Classification Prediction for a Bank's Marketing Outcome and Customer Acquisition Strategy

MQM 520Q Individual Final Project

Section A - Kangrui (Karry) Lu

Contents

Abstract

Financial institutions are highly invested in efficient marketing channels for customer acquisition and retention. The competition leads to the adaptation and utilization of machine algorithms for more effective marketing methodology and decision making. We developed a series of analysis to gain insights into the potential customers and previous campaign attempts, constructed a Random Forest Classification model to predict the outcome of current/upcoming campaigns. In this case, we specifically focused on whether target customers would enroll in the term-deposit that our client bank was trying to market. By a 10 folds validation, our model was able to cover a large range of information and achieved a prediction accuracy of 98.88% on average. Implementation of our model would greatly reduce marketing expenses and mitigate the risk of ineffective marketing attempts.

Business Understanding

The term deposit is one of the primary and most reliable resources for banks to profit. Unlike a regular deposit account that account holders may withdraw their money at any time in an unforeseeable manner, term deposit ensures a fixed period where the funds must stay in the bank. Therefore, the term deposit incurs a typically higher interest rate than a regular deposit account. However, with less uncertainty and higher incentives for customers to enroll (higher interest rate), banks will benefit from investing these funds in other financial products or lending to other clients with generally even higher rates. Consequently, banks have been putting effort into customer acquisition, which, in this case, is measured by whether the client will enroll/subscribe to a term deposit. In this project, the dataset limits customer acquisition as telemarketing, a direct and common way to reach/convert potential customers for a term deposit, given the target group varies in all attributes from age to education to marital and occupation.

To analyze the marketing strategy for a term deposit, we need first to define the most fundamental question: why some people choose term deposits over other investments? This piece of information will attach additional value to the customer background and campaign data we retrieved. The term deposit has a fixed rate of return or interest rate for a specified period, and governmental agencies, sometimes central banks typically back it. Meanwhile, it requires the lowest minimum deposit for investing, meaning the opportunities are most generous for many customers. Therefore, a term deposit is a risk-free, low barrier investment instrument that offers clear expectations for investors.

With the comprehensive information on the client, historical campaigns, and social, economic context, we constructed a model to predict if a specific client will subscribe, or putting it another way, the success of acquiring a potential customer. The model can instantaneously digest all available information and subsequently reduce risks and expenses by targeting 'wrong' customers. Thus, it will facilitate banks' marketing strategy, evaluating and improving customer acquisition.

Data Understanding and Preparation

The dataset was acquired from the UCI Machine Learning repository (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing). The data was collected initially from a Portuguese banking institution's direct marketing campaigns, which were based on telephone campaigns/telephone marketing records. The dataset was further enriched with the information on social and economic context attributes, which improved the robustness of the model and made the model adaptable for other markets/countries.

The data package comes with a training set (41188 records) and a test set (4119 records). There are 20 attributes/independent variables and one target variable in the dataset. The first seven attributes represent the clients' background information, such as age, type of job, education level, default records, etc. Attributes 8 to 15 relate to previous attempts for acquisition, such as contact method, previous contact month/day, the number of contacts performed for a client, the number of days from the last contact, etc. Attributes 16 to 20 are the five indicators for social and economic context, including the Consumer Price Index and Consumer Confidence Index, etc. The target variable is whether the client ends up subscribe or not (Binary: Yes/No).

Firstly, to have more flexibility, we combined the two datasets for data cleaning and exploratory analysis. Train/Test split will be conducted during model construction. Then, we transformed the dependent variable 'y' from Yes/No value to factor data type of 1/0 for analytical convenience. Also, we substituted the month name with the month number for wrangling and visualization easiness. The dataset was relatively clean, with no outstanding abnormal values. Additional calculations and columns' creation were performed during the exploratory analysis and modeling process, and details will be covered in those sections.

Exploratory Data Analysis

To gain in-depth understandings of the information, we utilized a top-down approach for exploratory data analysis, which proceeded from the general economic environment, the background profiling of customers, and the specific information regarding previous and current campaigns.

