Based on Machine Learning to Predict if the Client Will Subscribe the Term Deposit

Zhenyu Ni *

Department of statistic, University of California, Davis, United States

* Corresponding Author Email: zyni@ucdavis.edu

Abstract. This paper uses several approaches of machine learning to predict if the customers will buy the term deposit by telemarketing of the bank. The data is collected from the phone calls of a Portuguese banking institution from 2008 to 2012. The paper compared five models: logistic regression, k-nearest-neighbors, support vector machine, decision tree and random forest. The analysis revealed several key insights. The predictive significance of variables like duration, p-days, and previous was confirmed by their important effect on the desired outcome. The result of this study is focused on the accuracy, precision, recall, the F1-score, and the Area Under the Receiver Operating Characteristic (ROC) Curve. The logistic regression method has the best result among all the five models, because it has the highest accuracy and the largest Area Under ROC Curve (AUC). This shows a high level of ability to distinguish if the clients will subscribe to the term deposit. This study will give a valuable reference to the bank for precision marketing to their future customers.

Keywords: Machine learning; Logistic regression; Telemarketing.

1. Introduction

In recent years, telemarketing has become an important strategy for banks to promote financial products and attract potential customers. Through telemarketing, the bank can communicate directly to their potential customers without meeting face to face. The staff in the bank will introduce popular financial services and persuade the customers to consider various investment options on the phone calls. The centralized contact centers play an essential role in this process; they will streamline the coordination and management of marketing campaigns across the huge customer bases. These centers make their communication convenient through multiple remote channels, with telephone calls—whether mobile or fixed-line—remaining the most widely used and effective due to their immediacy and personal interaction [1].

Among the broad range of financial products offered by banks, the term deposits are the most popular one. This is probably because they provide customers with a stable and predictable return after a fixed period, making them a relatively low-risk investment option compared to more volatile financial instruments [2]. Term deposits are often favored by individuals finding secure and stable savings growth and those who have a long-term financial plan. Although the term deposit has many advantages, not all individuals contacted through telemarketing are willing to subscribe to such deposits. Customer decisions are influenced by many factors, like the education level of the customer, income stability, investment experience, convenience of the bank and the way the telemarketing staff communicates. James suggests that higher groups of educational level are more likely to take account of intrinsic attributes such as service features, the return rate and low service fees in their choice [3]. Omo reveals that the main reasons for customers' choice of banks are the funds safety and the availability of technology based service [4].

Therefore, it becomes essential for banks to develop a deeper understanding of customer characteristics and behavior. By analyzing customer preferences, needs, and past responses, banks can tailor their marketing approaches more effectively. This personalization can enhance customer satisfaction, improve target efficiency, and ultimately increase the success rate of term deposit subscriptions [5]. In the study which is closely related to this approach, Sergio analyzed propose generic social and economic indicators in addition to the more commonly used bank client and product attributes, in a total of 150 analyzed features. In the modeling phase, a semi-automated

process (based on business knowledge and a forward method) allowed to reduce the original set to 22 relevant features that are used by the data mining models [6].

Five approaches of machine learning are used to predict the success rate to make customers subscribe to the term deposit by the bank telemarketing. The data is collected from the phone calls of a Portuguese banking institution from 2008 to 2012. The five models are logistic regression, k-nearest-neighbors, support vector machine, decision tree and random forest. Accuracy, precision, recall, F1-score, and the AUC are the key factors this paper focused on. It is a binary classification task, and this study will give valuable reference to the bank for precision marketing.

2. Methodology

2.1. Exploratory Data Analysis

2.1.1. Data Validation

The data collected in the study is sourced from the dataset from the UCI Machine learning Repository. This dataset includes the direct marketing campaigns of one banking institution in Portugal. The campaigns of marketing were based on phone calls. In order to make sure if the term deposit of the bank would be ('yes') or not ('no') subscribed, over one contact to the same client was required to this study. The original data includes key indicators such as "Campaign", "Pdays", "Previous", and "Duration".

