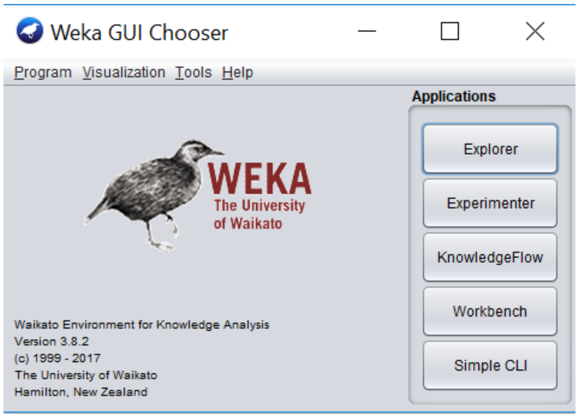
Documentation: Attribute Importance Analysis of Bank Marketing Data

Meghana Gaopande
meghanag@vt.edu

Pranavi Rambhakta
pranavi7@vt.edu

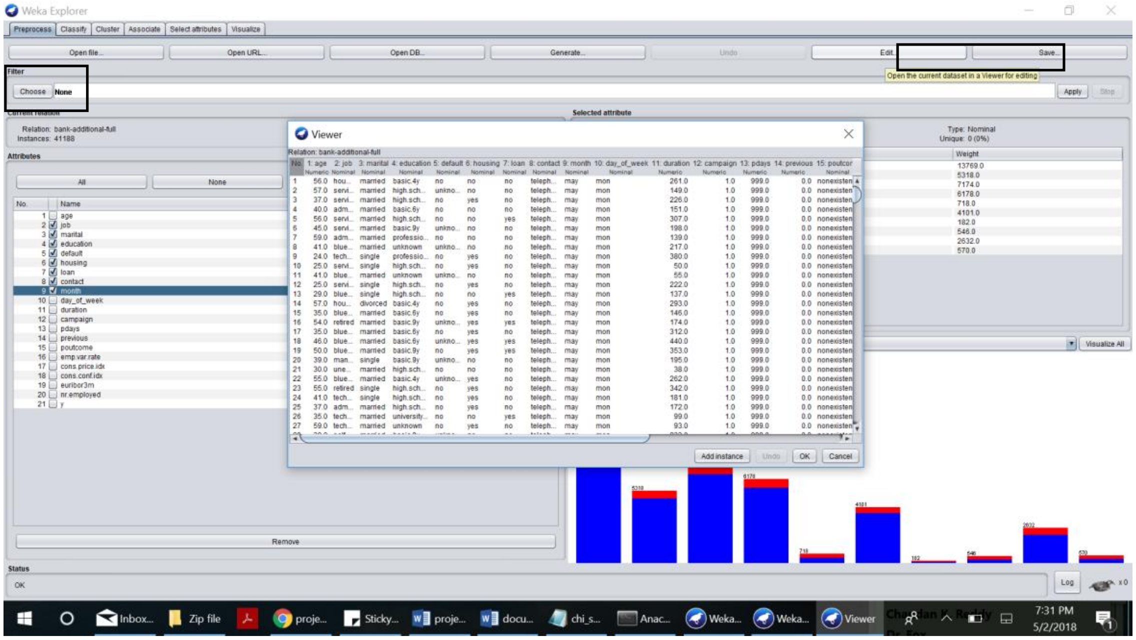
In the project, Weka software is used for data-preprocessing, model building and model evaluation of the data while Matlab and Python (Jupyter Notebook web application) is used for exploratory analysis and visualization of the data.

The Weka explorer can be accessed from the weka homepage by clicking on the "Explorer" option.