1. Monthly Economy Analysis with Deposit Enrollment

Banks need the information on whether the general economic environment would affect personal financial decisions, and they better prepare ahead for such uncontrollable fluctuations. To analyze and gain such insights, we calculated a data table (All previews for calculated tables are in Appendix) consisting of monthly economic indexes and the marketing campaign's enrollment information (There were no data for January and February). Firstly, the Consumer Price Index showed a steady level across the year (Figure - 1.A), signaling less volatility in price and tax level. Whereas the Consumer Confidence Index fluctuated throughout the year with a modest overall upward trend.

Looking at the term deposit enrollment information (Figure - 1.B), we can see that the enrollment rate, or in this case, the success of the marketing campaign slightly following the CCI index throughout the year. However, the bank has not invested in reasonable effort in the months that achieved a high enrollment rate (March, September, October, and December), which might be signaled by a relatively high CCI level. The bank outreached over 5000 target customers per month between April and August yet achieved their lowest enrollment rates of the entire year. Nevertheless, there should be reasons why the bank had such dramatically fewer attempts in particular months, bringing more insight if additional information is available. A possible explanation could be that the bank was preparing for the holiday season when people tend to withdraw rather than deposit their money for holiday expenses.

In the upcoming campaign outreaches, the bank should try to anticipate or predict the possible proxies for CCI, such as customer sentiment survey and major policy modifications, to adjust the volume of marketing attempts for more fruitful customer acquisition.

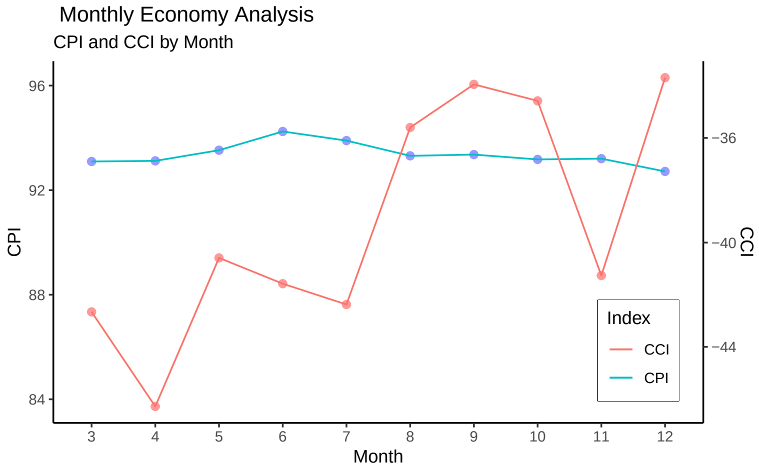

Figure − 1.A

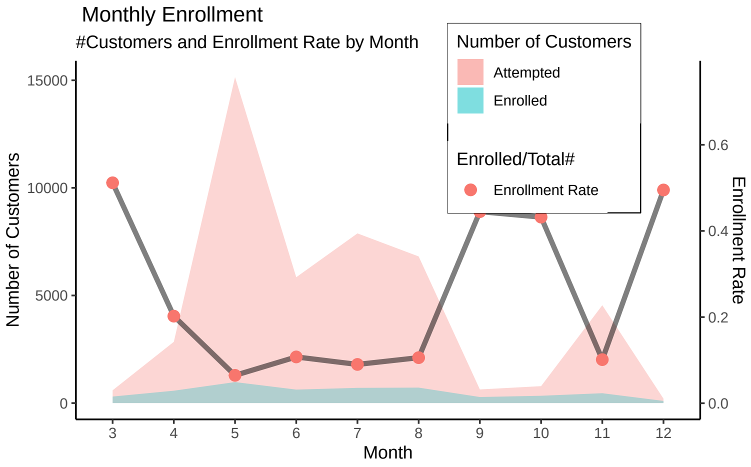

Figure − 1.B

2. Customer Profile - Background Information

Customer profiling is crucial for banks to understand their targets and subsequently be agile and precise in conducting marketing campaigns. We first built up the customer group's general picture and then proceeded deeper to relate their background information with the marketing campaign attempts and enrollment rate. Therefore, we could assess and make relevant suggestions for upcoming campaign strategies.

A. Basic Background Information Overview

In this section, we used customers' occupation as a benchmark attribute to gain insights into what kind of people we are facing and to adjust the strategy accordingly. The background information we have are age, education, and whether the customer has personal and housing loans (binary variables). We performed calculations on the information to get the average age, the proportion of people having loans by each job category. For education, we assigned education score from 0 to 6 according to the education level (i.e., illiterate: 0, basic.4y: 1, . . . , university.degree: 5, professional.course: 6), and assigned the average education score of 3 to the unknown group.

