Interpretable Data Mining Approaches to Predict Term Deposits Subscriptions

Qingyang Chen*

Weatherhead of Management, Case Western Reserve University, Cleveland, Ohio, United States

*Corresponding author: qxc130@case.edu

Abstract. Under the covid pandemic, informatization has accelerated. Customer behaviors have changed dramatically in almost every industry, which requires companies and organizations to analyze the latest customer features and develop specialized marketing contents. Banks play a major role in the financial markets, and term deposit subscription is one of the most important products of banks. In this study, a Portuguese retail bank's telemarketing data on selling term deposits has been analyzed which includes 16 input variables. All 16 variables have been converted to numeric inputs by using one-hot encoding method. For testing, the dataset has been divided into train and test groups by 80% and 20% respectively. To make this study more interpretable, two tree-structured data mining models, decision tree and random forest classifier, has been trained and tested. The input variable called "Duration" can only be known after the outcome is known so it has been dropped when creating predictive models. Two feature importance histograms and three confusion matrics are generated to visualize the customer features and evaluate the models. Area under Curve-Receiver Operating Characteristic Score (AUC-ROC scores) are also computed to compare the accuracy of models. Testing results have proved that "duration" has a severe relationship with clients' subscriptions and should not be included in predictive models. Random forest classifier is better than decision tree and can generate comparable accuracy to previous models which included "duration." Balance and age are the top two influential customer features on term deposit subscriptions.

Keywords: Telemarketing; Machine Learning; Feature Importance; Predictive models; Random Forest

1. Introduction

The bank always plays an important role in the social economy. It earns profits by providing various financial services, and term deposits is one of them. Term deposits refers to a client's fixed investment that is saved to the bank and can only be taken out after the maturity date. The majority of a bank's revenue often derives from its term deposit services, which have the potential to yield exceptional profits for the institution [1]. Though term deposits are short-period loans, clients might choose to keep making deposits after maturity. It is crucial for clients and the financial institution to maintain long-term connections. In order to increase product subscriptions, it is becoming increasingly crucial for banks to successfully reach out to their customers. In this case, building cooperative and collaborative relationships between the bank and its customers is essential to keep long-term relationships with the clients, and customer selectivity is one important facet of customer relationship management [2]. In today's world, financial market operates based on the digital revolution and the management of information [3]. This paper will predict the clients' behavior during bank's marketing campaign and analyze important clients' features to improve the efficiency and effectiveness of the bank's marketing campaign.

There are several constructive studies regarding bank's marketing campaign and term deposits subscriptions. Data mining and machine learning techniques are widely used in those studies. In [1], five algorithms' accuracy, sensitivity, and specificity are compared, and Logistic regression model was chosen to predict the clients' behaviors. However, there was not much analysis toward clients' features and the models could be further tested in this study. In [4], the authors applied five algorithms and more detailed analysis are performed based on the models' confusion matrices, evaluation metrics (i.e. accuracy, precision, recall, F-1 score, AUROC score), and ROC curves. However, in this model, "duration" had been included when building the algorithms. As a matter of fact, duration cannot be known until a call, or marketing campaign, is performed, and after a call, whether the customer subscribes to the term deposits is known. The variable "duration" should not be included in a predictive model since it is highly related to the outcome and cannot be known beforehand. In [5], the study used data mining to build five clients clusters to help the bank to deliver different marketing contents to specific groups of people. The authors' suggestions were constructive, but the model could be further improved by removing the input variable "duration," and more tests and algorithms could be used to generate more precise models. Ling et al. performed a comprehensive analysis using data mining and algorithms to generate more efficient plans for direct marketing, which is similar to customer selectivity [6]. The problem is that their studies were performed twenty years ago, and the marketing method was still based on mails. In today's world, clients' features could have changed dramatically.

Those previous studies have applied many insightful machine learning and data mining techniques, and many useful suggestions are made. However, the predictive models can be further improved and tested to give better and more precise predictions. Also, the variable "duration" can be dropped to get a more precise and convincing predictive model. Moreover, to solve the limitations mentioned above, this study presents two feature importance reports using the random forest and decision tree classifier. As the previous studies have already tested many classifiers, this study would focus on the interpretability, and that's why random forest and decision tree classifiers are chosen to give feature importance results. The AUC-ROC Scores of these two classifiers are also compared and two confusion matrices have been displayed to develop the better predictive model.

