two trees.

Fig. 1 Random Forest structure made of two trees



Random forest has meta-parameters like decision tree or bagging classifier. Fortunately, there is no need to combine a decision tree with bagging classifier and we can use random forest classifier. As stated earlier, regression problems can be solved using random forest, or in other words random forest regressor. Random Forest adds additional randomness to model as trees grow. The algorithm looks for the best features among a random set of features instead of searching for the most important features when splitting a node. This leads to more variety and a better model. Therefore, only a subset of features is considered by algorithm to split a node in random forest. The trees can be made even more random by added use of random threshold for each feature instead of searching for the best possible threshold (like what a normal decision tree does).

4. Difference between Decision Tree and Random Forest

As mentioned earlier, a random forest is a collection of decision trees, but there are differences between them. If an input dataset is given as input to algorithm with its features and labels, it will formulate some of rules to be used in making predictions. In comparison, the decision tree algorithm randomly selects observations, decides on features to build multiple decision trees, and then uses average results calculation. Another difference is that the deep decision tree may suffer from overfitting. Random forest often avoids overfitting through building a random sub-tree of features and building a smaller tree using this sub-tree. Then, it merges the random sub-trees. It is worth to note that the solution does not always work, and it slows down the calculation process depending on number of created random forests.

Random forest is a good algorithm for training in model development process to check how it works; therefore, making a bad random forest is practically harder than making a good model due to simplicity of this algorithm. Also, if you need to develop the model in a shorter period, this algorithm will be good option. In addition, the model provides a good indicator of importance that assigns to features.

5. Study Case and Results Analysis

The study case of this research is taken from Moro et al.'s study in 2011 (Moro et al., 2011). They study Guimarães Bank of Portugal in this research and the result is registration of customer information based on 17 different features. Information of 45,211 customers was registered in this research. Figures 2 - 5 show the employment, marital, educational status and communication (contact) method of customers with bank, respectively. Table 1 shows registered features of each customer. Also, the variable type under investigation is listed in this table which includes categorical or numerical. The 16 primary features presented in the bank database are the independent variables of forecasting process, and the seventeenth feature is dependent variable of present research.