First and foremost, apart from students and retired customers on the two extremes, the other occupation groups are generally between 40 and 50 years of age (Figure - 2.A). Then, technicians with the highest educational background and blue-collar and housemaids are on the other end. The midpoint was 52.5% for housing loans, meaning the red/orange groups have over 52.5% percent of people with housing loans. The unemployed, entrepreneur, and technicians appear to have relatively high housing loans but less personal loans. Meanwhile, students and people in administrative positions seem to have a somewhat high volume in both loans. We can conclude that people who are still actively climbing in their careers tend to possess housing and personal loans (The red ones), and the ones who have settled in their careers are clearing out their loans, especially housing loans (The blue ones).

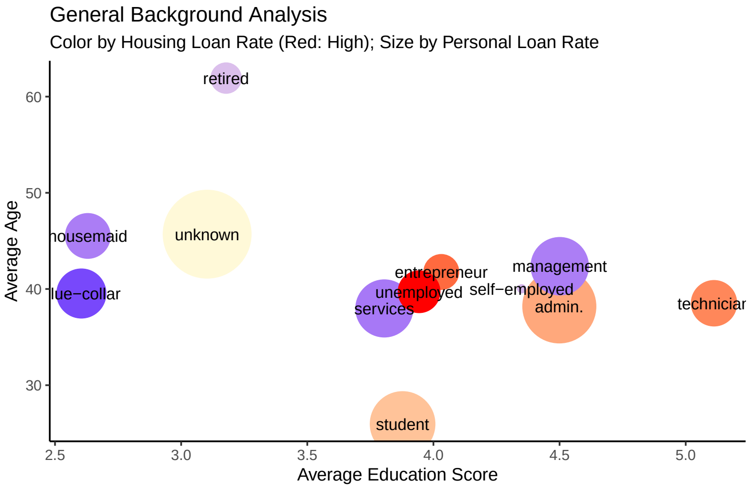

Figure − 2.A

B. Background Information Analysis with Enrollment Rate

After getting a primary picture of the customers, we then attached business value and insights from the marketing attempts and enrollment rate (Figure - 2.B). After calculating and visualizing loans against enrollment rate, we found no relations in-between, so we bring customers in the Marital status for further analysis.

The married group is dominating out of the people the bank targeted. Yet, single and divorced people appear to be more likely to enroll, possibly due to a sense of financial insecurity and in need of more reliable investment like term deposits rather than risky investments.

The bank has targeted heavily at people between 23 and 60 years old, and that is where most data records are. Yet, the enrollment rate in this age range was noticeably low compared with other age levels. There should be some underlying decisions for why the bank has not focused on people above 60 years old, since there was an abrupt drop in the number of market attempts around the age of 60, as shown in the Age graph. We assumed that people over 60 no longer need to put their money in the bank for a fixed period (and in many cases quite a considerable period) of time. Around the age of 60 or above, people may start to retire and generally begin to withdraw and enjoy whatever investment they made in their earlier careers. Thus, the bank has not treated this group as a potentially fruitful target. However, the enrollment rate was around three folds higher than the main target groups, so it might be worthwhile for the bank to re-package some short-term deposits for the elderlies and further test if this group is profitable.

The Occupation plot shows that the bank has targeted heavily blue-collar, self-employed, and administrative workers. These groups might have relatively unstable income sources, leading them to seek protective investments with low entrance standards. However, if this is sound logic, then the bank should have targeted unemployed and students on whom they have not yet spent significant effort. And surprisingly, these two groups, along with retired people, showed a high likelihood of term-deposit enrollment. The bank may concern the level of income and amount of deposit students, and unemployed people have available to put into the bank. These groups may either reply on their parents or live on paychecks. Therefore, the bank is recommended to acquire information on these customers further and cluster them into different groups and selectively approach more flexible offers.

The bank has focused on acquiring customers with relatively decent educations (from basic.9year to professional.course), and highly educated customers seem to bite the bait. This chart (Education) and the dataset do not tell us the income level, but we can conclude that financial literacy and investment awareness correlate with education level. Thus, we believed that the bank is in the right direction and may invest more in attracting people with advanced degrees and fulfilling their potential needs.

