Data Mining Project

Potential Deposit Subscription Customer Prediction

Ruixie Fang, Weidan He

Georgia State University

1. Introduction

Marketing campaign has been an effective domain in gaining subscribers and enlarging the company reputation. Meanwhile, facing the complaints from customers that they are bothered with irrelevant product calls. The main purpose of this project is to figure out the relevant factors of bank term deposit subscription and help the financial institution to have a greater effectiveness for future marketing campaigns by using data mining algorithms.

2. Data

The project is based on dataset called Bank Marketing Dataset from UCI website (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#). There is one dataset: bank-full.csv with all examples (45211) and 17 variables.

table 1: Description of variable

Variable	Type	Variable	Type	Variable	Type
Age	*Num*	housing	*Char*	campaign	*Num*
job	*Char*	loan	*Char*	pdays	*Num*
marital	*Char*	contact	*Char*	previous	*Num*
education	*Char*	day	*Num*	poutcome	*Char*
default	*Char*	month	*Char*	y	*Char*
balance	*Num*	duration	*Num*		

3. Procedure

3.1 Data preprocessing

Before we dive into modeling, we got an overview of the data to see what we'll be working on. The presence of “unknown” value of several variables and imbalanced data might affect the modeling performance and prediction results. So we solved these issues first.

We used Chi-square (for character variables) and correlation test (for numeric variables) to check the independence of each variable, and all results show that they all significantly related with “y”.

Then the stratified sampling method and ROC measurement were used to improve the performance. Stratification seeks to ensure that the mean response value is approximately equal in all the folds. And Roc measurement provide more accurate test rate. The dataset of 45211

observations were split into training and test data so that distribution of the outcome within training and testing datasets is preserved. The 70% (or 31649) of the observations was used for training the model, and 30% (or 13562) of observations was used to test the prediction outcome from the classifier model.

3.2 Methodology

Based on the summary of dataset, we learned that the subscription status (y) is represented as a binary variable, which makes it perfect for classification. So we selected Logistic Regression, SVM and Naive Bayes as our methods to analysis our data. And the confusion matrix and ROC curves was designed to measure the performance of training set and test set for all three methods.

3.2.1 Logistic Regression

Logistic regression method was used in this project to predict the relationship between deposit subscription and predictionary factors. Any variables with p-value >0.05 were removed in order to get the final logistic model. The final model's performance was displayed as follows:

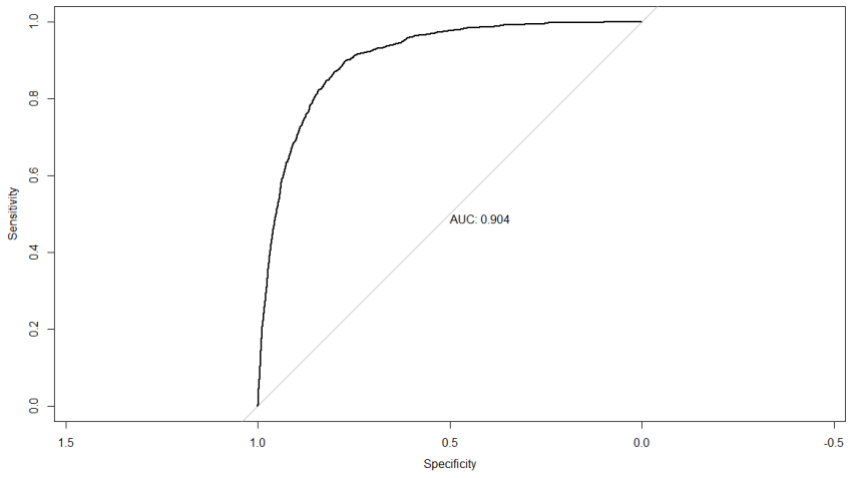
Confusion Matrix:

```
          actual
predicted    no   yes
      no  11671  1055
      yes   305   531

               Accuracy : 0.8997
```

After that, we put test data into the model.

ROC Curve(the area under curve is 0.904):



3.2.2 SVM

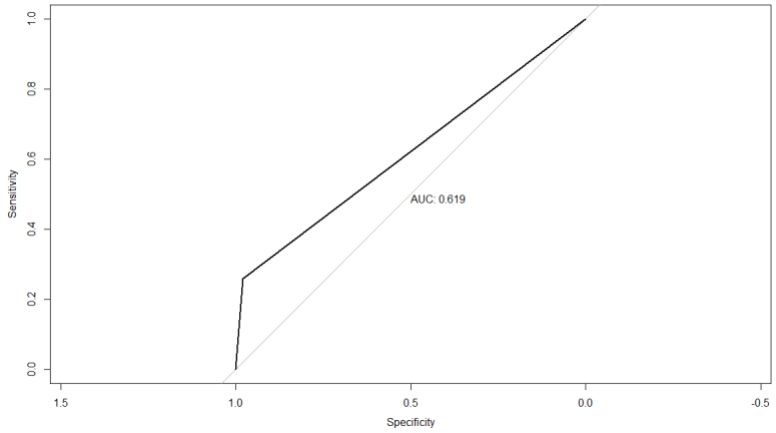
SVM is another classification method that can be used to predict if a client falls into either 'yes' or 'no' class. The performance of this algorithm were as follows:

Confusion Matrix:

```
          Reference
Prediction    no   yes
       no  11741  1177
       yes   235   409

               Accuracy : 0.8959
```

ROC Curve(the area under curve is 0.619):



According to the test result, SVM model doesn't fit the data well.

