Data Analysis for Bank Marketing

Final Project of Data Analytics

Harper He, Chenyan Huang, Chenying Jiang, Jingxian Sun

Agenda

- Introduction
- Data preprocessing
- Descriptive Analysis
- Data Analysis

Introduction

Introduction

- Dataset

  Bank Marketing Data Set: https://archive.ics.uci.edu/ml/datasets/bank+marketing#

  45,211 examples , 17 inputs

- Target Variable

  variable y: if the client will subscribe (yes/no) a term deposit

Introduction

- Business Problem

  whether or not the customer opted term deposit based on their demographic information and bank's marketing activities.

- Objective

  To find the best model to help bank to predict a new customer's subscription behavior based on customer's demographic information. And to seek the most helpful attributes, so that bank can pay more attention and put more effort on customers who consist with those factors.

Data preprocessing

Original dataset

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no

Original dataset

Attribute	Data Type	Missing Value
age	numeric	No
Job	nominal	Yes
marital	nominal	No
education	ordinal	Yes
default	nominal	No
balance	numeric	No
housing	nominal	No
loan	nominal	No

Attribute	Data Type	Missing Value
contact	nominal	Yes
day	numeric	No
month	nominal	No
duration	numeric	No
campaign	numeric	No
pdays	numeric	No
previous	numeric	No
poutcome	nominal	Yes

Info Gain

```
> infoGain
      default       marital     education          loan
"0.0004244884" "0.0030314170" "0.0037483854" "0.0037949345"
     campaign           day       balance           job
"0.0061513095" "0.0081193541" "0.0082252816" "0.0119228893"
      housing           age      previous       contact
"0.0139277448" "0.0173396802" "0.0178693026" "0.0196593713"
        month         pdays      poutcome      duration
"0.0351313059" "0.0371739348" "0.0424112545" "0.1037161265"
```

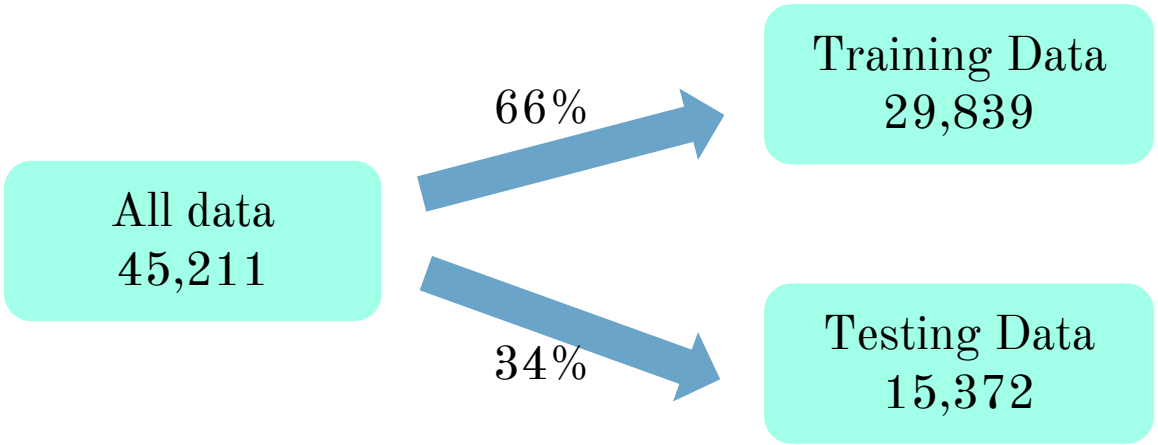
Attributes

age	pdays	balance		previous
19-24	not previously contacted	[-8020,0]	(448,701]	contacted 0 times
25-34	contacted in last 1 month	(0,22]	(701,1303]	contacted 1-3 times
35-44	contacted in last 3 months	(22,131]	(1303,1860]	contacted 3-10 times
45-54	contacted in last 6 months	(131,272]	(1860,3570]	contacted more than 10 times
55-64	contacted in last 1 year	(272,448]	(3570,102000]	
65+	contacted 1 year before			

