Report on
Bank Marketing Data Set

Course: CIND 119- Introduction to Big Data

The G. Raymond Chang School of Continuing Education
Ryerson University
Professor: Bilgehan Erdem

Project Members
Rebeca Furtado, 500641441
Sharlin Kahlon, 501124232

Table of Contents

Abstract

These days, banks and financial institutions leverage telemarketing strategy for easy and quick sale. In this report, we as data scientists are analyzing bank's previously collected data, strategy and predicting results based on dataset provided. Subscription to these long-term deposit accounts would further secure business growth as an investment or as a loan to other customers at higher rate. Our goal is to predict, whether the client would likely subscribe to a term deposit account or not based on the given information and which attribute has the maximum or least influence in our decision.

Data set is provided in two different formats- .csv and .arff and has 4521 customers recorded (instances) and 17 necessary attributes (columns) out of which 16 are independent variables and 1 dependent defined as binary data 'Yes' or 'No' aka successful or unsuccessful call. We have used various tools to explore given dataset and prepare data exploring outliers, distributions, and missing values. The different tools we are using here are WEKA (Waikato Environment for Knowledge Analysis), Statistical Analysis Software, Python and R programming. We have used three models J 48, naive bayes and random forest and selected cross-validation test option at 10 folds. We found random forest model has more accuracy and better True Positive (TP) /False Positive (FP) rate than decision tree and naive bayes.

Based on our analysis, our recommendation has focused 'duration' attribute, longer the duration on phone calls generate better results. And our keen focus is on customers who were contacted previously are more likely interested to subscribe for accounts.

Workload Distribution

Member Name	List of Tasks Performed
Sharlin	Abstract
Sharlin and Rebeca	Data preparation using R and Weka
Rebeca and Sharlin	Predictive Modeling/Classification Comparison using Weka
Rebeca	Conclusions and Recommendations
Both	Visualizations

Data Preparation

In this data set, there are 4521 observations, 7/17 attributes are quantitative (numerical) and 10 are qualitative (9 are nominal and 1 ordinal- “Education attribute”). Here dependent variable -Y is class attribute, binary nominal attribute and loan has nominal datatype. Dataset provided by Bank is complete as there are no missing values. Summary tool in R helps to define max, min, mean and standard deviation of numerical attributes. We have used R programming to verify structure and summary of data.

Attributes	**Datatype**
Age	Numeric
Job	Nominal
Marital	Nominal
Education	Ordinal
Default	Nominal- Binary (no,yes)
Balance	Numeric
Housing	Nominal-Binary (no,yes)
Loan	Nominal- Binary (no,yes)
Contact	Nominal
Day	Numeric
Month	Nominal
Duration	Numeric
Campaign	Numeric
Pdays	Numeric
Previous	Numeric
Poutcome	Nominal
Y	Binary (no,yes)