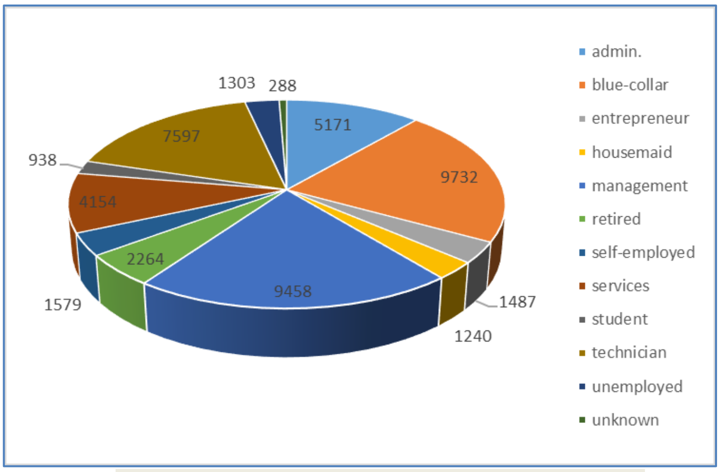


Fig. 2 Pie chart of customers' employment status

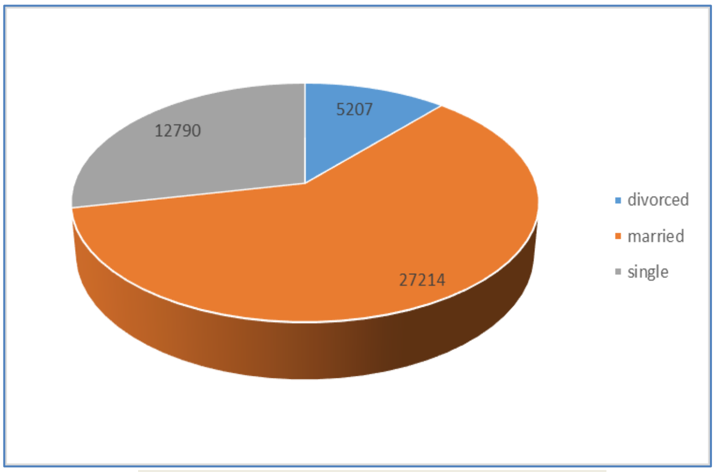


Fig. 3 Pie chart of customers' marital status

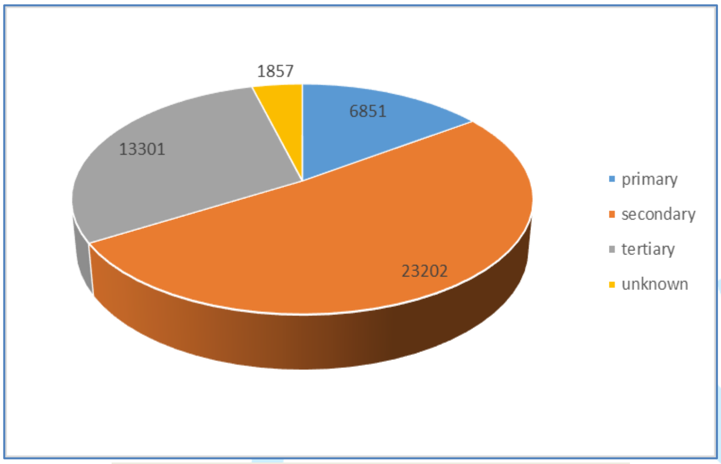


Fig. 4 Pie chart of customers' educational status

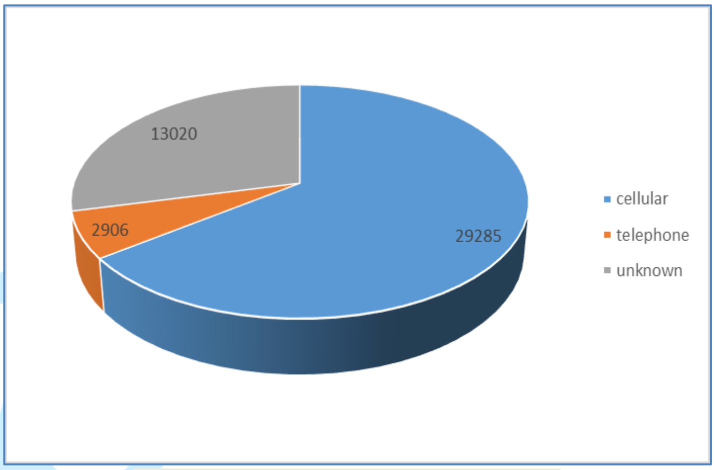


Fig. 5 Pie chart of customers' call status

Table 1 Customers' characteristics in database

Row	Feature name in data base	Feature type	Obtainable values in database	Description
1	Age	Numerical	-	Age
2	Job	Classified	✓ admin, ✓ unknown ✓ unemployed ✓ management ✓ housemaid ✓ entrepreneur ✓ student ✓ blue-collar ✓ self-employed ✓ retired ✓ technician ✓ services	Employment status
3	Marital	Classified	✓ married ✓ divorced ✓ single	Marital status
4	Education	Classified	✓ Unknown ✓ Secondary ✓ Primary ✓ Tertiary	Educational status
5	Default	Binary	0.1	Credit card

6	Balance	Numerical	-	Average annual balance (Euro)
7	Housing	Binary	0,1	Mortgage
8	Loan	Binary	0,1	Personal loan
9	Contact	Classified	✓ Unknown ✓ Telephone ✓ Cellular	Contact type
10	Day	Numerical	-	Last call day in month
11	Month	Classified	✓ Jan ✓ … ✓ dec	Last call month in year
12	Duration	Numerical	-	Last call duration
13	Campaign	Numerical	-	Number of calls in current period
14	P-days	Numerical	-	Number of days passed since last call
15	Previous	Numerical	-	Number of calls in previous period
16	P-outcome	Classified	✓ Unknown ✓ Other ✓ Failure ✓ Success	The result of previous marketing campaign
17	y	Binary	0,1	Opening long-term deposit

According to what we said on application of decision tree and random forest in classifying bank customers, the aim is to provide a model in order to predict the seventeenth feature of database introduced in Table 1. In this research, we intend to predict whether this customer will open a long-term deposit or not based on 16 features of each customer? In this regard, we will use decision tree and random forest approaches and compare and analyze the results of each method. We use the standard criteria of related research in evaluating the performance of classifiers. We must define the concept of confusion matrix before enumerating the evaluation criteria. This matrix shows how the classification algorithm works based on input data set and types of problem classes. Figure 6 presents a confusion matrix that includes two layers of "+" and "-". The problem aim is to detect records with positive classes of data that have not been observed before.

Fig. 6 Confusion matrix

		Estimated records	
		-Category	+Category
Real records	- Category	TN	FP
	+ Category	FN	TP

The concepts related to confusion matrix are defined as below:

Number of true negative (TN): number of records whose real category is negative and classification algorithm correctly recognizes them as negative.

Number of false positive (FP): number of records whose real category is negative but classification algorithm mistakenly recognizes them as positive.