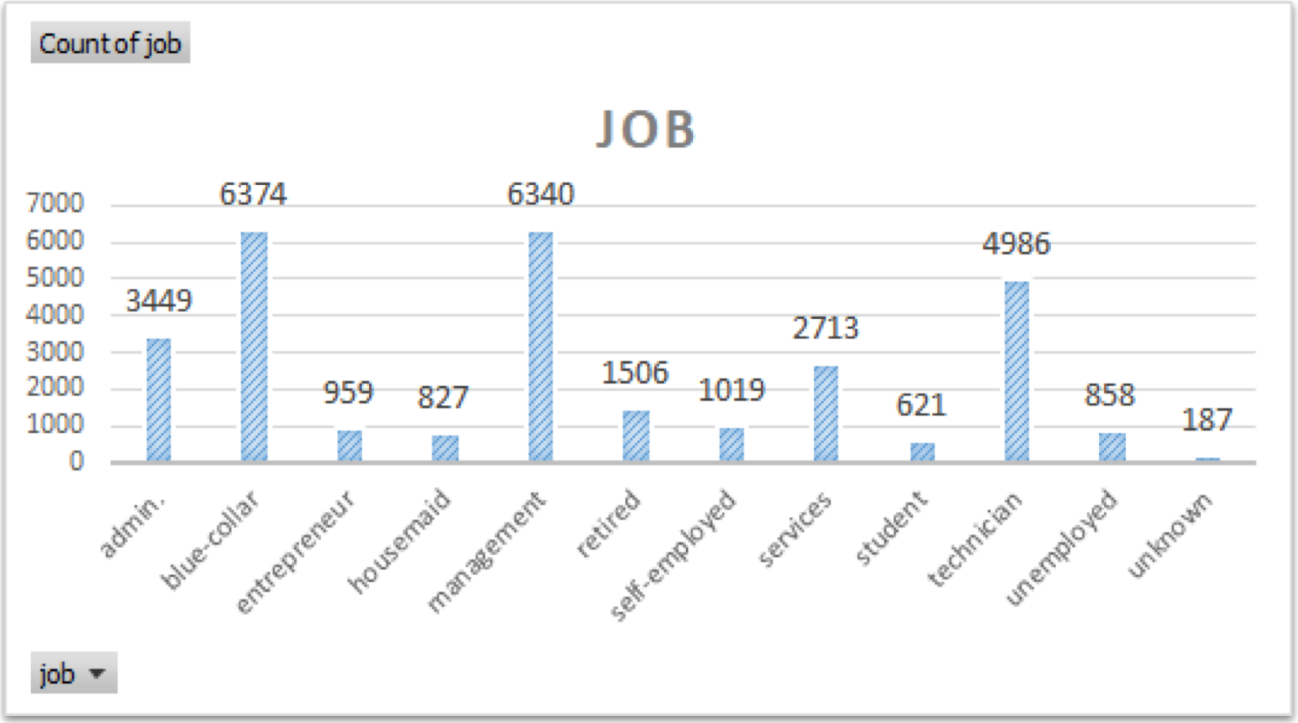
Attributes

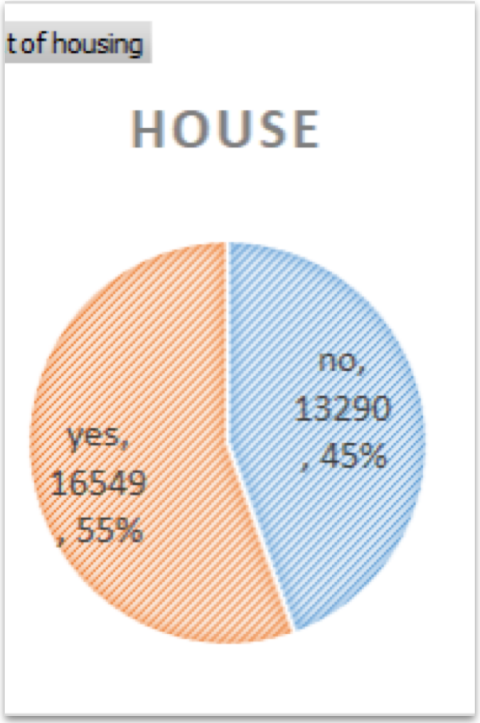
job	housing	month	y	age	balance	pdays	previous
unemployed	no	nov	no	45-54	[-8020,0]	not contacted	contacted 0 times
services	yes	may	no	35-44	(1860,3570]	not contacted	contacted 0 times
technician	yes	may	no	35-44	[-8020,0]	contacted in last 6 months to 1 year	contacted 1-3 times
technician	no	nov	no	35-44	(0,22]	contacted in last 3-6 months	contacted 1-3 times
blue-collar	yes	may	no	35-44	[-8020,0]	not contacted	contacted 0 times
blue-collar	no	nov	yes	35-44	(1860,3570]	not contacted	contacted 0 times
management	yes	may	no	45-54	[-8020,0]	contacted in last 6 months to 1 year	contacted 1-3 times
technician	yes	feb	no	35-44	(701,1303]	contacted in last 3-6 months	contacted 3-10 times
retired	yes	jul	no	45-54	(1860,3570]	not contacted	contacted 0 times

