Bank predictions for prospective long-term deposit investors using machine learning LightGBM and SMOTE

M A Muslim1,2*, Y Dasril2, A Alamsyah3 and T Mustaqim3

1 Postgraduate Student, Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia.
2 Faculty of Technology Management, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia.
3 Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

*Corresponding author: a212muslim@mail.unnes.ac.id

Abstract. Banks try to get profit from society in various ways. One way is to use long-term deposit investment offers. If the product offering process for potential investors is not carefully considered, it will waste resources. Therefore, this study analyzes the accuracy of the predictions of consumers who have a high chance of participating in this program. The dataset used is historical bank data provided by Kaggle. In previous research, accuracy prediction has been carried out, but the accuracy is still low because it does not use a method to balance the class. Better accuracy can be improved using LightGBM and SMOTE methods. The test results with the number of testing data as much as 6590 and training data as many as 32950 show the highest accuracy of 90.63%.

1. Introduction

Banks try to get profit from society in various ways. One of them is to use long-term deposit investment offers [1]. Long-term investment deposits are savings deposits that can only be taken according to the maturity date. The maturities ranged from 1, 3, 6, 12, and 24 months. The offering to consumers requires a variety of methods i.e. direct offering, email, cellular, and telephone [2]. If the process of approaching consumers is not right, it will waste bank resources. Banks need a special approach to consumers by paying attention to and analyzing previous data [3]. This will save resources and a more targeted approach.

Prediction of historical bank data can be made using data mining which is essentially formed by machine learning [4-8]. Bank historical data is formed in such a way that the patterns contained therein are found. The discovery of these patterns uses mathematical calculations by models from machine learning [9]. These patterns are found from bank history data features such as type of loan, marital status, occupation, age, and others [9]. The end result of pattern detection is the predictive level of consumers with the highest opportunities for long-term deposit investment. Several methods that can be used to make predictions are classical logistic regression [11-12, 14, 20], Random Forest [3, 13, 15-17, 21] and LightGBM [18-19, 29]. According to Sun et al. [22] Unequal class data in the dataset is balanced using SMOTE. This research aims to improve the prediction model that has been carried out with the neural network by Panigrahi and Patnaik [23] was resulted the highest accuracy of 90.02%. The method modification process is applying class data balancing using SMOTE in logistic regression, random forest, and LightGBM. The novelty shown in this study is that the level of prediction accuracy is higher than in previous studies.
2. Related work

Elsalamony [20] used several kinds of data mining methods such as MPLNN, TAN, LR, and C5.0 on the direct bank marketing dataset to classify deposit subscriptions. The results obtained are the c5.0 method has a better performance than MLPNN, LR, and TAN. Abbas [3] researched in 2015 to increase the efficiency of marketing procurement and assist the decision-making process using the Rough set theory and decision three. The results obtained, namely, the Rough set theory method, produces better data due to its pross feature reduction.

Ghosh and other [24] conducted research using several types of data mining methods such as MPLNN, Decision Tree, and SVM to improve the accuracy of user deposit predictions. The decision tree obtains the highest accuracy results compared to SVM and MPLNN. Musunur et al. [21] 2020 analyzed the use of features in bank datasets to increase accuracy. The data mining methods used are k-NN and Random Forest. The results of this study are that the highest level of accuracy is obtained by using the random forest as modeling and chi-square as its feature selection.

3. Methods

The research dataset is in the form of bank history data which can be downloaded publicly at Kaggle. This dataset contains 32,950 lines and 16 different features. The features available are age, job, marital, education, default, housing, loan, contact, month, day_of_week, duration, campaign, pdays, previous, income, y(label). Details of feature types are shown in Table 1.

| Features          | Features Type |
|-------------------|---------------|
| Age               | Numeric       |
| Job               | Categorical   |
| Marital           | Categorical   |
| Education         | Categorical   |
| Default           | Categorical   |
| Housing           | Categorical   |
| Loan              | Categorical   |
| Contact           | Categorical   |
| Month             | Categorical   |
| Day_of_week       | Categorical   |
| Duration          | Numeric       |
| Campaign          | Numeric       |
| Pdays             | Numeric       |
| Previous          | Numeric       |
| Outcome           | Categorical   |
| Y (label)         | Categorical   |

Prasetiyo, et al. [25-26] Data preprocessing means cleaning the dataset from data that does not support the analysis process. Data that do not support the analysis process are duplicate data, blank data, and abnormal data. According to Moro et al.[9-10], features of the dataset that are not relevant to the analysis process are deleted. While, Muslim et al. [27] said that the data modeling is the process of searching for patterns contained in the dataset. The pattern that is sought uses several mathematical calculations from the machine learning algorithms used. The resulting pattern is then tested on the test data to determine the resulting level of accuracy.

