Bank Marketing Prediction

DATA 1030 Project Proposal

Author: Guansu (Frances) Niu

GitHub Link: https://github.com/francesniu/Data-1030-Bank-Marketing-Project.git

Data Resource: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Project Goal:

The goal is to predict if clients will subscribe to a term deposit from the bank. The problem is classification, and the target variable is 'has the client subscribed to a term deposit' (binary). According to the dataset, 41,188 clients have already made the decision. By analyzing the correlation between those clients' data and their decision outcomes, the bank will be able to forecast the future client adaptability. Based on the result, the bank can improve the term deposit so that it will be more widely accepted by its clients.

Dataset description:

The data of direct marketing campaigns is provided by a Portuguese banking institution. The dataset consists of 20 attributes belonging to 41,118 clients. The 20 attributes can be classified into 4 categories: bank client data attributes (e.g. age, job, marital, etc.), current campaign attributes (e.g. contact communication type, month, duration, etc.), contact of previous campaign (e.g. outcome of previous campaign, etc.) and, social and economic context attributes (e.g. employment variation rate, consumer price index, etc.). 'Has the client subscribed to a term deposit' is the target variable.

Feature description:

Original data example:

BankMarketing

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
2	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
3	57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
4	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
5	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
6	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
7	45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no

Input variables:

Bank client data attributes:
1 - age (numeric)
2 - job: type of job (categorical: 'admin.','bluecollar', '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').

Current campaign attributes:
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.
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact).

Previous campaign attributes:
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means clients were 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).

Output variable:
21 - y - has the client subscribed to a term deposit? (binary: 'yes', 'no').

Public projects using the same dataset:

"A data-driven approach to predict the success of bank telemarketing"

This project proposes a data mining methodology to predict the success of telemarketing calls for selling bank long-term deposits. By integrating the dataset, four models including logistic regression, decision trees, neural network and support vector machine were tested. By comparing the two metrices, area of the receiver operating characteristic curve and area of the LIFT cumulative curve, that neural network presents the best result, allowing to reach 79% of the subscribers by selecting the half better classified clients.

"Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology"

The project describes a data mining approach to bank direct marketing campaigns using the dataset. It performed three iterations of the CRISP-DM and the best model, materialized by a Support Vector Machine, achieved high predictive performance. It also confirmed that open-source technology in the DM field is able to provide high quality models for real applications meanwhile allow a cost reduction of DM projects.

Preprocessing the dataset:

All of the 20 features are preprocessed, and they are classified into categorical features and continuous features. Under the categorical features, ‘Ordinal encoding’ is used for the ones that have specific order (e.g. education, month, day), and ‘Onehot Encoding’ is used for unordered ones (e.g. job, marital, etc.). Under the continuous feature, ‘Minmaxscaler’ is used for the one that has boundary. As mentioned in the feature description, the maximum value of 'pdays' is 999 meaning clients were not previously contacted. ‘Standardscalar’ is used for the unlimited numbers (e.g. age, etc.).

Preprocessed data example:

	education	month	day_of_week	x0_admin.	x0_blue-collar	x0_entrepreneur	x0_housemaid	x0_management	x0_retired	x0_self-employed	x0_services	x0_student	x0_technician	x0_unemployed	x0_unknown
0	0.0	2.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	3.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
2	3.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
3	1.0	2.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	3.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
5	2.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
6	5.0	2.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	7.0	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	5.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
9	3.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
10	7.0	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0