Observation

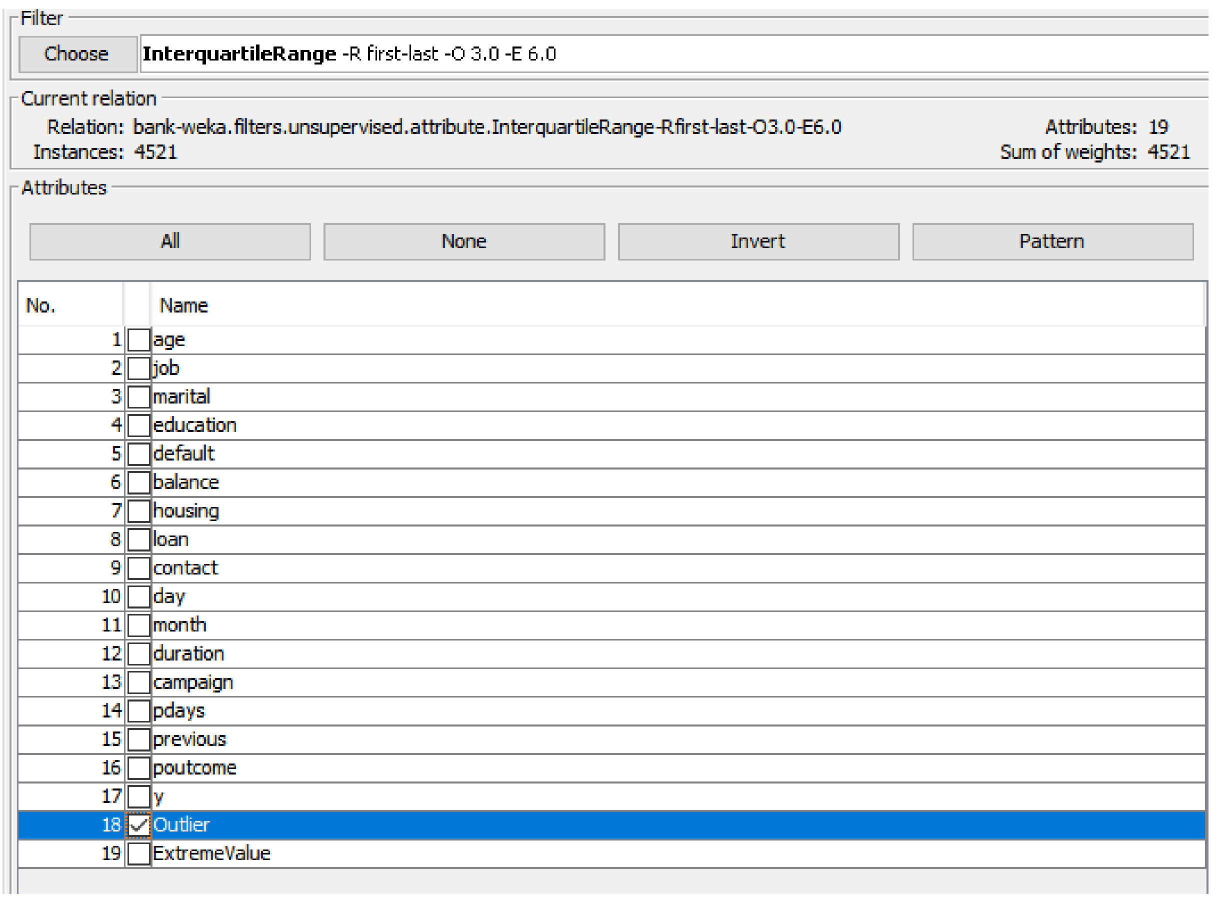
After summarizing the data, we see customers’ age is at an average of 41 and the minimum age being 18 and maximum age recorded is 95. Martial status shows 60% of customers are married whereas 30% are single and around 10% are divorced. Another attribute shows majority of people got secondary education followed by tertiary and primary. Also, in the class attribute we see there are 4000 in no and 521 in yes, which shows data set is imbalanced and is not normally distributed.

```
      age             job              marital
 Min.   :19.00   Length:4521        Length:4521
 1st Qu.:33.00   Class :character   Class :character
 Median :39.00   Mode  :character   Mode  :character
 Mean   :41.17
 3rd Qu.:49.00
 Max.   :87.00
   housing               loan              contact
 Length:4521        Length:4521        Length:4521
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character


     pdays           previous          poutcome
 Min.   : -1.00   Min.   : 0.0000   Length:4521
 1st Qu.: -1.00   1st Qu.: 0.0000   Class :character
 Median : -1.00   Median : 0.0000   Mode  :character
 Mean   : 39.77   Mean   : 0.5426
 3rd Qu.: -1.00   3rd Qu.: 0.0000
 Max.   :871.00   Max.   :25.0000


  education            default            balance
 Length:4521        Length:4521        Min.   :-3313
 Class :character   Class :character   1st Qu.:   69
 Mode  :character   Mode  :character   Median :  444
                                       Mean   : 1423
                                       3rd Qu.: 1480
                                       Max.   :71188
        day           month             duration        campaign
   Min.   : 1.00   Length:4521        Min.   :   4   Min.   : 1.000
   1st Qu.: 9.00   Class :character   1st Qu.: 104   1st Qu.: 1.000
   Median :16.00   Mode  :character   Median : 185   Median : 2.000
   Mean   :15.92                      Mean   : 264   Mean   : 2.794
   3rd Qu.:21.00                      3rd Qu.: 329   3rd Qu.: 3.000
   Max.   :31.00                      Max.   :3025   Max.   :50.000
      y
 Length:4521
 Class :character
 Mode  :character
```

Outliers

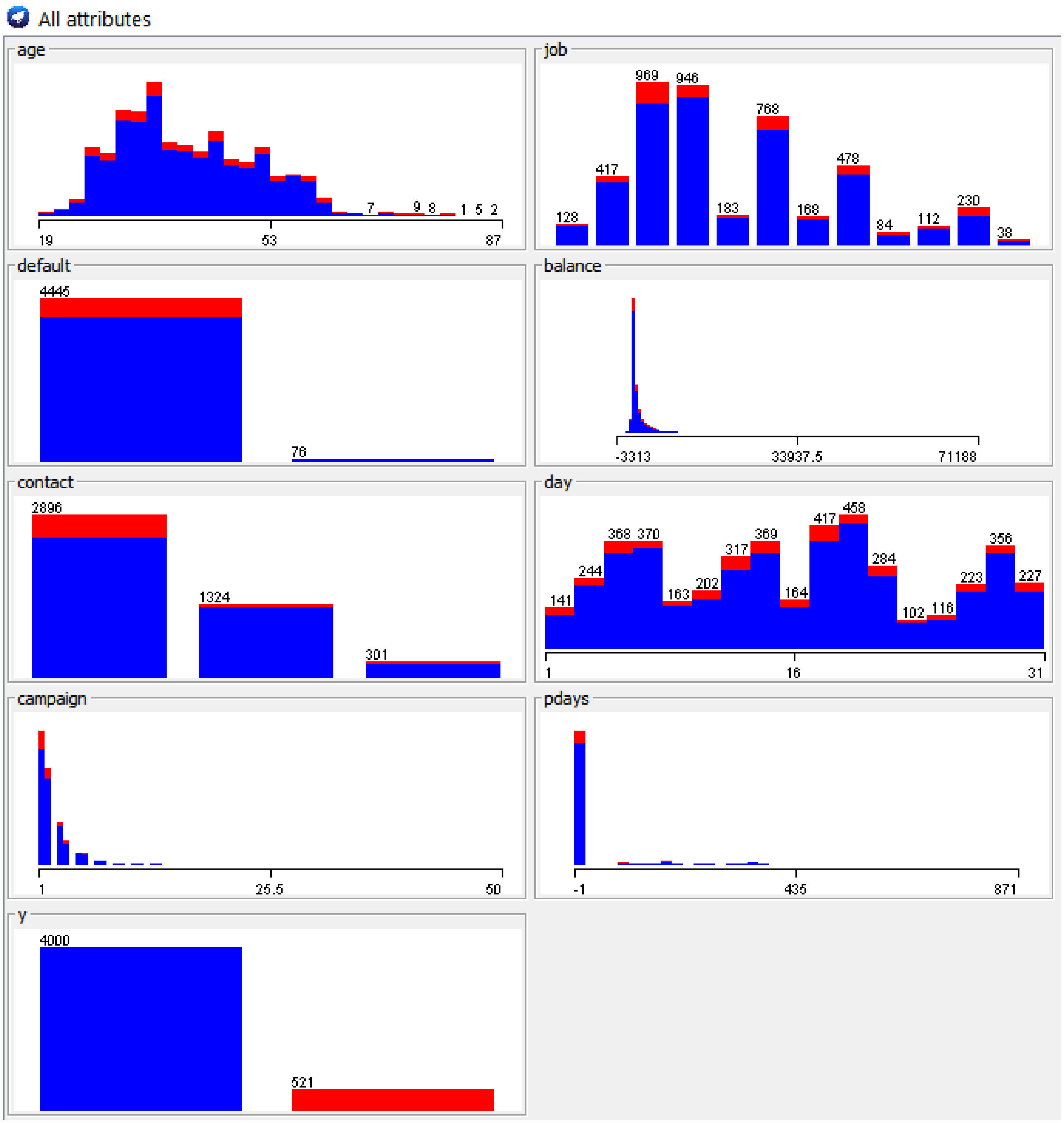
Here we are using WEKA tool for numerical attributes to find outliers. Box plot defines the attribute structure into minimum, Quartile 1(Q1), Median, Quartile 3(Q3) and maximum. Lower fence is calculated as Q1-1.5IQR and upper fence as Q3+ 1.5IQR.
Position of Q1 is given as .25(n+1) and Q3 as .75(n+1).
IQR= Q3-Q1

