Machine Learning Approaches for Retail Bank Marketing Practice

Yusen Wu [1, *]

[1]Questrom School of Business, Boston University, Boston, US

*Corresponding author: wuyusen@bu.edu

Abstract: The retail business is usually one of the most essential bank divisions that directly handles personal customers. As restrictions for COVID-19 are gradually phasing out globally, the market share of retail banking is also rapidly growing, which made improving current bank marketing practices necessary. However, most banks lack an objective and reliable approach to locating target customers. Meanwhile, they also need affordable and efficient ways to process the massive data generated in information systems to predict the purchase intent of their client. This paper introduced two machine learning algorithms to solve the issues mentioned above, which were developed on a dataset derived from the Portuguese bank's marketing campaign. Firstly, this study trained a Random Forest model to help select features to focus on when predicting client subscription on a term deposit. Secondly, an Artificial Neural Network model was implemented to evaluate the probability for an individual customer to purchase a bank product. The proposed Random Forest model successfully indicates that age and communication with banks are two of the most critical factors that drive customer subscription on term deposits. The proposed Artificial Neural Network also achieves an excellent accuracy of 91.82% in assessing whether a specific client will continue to purchase a specific financing product only at a loss of 18.88%. These results shed light on machine learning's promising applicability in retail bank marketing, which will vastly improve related costs and efficiency.

Keywords: Neural Network, Bank Marketing, Random Forest, Feature Importance.

1. Introduction

Retail banks are financial intermediaries targeting individual customers to provide services, including checking and savings accounts, debit and credit cards, personal loans, and so on [1]. Marketing is defined as the activities and processes carried out by a company to facilitate the purchase or sale of a specific product or service [2]. The global retail banking market is valued at USD 5,705.62 million in 2020 and is expected to expand at a compound annual growth rate of 10.28% over the next ten years [3]. In addition to the fast-growing Red Sea, retail businesses are also critical to the development of individual banks, often generating 50 to 60 percent of revenue for a typical large bank [4]. Since retail businesses target individual customers, personalized marketing for target clients is obviously indispensable. For most financial institutions, as the COVID-19 pandemic ends and personal consumption recovers, a diversified marketing strategy to drive personal business is now the cornerstone of their future development.

In most industries, including banks, the STP model is one of the most familiar strategic approaches in marketing practice to locate the target market. The model is developed by the well-known professor Phillip Kotler at Northwestern University, which follows the order of Segmentation Targeting, and Positioning [5]. Under this framework, commercial banks divide their customer pool into different segments based on 3 to 4 customer features, including gender, income, education, occupation, etc. Bankers will then focus marketing efforts on customers that satisfy all selected variables to advertise a specific product. However, the model is proven costly and inefficient in recent years [6]. Firstly, the selection of customer features to focus in STP is mainly based on experience and intuition, where important variables are usually ignored because of marketers' bias. Some researchers might examine if customers with specific attributes show more purchase history, but such testing is still non-exhaustive. Secondly, the STP model focuses only on customers that satisfy all selected attributes and assume that they have an identical purchase intent. People outside the pool are ignored, and personalized prediction of an individual's purchase likelihood is not given. The model is even more powerless facing the massive amount of data generated from banks' information systems.

In this regard, machine learning offers a promising solution for retail banks' marketing practice to locate target customers on an individual basis since it has been widely used in related customer data analysis [7-9]. Different machine learning algorithms offer different insights from the same datasets. Specifically for retail bank marketing, this study decided to adopt two machine learning algorithms to overcome the drawbacks of STP mentioned before. Firstly, Random Forest (RF) model can help analyze the importance of different customer features on an exhaustive and objective basis. Secondly, an Artificial Neural Network (ANN) developed based on these variables can predict the individual customers' preference for a specific product. Compared to the STP model, both algorithms can monitor customer behavior on a moving dynamic and produce measurable and reliable results. In the experiment on a dataset from a real commercial bank, the RF model successfully measures the importance of different variables compared to each other. The ANN model also generates an accuracy of 91% in predicting individual customer purchases on a specific bank product.