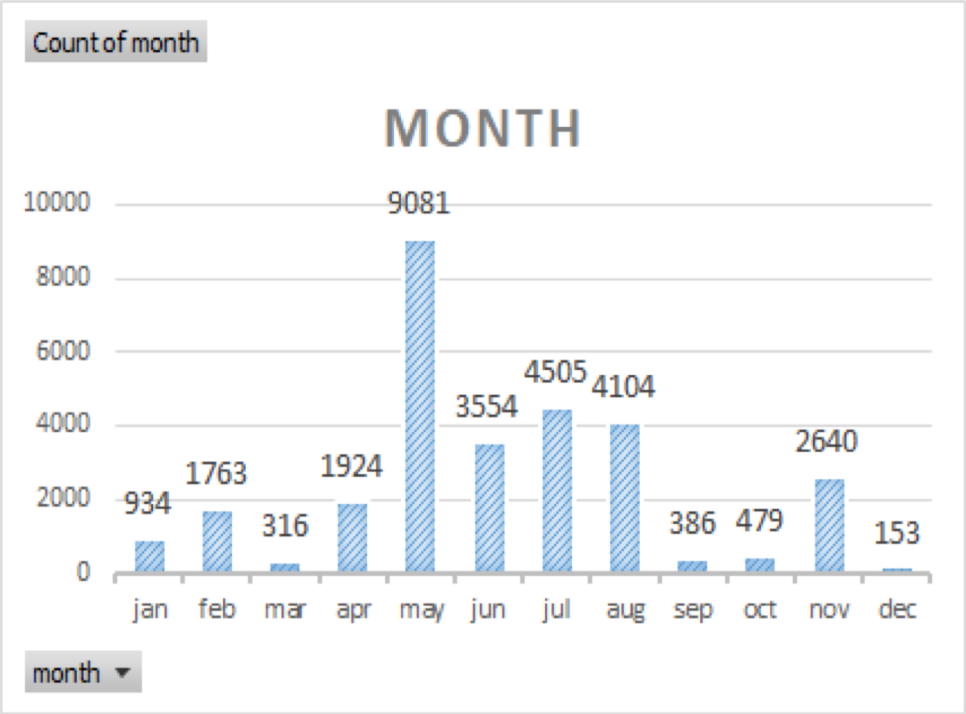
Instances

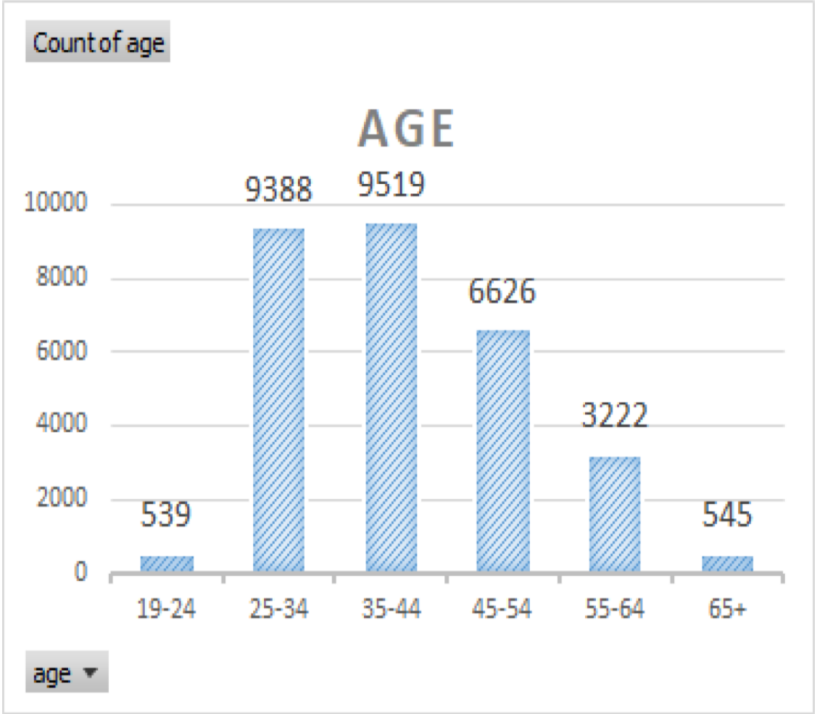


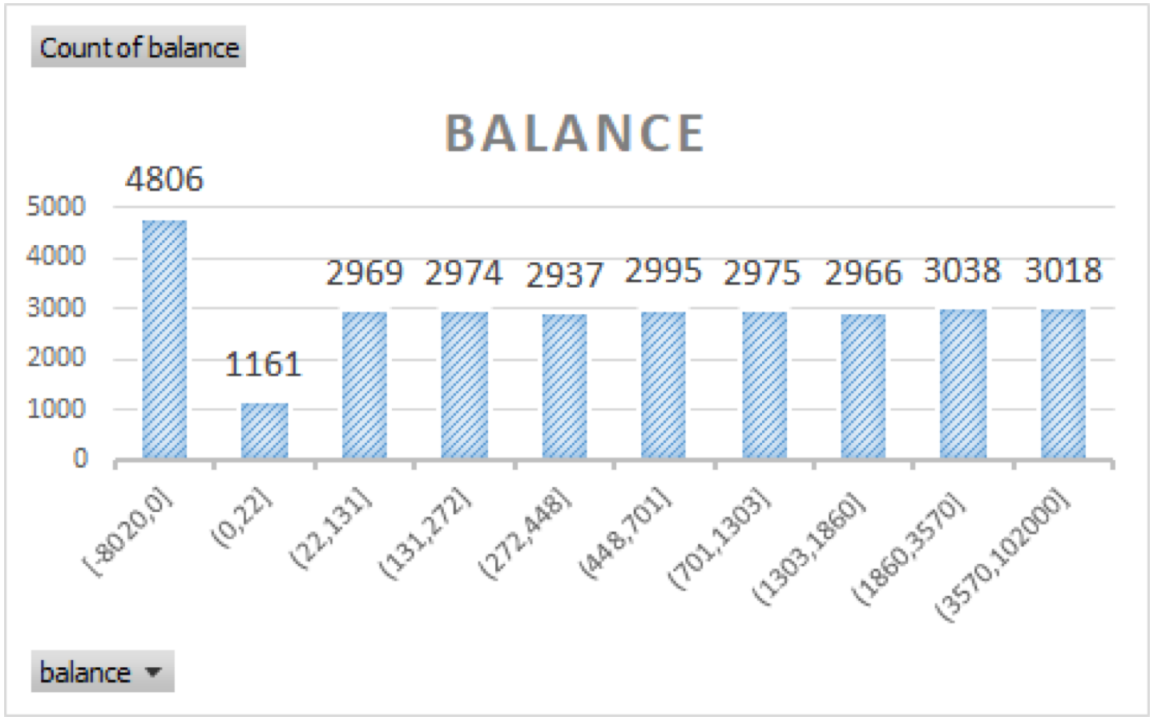