Waikato Environment for Knowledge Analysis
Version 3.8.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

There are direct navigation options available for Preprocess, Classify, Cluster, Associate, Select attributes and Visualize in which Preprocess page is the default. As shown below in the screenshot, the dataset can be loaded from the computer using the 'open file' option at the top left corner. The current dataset can be viewed for editing using the "Edit" option. The attributes are all listed and they can be selectively chosen individually and filters can be applied only on the attributes selected. The "filter" option is on the left top corner, just below the open file option. Additionally, the distribution of class labels for each attribute can be visualized in the form of histograms in the right bottom corner.



Filter: None

Current relation

Relation: bank-additional-full
Instances: 41188

Attributes

No.		Name
1	□	age
2	☑	job
3	☑	marital
4	☑	education
5	☑	default
6	☑	housing
7	☑	loan
8	☑	contact
9	☑	month
10	□	day_of_week
11	□	duration
12	□	campaign
13	□	pdays
14	□	previous
15	□	poutcome
16	□	emp.var.rate
17	□	cons.price.idx
18	□	cons.conf.idx
19	□	euribor3m
20	□	nr.employed
21	□	y

Viewer

Relation: bank-additional-full

No.	1: age Numeric	2: job Nominal	3: marital Nominal	4: education Nominal	5: default Nominal	6: housing Nominal	7: loan Nominal	8: contact Nominal	9: month Nominal	10: day_of_week Nominal	11: duration Numeric	12: campaign Numeric	13: pdays Numeric	14: previous Numeric	15: poutcor Nominal
1	56.0	hou...	married	basic.4y	no	no	no	teleph...	may	mon	261.0	1.0	999.0	0.0	nonexisten
2	57.0	servi...	married	high.sch...	unkno...	no	no	teleph...	may	mon	149.0	1.0	999.0	0.0	nonexisten
3	37.0	servi...	married	high.sch...	no	yes	no	teleph...	may	mon	226.0	1.0	999.0	0.0	nonexisten
4	40.0	adm...	married	basic.6y	no	no	no	teleph...	may	mon	151.0	1.0	999.0	0.0	nonexisten
5	56.0	servi...	married	high.sch...	no	no	yes	teleph...	may	mon	307.0	1.0	999.0	0.0	nonexisten
6	45.0	servi...	married	basic.9y	unkno...	no	no	teleph...	may	mon	198.0	1.0	999.0	0.0	nonexisten
7	59.0	adm...	married	professio...	no	no	no	teleph...	may	mon	139.0	1.0	999.0	0.0	nonexisten
8	41.0	blue...	married	unknown	unkno...	no	no	teleph...	may	mon	217.0	1.0	999.0	0.0	nonexisten
9	24.0	tech...	single	professio...	no	yes	no	teleph...	may	mon	380.0	1.0	999.0	0.0	nonexisten
10	25.0	servi...	single	high.sch...	no	yes	no	teleph...	may	mon	50.0	1.0	999.0	0.0	nonexisten
11	41.0	blue...	married	unknown	unkno...	no	no	teleph...	may	mon	55.0	1.0	999.0	0.0	nonexisten
12	25.0	servi...	single	high.sch...	no	yes	no	teleph...	may	mon	222.0	1.0	999.0	0.0	nonexisten
13	29.0	blue...	single	high.sch...	no	no	yes	teleph...	may	mon	137.0	1.0	999.0	0.0	nonexisten
14	57.0	hou...	divorced	basic.4y	no	yes	no	teleph...	may	mon	293.0	1.0	999.0	0.0	nonexisten
15	35.0	blue...	married	basic.6y	no	yes	no	teleph...	may	mon	146.0	1.0	999.0	0.0	nonexisten
16	54.0	retired	married	basic.9y	unkno...	yes	yes	teleph...	may	mon	174.0	1.0	999.0	0.0	nonexisten
17	35.0	blue...	married	basic.6y	no	yes	no	teleph...	may	mon	312.0	1.0	999.0	0.0	nonexisten
18	46.0	blue...	married	basic.6y	unkno...	yes	yes	teleph...	may	mon	440.0	1.0	999.0	0.0	nonexisten
19	50.0	blue...	married	basic.9y	no	yes	yes	teleph...	may	mon	353.0	1.0	999.0	0.0	nonexisten
20	39.0	man...	single	basic.9y	unkno...	no	no	teleph...	may	mon	195.0	1.0	999.0	0.0	nonexisten
21	30.0	une...	married	high.sch...	no	no	no	teleph...	may	mon	38.0	1.0	999.0	0.0	nonexisten
22	55.0	blue...	married	basic.4y	unkno...	yes	no	teleph...	may	mon	262.0	1.0	999.0	0.0	nonexisten
23	55.0	retired	single	high.sch...	no	yes	no	teleph...	may	mon	342.0	1.0	999.0	0.0	nonexisten
24	41.0	tech...	single	high.sch...	no	yes	no	teleph...	may	mon	181.0	1.0	999.0	0.0	nonexisten
25	37.0	adm...	married	high.sch...	no	yes	no	teleph...	may	mon	172.0	1.0	999.0	0.0	nonexisten
26	35.0	tech...	married	university...	no	no	yes	teleph...	may	mon	99.0	1.0	999.0	0.0	nonexisten
27	59.0	tech...	married	unknown	no	yes	no	teleph...	may	mon	93.0	1.0	999.0	0.0	nonexisten