2. Methods

2.1 Dataset and preprocessing

The experiment is conducted on a dataset derived from a direct marketing campaign of a Portuguese banking institution. The bank has seen its revenue declining recently and would determine the action to take. After investigation, they found that the reason was that their clients did not invest enough in long-term deposits. Therefore, the bank would like to locate customers who prefer long-term deposits and increase marketing spending on these customers. In this paper, we would like to determine important features of the target client and predict if a specific customer will subscribe to a term deposit. The dataset contains 32, 950 observations, which are sourced from May 2008 to October 2010. Sixteen qualitative and quantitative features, including the target variables, are recorded like age, job, marital status, education, marital status, housing loan status, etc. Table.1. shows us five randomly selected samples from the original dataset, and Table.2. is the complete list of variables in the dataset. The original dataset can be accessed and downloaded from the UCI machine learning repository [10]. Since there is no missing value in the dataset, no additional data cleaning is needed. However, out of the 15 input variables, 10 are qualitative variables stored as strings like job and education. Therefore, for each string variable with N levels, this study created N dummy variables correspondingly as RF algorithms can process only numeric variables. After the normalization of these string attributes, 58 features in total are available as input values for the models. To further assess the reliability and accuracy of the proposed model, this study split the entire dataset into train and test sets on a ratio of 0.8 to 0.2. The train set would help train the model, while the test set would test the accuracy and applicability of the model facing real-world data.

Table.1. Samples from original dataset

age	**job**	**marital**	**education**	**default**	**housing**	**loan**	**contact**	**month**	**day_of_week**	**duration**	**campaign**	**pdays**	**previous**	**poutcome**	**y**
49	blue-collar	married	basic.9y	unknown	no	no	cellular	nov	wed	227	4	999	0	nonexistent	no
37	entrepreneur	married	university.degree	no	no	no	telephone	nov	wed	202	2	999	1	failure	no
78	retired	married	basic.4y	no	no	no	cellular	jul	mon	1148	1	999	0	nonexistent	yes
36	admin.	married	university.degree	no	yes	no	telephone	may	mon	120	2	999	0	nonexistent	no
59	retired	divorced	university.degree	no	no	no	cellular	jun	tue	368	2	999	0	nonexistent	no

Table.2. Description and types of variables in the dataset

Variable	Description	Type
age	age of client	numeric
job	job of client	string
marital	marital status of client	string
education	education status of client	string
default	has default credit history or not	string
housing	has housing loan or not	string
loan	has personal loan or not	string
contact	Contact communication type	string
month	Month of year for last contact	string
Day_of_week	Day of week for last contact	string
duration	Duration of last contact	numeric
campaign	Number of contacts during the campaign	numeric
pdays	Number of days after the client was contacted from previous campaign	numeric
previous	Number of contacts performed before the campaign	numeric
poutcome	Outcome of last campaign	string
y	Has the client subscribed a term deposit or not	string

2.2 Random Forest algorithm to assess important customer feature

Random Forest is an ensemble supervised machine learning algorithm widely used in classification and regression problems. The algorithm grows multiple decision trees based on different subsets of the original data. Variables from the original dataset are randomly selected to form the nodes of each decision tree. All the decision trees inside the Random Forest model can produce a prediction separately, and the final output will be measured based on the average of all these predictors. The randomness in both splitting dataset and choosing variables reduces the model's variance and avoids overfitting problems that are common in a single decision tree model, thus giving reliable results. However, a more critical application of the Random Forest model in this paper is its ability to assess feature importance in the dataset, which the Gini-impurity of each variable will measure. Gini-based feature importance is measured by Gini decrease across all trees when a specific variable is selected to split each node. The sum of the Gini decrease of this variable across all trees will be recalculated each time it was chosen. An average Gini impurity will be given after the sum has been divided by the total number of trees in the model. The higher the Gini impurity is, the more frequently it was chosen when constructing the model, and the more important a feature is for predicting the target variable. The RF model consists of 100 independent decision trees, and its Gini-based feature importance score can help assess essential variables to consider when predicting client purchases of bank products.

