

Marketing Campaign for Banking products – Milestone 2

Sri R Sankaranarayanan

DSC680

<https://github.com/rengsankar1986/DSC680>

Business Problem :

Finding out the factors and characteristics that are helping Bank to make customers successfully subscribe for their products, which helps in increasing campaign efficiently and selecting high value customers/target customers.

The vast benefit of determining the success factors will enable company to effectively utilize its resources (time, money, hr, etc) and identifying potential prospects in the most optimal and efficient manner.

Background/History :

Marketing Campaign for Banking products is one the pioneer applications of employing Data Science. Asset customers are the backbone of Banking Industry more than non-asset customers. This study aims at creating generic reusable model for many banking product campaigns effectiveness. It is very important to analyze each campaigns effectiveness so that will help the entire product life cycle within the company on what improvisations can be made to which specific areas.

Data explanation:

I will be using data from UCI case study for predicting the success rate of marketing campaigns.

<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

bank client data based of Telemarketing phone calls:

1 — age (numeric)

2 — job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')

3 — marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)

4 — education (categorical:

'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')

5 — default: has credit in default? (categorical: 'no', 'yes', 'unknown')

6 — housing: has housing loan? (categorical: 'no', 'yes', 'unknown')

7 — loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

related with the last contact of the current campaign:

8 — contact: contact communication type (categorical: 'cellular', 'telephone')

9 — month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

- 10 — day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- 11 — duration: last contact duration, in seconds (numeric).

Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

other attributes:

- 12 — campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)**
- 13 — pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)**
- 14 — previous: number of contacts performed before this campaign and for this client (numeric)**
- 15 — poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')**

social and economic context attributes

- 16 — emp.var.rate: employment variation rate — quarterly indicator (numeric)**
- 17 — cons.price.idx: consumer price index — monthly indicator (numeric)**
- 18 — cons.conf.idx: consumer confidence index — monthly indicator (numeric)**
- 19 — euribor3m: euribor 3 month rate — daily indicator (numeric)**
- 20 — nr.employed: number of employees — quarterly indicator (numeric)**

Data Science Methods:

I am planning to employ EDA on my dataset, followed by correlation matrix between different types of variables (after converting age into categorical variable), running Logistic regression tests using SVMs, performing accuracy using Random Forest, XGBoost, and Adaboost. Also will be running performance metrics using AUROC/SMOTE to deal with imbalanced dataset, Macro-F1 score (only if needed).

Data Science Analysis :

I am planning on comparing various models and check which one should be used for deployment.

Challenges/Potential Issues:

I am expecting issues on imbalanced datasets which is going to bias my results based on majority classes, but I am planning to counter them using AUROC and SMOTE.

Assumptions :

This will be the first ever marketing campaign for my Bank and we are assessing its effectiveness.

Ethical assessments :

I have made my best effort to employ basic principles of ethical considerations when working with Customer data by mocked up/cleaned datasets wherever possible. For the final outcome, we will be making decisions (if it is binary) by not conflicting with any moral societal values affecting target campaigns.

Limitations:

1. We can impute pdays which have around 70% of values with 999 by using different imputation methods – probably perform data transformation for some real values will help.
2. For balancing of dataset, we have used `class_weight='balanced'`, in milestone 3, I will plan to conduct SMOTE and AUROC which might further improve our AUC score.

Implementation Plan:

I am planning on writing function for the following :-

1. Read the input dataset
2. Convert input to Dataframe
3. Load the best fit model based on model comparison
4. Preprocessing dataframe
5. Run the model and return the prediction

Conclusion :

This analysis will help save lot of resources, increase the overall revenue and the conversion rate of current and future campaigns.

References :

<https://towardsdatascience.com/4-analytics-use-cases-in-banking-and-what-data-you-need-5f654235bbf> --> Analytic Case Study using propensity models to predict likely Customers who will subscribe to a product.

<https://towardsdatascience.com/machine-learning-case-study-a-data-driven-approach-to-predict-the-success-of-bank-telemarketing-20e37d46c31c> --> Data Driven approach to predict the success of Bank Marketing.

<https://www.kaggle.com/code/aleksandradeis/bank-marketing-analysis/notebook> --> This Kaggle dataset contains one of the best categorical columns exploration and I will use this inference when conducting the propensity models.

<https://cio.economictimes.indiatimes.com/news/big-data/heres-how-this-bank-is-using-predictive-analytics-for-marketing-campaigns/84995037> --> this article details RoMI (Return on Marketing Investment) accuracy on Banking Marketing campaigns.

<https://medium.com/mlearning-ai/data-driven-approach-to-predict-success-of-bank-marketing-31791cad8f81> --> Another data driven approach to predict the success of Marketing campaign.

<https://www.absentdata.com/pandas/pandas-cut-continuous-to-categorical/> → This is used for converting age and other possible categorical variables for easy analysis.

<https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/> → I will be using this reference to counter imbalanced dataset classification problems using SMOTE and penalized data sampling.