Bank Marketing Analysis ---- whether the client will subscribe a term deposit

DATA1030 Mid-Term Report----Qingyan Guo

1. Introduction

Over the course of the latest technology revolution and increasing globalization, one of the industries that is being influenced the most by the escalation of volume and variety of data is bank marketing. How to develop a feasible and optimized data pipeline for providing better profitability always cause huge pressure for banks due to intensive market environment and frightful industry competition. Alternatively, bank industry begins to implement the direct marketing campaigns (phone calls) to attract clients' attention to specific products or services.Bank industry comes up with an effective strategy to offer term deposit with appreciable interest rate via direct marketing campaigns (phone calls), however, it is necessary to improve the efficiency: lesser direct campaign should be tolerated, but a considerable number of successes (subscribing a term deposit) should be kept.

This project aims to do a binary classification task to predict whether the client will subscribe a bank term deposit based on the data of bank marketing (direct marketing campaigns using phone calls). The dataset consists of 4,119 data points with 20 independent variables and 1 target variables. The 20 features can be divided into basic client features, campaign related features, social and economic features, and other features. A data mining (DM) approach has been proposed to predict the success of bank telemarketing [1] where the researchers come up with the data-driven approach using the similar dataset as the one in our project. The researchers compare four DM models and summarize that the best result is presented by neural network (NN) with an AUC score of 0.8 and an ALIFT of 0.7.

Several works have used machine learning methods to improve the bank marketing campaigns. The authors introduce the concepts of 'confidence measurement' and use 'lift' as evaluation criterion to solve the problem of inadequacy of binary classification algorithms [2]. Machine learning algorithms are suitable to work on a classification problem where the task is to predict the label of a data point into target classes (i.e., "yes" or "no" in our project).

2. Exploratory Data Analysis

The dataset used in this project has 20 variables out of which 10 are categorical features and 10 are continuous features. We first investigate the target variable 'deposit', and we find that the data is imbalanced (figure 1). Then we explore 'duration' variable which is highlighted in description file (figure 2). As suggested in the file, we will drop this variable in our project. Then we have some interesting findings in both categorical features and continuous features.

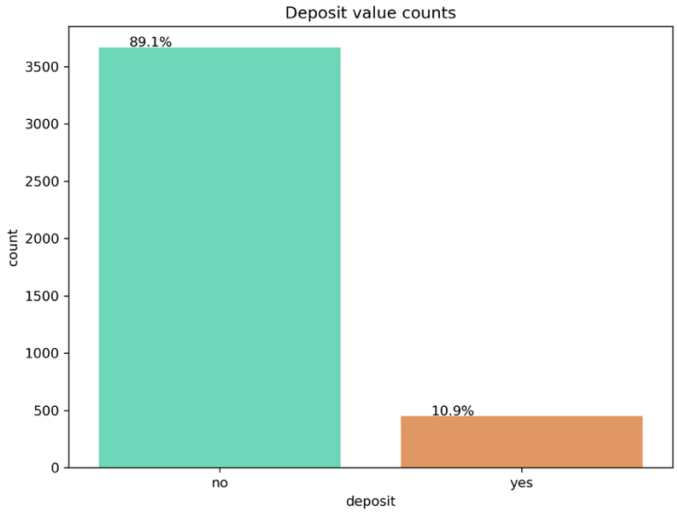

Figure 1. Counts for Deposit Value

Figure 1. This bar plot shows the distribution of target variable y (deposit). We can see from the above plot that the dataset is imbalanced, where the number of people who will not subscribe a term deposit is almost 8 times the number of ones who will subscribe a term deposit.

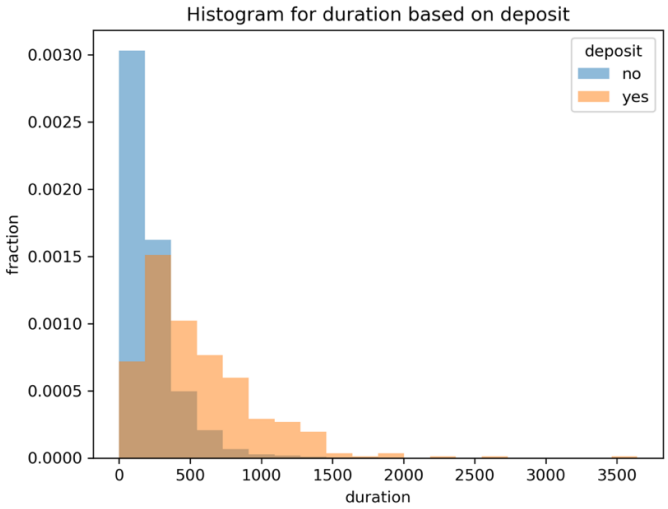

Figure 2. Histogram for Duration Based on Deposit

Figure2. This histogram denotes the duration of the last contact. As mentioned in the description of dataset, this attribute highly affects the target variable. As we can see that if the duration=0, then deposit='no'. Duration is not known before a call is performed. Also, after the end of the call, the result of subscribing a term deposit is determined. Thus, this input should be discarded if the intention is to have a realistic predictive model as the two are highly correlated and have direct causal relationship.