2.3 Artificial Neural Network algorithm to predict customer subscription

Other algorithms like Decision Tree and Naïve Bayes have been developed on this dataset in other papers to predict client subscription, but most show limited accuracy. Therefore, this study deiced to build an ANN model to test if it is more predictive for individual customer purchases. Artificial Neural Network is a machine learning algorithm prevalent in classification and regression problems. The algorithm is inspired by animal brain network that is constructed by numerous connected neurons. Inside the model, inputted signals-real numbers-will be processed by the artificial neurons or nodes, and the output will serve as the input of the following connected neurons. The transformation from input to output through a neuron is calculated based on some non-linear functions. These connected neurons are usually accumulated as different layers that can receive input and transmit output from connected layers. The ANN model consists of 4 layers. An input layer consisting of 58 nodes will receive the 58 corresponding columns from the dataset to describe the customers. As followed, two hidden layers activated by a Relu function will learn from every aspect of the data while minimizing the cost function. The first hidden layer is constructed by 50 neurons, while the second consists of 25, where all computation will be completed. The final output layer activated by a sigmoid function will summarize all signals received from the previous layers. This layer consists of only one neuron that gives the final prediction on whether the customer will continue to subscribe to a term deposit.

3. Results and Discussion

3.1 Analysis of Random Forest

Based on the RF model, the first six most predictive variables on customer subscription on term deposit are duration, age, and campaign. The relative importance of all variables is portrayed in Figure 1. Clients' communication with banks apparently plays a more important role than their intrinsic feature. Duration has importance of 29.29% to predict customer subscription: the longer the last contact made, the more likely the customer is to subscribe on a term deposit. This is also reflected in Figure 2. The probability of a client subscribing to a term deposit also grows as the number of contacts increases during the campaign, which takes up a 4.72% significance in the model. Customers' communication with bankers can increase their subscription intent as they are more informed with the information gained from the bank. In addition, age of 11.07% importance is the most significant intrinsic customer feature to predict term deposit purchases. As shown in Figure 3, older people are more likely to subscribe to a term deposit. A possible reason is that the elderly has more purchasing power and have more robust demand for financing their excess capital. However, the situation might be different for other bank products as younger people might prefer financing products with a higher return rate than term deposits.