Descriptive Analysis

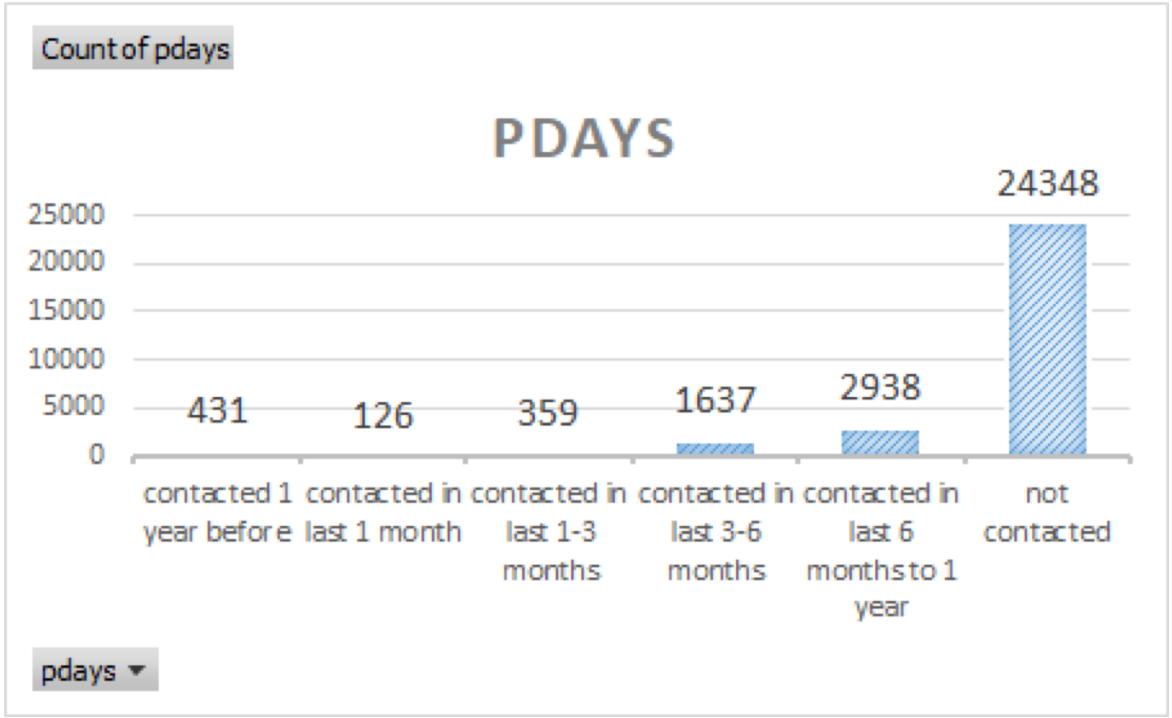

Count of job
JOB
7000
6000
5000
4000
3000
2000
1000
0
3449
6374
959
827
6340
1506
1019
2713
621
4986
858
187
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
unknown
job