Selected attribute

Type: Nominal
Unique: 0 (0%)

Weight
13769.0
5318.0
7174.0
6178.0
718.0
4101.0
182.0
546.0
2632.0
570.0

Status

OK

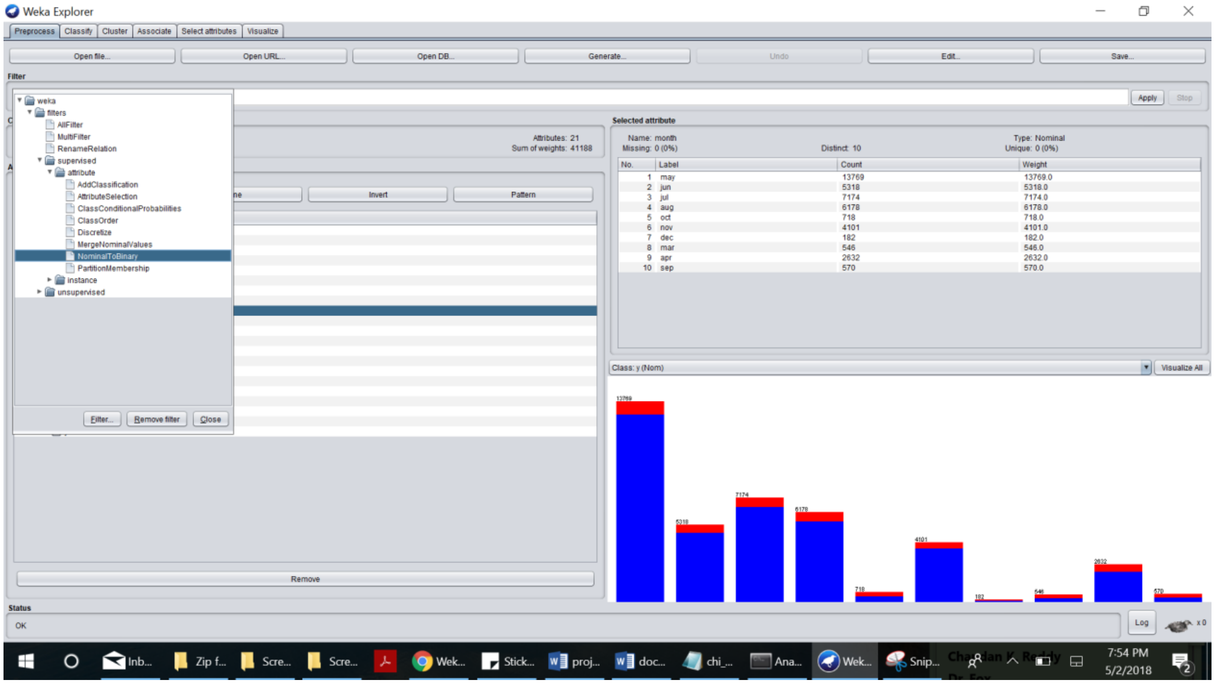
7:31 PM
5/2/2018

Data Preprocessing:

Unordered categorical attributes can be selected and to perform One Hot Encoding on them, we used the filter: weka->supervised->NominalToBinary

Integer Encoding on ordinal categorical data can be performed by using the filter: weka->unsupervised->OrdinalToNumeric

Though weka is used for One Hot Encoding and Integer encoding, it is also coded in python (The code for is available in “data_exploration.py”- code folder)



Attributes: 21
Sum of weights: 41188

Selected attribute

Name: month
Missing: 0 (0%)
Distinct: 10
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	may	13769	13769.0
2	jun	5318	5318.0
3	jul	7174	7174.0
4	aug	6178	6178.0
5	oct	718	718.0
6	nov	4101	4101.0
7	dec	182	182.0
8	mar	546	546.0
9	apr	2632	2632.0
10	sep	570	570.0

Status

OK

7:54 PM
5/2/2018

The percentage of the missing values for each attribute in the data is shown in the form of a bar plot (The code for is available in “data_exploration.py”- code folder)

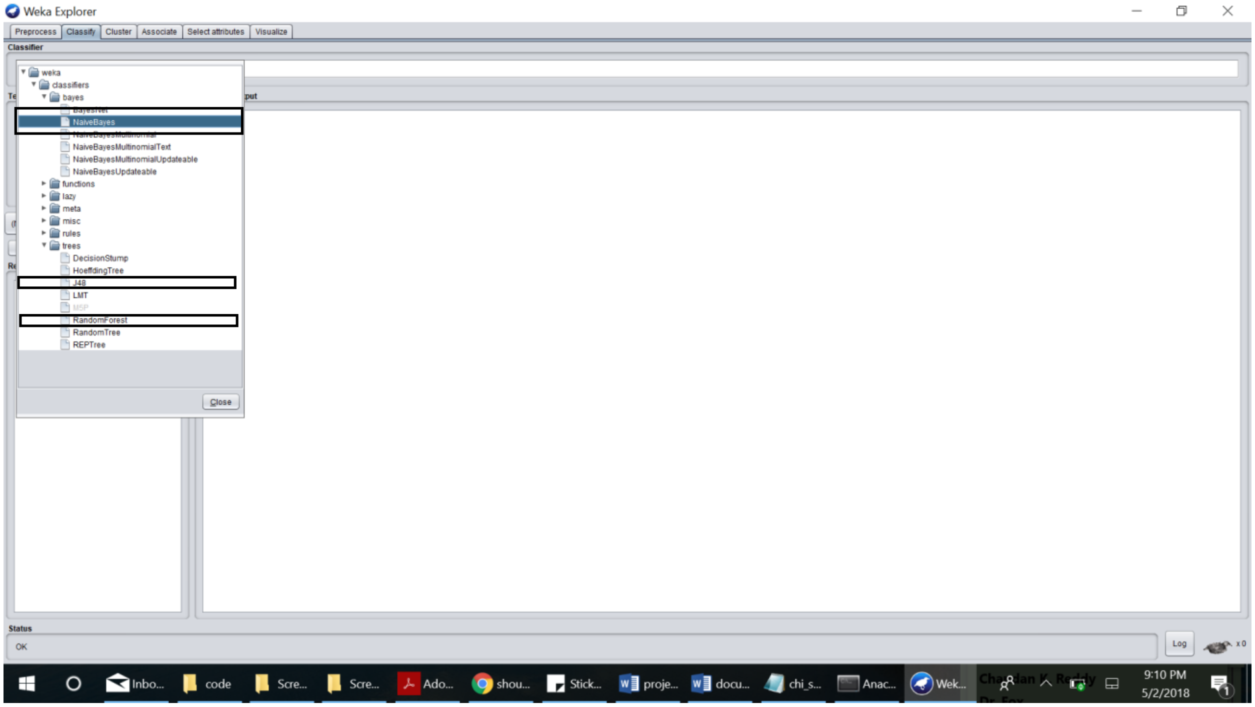
Data Exploration:

The exploratory analysis and visualization of numeric data is shown in the form of correlation matrices and boxplots (The code is available in “box_and_corr_additional.m” and “box_and_corr_full.m”- code folder).

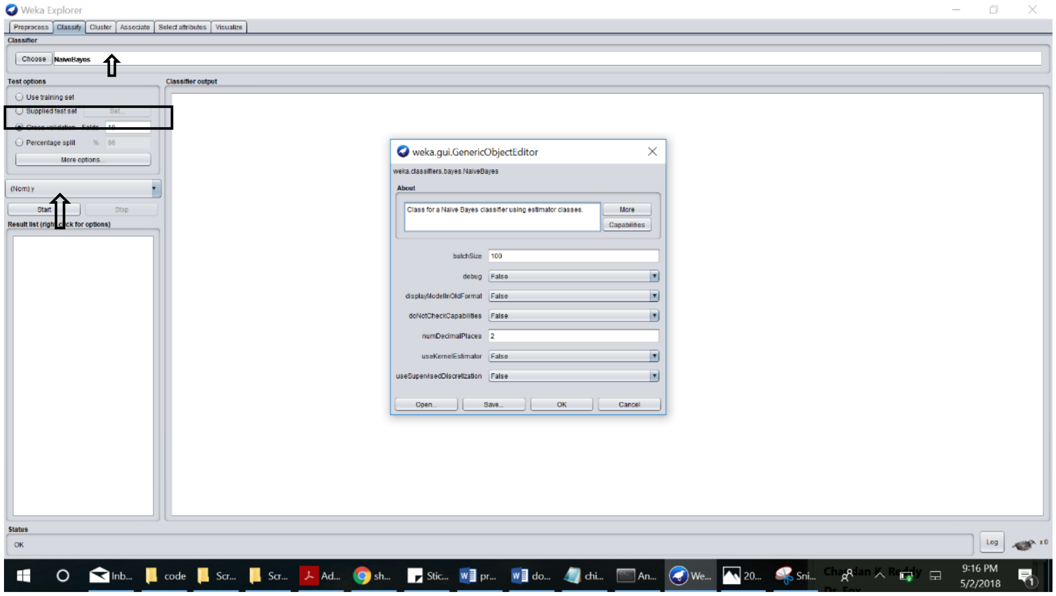
Additionally, the relation between nominal attributes is shown in the form of scatters plots and correlation plots. Chi-squared test is also performed to further analyze their correlation. (The code is available in “data_exploration.py”- code folder)

Model Building:

Classification models can be chosen from the option “choose” in classify page. The drop down has all the model from which one can be selected. We have used three classification models: NaïveBayes, J48 and RandomForest. Decision tree is J48 in weka.



The parameters of the classification model can be specified or modified by clicking on the chosen model as shown below. Further, datasets for training and testing can also specified. We have used the default parameters and have chosen 10-fold cross validation. On clicking on the start, the classifier model starts running to produce output which will be displayed in the classifier output section. The result buffers are stored in the result list.