SMOTE (Synthetic Minority Over-sampling Technique) is a development method of the oversampling concept whose work process produces synthetic samples based on object characteristics and k-NN. SMOTE increases the number of minor class data to balance the major class. Sun et al. [22] said that the function of
SMOTE is to make the data class balanced and improve the quality of the dataset for modeling purposes. The SMOTE algorithm can be described in pseudocode as shown in Figure 1 [27].

Asare-Frempong and Jayabalan [29] said that The logistic regression algorithm works to find the probability of data from previously known data. Logistic regression has the advantage of being able to find a relationship between the dependent variable and the explanatory variable. The process of finding the probability value starts with the initialization of previously known data. Musunur et al. [21] said Random forest is a machine learning method whose work process combines several decision tree algorithms. The process of determining the results of the random forest algorithm is by using the voting method. The voting method is taken from the dominant results generated from each decision tree algorithm. Ke et al. [30], the LightGBM machine learning algorithm is tree-based and is included in the gradient boosting framework. This algorithm has the advantage that it can be used to analyze data with high dimensions, large datasets, and a more efficient training process. The LightGBM is a boosting type that has three steps as follows [31]:

For simplicity, \( X \) is given as a pre-processed streaming data set.

Step 1: Initialize the weak learner by using Eq. (1)

\[
f_0(x) = \text{argmin}_c \sum_{i=1}^{n} L(y_i, c)
\]

where:

\( f_0(x) \) as the weak learner basis function,

\( L(y_i, c) = L(y, f(x)) = (y - f(x))^2 \) as the function of loss,

\( n \) as the amount of samples.

Step 2: Calculate weak learners \( M \) times, iteratively.

a) For the sample \( x_i \in X \) (\( \forall i = 1,2,\cdots,n \)), calculate the negative gradient of loss function evaluated in the existing model as follows

\[
r_{mi} = -\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}|_{f(x)=f_{m-1}(x)}
\]

where \( r_{mi} \) as negative gradient of the loss function

b) The residual \( r_{mi} \) resulted is taken as sample new real value. Fit a regression tree for \( \{ (x_1, r_{m1}), \cdots, (x_n, r_{mn}) \} \) and make a new regression tree \( f_m(x) \)

c) Calculate the best-fit value of the leaf area \( j = 1,2,\cdots,M \). By using \( c_{mj} \) as linear search to predict leaf node region value for minimizing the loss function.

\[
c_{mj} = \text{argmin}_c \sum_{x_i \in R_{mj}} L(y_i, f_{m-1}(x_i + c)), \quad i = 1,\cdots,M
\]

d) Update the robust learner.

\[
f_m(x) = f_{m-1}(x) + \sum_{i=1}^{l} c_{mj} I(x \in R_{mj})
\]
where:
\( f_m(x) \) as the existing weak leaner,
\( f_{m-1}(x) \) as a pre-weak leaner,
\( I \) as the indicator function with 
\[
I = \begin{cases} 
1, & x \in R_{mj} \\
0, & x \notin R_{mj} 
\end{cases}
\]

Step 3: Determine the final regression tree by using Eq. (2)
\[
F(x) = \sum_{m=1}^{M} \sum_{j=1}^{C_m} c_{mj} I(x \in R_{mj})
\]

The significance of a feature is calculated as the normalized total reduction of criterion brought by that feature. It is also known as the Gini significance. Gini is denoted by \( Gini(p) \) in Eq (3).
\[
Gini(p) = \sum_{l=1}^{L} p_l (1 - p_l) = 1 - \sum_{k=1}^{L} p_k^2
\]

where:
\( L \) as the number of labels
\( p_k \) as the weight of \( l \)-label

4. Results and discussion

The testing process is carried out using google colab. This is a data analysis tool based on python language and powered by google. The data that has been taken from Kaggle is analyzed by exploratory data analysis. The goal is to find out the description of the data. The descriptions in question are data types, data diversity, unique data values, and missing data checks. The exploratory data analysis results were no missing data was found, and the amount of data was 32,950 and 16 features. The dataset label is in Features "y" showing the output of the previous data.

