Customer Quality Prediction

Name: Yukang Xu

Ⅰ. Introduction

ProcessMiner is commissioned by an insurance company to develop a tool to optimize their marketing efforts. My objective is to determine which set of customers the marketing firm should contact to maximize profit.

The insurance company has provided us with a historical data set. The company has also provided us with a list of potential customers to whom to market. From this list of potential customers, I need to determine yes/no whether you wish to market to them.

Ⅱ. Data Analysis

A. Identify the data

After importing data into data frame, I identified the type of variables and the variables with missing value.

Numeric Variable (13): custAge, compaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed, pmonths, pastEmail, profit, id

Categorical Variable (11**):** profession, marital, schooling, default, housing, loan, contact, month, day_of_week, poutcome, responded

Variables with missing values (3): custAge, schooling, day_of_week

Variables with unknown (4): profession, marital, housing, loan

There are three different methods to deal with missing value or unknown value (removing, filling in or doing nothing). For the next step, let us explore these variables and find some evidence to deal with these missing values.

B. Explore the data

1.Age VS Marital

After exploring the mean and median of age in different Marital levels, I put what I got in table B.1 as below. From this table, people with different marital status are in fairly different age. Therefore, I used median of age in different marital level to fill in missing values for age variable.

2. Default variable exploration

Default variable has two levels ('yes' and 'no'). After exploring the rate of these two levels, I put what I got in table B.2 as below. I find that they are too unbalanced. It means that this variable may not provide useful information about our target feature. I will drop default variable.

	default variable
yes	1
no	19628

(table: B.1)

	mean	median
single	32.901760	32.0
married	42.271783	41.0
divorced	45.267806	45.0

(table: B.2)

	people who were never contacted
poutcome	7060
previous	7060
pmonths	7922
pdays	7922
campaign	0
pastEmail	7219

(table: B.3)

3. people who were never contacted

The information provided by different variables about 'people who were never contacted' are different. (table B.3)

To assure the consistence of our data, I dropped pmonths, pdays, campaign, pastEmail which may give us misleading information.

C. Clean the data

Based on what we analyzed above, we will deal with our data by five steps as follows.

1. Completing: fill in the missing value

2. Creating: create a new variable (put multiple variables together or extract information from existed variable)

3. Converting: convert ordinal, nominal data to dummy variables

4. Correcting: detect the outliers

5. Dropping: drop some variables which have nothing to do with our target

(four rules for dropping: too many missing values/ cannot find related variables to fill in/ have nothing to do with target variables)

What we will deal with for every variable are attached below.

Variable	Handling method	Details
custAge	Completing	Completing based on related variable: marital
profession	Dropping; Converting to three levels as right	Remove observations with unknown value; 0: unemployed, student, retried, housemaid 1: blue-collar, services, technician 2: admin, entrepreneur, management, self-employed
marital	Converting to three levels as right; Dropping	Remove observations with unknown value 0: divorced 1: married 2: single
schooling	Dropping	Fail to find related variables
default	Dropping	Did not provide useful classification information
housing, loan	Creating new feature; Completing	Creating new feature which is called loan_Code with two levels 0: without loan 1: loan
contact	Converting	0: cellular 1: telephone
month, day_of_week	Dropping	Has nothing to do with target variables
poutcome	Converting to three levels as right	0: nonexistent 1: failure 2: success
campaign, pdays, pmonths,pastEmail	Dropping	Conflicting with other data
id	Dropping	Random data
responded	Converting to two levels as right	0:no 1:yes

D. Model Development

I chose eight machine learning algorithms to train our models and rank their accuracy.

Table.D.1. Decision Tree with 97.66% will be our best choice.

	Model	Score
8	Decision Tree	97.66
3	Random Forest	97.63
1	KNN	92.00
0	Support Vector Machines	91.09
2	Logistic Regression	89.28
5	Perceptron	88.81
7	Linear SVC	88.81
4	Naive Bayes	82.21
6	Stochastic Gradient Decent	11.19

(table:D.1)

(1)custAge	0.294483
(2)previous	0.263370
(3)emp.var.rate	0.067146
(4)cons.price.idx	0.062761
(5)cons.conf.idx	0.058188
(6)euribor3m	0.050370
(7)nr.employed	0.042736
(8)poutcome_Code	0.040588
(9)contact_Code	0.035788
(10)profession_Code	0.035522
(11)marital_Code	0.028907
(12)loan_Code	0.020141

(table:D.2)

	Model	Score
8	Decision Tree	96.99
3	Random Forest	96.98
1	KNN	91.99
0	Support Vector Machines	91.06
2	Logistic Regression	89.27
5	Perceptron	88.81
6	Stochastic Gradient Decent	88.81
7	Linear SVC	88.81
4	Naive Bayes	82.20

(table:D.3)

E. Model Evaluation

Right now, let us think about how to improve accuracy of our model. I tried to use random forest to rank the importance of our features (table:D.2). Then I dropped the loan_Code with lowest importance and run all of algorithms again (table:D.3). It turns out that that the highest accuracy is going down. So, I choose not to remove loan_Code from our analysis.

Ⅲ. Prediction

Based on the analysis above, decision tree is the champion model. The prediction accuracy of it is 97.66%.

I added one more column in ‘testingCandidate.csv’ which is call ‘responded’. ‘0’ means that we should not market to the customer while ‘1’ means we should.