2.1.2. Data Distribution

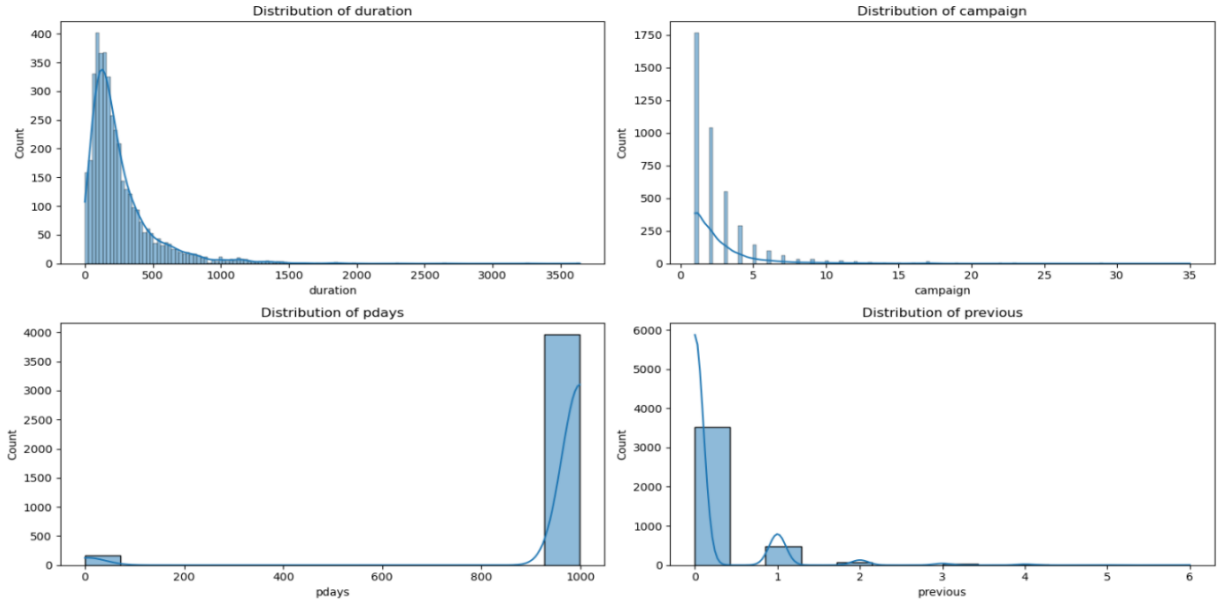


Fig. 1 Distribution of duration, campaign, pdays and previous.

The distribution of duration in figure 1 exhibits a pronounced right-skewness, with the majority of observations concentrated below 500 seconds. A small portion of the calls lasted longer than 3,000 seconds, which means a few conversations took much more time than usual. This shows that most marketing calls were fairly short, while the longer ones might have involved more engaged customers or more successful interactions. In Figure 1, the campaign variable shows a positively skewed pattern. Most customers were contacted by the bank only once or twice, and only a small number of them were contacted more than ten times in the figure. This plot suggests that calling the same customer more than once is not often happen and possibly leads to bad results. The distribution of pdays (the number of days since a customer was last contacted) in Figure 1 looks like bimodal. Many values of pdays are very close to zero, which means some customers were contacted recently, while another big group of values show 999, which represents that the customers who were never contacted by the bank before. The difference states that although some customers have a history of engagement, others

are completely new to the campaign. The previous variable, which means the number of earlier contacts before the current campaign, is also clearly right-skewed. Most customers' value is zero, which means that they are the first time and have had no previous interactions. Only a few customers have been contacted multiple times before.