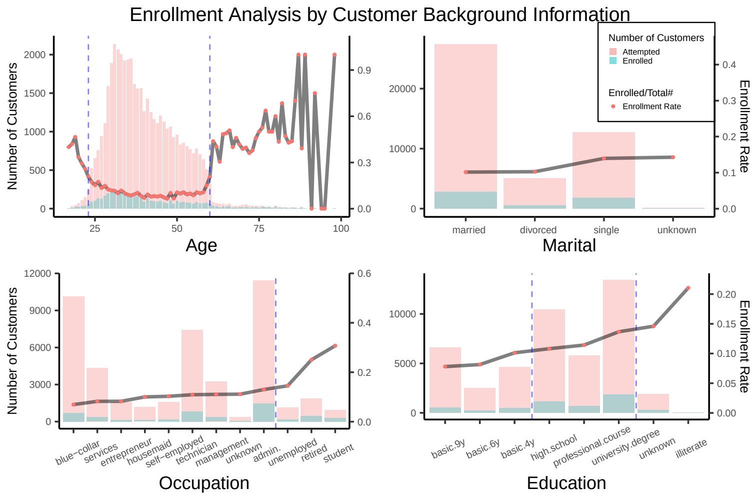

Figure − 2.B

3. Marketing Campaign Analysis

We have some information on previous and current campaigns, including the number of contacts performed for previous and current campaigns, the duration of the current campaign's last contact, and the previous campaign's outcome. Several interesting insights were drawn from this information.

According to the data source, duration may directly relate to the marketing outcome, given an investment decision is likely made after a long call. Therefore, we believed there are insights to be extracted from the duration of the last contact. If we can find something that positively correlates with duration as a proxy, we may indirectly find its impact on enrollment.

Firstly, we can see from the upper plot (Figure - 3) that successful acquisitions mostly have long durations but fewer contacts (mainly under eight times). Again, if account executive or relation managers can keep their prospective customers interested and on the phone for a longer time, they may have a higher chance of breaking through. Then, we further broke down the successful cases to see the duration distribution by previous campaign outcomes. It was intuitive that it made the bank's life easier to retain their old customers, who were acquired from the last campaign. The trust was already built up, and the customer's knowledge about this bank was sufficient. Therefore, calls for this group need not to be long. Yet, those who said no to the last campaign are hard for the bank to crack down. People who rejected last time rejected even faster this time (Red center of distribution is farther left in the 'no' category than 'yes'), and calls relatively took longer for them to enroll. Then, what about the new customers? Relation managers need to start from scratch for this group of customers, building up knowledge and confidence. For those who had no patience, the cold calls could not survive 20 seconds; and for those who had absolutely no need and interest, the calls generally ended within one minute. Again, for those who cared to listen or were actively looking for some sort of safe investments, the representatives would walk them through the precise terms and rules and possibly reach a deal at the end after a lengthy call.

We can see now how previous campaigns bring insights into how to distribute attention and workforce for different customers. Some experienced representatives who are highly skilled in elevator pitch better work on new and previous 'no' customers. To deal with those who turned down the offer last time, the banks should go back and visit their file and background information to see how they should go around some topic and switch to another approach. To further maintain the loyalty and strengthen the relations with old customers, the bank should lower the barrier by offering some promo deals for them during the new campaign.