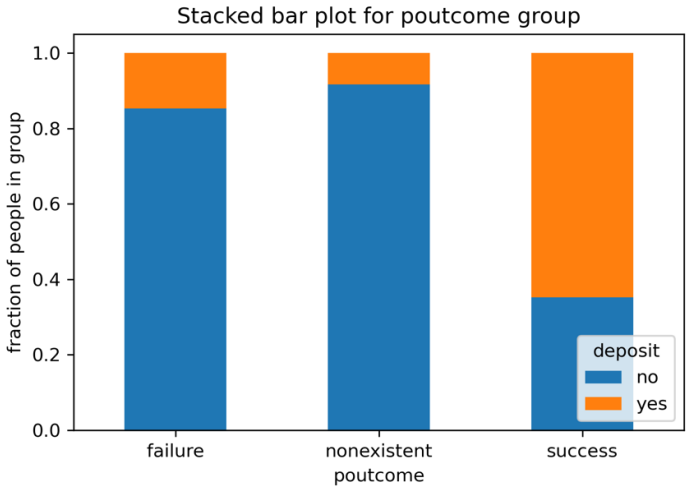

Figure 3. Stacked Bar Plot for Previous Marketing Campaign Outcome

Figure 3. For most of the customers, the previous marketing campaign outcome does not exist. It means that most of the clients are new clients who have not been contacted earlier. There is one thing to note here that, for the clients who had a successful outcome from the previous campaign, majority of those clients did subscribe for a term deposit.

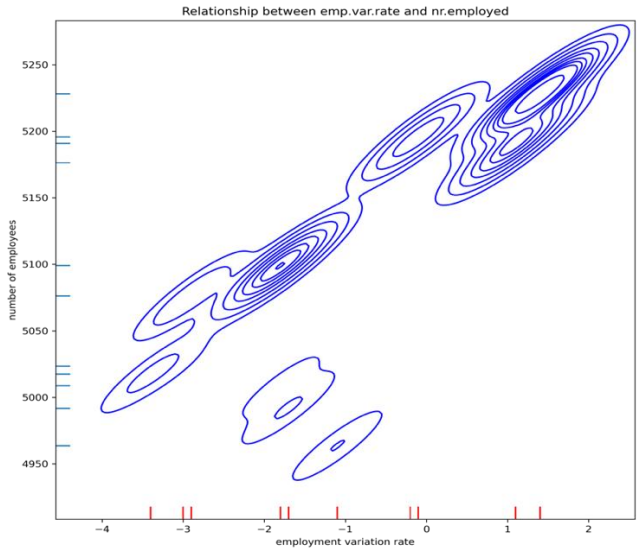

Figure 4. Relationship between Employment Variation Rate and Number of Employees

Figure 4. This density plot can determine the interdependency between 'emp.var.rate' and 'nr.employed'. The features employment variation rate and number of employees are both numerical variables and they are both related to employment. It is not surprised to find they have high correlations.

3. Machine learning methodology

3.1 Data Preprocessing

As each data point represents the basic information, campaign results and social and economic influences of one client, the bank market analysis dataset is independent and identically distributed without group structure and time-series properties. Adding the fact that the dataset is imbalanced, so we should apply stratified split to the whole dataset. In the stratified splitting step, we allocate 20% data points to the test set and the left 80% data points will be used in stratified k folds step to do the 5-fold cross validation. In each cross validation, the preprocessor will fit and transform on the training set first. As a result of exploratory data analysis, we drop the variable 'duration'. The variable 'duration' directly influences the label of target variable, or in other words, it can also be seen as the target variable. The 'duration' variable is just the duplication of the target variable 'y' (deposit). So currently, 10 categorical features and 9 continuous features left. There are no missing values and duplication data in the dataset, so we can just consider the 'unknown' category as a new category for corresponding categorical features. The feature 'education' is ordinal and categorical, so we will use OrdinalEncoder to it. For the left 9 categorical features, we choose to use OneHotEncoder. We use StandardScaler to all 9 continuous features. After we do these preprocessing works, there are 55 features in total. Also, we replace 'no' with 0 and 'yes' with 1 for the target variable 'y' (deposit) as it only has two labels. After all data preprocessing works all set, we visualize the high dimensional data in 2D (figure 5).

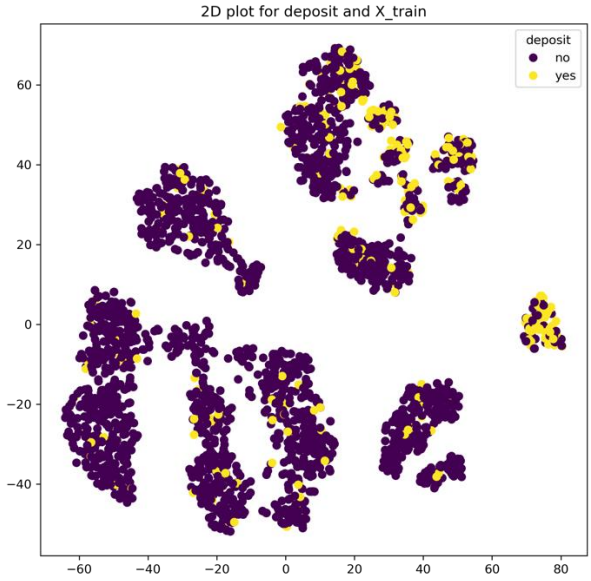

Figure 5. 2D Visualization Plot for Training Set

Figure 5. Use the training set as an example. Even though there seems to be some overlap in the data, there is also some distinction between the two classes in deposit.

Reference:

[1] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014), doi:10.1016/j.dss.2014.03.001.

[2] Ling, X. and Li, C., 1998. "Data Mining for Direct Marketing: Problems and Solutions". In Proceedings of the 4th KDD conference, AAAI Press, 73-79.

Github Repository: https://github.com/gfreya/DATA1030-Bank-Marketing-Analysis