In WEKA, we have chosen filter caller Interquartile Range and once apply, it creates new attribute in dataset called “Outlier”. This selected attribute shows for all instances there are 355 outliers over 4166 -not outliers. Ratio comparatively is quite low- 0.085%. We see that there are not many outliers, hence we decided to keep data as it is as collected originally.

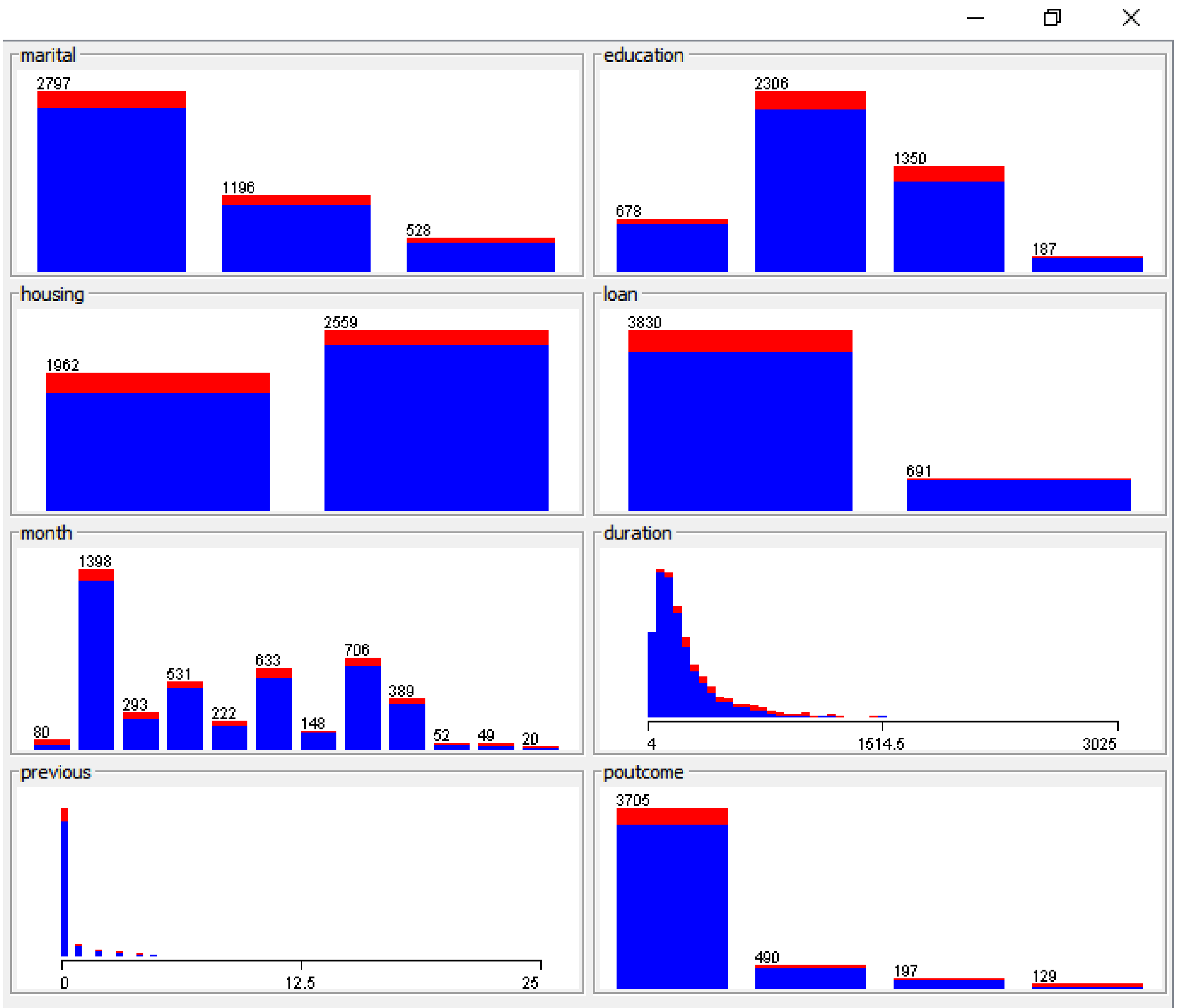


Histogram of all Attributes

Using Weka , histogram of all attributes are given below. We see no attribute is normally distributed and shows imbalancing.