3.2.3 Naive Bayes

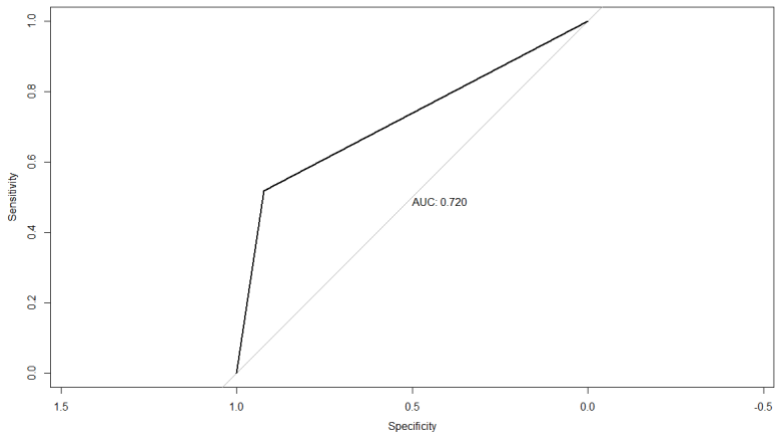
The next method used to predict was Naive Bayes method. The Naive Bayes method assumed independence among each variable, i.e. the algorithm assumes that attributes such as job and education are independent from each other in predicting whether a customer will open a bank account or not. The results were shown below:

Confusion Matrix:

```
          Reference
Prediction    no   yes
       no  11054   765
       yes   922   821

               Accuracy : 0.8756
```

ROC Curve(the area under curve is 0.720):



According to the test result, Naive Bayes model doesn't fit the data as well.

4. Conclusion

	Logistic regression	SVM	Naive Bayes
Accuracy	0.8997	0.8959	0.8756
AUC	0.904	0.619	0.720

In the light of overall test accuracy and ROC Curve, the best model is Logistic Regression. It has the most powerful prediction ability. Next, we find out which factors are most important and how these factors influence customers' decision. In addition, we have the importance of variables table and the coefficient table, as shown below:

Table2.The importance of Var

importance of variable			
duration	54.6224691	jobself-employed	2.7413202
poutcomesuccess	24.5896585	jobtechnician	2.7353676
contactunknown	18.3606934	balance	2.5441788
housingyes	13.6766575	maritalmarried	2.5228856
monthmar	10.1575898	jobstudent	2.4419068
monthjul	9.3777756	jobretired	2.4223101
monthjan	9.0324118	monthdec	2.4205803
monthaug	8.7730083	jobunknown	2.3914525
monthnov	8.6001078	poutcomeother	2.355932
campaign	6.8423986	jobservices	2.3369792
loanyes	6.5948009	monthfeb	2.3054276
monthoct	6.4781716	educationsecondary	2.1352706
monthmay	5.8413904	day	2.0759602
monthsep	5.1868441	jobmanagement	1.8159763
jobblue-collar	4.7652037	jobunemployed	1.7934088
jobhousemaid	3.6506456	educationunknown	1.7354077
educationtertiary	3.6303495	maritalsingle	1.4615783
monthjun	3.5295646	contacttelephone	0.8231976
jobentrepreneur	2.8501106	poutcomeunknown	0.6792136

Table3.The coefficient of regression

coefficient			
poutcomesuccess	2.363132	jobmanagement	-0.1573182
monthmar	1.463581	maritalmarried	-0.1784443
monthoct	0.8290982	jobtechnician	-0.2248554
monthsep	0.7383303	jobservices	-0.2326186
monthdec	0.5078504	jobunemployed	-0.2394225
monthjun	0.3939454	monthfeb	-0.2450946
educationtertiary	0.3233015	jobself-employed	-0.3731303
jobstudent	0.3094548	jobblue-collar	-0.4171753
poutcomeother	0.2556536	jobentrepreneur	-0.4185463
jobretired	0.2514094	loanyes	-0.4770037
educationunknown	0.215917	monthmay	-0.5043537
educationsecondary	0.1648153	jobhousemaid	-0.5928374
maritalsingle	0.1109926	housingyes	-0.7127624
day	0.006223691	jobunknown	-0.7552275
duration	0.004187739	monthaug	-0.820606
balance	1.55729E-05	monthnov	-0.8458536
poutcomeunknown	-0.04699736	monthjul	-0.8564603
contacttelephone	-0.07167316	monthjan	-1.29443
campaign	-0.08293811	contactunknown	-1.6301

According to the importance ranking of the variables, we can tell that the most influential variables were duration (The reason may be the longer the conversations on the phone, the higher interest the customer will show to the term deposit.) and outcome of the previous marketing campaign(clients' decision would be influenced by the successful outcome in the previous campaign). And based on the coefficients of the logistic regression model, month is also a critical factor (Stable marketing economy environment or higher bank interest rate might affect the performance of campaigns in these months) .

Therefore, if banks want to improve campaign efficiency, they should improve the quality of conversation on the phone, focus on the potential clients and follow the good timing to run their campaigns.