Number of false negative (FN): number of records whose real category is positive, but classification algorithm mistakenly recognizes them as negative.

Number of true positive (TP): number of records whose real category is positive but classification algorithm correctly recognizes them as positive.

The most important criterion in determining the efficiency of classification technique is accuracy criterion. The accuracy criterion calculates total classification. The criterion indicates the fact that what percentage of total test record set is correctly classified by designed classifier. The classification accuracy is obtained using equation **1.** Two TP and TN values are the most important values that should be maximized in a binary problem. The main problem is data imbalance and significant difference of each category samples that makes a model shows high overall accuracy with large number of categories. Therefore, we need a more accurate criterion in measuring accuracy and efficiency of proposed classification algorithms that is shown in equation **2.** Sometimes our rereading is high due to weakness of proposed model. We must measure this weakness with another criterion. In addition to rereading criterion, we define another criterion called precision, equal to number of true positive samples to total number of positive samples (equation 3) to solve this problem and take into account false positives.

$$\text{(1)} \qquad \text{Accuracy}= \frac{tp+tn}{tp+tn+fp+fn}$$

$$\text{(2)} \qquad \text{Recall}= \frac{tp}{tp+fn}$$

$$\text{(3)} \qquad \text{Precision}= \frac{tp}{tp+fp}$$

Table 2 Evaluate performance of decision tree and random forest methods in customer classification

Classification method	Accuracy	Recall	Precision	Computing time (seconds)
Decision tree	0.8870	0.7153	0.7157	7
Random forest	0.9073	0.6827	0.7835	45

According to Table 2, random forest has more accuracy and validity than decision tree in classifying customers in order to predict the opening long-term deposit account, but calculation time and number of rules for classification process of decision tree is less than random forest. Therefore, it can be concluded that random forest has higher accuracy and more complexity than decision tree for bank customers' classification. It should be noted that a combination of 100 linked decision trees was used in this research in order to apply the random forest algorithm in classifying bank customers. Figure 7 shows the increased random forest accuracy process per increased number of combined trees with each other.

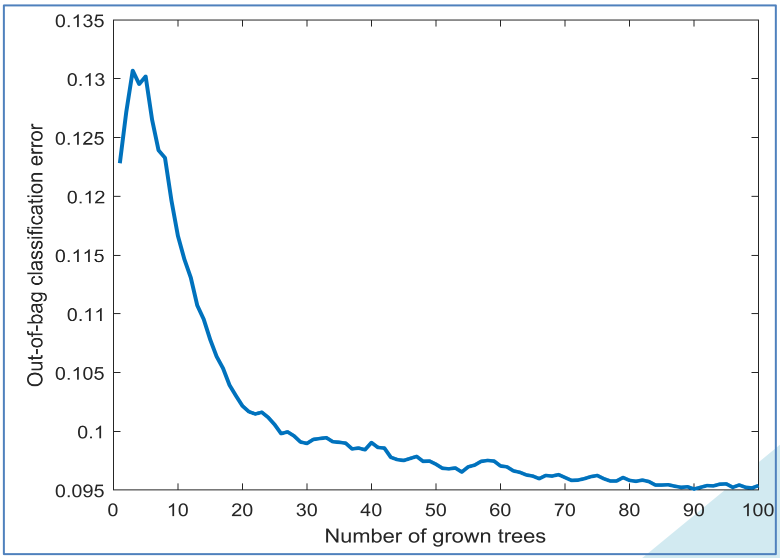


Fig. 7 Random Forest performance process per number of grown trees

6. Conclusion

The application of data mining techniques in the banking industry has resulted in the improvement of banking processes in advanced countries. However, there are many potential fields in applying this knowledge in our country's banks and financial institutions, making it essential to create a foundation for familiarizing banking experts and specialists with data mining techniques and their applications. Holding training courses in this field and the practical application of this science in the country's banks and other financial institutions is vital for their development and growth.

The study of Guimarães Bank of Portugal's data showed that the registered features of customers have a significant relationship with their future decisions in the banking system, such as opening a long-term deposit. The decision tree and random forest classification methods were used to predict the opening of long-term deposits based on these registered features of customers in the bank's database, and both methods achieved relatively acceptable accuracy. The comparison of the random forest algorithm with the traditional decision tree classification method's performance provided valuable insights for future research in bank customers' classification.