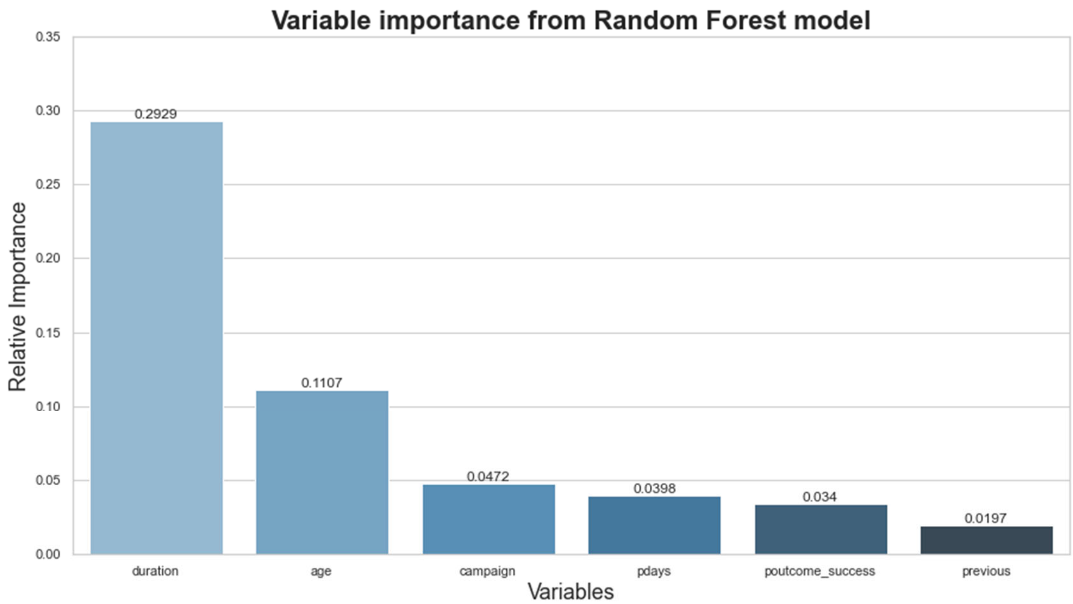

Figure 1. Variable importance from Random Forest model

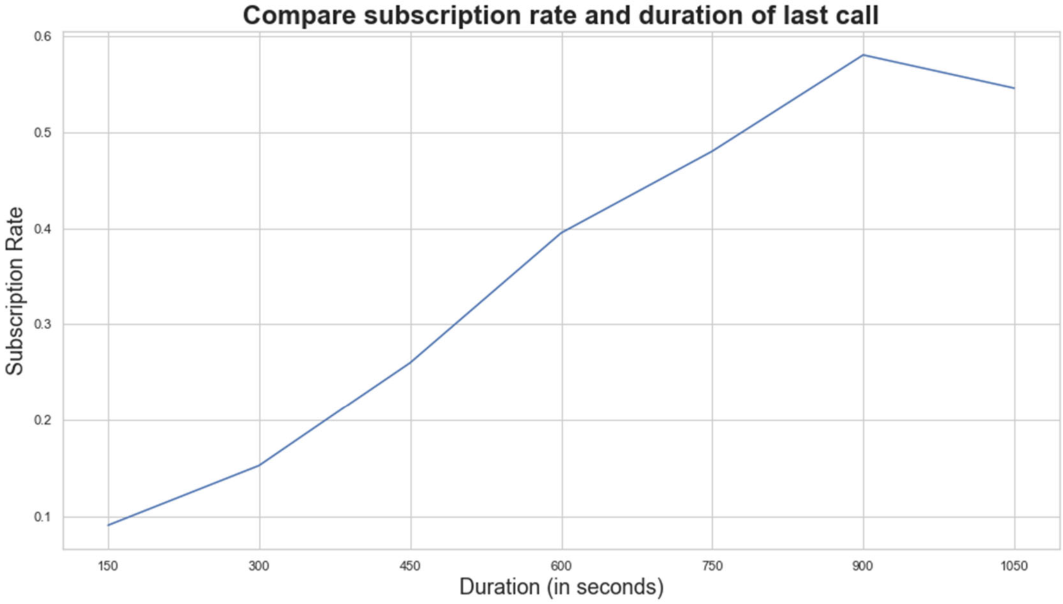

Figure 2. Compare subscription rate with duration of last contact

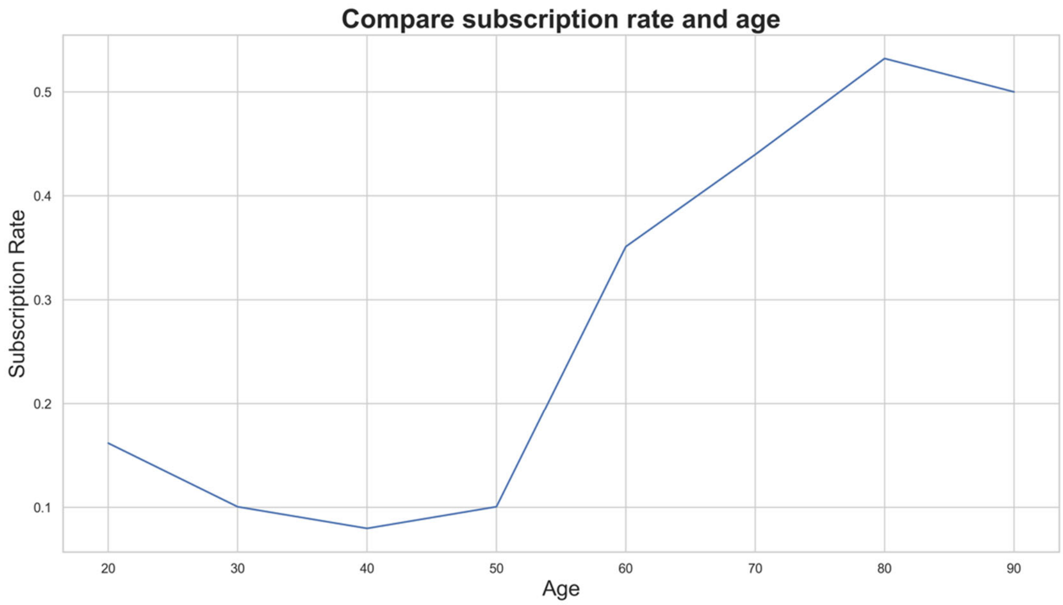

Figure 3. Compare subscription rate with age

3.2 Analysis of Artificial Neural Network

To measure the performance of the ANN model, this study selected Binary Crossentropy Algorithms as the loss function and Adam function as the optimizer. A total of 30 epochs are completed to train the model. The results show ANN's exceptional performance in predicting whether an individual customer will subscribe to a term deposit. The highest accuracy achieved by ANN on the test dataset is 91.82% out of these epochs, only at a loss of 18.88%. The model also reaches an accuracy of 91.02% on the training dataset, with a loss of 20.08%. In Table.3., this paper compares the performance of ANN with the model used in other papers trained on the same dataset, including