All attributes
age
19
53
87
7
9
8
1
5
2
job
128
417
969
946
183
768
168
478
84
112
230
38
default
4445
76
balance
-3313
33937.5
71188
contact
2896
1324
301
day
141
244
368
370
163
202
317
369
164
417
458
284
102
116
223
356
227
1
16
31
campaign
1
25.5
50
pdays
-1
435
871
y
4000
521



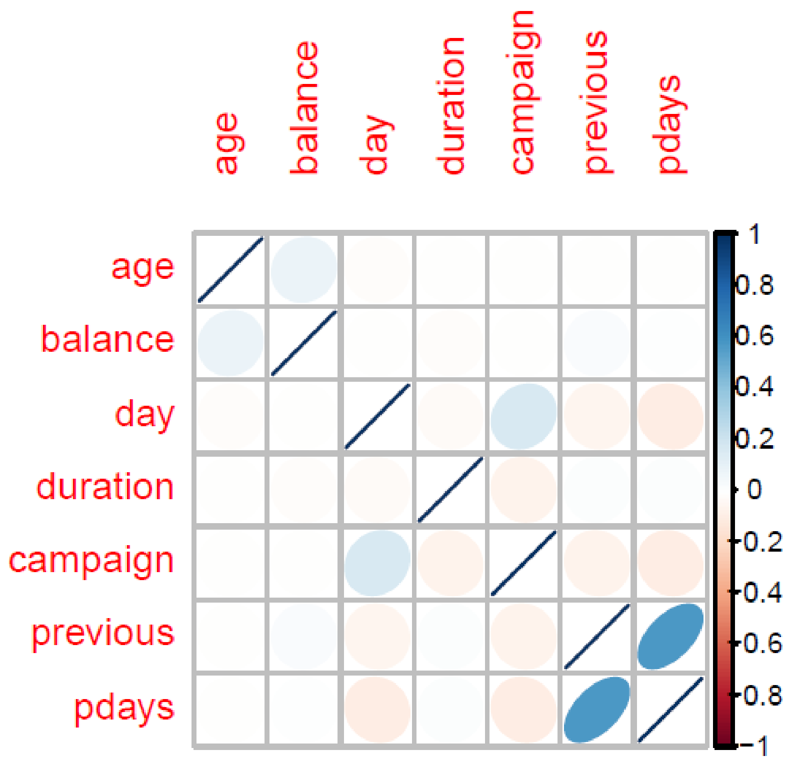
Correlation

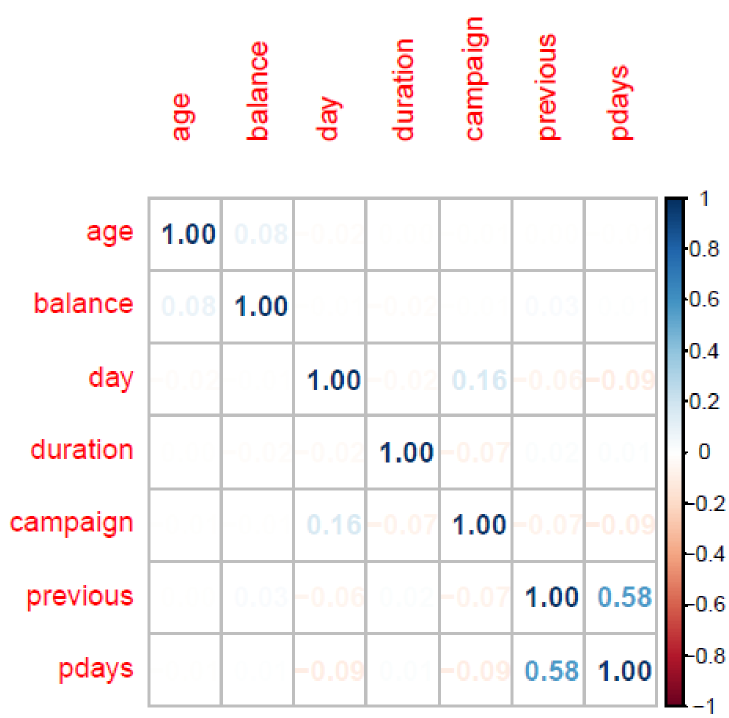
<u>Observation: -</u>

In R, we executed the code given below and observed no attributes (numerical) are strongly negatively correlated. All are weakly or strongly positively correlated. There is no much corelation among campaign, duration, balance and age.

Pdays and previous are strongly correlated to each other.

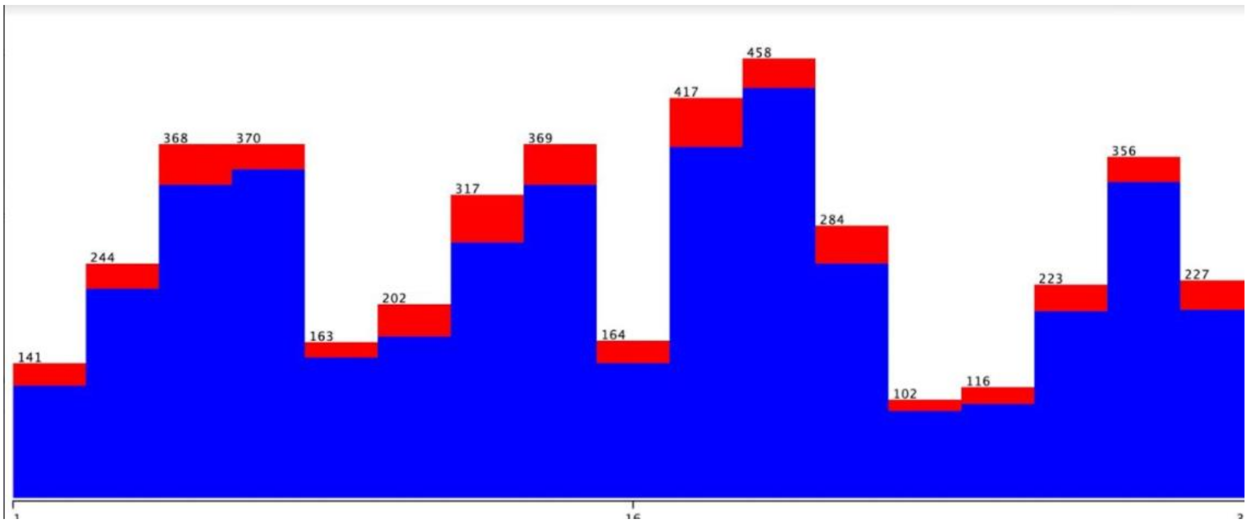
```
library(corrplot)
install.packages("corrplot")
bd <- bankdata[c('age','balance','day','duration','campaign','previous','pdays')]
M <- cor(bd)
corrplot(M, method= "ellipse")
```


age
balance
day
duration
campaign
previous
pdays
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

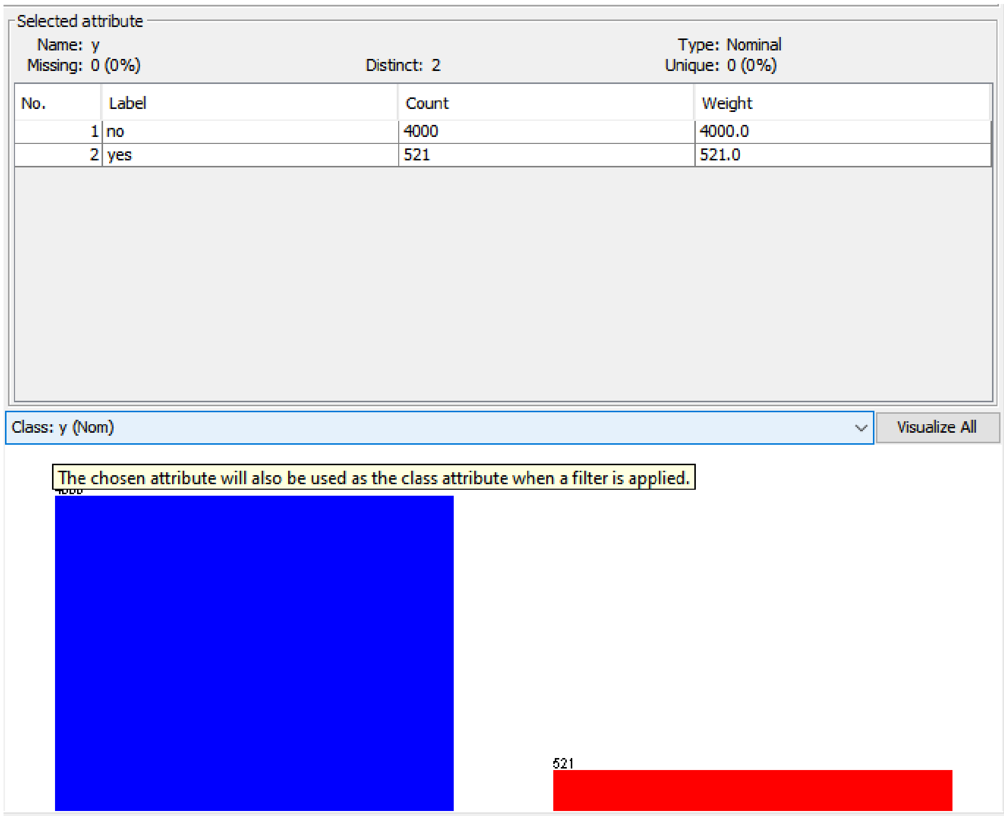

age
balance
day
duration
campaign
previous
pdays
1.00
0.16
0.58
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Removal of attribute and balancing of dataset

Based on coorelation matrix above, we observed pdays and poutcome are strongly correlated, as a result chossing one of themwould have have same influence on our prediction. Based on the disribution below, we seein spite of varying distribution of day it doesnt have much influence on class attribute. Hence based on our observations, ‘pday’ could be removed from data set.



Using WEKA, we found our dependent variabe determines whether customer would suscribe to term deposit accounts or not has imbalanced class distribution ‘no’ count at 4000 and ‘yes’ at 521. . We are not balancing data set in this report.


Selected attribute
Name: y
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)
No. Label Count Weight
1 no 4000 4000.0
2 yes 521 521.0
Class: y (Nom)
Visualize All
The chosen attribute will also be used as the class attribute when a filter is applied.
521

Predictive Modeling/Classification

Predictive Modeling

Learning Method: Classification Tree

We favoured the Classification Tree since it is easy to comprehend and quickly visualize where it is more beneficial to invest time and resources on. We believe this will lead to the marketing team to make a better educated decision.

We used three different classifiers: Naïve-Bayes, J48 and Random Forest.

Detailed Accuracy by Class (Yes)

Classifier	Correctly Classified Instances Rate	TP Rate for "yes" Class	FP Rate for "yes" Class
Naïve-Bayes	86.88%	0.509	0.084
J48	88.98%	0.355	0.041
Random Forest	89.80%	0.269	0.020

J48

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances   4023   88.9847 %
Incorrectly Classified Instances   498   11.0153 %
Kappa statistic   0.368
Mean absolute error   0.1448
Root mean squared error   0.2977
Relative absolute error   70.9698 %
Root relative squared error   93.2371 %
Total Number of Instances   4521

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.960	0.645	0.920	0.960	0.939	0.377	0.762	0.935	no
	0.355	0.041	0.533	0.355	0.426	0.377	0.762	0.387	yes
Weighted Avg.	0.890	0.575	0.875	0.890	0.880	0.377	0.762	0.871	

=== Confusion Matrix ===

a   b   <-- classified as

3838 162 | a = no
336 185 | b = yes

We are using here classification algorithm J48 and examine the output generated using this algorithm. Selecting all 16attributes, how accurately it can predict 17th one. J48 doesn't work with numeric classes and works with nominal variables. Choosing nominal attribute "Y" which is our dependent variable and using 10 fold cross validation as test option out of other 4 test options (test/train is other method). Correctly classified instances are 4023 in total that makes accuracy at 88.98% close to 90%.

Confusion matrix also proves 498 was classified incorrectly and rest are classified correctly. Higher accuracy and likelihood of true positive and true negative would help us to determine the algorithm with higher accuracy and correctly classified instances.

Naïve-Bayes

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 3928 86.8834 %
Incorrectly Classified Instances 593 13.1166 %
Kappa statistic 0.3975
Mean absolute error 0.1625
Root mean squared error 0.3233
Relative absolute error 79.6454 %
Root relative squared error 101.2447 %
Total Number of Instances 4521

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.916	0.491	0.935	0.916	0.925	0.399	0.845	0.972	no
	0.509	0.084	0.440	0.509	0.472	0.399	0.845	0.403	yes
Weighted Avg.	0.869	0.444	0.878	0.869	0.873	0.399	0.845	0.906	

=== Confusion Matrix ===

a b <-- classified as
3663 337 | a = no
256 265 | b = yes

Random Forest

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	4060	89.8031 %
Incorrectly Classified Instances	461	10.1969 %
Kappa statistic	0.3322	
Mean absolute error	0.1421	
Root mean squared error	0.2649	
Relative absolute error	69.6295 %	
Root relative squared error	82.954 %	
Total Number of Instances	4521	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.980	0.731	0.911	0.980	0.944	0.369	0.908	0.986	no
	0.269	0.020	0.636	0.269	0.378	0.369	0.908	0.537	yes
Weighted Avg.	0.898	0.649	0.880	0.898	0.879	0.369	0.908	0.934	

=== Confusion Matrix ===

a	b	<-- classified as
3920	80 |	a = no
381	140 |	b = yes

Conclusion

Almost three out of four community banks and credit unions admit they do not have a formal data analytics strategy, but for those that can get over the organizational hurdles of implementing a data strategy, the competitive advantages are significant. (Koechlein, 2016)

Direct marketing is becoming a prevalent application of data mining. With benefits that range from identifying prospective customers to test new product adoption to customer retention.

Financial services institutions such as banks have been dropping mass marketing to invest in direct marketing, which provides higher ROI.

By understanding the attributes that most effectively influence the clients' decision to subscribe to term deposit, it is our hope our work will help the bank in choosing the right approach.

In our assessment we analyzed the data of 4521 clients using the tree classification method, using a tree decision and random forest and then tested the model we created. The calculations suggest an accuracy percentage of 89%.

Recommendations

We recommend the bank's marketing team to try to engage the clients via telephone calls, taking the time to understand their needs. The data suggests that call Duration is a good predictor of the client signing for the term deposit. Therefore, spending the necessary time and by conveying confidence in the investment, explaining the value of the proposition, avoiding jargons, especially due to the fact we have a client base with all sorts of education level, the marketing team increases the odds of getting the client to sign.

It is important to ensure the clients understand the offering proposed by the marketing team so that it gives them the assurance it is a good investment in which they want to take part in.

Appendix

J48 pruned tree

```
------------------

duration <= 211: no (2548.0/73.0)
duration > 211
|   duration <= 645
|   |   poutcome = unknown
|   |   |   age <= 60
|   |   |   |   contact = cellular
|   |   |   |   |   month = oct
|   |   |   |   |   |   balance <= 2469: yes (7.0)
|   |   |   |   |   |   balance > 2469: no (5.0/1.0)
|   |   |   |   |   month = may: no (102.0/14.0)
|   |   |   |   |   month = apr
|   |   |   |   |   |   day <= 20: no (58.0/5.0)
|   |   |   |   |   |   day > 20
|   |   |   |   |   |   |   duration <= 238: no (3.0)
|   |   |   |   |   |   |   duration > 238: yes (13.0/1.0)
|   |   |   |   |   month = jun: yes (21.0/7.0)
|   |   |   |   |   month = feb
|   |   |   |   |   |   day <= 7
|   |   |   |   |   |   |   balance <= 1: yes (3.0/1.0)
|   |   |   |   |   |   |   balance > 1: no (37.0)
|   |   |   |   |   |   day > 7: yes (9.0/3.0)
|   |   |   |   |   month = aug: no (153.0/19.0)
|   |   |   |   |   month = jan: no (32.0/1.0)
|   |   |   |   |   month = jul: no (192.0/7.0)
|   |   |   |   |   month = nov: no (86.0/9.0)
|   |   |   |   |   month = sep: no (7.0/1.0)
|   |   |   |   |   month = mar
|   |   |   |   |   |   housing = no
|   |   |   |   |   |   |   duration <= 312: no (2.0)
|   |   |   |   |   |   |   duration > 312: yes (2.0)
|   |   |   |   |   |   housing = yes: yes (4.0)
|   |   |   |   |   month = dec: no (3.0/1.0)
|   |   |   |   contact = unknown: no (464.0/16.0)
|   |   |   |   contact = telephone: no (50.0/10.0)
|   |   |   age > 60
|   |   |   |   age <= 68: yes (21.0/8.0)
|   |   |   |   age > 68: no (16.0/7.0)
|   |   poutcome = failure
|   |   |   pdays <= 373: no (163.0/26.0)
|   |   |   pdays > 373
|   |   |   |   balance <= 2581: yes (9.0)
```

```
|   |   |   |   balance > 2581: no (2.0)
|   |   poutcome = other
|   |   |   month = oct: yes (3.0/1.0)
|   |   |   month = may: no (20.0/1.0)
|   |   |   month = apr
|   |   |   |   campaign <= 5: no (11.0/1.0)
|   |   |   |   campaign > 5: yes (2.0)
|   |   |   month = jun: yes (7.0/2.0)
|   |   |   month = feb: no (4.0/1.0)
|   |   |   month = aug: yes (6.0/1.0)
|   |   |   month = jan: no (6.0)
|   |   |   month = jul: no (1.0)
|   |   |   month = nov
|   |   |   |   age <= 32: yes (4.0)
|   |   |   |   age > 32: no (6.0)
|   |   |   month = sep: no (3.0/1.0)
|   |   |   month = mar: no (1.0)
|   |   |   month = dec: no (0.0)
|   |   poutcome = success: yes (76.0/16.0)
|   duration > 645
|   |   marital = married
|   |   |   default = no
|   |   |   |   contact = cellular
|   |   |   |   |   job = unemployed
|   |   |   |   |   |   balance <= 640: no (3.0)
|   |   |   |   |   |   balance > 640: yes (2.0)
|   |   |   |   |   job = services
|   |   |   |   |   |   loan = no: yes (6.0)
|   |   |   |   |   |   loan = yes: no (4.0/1.0)
|   |   |   |   |   job = management
|   |   |   |   |   |   poutcome = unknown
|   |   |   |   |   |   |   month = oct: no (0.0)
|   |   |   |   |   |   |   month = may: yes (3.0/1.0)
|   |   |   |   |   |   |   month = apr: no (1.0)
|   |   |   |   |   |   |   month = jun: no (1.0)
|   |   |   |   |   |   |   month = feb: yes (1.0)
|   |   |   |   |   |   |   month = aug: yes (5.0/1.0)
|   |   |   |   |   |   |   month = jan: no (1.0)
|   |   |   |   |   |   |   month = jul
|   |   |   |   |   |   |   |   age <= 43: no (8.0/2.0)
|   |   |   |   |   |   |   |   age > 43: yes (2.0)
|   |   |   |   |   |   |   month = nov
|   |   |   |   |   |   |   |   age <= 50: yes (4.0/1.0)
|   |   |   |   |   |   |   |   age > 50: no (3.0)
|   |   |   |   |   |   |   month = sep: no (0.0)
|   |   |   |   |   |   |   month = mar: no (0.0)
|   |   |   |   |   |   |   month = dec: no (0.0)
|   |   |   |   |   |   poutcome = failure
```

```
|   |   |   |   |   |   |   education = primary: no (0.0)
|   |   |   |   |   |   |   education = secondary: yes (2.0)
|   |   |   |   |   |   |   education = tertiary: no (3.0)
|   |   |   |   |   |   |   education = unknown: no (0.0)
|   |   |   |   |   |   poutcome = other: yes (2.0)
|   |   |   |   |   |   poutcome = success: yes (1.0)
|   |   |   |   |   job = blue-collar
|   |   |   |   |   |   housing = no
|   |   |   |   |   |   |   campaign <= 3: no (7.0)
|   |   |   |   |   |   |   campaign > 3: yes (2.0)
|   |   |   |   |   |   housing = yes
|   |   |   |   |   |   |   previous <= 1
|   |   |   |   |   |   |   |   campaign <= 1: no (7.0/1.0)
|   |   |   |   |   |   |   |   campaign > 1
|   |   |   |   |   |   |   |   |   duration <= 1073: yes (4.0)
|   |   |   |   |   |   |   |   |   duration > 1073: no (2.0)
|   |   |   |   |   |   |   previous > 1: yes (3.0)
|   |   |   |   |   job = self-employed
|   |   |   |   |   |   campaign <= 3: no (2.0)
|   |   |   |   |   |   campaign > 3: yes (3.0)
|   |   |   |   |   job = technician: yes (10.0/2.0)
|   |   |   |   |   job = entrepreneur: no (7.0/2.0)
|   |   |   |   |   job = admin.: no (5.0/1.0)
|   |   |   |   |   job = student: no (2.0/1.0)
|   |   |   |   |   job = housemaid: no (2.0)
|   |   |   |   |   job = retired
|   |   |   |   |   |   pdays <= 28: no (3.0)
|   |   |   |   |   |   pdays > 28: yes (2.0)
|   |   |   |   |   job = unknown: yes (3.0/1.0)
|   |   |   |   contact = unknown: no (67.0/18.0)
|   |   |   |   contact = telephone: no (19.0/7.0)
|   |   |   default = yes: yes (3.0)
|   |   marital = single: yes (101.0/39.0)
|   |   marital = divorced
|   |   |   poutcome = unknown
|   |   |   |   duration <= 924
|   |   |   |   |   job = unemployed: no (2.0)
|   |   |   |   |   job = services: no (2.0/1.0)
|   |   |   |   |   job = management: no (4.0/1.0)
|   |   |   |   |   job = blue-collar
|   |   |   |   |   |   balance <= -145: yes (2.0)
|   |   |   |   |   |   balance > -145: no (5.0)
|   |   |   |   |   job = self-employed: yes (2.0)
|   |   |   |   |   job = technician: no (0.0)
|   |   |   |   |   job = entrepreneur: no (2.0)
|   |   |   |   |   job = admin.: yes (4.0/1.0)
|   |   |   |   |   job = student: no (0.0)
|   |   |   |   |   job = housemaid: yes (1.0)
```

```
|   |   |   |   |   |   job = retired: yes (2.0)
|   |   |   |   |   |   job = unknown: no (0.0)
|   |   |   |   |   duration > 924: yes (18.0/1.0)
|   |   |   |   poutcome = failure
|   |   |   |   |   duration <= 834: yes (2.0)
|   |   |   |   |   duration > 834: no (3.0)
|   |   |   |   poutcome = other: yes (3.0/1.0)
|   |   |   |   poutcome = success: no (1.0)
```