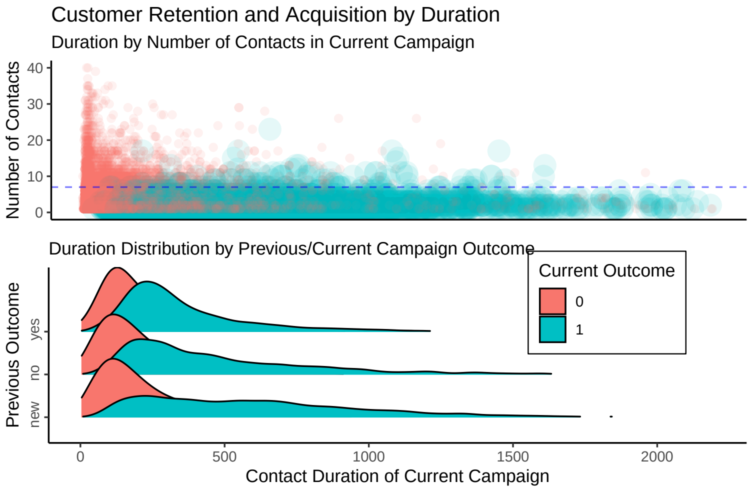

Figure − 3

Modeling

Each attribute for the customers presents some advice for adjusting marketing strategy, yet a firm needs to consider multiple dimensions in making cost-effective decisions comprehensively. Given our goal is to identify or predict whether a prospective customer will enroll, the target variable is binary (0: no, 1: yes). Therefore, we chose to implement Classification models in this project. We trimmed our dataset to eliminate variables with obvious multi-collinearity or potential selection bias, and modeled the target variable on 13 other variables.

1. Model Selection

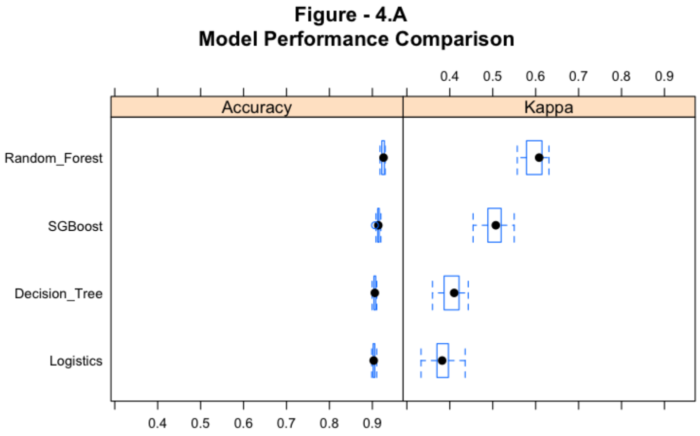

Figure 1: 4.A

Firstly, we constructed four models on the dataset, each with 10-Folds cross-validation to ensure a high and consistent performance level. The pack included Logistics Regression, Decision Tree, Stochastic Gradient Boosting, and Random Forest (tentatively 100 trees). We also considered Kernel SVM (radial), yet it did not work well given we have a large dataset with 13 predictors. By comparing the prediction accuracy and kappa statistics between models (Figure - 4.A), we settled on the Random Forest Model for its robustness.

2. Random Forest

To build the Random Forest model, we first performed a train/test split by an 8:2 ratio and then tuned the number of trees to an optimal prediction result, which in this case is 200. This model achieved an accuracy of 92.39%, an error rate of 7.61%, and an AUC of 75.4% (partial AUC of 52.9%).

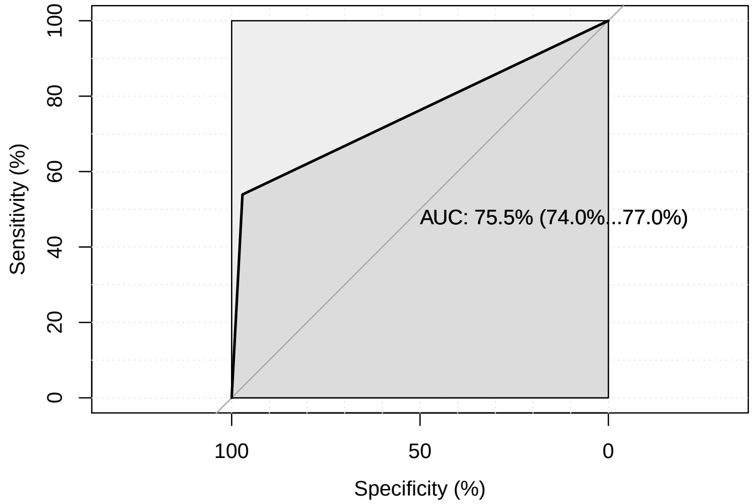

Table 1: K-Folds Cross Validation of Random Forest (Accuracy)

Fold01	Fold02	Fold03	Fold04	Fold05	Fold06	Fold07	Fold08	Fold09	Fold10
0.99	0.99	0.98	0.99	0.99	0.98	0.99	0.99	0.99	0.99

Using the 10-folds cross-validation to solidify the model further, we achieved 98.88% accuracy in prediction.