Naïve Bayes, Decision Tree, and Random Forest [11]. Neural Network has the highest accuracy compared to all these models, which is about 0.7% more accurate than the second-highest model (Random Forest). ANN's strength in dealing with non-linear data points might contribute to its excellence in predicting customer behavior. Some of the comparative algorithms mentioned before, like Naïve Bayes, usually rely heavily on a linear decision boundary, which is only suitable for linear-separable problems. However, each neuron in ANN can process data with some non-linear functions learned from training data, explaining why it outperforms other algorithms on this dataset.

Table.3. Accuracy results of different models.

Models	Accuracy Rate	Error Rate
Decision Tree	0.9002	0.0998
Naïve Bayes	0.8846	0.1154
Random Forest	0.9113	0.0887
ANN	0.9182	0.0818

4. Conclusions

This study proposed machine learning algorithms instead of the traditional STP model to find target customers for a specific product in a retail bank's marketing practice. This paper developed a Random Forest model to select essential factors to consider when predicting customers' subscriptions to a specific bank product. This study also implemented an Artificial Neural Network to predict an individual's related probability of continuing to purchase that bank product. The RF model indicates that customer communication with banks and their age can help predict whether they will subscribe to a term deposit. The ANN also enables us to predict their purchase intent on an individual basis with an accuracy of 91.82%, which outperforms all comparable machine learning algorithms. Both algorithms provide an objective and reliable way for banks to process their massive data to find a target market, which is costly or even impossible for STP to finish. However, a possible drawback in this paper is the lack of covariances analysis between different features, which might lead to a possible bias on variable importance. In the future, further study will try to improve the employed Random Forest model to conduct additional analyses on influences between different variables.

References

[1] Majaski, Christina. Retail Banking. Investopedia [R], 26 Aug. 2021, https://www.investopedia.com/terms/r/retailbanking.asp.

[2] Twin, Alexandra. Marketing. Investopedia [R], 10 Nov. 2021, https://www.investopedia.com/terms/m/marketing.asp.

[3] Verified Market Research. Retail Banking Market Size and Forecast [R]., Nov. 2021, https://www.verifiedmarketresearch.com/product/global-retail-banking-market-size-and-forecast-to-2025/.

[4] Clark, Timothy, et al. The Role of Retail Banking in the U.S. Banking Industry: Risk, Return, and Industry Structure [R], 1 Dec. 2007, https://www.newyorkfed.org/medialibrary/media/research/epr/07v13n3/0712hirt.html.

[5] Shi Biao, Li Yuxia, Yu Xhua, Yan Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model [J]. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[6] Hanlon, Annmarie. The Segmentation, Targeting, Positioning (STP) Marketing Model [R]. Smart Insights, Smart Insights, 17 Jan. 2022, https://www.smartinsights.com/digital-marketing-strategy/customer-segmentation-targeting/segmentation-targeting-and-positioning/.

[7] Martínez, A., Schmuck, C., Pereverzyev Jr, S., Pirker, C., & Haltmeier, M. A machine learning framework for customer purchase prediction in the non-contractual setting [J]. European Journal of Operational Research, 281(3), 588-596, 2020.

[8] Qiu, Y., Chen, P., Lin, Z., Yang, Y., Zeng, L., & Fan, Y. Clustering Analysis for Silent Telecom Customers Based on K-means++ [C]. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (Vol. 1, pp. 1023-1027). IEEE, 2020.

[9] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. A comparison of machine learning techniques for customer churn prediction [J]. Simulation Modelling Practice and Theory, 55, 1-9, 2015.

[10] Bank Marketing Dataset. https://archive.ics.uci.edu/ml/datasets/bank+marketing, 2014.

[11] Alexandra Jennifer, and Kristina Pestaria Sinaga. Machine Learning Approaches for Marketing Campaign in Portuguese Banks. 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS). IEEE, 2021.