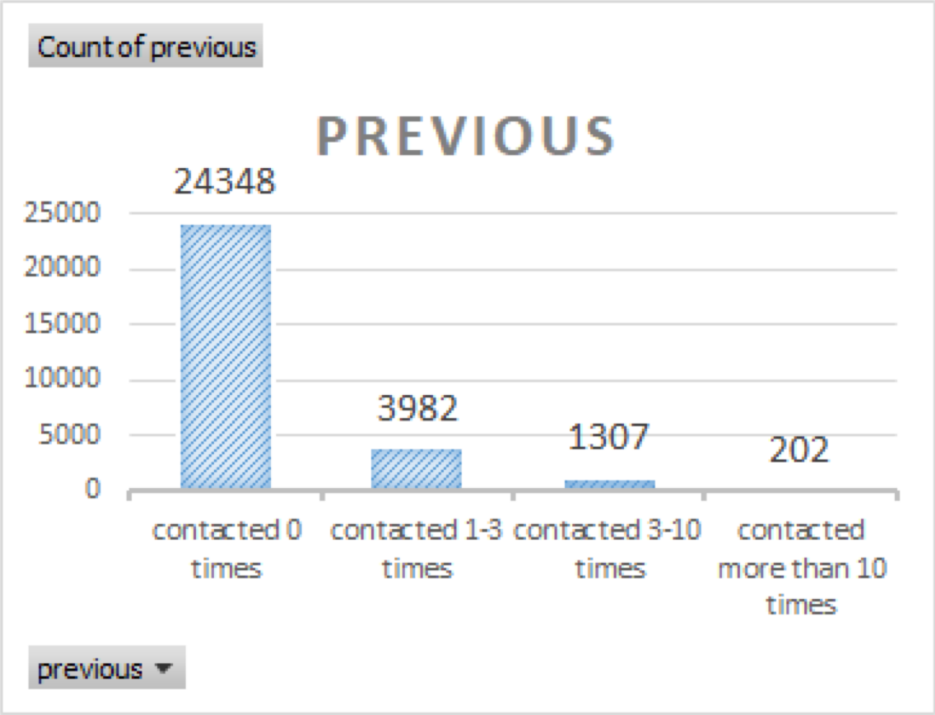

t of housing
HOUSE
no, 13290 , 45%
yes, 16549 , 55%

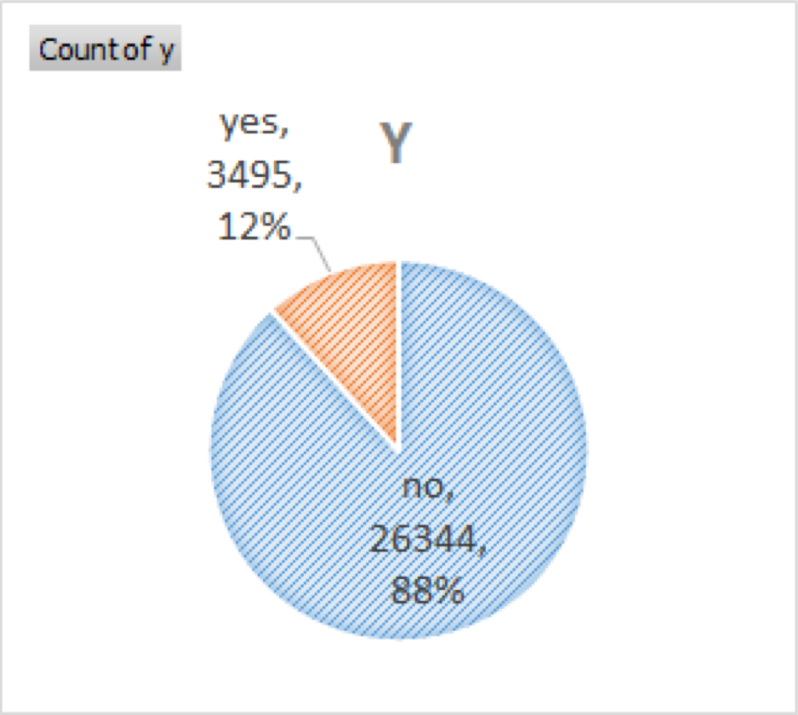

Count of month
MONTH
10000
8000
6000
4000
2000
0
934
1763
316
1924
9081
3554
4505
4104
386
479
2640
153
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
month


Count of age
AGE
10000
8000
6000
4000
2000
0
539
9388
9519
6626
3222
545
19-24
25-34
35-44
45-54
55-64
65+
age


Count of balance
BALANCE
4806
1161
2969
2974
2937
2995
2975
2966
3038
3018
5000
4000
3000
2000
1000
0
(-8020,0]
(0,22]
(22,131]
(131,272]
(272,448]
(448,701]
(701,1303]
(1303,1860]
(1860,3570]
(3570,102000]
balance


Count of pdays
PDAYS
25000
20000
15000
10000
5000
0
431
126
359
1637
2938
24348
contacted 1 year before
contacted in last 1 month
contacted in last 1-3 months
contacted in last 3-6 months
contacted in last 6 months to 1 year
not contacted
pdays


Count of previous
PREVIOUS
24348
3982
1307
202
25000
20000
15000
10000
5000
0
contacted 0 times
contacted 1-3 times
contacted 3-10 times
contacted more than 10 times
previous