The process of checking data balance is carried out to determine the comparison between the Features label's data. Data balance is needed in the modeling process to avoid invalid results. The process of checking the data balance resulted in 29,238 labeled "no" and 3,712 labeled "yes". There is a large gap between the labels "yes" and "no" with a ratio of 89:11. This can clearly affect the resulting level of accuracy in the modeling process.

SMOTE is used in this study to balance the data class. The results of the SMOTE process consisted of 50,692 "yes" and "no" labels, respectively. This new data label can be used in the modeling process.

Classes that have been balanced are then divided into 80% training data and 20% testing data. The process of sharing the data was taken randomly with 42 random seeds. Several models of machine learning methods were used in this study as a comparison and reinforcement of the final result. The methods include logistic regression and random forest. Logistic regression uses parameter C of 1, random forest uses criterion parameter in the form of Gini, and lightGBM uses a parameter of boosting in the form of gbdt. The process of determining parameters using the GridSearchCV python library is equipped with cross-validation. The final results of each test data testing accuracy of 6590 are shown in Table 2.

| Algorithm         | Accuracy  |
|-------------------|-----------|
| Logistic Regression| 88.89%    |
| Random Forest     | 90.34%    |
| LightGBM          | 90.63%    |

From the results of Table 2, it is shown that the highest accuracy is obtained by LightGBM of 90.63% with the number of correct classifications of 5973. The accuracy level of the random forest has a slight difference with LightGBM with a range of 0.29% with the number of correct classifications of 5954. The use of logistic regression with a difference from LightGBM of 1.74% with the number of correct classifications of 5858. Each number of wrong classification results, namely logistic regression of 732 with an accuracy of 11.11%, random forest of 636 with an accuracy of 9.66%, and LightGBM a total of 617
with an accuracy of 9.37%. The results of precision, recall, and f1-score are shown in Figure 2 and Figure 3.

Figures 2 and 3 show the accuracy results from the precision, recall, and f1-score of each machine learning algorithm. There are several different results such as the precision label "no" LightGBM has the highest accuracy and the highest precision label "yes" is obtained by random forest. The highest accuracy result of recall label "no" is obtained by logistic regression and the recall label "yes" is obtained by LightGBM. The highest accuracy F1-score on the “no” label is indicated in series by LightGBM and Random forest then the F1-score for the label “yes” is indicated by LightGBM. On average, the LightGBM algorithm has the best level of accuracy from logistic regression and Random Forest. The results will be strengthened by the appointment of the Confusion matrix table in Table 3, Table 4, and Table 5.

Analysis of the data in Tables 3, 4 and 5 shows the highest prediction rate shown by the LightGBM algorithm with the number of "True Yes" as much as 355 and "True No" as much as 5618. In figure 2 shows the accuracy analysis of the label "no" with LightGBM superior to precision, low on recall, and draw on f1-score measurements. Figure 3 shows the accuracy analysis of the label "yes" with low LightGBM on the precision with a slight difference of 3 points with random forest but superior in recall and f1-score measurements.

![Figure 2. Detail measurement accuracy label "no"](image)

![Figure 3. Detail measurement accuracy label "yes"](image)

Table 3. Confusion matrix logistic regression

|       | True No | True Yes |
|-------|---------|----------|
| Pred. No | 5679    | 145      |
| Pred. Yes | 587     | 179      |

Table 4. Confusion matrix random forest

|       | True No | True Yes |
|-------|---------|----------|
| Pred. No | 5662    | 162      |
| Pred. Yes | 474     | 292      |

Table 5. Confusion matrix LightGBM

|       | True No | True Yes |
|-------|---------|----------|
| Pred. No | 5618    | 206      |
| Pred. Yes | 411     | 355      |

5. Conclusion
In this study, a new method that combines the LightGBM and SMOTE algorithms have been proven to increase the accuracy of bank predictions for potential long-term deposit investors. From the results of the accuracy of several algorithms tested, logistic regression resulted in 88.89%, random forest resulted in 90.34%, and LightGBM resulted in 90.63%.
Acknowledgments
The authors would like to express their gratitude and appreciation to the Universiti Tun Hussein Onn Malaysia (UTHM) through the research grant TIER 1 (H777).