2. Method

2.1 Dataset description and preprocessing

In this research, predictive models and charts will be built through Jupyter Notebook in Python. Several python packages have been applied in this study such as pandas, numpy, matplotlib. The source of the dataset used in this study came from UCI Machine Learning Repository [7]. The dataset originally collected long-term deposit selling data by a Portuguese retail bank from 2008 to 2013. There are 11,162 lines, 16 customer feature variables, and one output variable in this dataset. The output variable is whether the customer subscribed to term deposits. There are eight input variables related to customer personal information: age, job, marital, education, default (does the client has a default or not), housing, loan and balance. Other eight variables are information about the bank's last marketing campaign: contact, month, day, duration, campaign, pdays (number of days that passed by after the client was last contacted from a previous campaign), previous (number of contacts performed), poutcome (outcome of the previous marketing campaign).

In order to make this research has more interpretability, exploratory data analysis and several analyses from different aspects are performed to discover and explain the dataset features. To build predictive models, one-hot encoding method is used to transform the categorical variables into numerical inputs so that these variables can be processed by the model. Moreover, to test and compare the performance of the models, the dataset has been split into train and test sets by 80% and 20%. The variable "duration" has been eliminated when conducting the models. For "Duration", it is highly related to the output Y, since if duration = 0, then the output will be "no." The longer the bank talks to the customer, the customer is more likely to make term deposits since a longer duration represents higher interest. Also, the duration cannot be known until a call is performed, and after a call, whether the customer subscribes to the term deposits is known. Therefore, "duration" should be excluded when performing predictive model.

2.2 Decision tree

Decision tree classifier is a data mining method used to build classification system, and this method classifies the dataset into branch-like segments, which construct an inverted tree model [8]. A decision tree method approximates the outcome Y with a set of steps or stages. It divides the independent variable space into boxes and places a step above each box at the height of the average of the dependent variables y in that box. In the next stage, the decision tree will assign a score to each box. In this research, decision tree classifier will be performed through the "sklearn" python package. It will be applied to both train and test sets. To evaluate the performance of this classifier, the histogram of the feature importance, confusion matrix and Area under Curve-Receiver Operating Characteristic Score (AUC-ROC score) are presented. AUC-ROC score is the AUC, which represent the area percentage under the ROC curve, and it is generated using sklearn functions so that no extra computation is needed. This metric is used to measure the accuracy of the model.

2.3 Random forest

In addition to the decision tree method, random forest classifier will also be applied in this research. Random forest classifier is also a tree structure method. It is a collection of many strong trees, and it process each tree separately. The final prediction outcomes of the random forest classifier will be combined by averaging or voting [4]. In this research, random forest method is also presented through the "sklearn" python package. The model will be built by the same process as building decision tree classifier. To better evaluate the feature, duration has been added to draw the feature importance histogram, and two confusion matrices, one with duration and one without, have been created. AUC-ROC score will also be presented.

3. Result and discussion

3.1 Performance of models

Fig. 1 represents decision tree classifier's confusion matrix on the testing set.

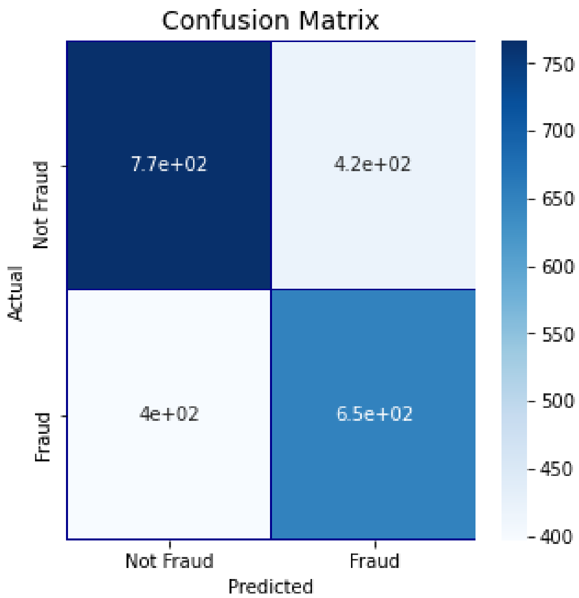

Fig. 1 Decision tree's confusion matrix

In the confusion matrix, the two areas marked blue are the correct predictions, which is true positive and true negative. The AUC-ROC score is 63.37%, which suggests that the decision tree classifier predicts 63.37% of the clients' behavior correctly. The model has a moderate accuracy rate compared to previous study's predictive model, whose accuracies were around 75% to 87%. This difference is as expected since the most influential input variable "duration" has been dropped in this predictive model.

Fig. 2 and Fig. 3 are the confusion matrices of the random forest classifier under two conditions (i.e. with and without 'duration' feature).