Classifier: NaiveBayes

(Nom) y

weka.gui.GenericObjectEditor

weka.classifiers.bayes.NaiveBayes

About

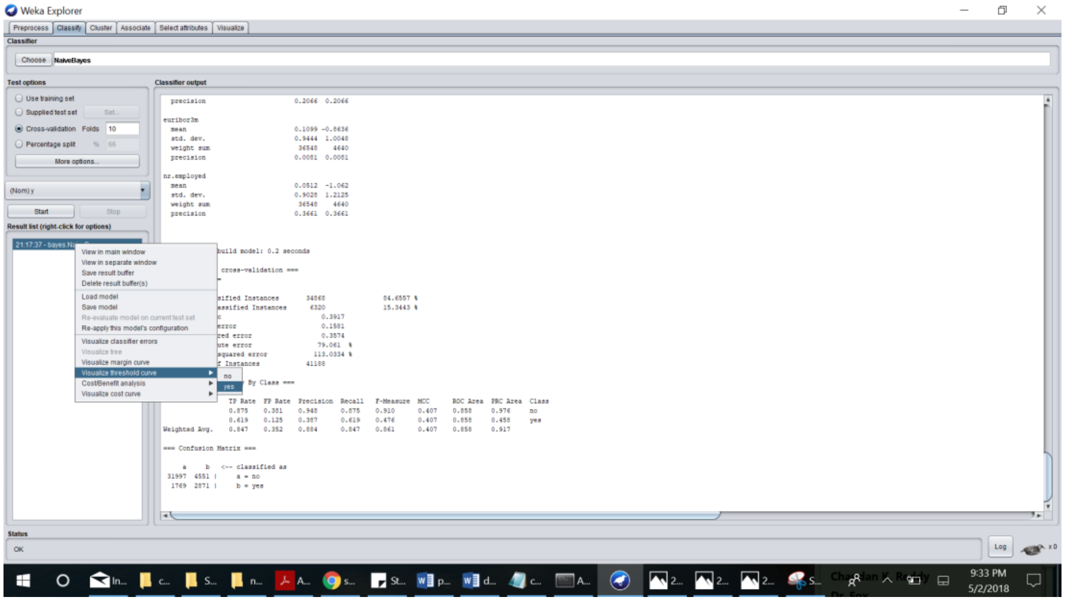
Class for a Naive Bayes classifier using estimator classes.

batchSize	100
debug	False
displayModelInOldFormat	False
doNotCheckCapabilities	False
numDecimalPlaces	2
useKernelEstimator	False
useSupervisedDiscretization	False

Status: OK

9:16 PM 5/2/2018

Apart from the output, results can be visualized by right-clicking on the respective result buffer and choosing from the options as shown below.