References
[1] Jianguo C, Sai Z, Yongfan L and Kai L 2020 J. Serv. Sci. Manag. 13 435
[2] Marek G and Kornelius K 2020 J. Bank. Fin. 121 105961
[3] Abbas S 2015 Int. J. Comput. Appl. 110 1
[4] Mokrane S 2020 Int. J. Eco. Manag. Eng. 14 4
[5] Yaw M M and Ussiph N 2020 Adv. in Fuzzy Sys. 2020 1
[6] Jun W, Tao Y and Zhe Z 2021 Complexity 2021 1
[7] Ho S, Wong K, Yau Y and Yip C 2019 Machine Learning and Cognitive Science Applications in Cyber Security (Canada: University of Manitoba)
[8] Amruta S, Aphale, Sandeep R and Shinde 2020 Int. J. Eng. Resea. Tech. 09 991
[9] Moro S, Cortez P and Rita P 2014 Neural Comput. Appl. 26 131
[10] Moro S, Cortez P and Rita P 2014 Decision Supp. Sys. 62 22
[11] Viswanathan P K, Srinivasan S and Harirhan N 2020 J. Emer. Mark. Fin. 19 226
[12] Petropoulos A, Siakoulis V, Stavroulakis E and Vlachogiannakis N E 2020 Int. J. Forecas. 36 1092
[13] Rahman M and Kumar V 2020 Int. Conf. Elec. Comm. Aero. Tech. 4 1196
[14] Le H H and Viviani J L 2018 Rese. Int. Bus. Fin. 44 25
[15] Rustam Z and Saragih G S 2018 Inter. Work. Big Data Infor. Sec. 12 81
[16] Hartini S, Rustam Z, Saragih G S and Vargas M J S 2018 Int. J. Artific. Intell. 10 1
[17] Carbo-Valverde S, Cuadros-Solas P and Rodriguez-Fernandez F 2020 PLOS ONE 15 e0240362.
[18] Ilyas S, Zia S, Nisa Z U, Butt U M, and Letchmunan S 2020 Int. J. Adv. Comp. Sci. Appli. 11 286
[19] Machado M R, Karray S and de Sousa I T 2020 Int. Conf. Comp. Sci. Edu. 14 1111
[20] Elsalamony A H 2014 Int. J. Comput. Appl. 85 12
[21] Musunuru S, Mukkamala M, Kunaparaju L and Ganapathi R N V 2020 A E3S Web Conf. 184 13
[22] Sun J, Lang J, Fujita H, and Li H 2018 Inf. Sci. 425 76
[23] Panigrahi A and Patnaik M C 2020 Int. J. Appl. Eng. Res. 15 253
[24] Ghosh S, Hazra A, Choudhury B, Biswas P and Nag A 2018 Lect. Notes Comput. Sci. 10934 268
[25] Prasetiy o B, Alamsyah, Muslim M A, Subhan, and Baroroh N 2020 J. Phys.: Conf. Ser. 1567 032022
[26] Prasetiy o B, Alamsyah and Muslim M A 2019 J. Phys. Conf. Ser. 1321 032016
[27] Faris H, Abuikurma R, Almanaseer W, Saadeh M, Mora A M, Castillo P A and Aljarah I 2019 Prog. in Artifical Intell. 9 31
[28] Muslim M A, Nurzahputra A and Prasetiy o B 2018 Int. Conf. Info. Comm. Tech. 2018 141
[29] Asare-Frempong J and Jayabalan M 2017 Int. Conf. Eng. Technol. Technoprener. 2018 1
[30] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma M, Ye Q and Liu T Y 2017 Adv. Neu. Infor. Proces. Sys. 30 3155
[31] Chu Z, Yu J and Hamdulla A Inf. Sci. 535 129