Deployment and Business Value

The performance metrics proved that our model can be utilized as an efficient tool for the bank to predict whether the targeted customer will enroll in the term deposit. Meanwhile it can also assess the effectiveness and efficiency of the marketing campaign and provides valuable feedback on how to further improve and adjust. One of the strengths of this model is the comprehensiveness of information it incorporated: not only for customer information, but also for the previous campaign record. It could subsequently contribute to the business sustainability and data management.

For deployment scenarios. Suppose we are given a new set of customers that contains both old and new customers. And we assume the bank collects their information with the same fashion as always, upon which we built the Classification model. We will be able to identify which customer to chase after given the likelihood of retention or acquisition. On top of that, our exploratory analysis can break down the dataset in details. Although we cannot modify the background information of the given customers, we can adjust the campaign approaches, such as the frequency of contacts, time of the year to attempt, according to the prediction and other information. Thus, the model can also facilitate in adjusting campaign strategies.

The beauty of such predictive model is that it can simultaneously take on multiple dimensions of information and make judgments within seconds. It saves tremendous expenses in human capital and opportunity cost, and more importantly, it mitigates risk for making unfavorable decision and unsuccessful marketing attempts.

Conclusion and Discussion

Marketing campaigns are becoming increasingly competitive, and gradually extending to multiple platforms. Banks rushing to create various deals and packages in order to acquire or retain their customers. How to become even more efficient in the early stage of analyzing customer information and narrowing down the scope is exactly where our model comes into play. We will be able to optimize both marketing strategy and financial operations.

There are several drawbacks and areas for improvements both in data acquisition and model building, which we would like to discuss for future references. Firstly, the data collection, which, in this case, was conducted by the bank, can be more diverse and precise in terms of customer profiling and previous campaign records. If we were able to expand the dimensions of data attributes, we may be able to dig deeper into the campaign strategy, and potentially improve our model. Moreover, our choice of Classification model is based on the target variable being binary. The shortcoming of such model is that we are not able to see exactly the variable selection and the marginal effect of each variable, meaning we may find cumbersome adjust each variable based on its impact on the outcome. Therefore, we can potentially try gathering or calculating the likelihood of customer enrollment and treat it as the target variable. By doing so, we may be able to construct regression models to assess the marginal effects for more detailed strategy tuning.

Appendix

EDA: Figure - 1 (Monthly Index)

Table 2: Table Preview for Figure - 1

month	avg_price	avg_conf	avg_emp	base_num	enroll_num	enroll_rate
3	93.09714	-42.65589	-1.8000000	594	304	0.5117845
4	93.11951	-46.27457	-1.8000000	2847	575	0.2019670
5	93.52762	-40.59078	-0.1685944	15147	976	0.0644352

EDA: Figure - 2.A (General Background)

Table 3: Table Preview for Figure - 2.A

job	base_num	mean_age	mean_edu	loan_rate	hous_rate
admin.	11434	38.19197	4.498951	0.1637222	0.5348085
blue-collar	10138	39.53048	2.604557	0.1489446	0.5078911
entrepreneur	1604	41.76746	4.031172	0.1433915	0.5411471

EDA: Figure - 2.B (Background with Enrollment)

Table 4: Table Preview for Figure - 2.B

age	base_num	enroll_num	enroll_rate
17	5	2	0.4000000
18	31	13	0.4193548
19	43	20	0.4651163

job	base_num	enroll_num	enroll_rate
admin.	11434	1485	0.1298758
blue-collar	10138	699	0.0689485
entrepreneur	1604	132	0.0822943

marital	base_num	enroll_num	enroll_rate
divorced	5058	519	0.1026097
married	27437	2784	0.1014688
single	12721	1775	0.1395331

education	base_num	enroll_num	enroll_rate
basic.4y	4605	466	0.1011944

education	base_num	enroll_num	enroll_rate
basic.6y	2520	205	0.0813492
basic.9y	6619	516	0.0779574

EDA: Figure - 3 (Campaign)

Table 8: Table Preview for Figure - 3

month	avg_duration	med_duration
3	248.2879	188.5
4	293.0566	218.0
5	261.0524	191.0

Model: Confusion Matrix - RF(200-TT)

Table 9: Confusion Matrix for RF (n=200 Train/Test)

	0	1
0	8401	248
1	488	571