The results of this study have practical implications for the banking industry, as they can improve the decision-making process for banks and financial institutions. By utilizing data mining techniques, banks can extract valuable insights from their databases, which can help them better understand their customers' behavior and preferences. This information can be used to develop better products and services tailored to their customers' needs, leading to increased customer satisfaction and retention. Overall, this research contributes to the field of data mining in the banking industry, providing a foundation for further research and practical application in the future.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Conflict

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

References

1. Botshekan, M. (2000). *Predict stock price using neural-fuzzy networks and compare with linear prediction models by Rai and Mehrgan* Tehran University

2. Chen, Y., Peng, L., & Abraham, A. (2006). Stock index modeling using hierarchical radial basis function networks. In Knowledge-Based Intelligent Information and Engineering Systems. 10th International Conference, KES 2006, Bournemouth, UK, October 9-11, 2006. Proceedings, Part III 10, Bournemouth.

3. Dinçer, H., Yüksel, S., Canbolat, Z. N., & Pınarbaşı, F. (2 .(020*Data Mining-Based Evaluating the Customer Satisfaction for the Mobile Applications: An Analysis on Turkish Banking Sector by Using IT2 Fuzzy DEMATEL* Tools and Techniques for Implementing International E-Trading Tactics for Competitive Advantage ,

4. Hadipour, H. (2003). *Determine the best model for forecasting stock prices in food and beverage industry group of Tehran Stock Exchange*

5. Han J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. *Morgan Kaufmann Publishers*, *4*, 285-464 .

6. Hassani, H ,.Huang, X., & Silva, E. (2018). Digitalisation and big data mining in banking. *Big Data and Cognitive Computing*, *2*(3), 18-20 .

7. Ince, H., & Trafalis, T. B. (2006). A hybrid model for exchange rate prediction. *Decision Support Systems*, *42*(2), 1054-1062 .

8. Kamath, D., Pavithra, K., & Pujari, K. (2019). Data mining techniques applied in banking sector-A review. *International Journal of Social and Economic Research*, *9*(3), 358-365 .

9. Kanchana, M., Chavan, P., & Johari, A. (2020). Detecting Banking Phishing Websites Using Data Mining Classifiers. *EasyChair*, *2855* .

10. Kaur, M., Bhaddal, P., & Singh, G. (2019). Calculation of client credit risk prediction in banking sector using data mining. *International Journal of Advance Research, Ideas and Innovations in Technology*, *5*(3), 1341-1343 .

11. Kerkaus E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements, . *Expert Systems with Applications,*, *32*, 995-1003 .

12. Le Khac, N. A., & Kechadi, M. T. (2010). Application of data mining for anti-money laundering detection: A case study. 2010 IEEE international conference on data mining workshops ,

13. Mirfeizi, F. A. (2009). *Forecast five-year deposits of Tejarat Bank based on ARMA model by Maryam Khalili Eraqi and Kambiz Peikarjou* Islamic Azad University .[

14. Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. European Simulation and Modelling Conference-ESM'2011 ,

15. Phua, C., Lee, V., Smith, K & ,.Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv*, *2*, 1009.6119 .

16. Qavamzadeh, M. (1997). *orecasting in organized trading markets" by guidance of Crolox* Tehran University .[

17. Rahnamaye Roudpashti ,F., & Mousavi, S. R. (2012). Comparing the power of classic time series models and artificial neural network in forecasting banks' liquidity needs. *Financial Engineering and Portfolio Management*, *3*(12), 17-37 .

18. Rajabzadeh Qatarmi, A. (1998). *Combined evaluation of forecasting methods and presentation of optimal model for stock price forecasting in stock exchange* Tarbiat Modares University .[

19. Sarafaraz, L., & Afsar, A. (2005). Investigate factors affecting gold price and present forecast model based on fuzzy neural networks. *Economic Research Magazine*, *7*, 35-42 .

20. Shahbazi, F. (2020). Using Decision Tree Classification Algorithm to Design and Construct the Credit Rating Model for Banking Customers. *IOSR Journal of Business and Management*, *21*(3), 24-28 .

21. Sinaii, H. A., Mortazavi, S., & Teimouri Asl, Y. (2005). Forecast Tehran Stock Exchange Index Using Artificial Neural Networks. *Economic Research Magazine*, *7* .(32-25)

22. Stoviček, K. (2007). *Forecasting with ARMA Models: The case of Slovenian inflation*. Elsevier .

23. Wang, T.-C., & Lee, H.-D. (2006). Constructing a fuzzy decision tree by integrating fuzzy sets and entropy. *WSEAS Transactions on Information Science and Applications*, *5*(2), 25-32 .

24. Yu, L., Lai, K. K., & Wang, S. (2008). Multistage RBF neural network ensemble learning for exchange rates forecasting. *Neurocomputing*, *71*(16-18), 3295-3302 .

25. Zhang, D., & Zhou, L. (2004). Discovering Golden Nuggets: Data Mining in Financial Application. *IEEE Transactions on Systems*, *34*(2004), 513-522 .