Count of y
Y
yes, 3495, 12%
no, 26344, 88%

Data Analysis

SVM

Accuracy		Kernel					
		polydot	tanhdot	vanilladot	laplacedot	besseldot	rbfdot
Cost	3	88.22%	79.47%	88.22%	88.37%	88.41%	88.38%
	5	88.15%	/	88.20%	**88.37%**	88.34%	88.15%

Precision		Kernel					
		polydot	tanhdot	vanilladot	laplacedot	besseldot	rbfdot
Cost	3	0.50	0.11	0.50	0.88	0.69	0.78
	5	0.50	/	0.50	0.92	0.72	0.80

Recall		Kernel					
		polydot	tanhdot	vanilladot	laplacedot	besseldot	rbfdot
Cost	3	0.05	0.11	0.05	0.24	0.15	0.20
	5	0.05	/	0.05	0.36	0.18	0.24

F-measure		Kernel					
		polydot	tanhdot	vanilladot	laplacedot	besseldot	rbfdot
Cost	3	0.08	0.11	0.08	0.38	0.24	0.31
	5	0.08	/	0.08	**0.52**	0.29	0.36

Random Forest

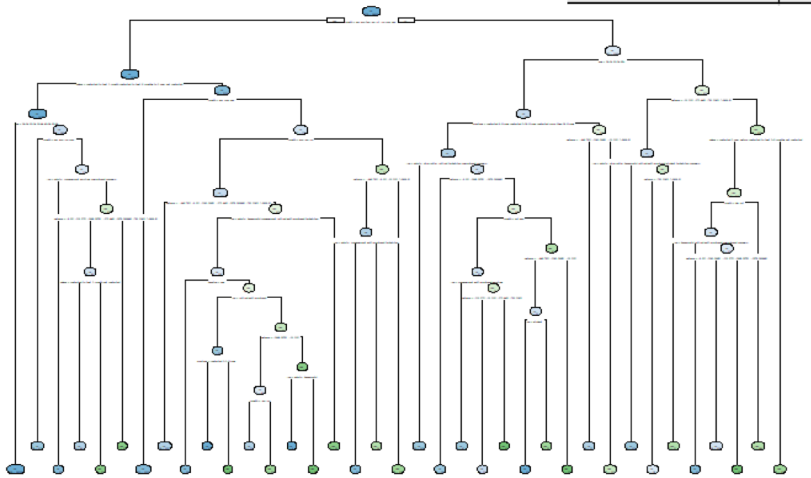
		Accuracy	Precision	Recall	F-measure
Trees	100	88.25%	0.95	0.31	0.47
	500	**88.35%**	0.95	0.32	**0.48**

kNN

		Accuracy	Precision	Recall	F-measure
K	3	**91.40%**	0.94	0.70	**0.80**
	5	87.65%	0.97	0.69	0.78

Decision Tree

		Accuracy	Precision	Recall	F-measure
Confidence	0.3	88.52%	0.65	0.11	0.19
	0.5	**88.17%**	0.74	0.26	**0.38**



Naive Bayes

The accuracy of model is 87.2616 %. Precision for model is 0.4413432, Recall is 0.2895565 and F Measure is 0.349689.

Association Rules

From the result of arules of yes-subscription, we can tell that people who have been contacted 1-3 times, do not have housing loan or have a job as management are more likely to subscribe the term deposit.

[1] {previous=contacted 1-3 times} => {y=yes}

[2] {housing=no} => {y=yes}

[3] {job=management} => {y=yes}

Association Rules

From the result of arules of no-subscription, we can tell that people who have not been contacted in the last campaign, have housing loan or were contacted in May are more likely to not subscribe the term deposit.

[1] {housing=yes,month=may} => {y=no}

[2] {housing=yes,pdays=not contacted} => {y=no}

[3] {housing=yes,previous=contacted 0 times} => {y=no}

Evaluation--Decision Tree

The accuracy is 88.42701%. The precision is 0.5128645. The recall is 0.1666667. The f-measure is 0.2515776.