2.1.3. Data Relationships

The correlation heatmap in Figure 2 shows how strongly the numerical and categorical variables in the dataset are related to each other. The range from the correlation values is from −1 to +1. Positive numbers mean the variables move in the same direction and have a positive relationship, while negative numbers mean they move in opposite directions and have a positive relationship. One clear pattern is the strong positive correlations among the economic indicators. Variables like emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, and nr.employed are all closely connected. For example, emp.var.rate (employment variation rate) has a very high correlation with euribor3m (0.97) and nr.employed (0.96). This suggests that these economic factors tend to rise or fall together. The duration variable shows only weak negative correlations with pdays (−0.09) and previous (−0.05). Although the relationship is small, it shows that customers who have been contacted more often or more recently might spend less time on the phone. There is also a strong negative correlation between pdays and previous (−0.59). This makes sense based on how the dataset is defined—clients who were contacted recently (lower pdays) would not usually have a high number of past contacts, so the two values naturally move in opposite directions. Additionally, several economic indicators display weak negative correlations with campaign-related variables, implying limited direct influence of macroeconomic conditions on individual campaign activities.

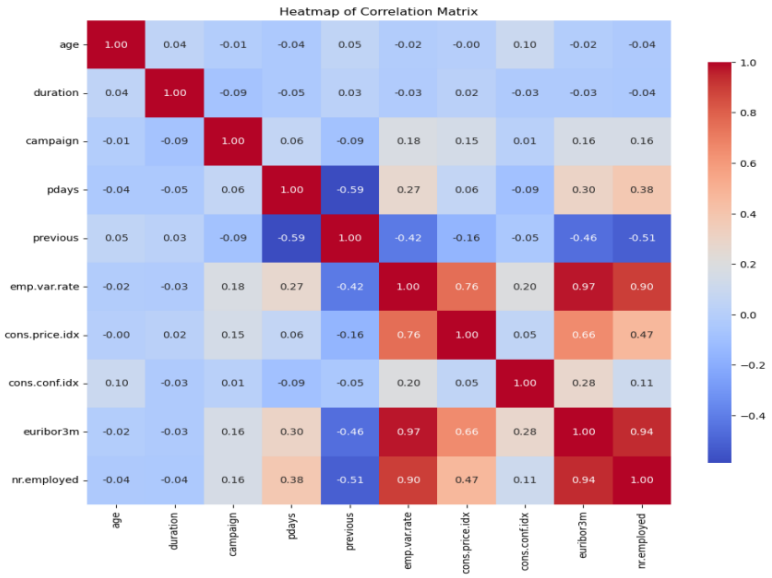


Fig. 2 Correlation heatmap.

2.2. Analysis of the Models

2.2.1. Models Selection

Logistic Regression models the probability that an observation belongs to a particular class by applying the logistic function to a linear combination of input features. This makes it particularly suitable for the tasks of binary classification such as predicting customer subscription behavior. Moreover, logistic regression allows direct interpretation of feature coefficients, providing insight into the magnitude and direction of influence of each predictor variable on the target outcome. K-nearest-neighbors method is to classify observations which are unlabeled by assigning them to the class of the most similar and closest examples [7]. Support vector machine is one of the classical machine learning approaches that may still help address huge data categorization challenges.

Especially, it can aid the multi domain applications in a huge data environment [8]. Support vector machine notably provides balanced prediction performance, even in investigations where sample numbers may be small, because it has relative simplicity and versatility for solving a range of classification difficulties [9]. Decision tree approach is a frequently used data mining method for constructing classification systems based on many variables or for developing prediction algorithms for a target variable. This approach classifies a population into branch-like segments that form an inverted tree with a root node, internal nodes, and leaf nodes [10]. Random forests are a mixture of tree predictors so that each tree depends on the values of a random vector selected separately and with the same distribution for all trees in the forest. As the number of trees in a forest increases, the generalization error for forests converges asymptotically to a limit [11].

2.2.2. Models Evaluation

Figure 3 illustrates the ROC curve for the logistic regression model. The ROC curve in figure 3 draws the True Positive Rate (TPR) against the False Positive Rate (FPR) across the various thresholds of classification. The value of 0.935 of AUC demonstrates that the model has a high level of ability which can effectively distinguish between customers (yes') who subscribed to the term deposit and the customers(no') who did not.