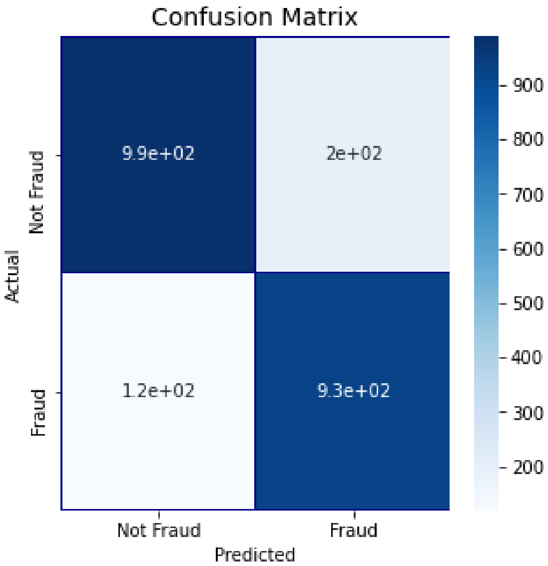

Fig. 2 Random forest’s confusion matrix (With duration)

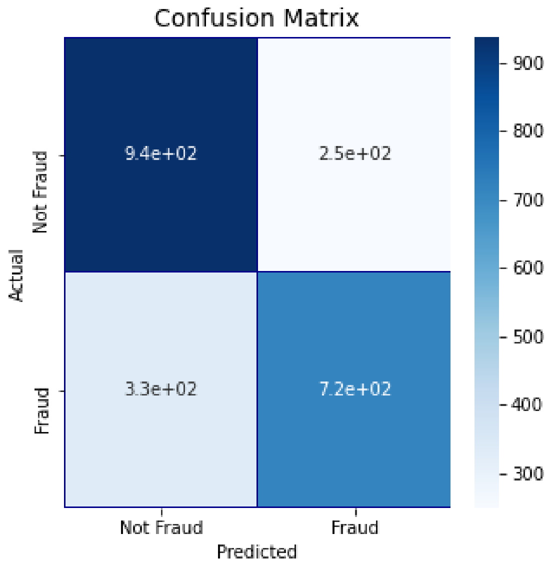

Fig. 3 Random forest’s confusion matrix (Without duration)

From the two confusion matrices, accuracy is computed to visualize the influence of duration. Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative). The accuracy is 85.7% when duration is considered and 74.1% after dropping duration. The AUC-ROC score for the random forest model, without duration, is 73.7%. From the differences of accuracies in two conditions, “duration” has a very strong relationship with the clients’ subscription, which proves that high accuracy from the previous predictive models might result from “duration.” Random forest classifier outperforms decision tree method in AUC-ROC score, which suggests that random forest classifier is a better predictive model for this dataset. Though 73.7% AUC score is still lower than previous models, this predictive model is more persuasive and applicable in predicting consumer behaviors in future marketing campaigns.

3.2 Feature importance

Fig. 4 and Fig.5 below are the feature importance result built by employed models.