TRAINING DATA: The accuracy of model is 88.1698%. The precision is 0.7394541. The recall is 0.255794. The F-measure is 0.380102.)

	predictions		
targets		No	Yes
	No	13294	1495
	Yes	284	299

Evaluation--Naive Bayes

The accuracy is 87.01535%. Precision for model is 0.4117133, Recall is 0.2625418 and F Measure is 0.3206263.
TRAINING DATA: The accuracy of model is 87.2616 %. Precision for model is 0.4413432, Recall is 0.2895565 and F-Measure is 0.349689.

	predictions		
targets		No	Yes
	No	12905	1323
	Yes	673	471

Evaluation--KNN(k = 3, l = 2)

The accuracy is 97.72934%. Precision for model is 0.9870968, Recall is 0.6181818 and F Measure is 0.7602484.
TRAINING DATA: The accuracy is 91.3972%. The precision equals to 0.9399, the recall is 0.6958 while the F-measure is 0.7996.

	predictions		
targets		No	Yes
	No	13294	1495
	Yes	284	299

Evaluation--KNN(k = 5, l = 4)

The accuracy is 98.58556%. Precision for model is 0.9870968, Recall is 0.6181818 and F Measure is 0.7602484.
TRAINING DATA: The accuracy is 87.65039%. Precision for k = 5 is 0.9708589, Recall is 0.6519053, F-Measure is 0.780037.

	predictions		
targets		No	Yes
	No	13360	291
	Yes	34	628

Evaluation--Random Forest(ntree = 500)

The accuracy is 88.28%. Precision for model is 0.4922179, Recall is 0.1410256, F Measure is 0.2192374.

TRAINING DATA: The accuracy is 88.35%, the precision for 500 trees is 0.9526227, Recall is 0.3221745, F Measure is 0.4815052.

	predictions		
targets		No	Yes
	No	13317	1541
	Yes	261	253

Evaluation--Random Forest(ntree = 100)

The accuracy is 88.23%. Precision for 100 model is 0.4841897, Recall is 0.1365663, F Measure is 0.2130435.
TRAINING DATA: The accuracy is 88.25%, the precision for 100 trees is 0.9468641, Recall is 0.3110157, F-Measure is 0.4682317.

	predictions		
targets		No	Yes
	No	13317	1549
	Yes	261	245

Evaluation--SVM

The accuracy is 88.43352%. Precision for model is 0.5223464, Recall is 0.1042363, F Measure is 0.1737918.
TRAINING DATA: The accuracy is 88.3709%. The precision is 0.8836478. The recall is 0.2412017. The f-measure is 0.3789616.

	predictions		
targets		No	Yes
	No	13407	1607
	Yes	171	187

Find the best model

This model is KNN with K=5 and l=4. The accuracy is 99.16%. Precision for model is 0.9911788, Recall is 0.7907869, F-Measure is 0.8797153.

	predictions		
targets		No	Yes
	No	38887	327
	Yes	11	1236

Find the best model

This model is KNN with K=3 and l=2. The accuracy is 98.53%. Precision for model is 0.973384, Recall is 0.8207759, F-Measure is 0.8905897.

	predictions		
targets		No	Yes
	No	39486	559
	Yes	70	2560

Challenge

- The RWeka package has limited heap space.

After searching other methods, this project used other classifier package to solve the challenges.

- There was no specific direction for setting the parameters to move on.

The solution is to focus on several models which had better initial performances and then dig deeper on parameters setting of these models in all aspects. After doing large amount of exploration, the project successfully got the best model.

Conclusion

- This project processed the whole original dataset by eliminating missing values and setting groups.
- As the models developing needed, data were transformed into factors and divided into training dataset (66%) and testing dataset (34%) .
- The project had explored 6 models, included Support Vector Machine, Random Forest, kNN, Decision Tree, Naive Bayes, Association Rules.
- By using association rules, the project suggested the bank marketing team focus on customers who had been contacted 1-3 times without housing loan and not working as “management”.
- Among 5 predicting models, the best is kNN model (K=5, l=4) with an accuracy of 99.16% and a precision of 0.9911788, while the Recall is 0.7907869 and the F-Measure is 0.8797153.