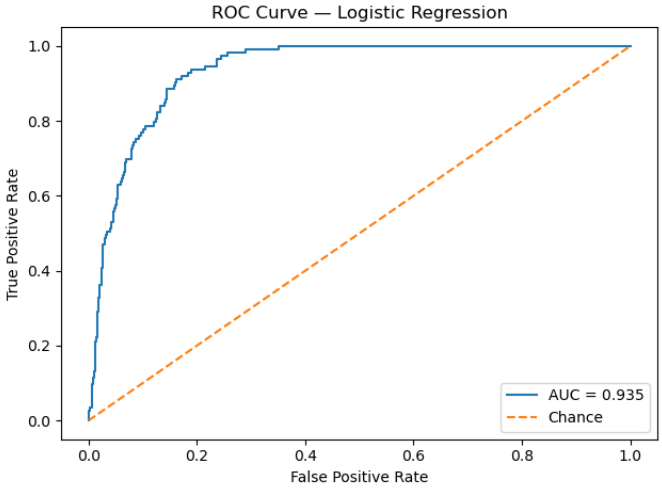


Fig 3. ROC curve of logistic regression model.

To check how well the prediction system worked, several classification models were built and tested using the same training and testing datasets. The models included logistic regression, decision tree, random forest, k-nearest-neighbors, and support vector machine. Their performance was measured using common evaluation outputs such as accuracy, precision, recall, the F1-score, and AUC. Among all the models, the logistic regression model showed the best results, had the AUC of 0.935 (Figure 3). This means that it performs well at telling apart clients who subscribed to the term deposit from those who did not. The random forest model produced slightly higher accuracy because it combines many trees predictors, but it is more difficult to interpret compared to the logistic regression model. The decision tree model often overfitted the training data, leading to high variance and weaker performance on the test set. The SVM model also performed well, but it needed a lot of parameters tuning and more computational power, which made it less convenient to use.

3. Conclusion

This paper compares five models of machine learning to predict if the customers will subscribe to the term deposit by telemarketing of the bank staff. After deleting the missing value and changing the categorical variables to numerical, the dataset was cleaned. Then the dataset is split to the training and test dataset. Although the other four methods have good performance, the logistic regression model performs the best with both the highest accuracy and AUC of 0.935. This shows that it can

clearly tell the difference between customers who subscribed to the term deposit and those who did not. The correlation heatmap also found important factors like duration, pdays, and previous which have effective impact of the result.

There are still many spaces to improve this study in future research. More advanced machine learning methods, such as gradient boosting models (XGBoost or LightGBM) and neural networks, could be added to increase prediction accuracy even further. These models may also help capture more complex, nonlinear patterns that logistic regression might miss.

References

[1] Yi F, Yin Y Q, Wang D J, and Dhamotharan L. A dynamic ensemble selection method for bank telemarketing sales prediction. Journal of Business Research 139: 368–82. 2022.

[2] Jiang Y B. Using Logistic Regression Model to Predict the Success of Bank Telemarketing. Journal of data science, 4, 35. 2018.

[3] Devlin J F. Customer knowledge and choice criteria in retail banking. Journal of Strategic Marketing, 10(4), 273–290. 2002. https://doi.org/10.1080/0965254022000014541

[4] Aregbeyen O. The determinants of bank selection choices by customers: Recent and extensive evidence from Nigeria. International journal of business and social science, 2(22), 276-288. 2011.

[5] Singh D K, & Goel N O O P U R. Customer relationship management: Two dataset comparison in perspective of bank loan approval using machine learning techniques. Journal of Theoretical and Applied Information Technology, 101(19). 2023.

[6] Moro S, Cortez P, & Rita P. A data-driven approach to predict the success of bank telemarketing. Decis. Support Syst., 62, 22-31. 2014.

[7] Zhang Z. Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11), 218. 2016. https://doi.org/10.21037/atm.2016.03.37

[8] Suthaharan, S. Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems, vol 36. Springer, Boston, MA. 2016. https://doi.org/10.1007/978-1-4899-7641-3_9

[9] Pisner D A, & Schnyer D M. Support vector machine. In Machine learning (pp. 101-121). Academic Press. 2020.

[10] Song Y Y, & Lu Y. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry. 2015.

[11] Breiman L. Random forests. Machine learning, 45(1), 5-32. 2001.