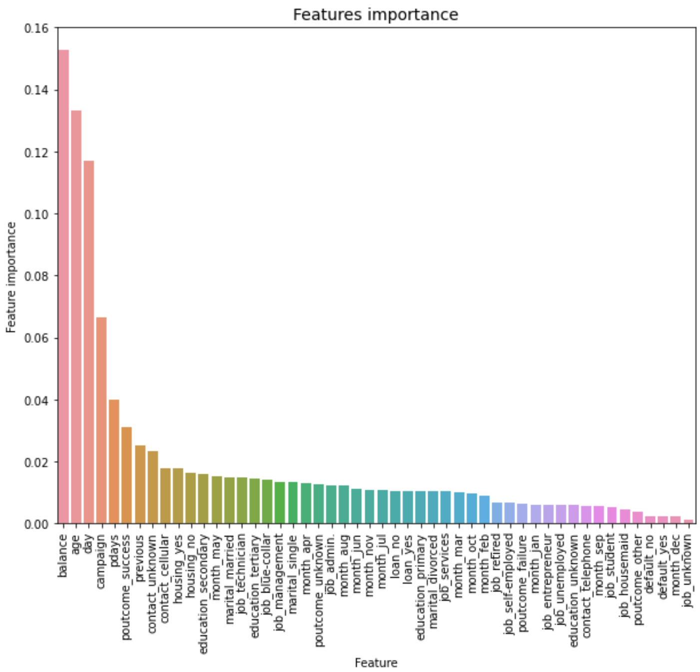

Fig. 4 Decision tree's feature importance

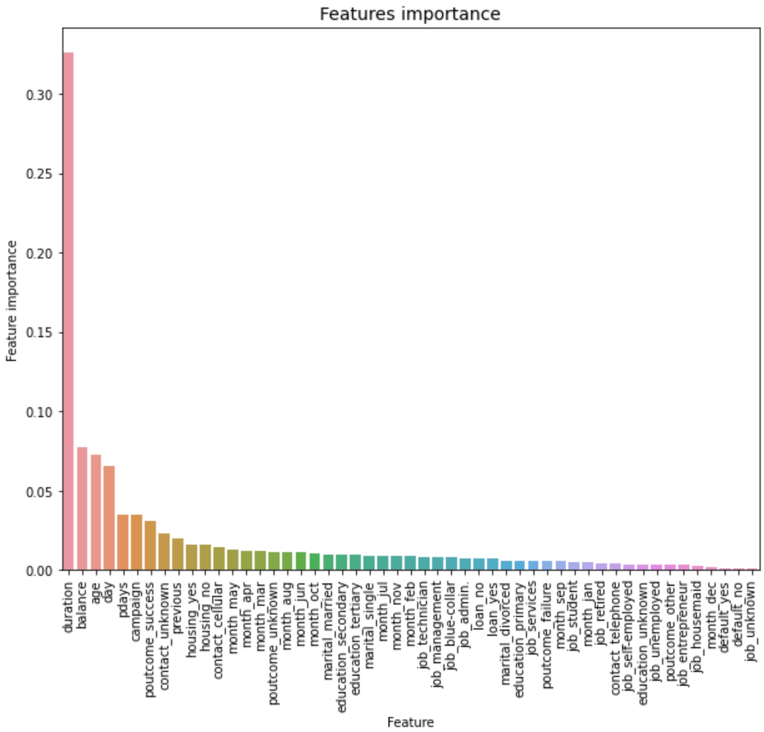

Fig. 5 Random forest's feature importance

From Fig.4, “balance”, “age”, and “day” are the top three influential features towards clients’ decisions on term deposits. Despite duration, random forest and decision tree classifiers present similar feature importance prediction. “Duration” is the most essential influencer when included. As Sheth stated, “another important facet of CRM is ‘customer selectivity [2].” The company therefore must be selective in tailoring its program and marketing efforts by segmenting and selecting appropriate customers for individual marketing programs. Balance and age information of clients may serve as initial indicators in future marketing campaigns. When performing Exploratory Data Analysis (EDA), clients with average balances were more likely to make deposits. Also, the previous marketing campaign targeted more clients with low balances. Moreover, through EDA, the youngest and oldest potential clients had higher possibilities to make term deposits compared to 30 to 50 years old clients. It is reasonable to target more potential clients in their 20s or younger and 60s or older in the next marketing campaign. Potential clients in this age range may share a higher possibility to agree to make term deposits. In addition, some more advanced machine learning models (e.g. neural network models) may also be considered in this case in the future for further improving the performance due to their excellent data processing capability in many tasks [9, 10].

4. Conclusion

In this study, two tree-structured classifiers have been used to create predictive models targeting bank’s telemarketing campaign with the analysis of potential clients’ behaviors. Feature importance histograms were generated to make the model more interpretable, and this study evaluated two models by generating confusion matrices and AUC-ROC scores. Through comparations between confusion matrices and feature importance histograms, this study proved the influences of “duration” on predictive models. The results showed that this predictive model created using the random forest classifier is applicable in the bank’s future marketing campaigns. Moreover, as duration, balance, and age are the top three influencers, the bank’s marketing department could consider methods to elongate the duration of marketing calls to increase clients’ subscriptions. Models creating customer clusters, and more related feature, could be generated in the future. It will be helpful for the bank to allocate marketing resources and focuses to generate loyal clients and improve their marketing campaign.

References

[1] M. A. T. Rony, M. M. Hassan, E. Ahmed, A. Karim, S. Azam and D. S. A. A. Reza, Identifying Long-Term Deposit Customers: A Machine Learning Approach, 2021 2nd International Informatics and Software Engineering Conference (IISEC), 2021, pp. 1-6, doi: 10.1109/IISEC54230.2021.9672452.

[2] Parvatiyar A, Sheth J N. Customer relationship management: Emerging practice, process, and discipline. Journal of Economic & Social Research, 2001, 3(2).

[3] Kotler, Philip, and Kevin Lane Keller. A framework for marketing management. Boston, MA: Pearson, 2016.

[4] Borugadda P, Nandru P, Madhavaiah C. Predicting the success of bank telemarketing for selling long-term deposits: An application of machine learning algorithms. St. Theresa Journal of Humanities and Social Sciences, 2021, 7(1): 91-108.

[5] Zhuang Q R, Yao Y W, Liu O. Application of data mining in term deposit marketing. Proceedings of the International Multi Conference of Engineers and Computer Scientists. 2018, 2.

[6] Ling C X, Li C. Data mining for direct marketing: Problems and solutions, Kdd. 1998, 98: 73-79.

[7] Moro et al., S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, 2014

[8] Song Y Y, Ying L U. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 2015, 27(2): 130.

[9] Yu Q, et al. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control 72, 2022, 103323.

[10] Patel Jigneshkumar L., and Ramesh K. Goyal. Applications of artificial neural networks in medical science. Current clinical pharmacology 2.3, 2007, 217-226.