Number of Leaves :   104

Size of the tree :   146

Time taken to build model: 3.53 seconds

Naive Bayes Classifier

```
---------------------------------------------------------------
                 Class
Attribute           no       yes
                (0.88)   (0.12)
=====================================
age
  mean            41.0264   42.5015
  std. dev.       10.2053    13.145
  weight sum         4000       521
  precision        1.0303    1.0303

job
  unemployed        116.0      14.0
  services          380.0      39.0
  management        839.0     132.0
  blue-collar       878.0      70.0
  self-employed     164.0      21.0
  technician        686.0      84.0
  entrepreneur      154.0      16.0
  admin.            421.0      59.0
  student            66.0      20.0
  housemaid          99.0      15.0
  retired           177.0      55.0
  unknown            32.0       8.0
  [total]          4012.0     533.0

marital
  married          2521.0     278.0
  single           1030.0     168.0
  divorced          452.0      78.0
  [total]          4003.0     524.0
```

education

primary	615.0	65.0
secondary	2062.0	246.0
tertiary	1158.0	194.0
unknown	169.0	20.0
[total]	4004.0	525.0

default

no	3934.0	513.0
yes	68.0	10.0
[total]	4002.0	523.0

balance

mean	1403.0309	1572.1066
std. dev.	3075.0271	2442.0069
weight sum	4000	521
precision	31.6756	31.6756

housing

no	1662.0	302.0
yes	2340.0	221.0
[total]	4002.0	523.0

loan

no	3353.0	479.0
yes	649.0	44.0
[total]	4002.0	523.0

contact

cellular	2481.0	417.0
unknown	1264.0	62.0
telephone	258.0	45.0
[total]	4003.0	524.0

day

mean	15.9488	15.6583
std. dev.	8.2487	8.2272
weight sum	4000	521
precision	1	1

month

oct	44.0	38.0
may	1306.0	94.0
apr	238.0	57.0
jun	477.0	56.0
feb	185.0	39.0
aug	555.0	80.0
jan	133.0	17.0

```
  jul             646.0     62.0
  nov              351.0     40.0
  sep              36.0     18.0
  mar               29.0     22.0
  dec              12.0     10.0
  [total]         4012.0    533.0

duration
  mean            226.3607  552.8113
  std. dev.        210.294  389.9421
  weight sum          4000       521
  precision         3.4565    3.4565

campaign
  mean               3.023    2.4999
  std. dev.         3.1626    2.0411
  weight sum          4000       521
  precision         1.5806    1.5806

pdays
  mean             36.8015    69.243
  std. dev.        95.8516  121.4075
  weight sum          4000       521
  precision         2.9966    2.9966

previous
  mean              0.4943    1.1558
  std. dev.         1.6518    2.1173
  weight sum          4000       521
  precision          1.087     1.087

poutcome
  unknown          3369.0    338.0
  failure          428.0     64.0
  other            160.0     39.0
  success           47.0     84.0
  [total]         4004.0    525.0
```

Time taken to build model: 0.06 seconds

Random Forest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 4.67 seconds

References

Koechlein, F. (December, 2016). Maximizing Marketing ROI With Data Analytics. The Financial Brand. https://thefinancialbrand.com/62466/marketing-data-analytics-banking/