Classifier: NaiveBayes

Test options: ● Cross-validation Folds 10

(Nom) y

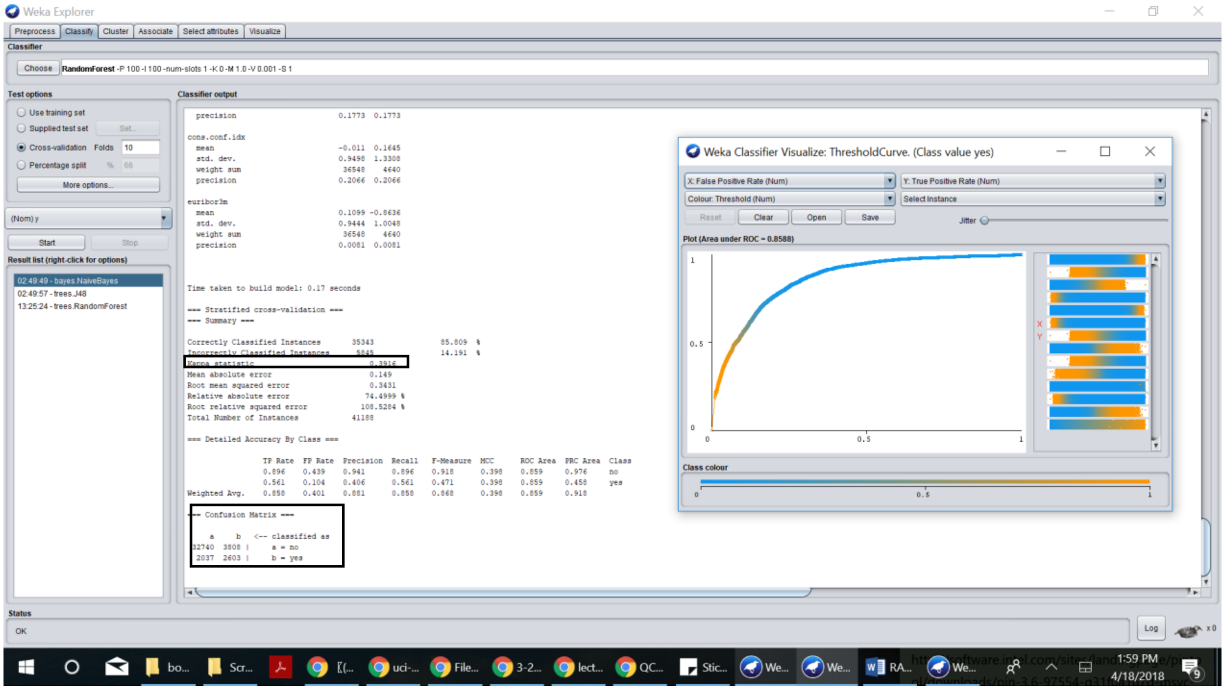
21:17:37 - bayes.Na...

```
  precision                  0.2066  0.2066

euribor3m
  mean                       0.1099 -0.8636
  std. dev.                  0.9444  1.0048
  weight sum                  36548    4640
  precision                  0.0081  0.0081

nr.employed
  mean                       0.0512  -1.062
  std. dev.                  0.9028  1.2125
  weight sum                  36548    4640
  precision                  0.3661  0.3661



build model: 0.2 seconds

 cross-validation ===
=

sified Instances       34868               84.6557 %
assified Instances      6320               15.3443 %
c                                0.3917
error                            0.1581
red error                        0.3574
ute error                       79.061  %
squared error                  113.0334 %
f Instances                41188

  By Class ===

            TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
            0.875    0.381    0.948      0.875    0.910      0.407    0.858     0.976     no
            0.619    0.125    0.387      0.619    0.476      0.407    0.858     0.458     yes
Weighted Avg.    0.847    0.352    0.884      0.847    0.861      0.407    0.858     0.917

=== Confusion Matrix ===

     a     b   <-- classified as
 31997  4551 |     a = no
  1769  2871 |     b = yes
```

Status: OK

9:33 PM 5/2/2018

The results have Kappa statistic, confusion matrix and many other evaluation metrics listed. We have used Kappa statistic as our evaluation metric.



Choose RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) y

02:49:49 - bayes.NaiveBayes
02:49:57 - trees.J48
13:25:24 - trees.RandomForest

Status: OK

```
  precision                   0.1773  0.1773

cons.conf.idx
  mean                        -0.011  0.1645
  std. dev.                   0.9498  1.3308
  weight sum                   36548    4640
  precision                   0.2066  0.2066

euribor3m
  mean                        0.1099 -0.8636
  std. dev.                   0.9444  1.0048
  weight sum                   36548    4640
  precision                   0.0081  0.0081



Time taken to build model: 0.17 seconds

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

Correctly Classified Instances       35343               85.809  %
Incorrectly Classified Instances      5845               14.191  %
Kappa statistic                          0.3916
Mean absolute error                      0.149
Root mean squared error                  0.3431
Relative absolute error                 74.4999 %
Root relative squared error            108.5284 %
Total Number of Instances            41188

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.896    0.439    0.941      0.896    0.918      0.398    0.859     0.976     no
                 0.561    0.104    0.406      0.561    0.471      0.398    0.859     0.458     yes
Weighted Avg.    0.858    0.401    0.881      0.858    0.868      0.398    0.859     0.918

=== Confusion Matrix ===

     a     b   <-- classified as
 32740  3808 |     a = no
  2037  2603 |     b = yes
```

Weka Classifier Visualize: ThresholdCurve. (Class value yes)

X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance