# DEVISING A TELEMARKETING STRATEGY

## PORTUGUESE BANK CASE STUDY: INCREASING SUBSCRIBERS TO LONG-TERM DEPOSIT ACCOUNTS

CIND119 Section 6A0 Team
YoonSeok Kim, Manoj Bhattarai, Surender Kumar, Xiaohan Li, Sonya David

**Devising a Telemarketing Strategy**: Increasing Subscribers to Long-Term Deposit Accounts

## Members

Xiaohan Li, Xiaohan.li@ryerson.ca

YoonSeok Kim, Yoonseok.kim@ryerson.ca

Manoj Bhattarai, Manoj.bhattarai@ryerson.ca

Surender Kumar, Surender1.kumar@ryerson.ca

Sonya David, Sonya.david@ryerson.ca

## Summary:

Telemarketing is one of the most widely used and successful tools for selling banking and financial products around the globe [Ref-0]. A large set of socioeconomic data, based on potential consumers, which has over 4500 instances corresponding to 16 different attributes was analyzed. Prior knowledge of the class exists, and classes are pre-defined, Classification (supervised learning) was used for analysis in order to determine the best classifier algorithm. Discretization was applied to the numeric attributes in order to narrow the outliers and organize the data into bins, discretizaton error accepted. After using evaluators (in Weka), including cfsesubsetval, infogain and gainratio all showed Age, Contact, Loan, Day, and Duration most highly correlated to the class in a supervised learning environment, as such these attributes are used for training and testing and further predictive modeling. Among the three algorithms used, i.e. supervised learning J48, Naïve Bayes, and Random Forest; Naive Bayes gave the most accuracy of the model while doing tenfold cross validation; and also considering 60-40 training and test data set. All of these analysis, provided a model that is valuable to the telemarketing department to consider while making campaign and marketing strategies.

## Workload Distribution

| Member Name | List of Tasks Performed |
|---|---|
| Xiaohan Li | |
| Manoj Bhattarai | Part I: Data Preparation & descriptive statistics, Part II: Predictive Modeling, references, conclusion. |
| Surender Kumar | Part I: Data Preparation Part II: Predictive Modeling |
| YoonSeok Kim | Part I: Data Preparation + box plot with R, Part II: Predictive Modeling, Part III: Post-Predictive Analysis. |
| Sonya David | Document formatting and presentation, Intro (BPF), Part I: Data Preparation, Part II: Predictive Modeling + model comparison, Discussion. |

# Contents

# Business Problem Framing

## Business Problem

The financial product market is increasingly competitive and due to greater customer awareness and sheer amount of available alternative. The strategization and implementation of financial marketing and campaigns are challenging. Companies use direct marketing, like telemarketing, to increase sales and provide sufficient amount of information for buying decision. Marketing operationalized through a contact call center is called telemarketing due to the remoteness characteristic and delivery method [1]. Telemarketing used in this bank is not able to turn the prospect into real customers, so there is the need for an effective telemarketing strategy which will increase the number of customers who are more likely to subscribe to a long-term deposit accounts.

## Stakeholders

Stakeholders are "denoted a type of organization or system in which all the members or participants are seen as having an interest in its success"[2]. In this study, internal and external stakeholders are considered those who are contributing partners in this project or if not at least influenced by the project. The internal and external stakeholders are as listed:

I. The Portuguese bank (finance department, marketing department, IT department, telemarketing department)
II. Customers or clients of the bank
III. Government agencies (bonds)
IV. Outsourcing agencies (promotions, international employees)
V. Data Scientists

## Analytical Problem

The dataset in this study is historical and a combination of nominal and numeric. Quantitative or numerical variable are variables in which numbers serve as values, whereas qualitative variables do not have number values, and remain as categorical or ordinal state [3]. Overall, there were 16 different variables used in this study (7 numeric attributes and 9 qualitative attributes), but upon attribute selection, the variables are narrowed down to 5 attributes. Further, the features of a data set are as below:

1. Clean data, no missing variables
2. Prior knowledge on pre-defined classes (Y/N)
3. Imbalanced dataset (with more instances leading to 'No' class (4000 records)

**Inputs and Drivers**

Inputs

Data collected →

see data type

Data granularity →

relatively low granularity

Effective Telemarketing Strategy

than 'Yes' class (521 records)

Imbalanced datasets are a special case where the class distribution is not uniform among the classes. Typically, they are composed of two classes (binomial): the majority (negative) class and the minority (positive) class [4]. Cross-validations and partitioning of training and test sets are used for overcoming this issue. Further, to analyze the data we have performed data preparation, predictive modeling/classification, post-predictive modeling using clustering methodology.

## Feasibility of Solution

The methodology used in this research project is the historical research design. Historical research has been defined as the systematic evaluation and synthesis of evidence in order to establish facts and draw conclusions about past events using historical data [5]. The conclusion drawn from this study are based on the historical data analysis with the assumption of historical repetition via classification. This study provides a set of attributes which are most highly correlated to subscription success and a strategy to accompany.

# Data Preparation

## Descriptive Statistics

The results of descriptive analysis indicates that most people called were middle aged, management professionals, and educated, who don't have any credit default history. The yearly balance of these people has a wider range with an average of 71000. Most of the called people have no housing loan and personal (meaning that they have greater investment potentiality). The calls were made throughout the year, but in the month of May most calls were made. Second week of every month was comparatively busy days to call as well. The average time to call was about 4 and half minutes. Based on descriptive statistics, there are some outliers associated with some attributes and that will be taken care by further "data preparation".

| Attributes | | Min | Max | Mean | Standard Deviation | Comparison |
|---|---|---|---|---|---|---|
| **Age** | Numeric | 19 | 87 | 41.17 | 10.58 | Mainly middle aged |
| **Job** | Qualitative | 38-unknown | 969-management | Nominal | Nominal | Largest % =management, technician and blue collar |
| **Marital** | Qualitative | 528-divorced | 2797-married | Nominal | Nominal | Most married |
| **Education** | Qualitative | 187-unknown | 2306-secondary | Nominal | Nominal | Most are secondary educated |
| **Default** | Qualitative | 76-yes | 4445-no | Nominal | Nominal | Most don't |

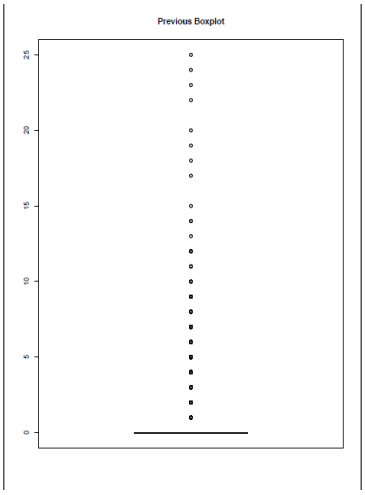
| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | have credit default |
| **Balance** | Nnumeric | -3313 | 71188 | 1422.66 | 3009.64 | Greater difference in balance |
| **Housing** | Qualitative | 1962-no | 2559-yes | Nominal | Nominal | Most have no housing loan |
| **Loan** | Qualitative/categorical | 691-yes | 3830-no | Nominal | Nominal | Most have no personal loan as well |
| **Contact** | Qualitative | 301-telephone | 2896-cellular | Nominal | Nominal | Mostly by cellular |
| **Day** | Numeric | 1st day | 31st day | 15.92th day | 8.25 day | Most calls were made in the middle of the month |
| **Month** | Qualitative | 20-Dec | 1398-May | Nominal | Nominal | May was the most contacted month |
| **Duration** | Numeric | 4 seconds | 3025 seconds | 263.96 seconds | 259.86 seconds | Difference in call periods |
| **Campaign** | Numeric | 1 | 50 | 2.8 | 3.11 | Difference in doing contact |
| **Pdays** | Numeric (-1 means client was not previously contacted) | -1 (no contact) | 871 contacts | 39.77 contacts | 100.12 | Most clients were contacted |
| **Previous** | Numeric | 0 | 25 | 0.543 | 1.70 | |
| **Poutcome** | Qualitative | 129-success | 3705-unknown | Nominal | Nominal | Most outcomes were unknown |
| **Y (Class)** | Class Attribute (Binary) | 521-yes | 4000-no | Nominal | Nominal | Most did not subscribe |

~ (7.7 No : 1 Yes)

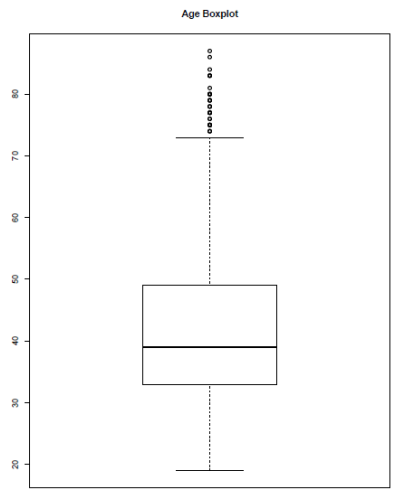
## Box Plot Analysis

Box plots show overall patterns of response for a group. They provide a useful way to visualise the range and other characteristics of responses for a large group. It also shows the outliers with the maximum and minimum values.17 Attributes / 7 numeric

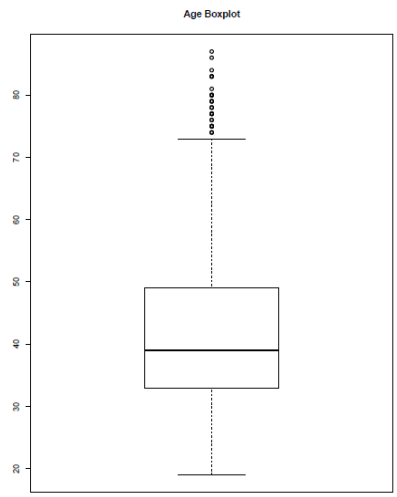
Previous Boxplot



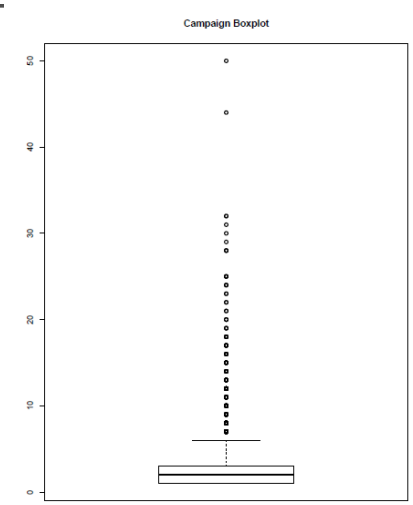
Age Boxplot



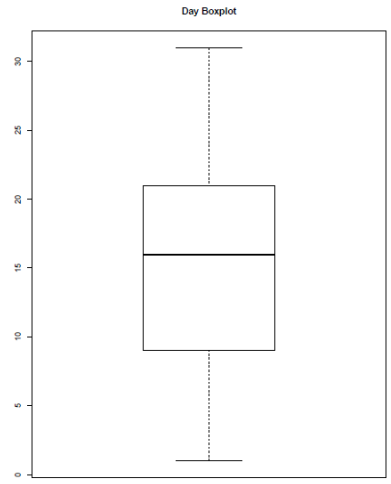

Balance Boxplot



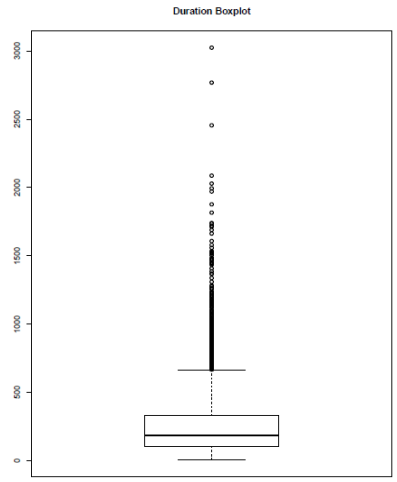

Campaign Boxplot



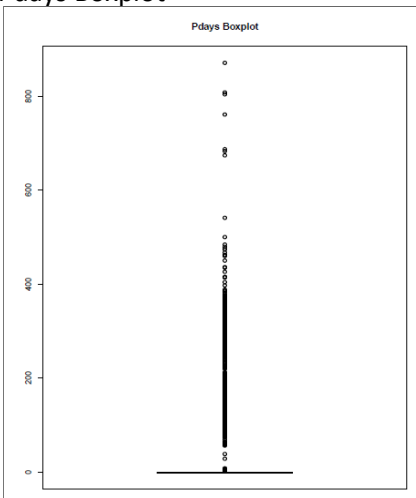

Day Boxplot




Duration Boxplot




Pdays Boxplot




## Determining Outliers in WEKA

For numeric values

Age – outliers detected

Balance – outliers detected

Day – no outliers detected

Duration – outliers detected

Campaign - outliers detected

Pdays – outliers detected

WEKA is used to determine the outliers. A filter for detecting outliers and extreme values based on interquartile ranges is used with the following steps. The filter skips the class attribute [6].

Filters >unsupervised>attributes >IQR

355 records were identified as outliers

Name: Outlier Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | no | 4166 | 4166.0 |
| 2 | yes | 355 | 355.0 |

Extreme values

953 records identified as extreme

**Selected attribute**

Name: ExtremeValue Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | no | 3568 | 3568.0 |
| 2 | yes | 953 | 953.0 |

Opening the saved file in Notepad++ highlights all the instances which have outliers. Having added two attributes outliers and extreme values, outliers are now the 18th data point in an instance. Each 18th data point with a 'YES' is an outlier.

Example:

```
26,housemaid,married,tertiary,no,543,no,no,cellular,30,jan,169,3,-1,0,unknown,no,no,no
41,management,married,tertiary,no,5883,no,no,cellular,20,nov,182,2,-1,0,unknown,no,yes,no
55,blue-collar,married,primary,no,627,yes,no,unknown,5,may,247,1,-1,0,unknown,no,no,no
67,retired,married,unknown,no,696,no,no,telephone,17,aug,119,1,105,2,failure,no,no,yes
```

To remove the outliers or extreme values:

Choose filter> unsupervised > instance > remove with values > in the attributes list, it is based on the 18th data set. Set attribute index to 18 > nominal indices: last to remove outlier.

To remove extreme values the same steps are performed as above.

Total values remaining after outlier removal:

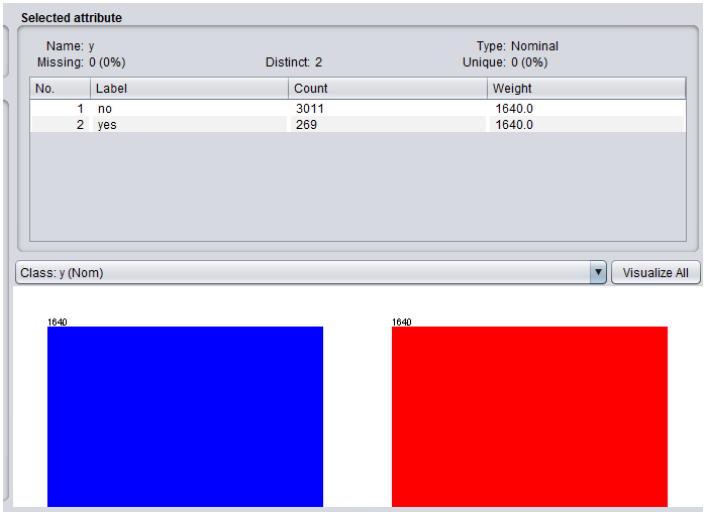
3011 No – Y Class

269 Yes- Y Class

**Total of 3280 instances**

## Balancing the Class

### Choose Filter > Supervised > Instance > Class Balancer

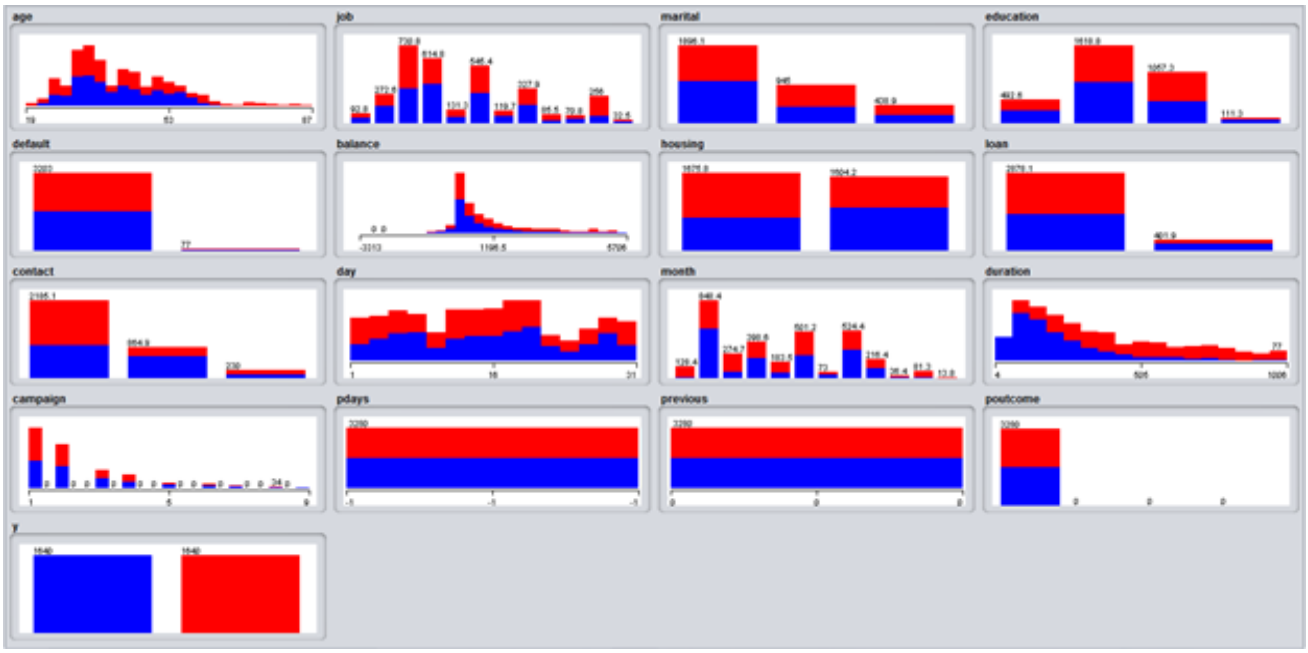



The data has an imbalanced class distribution. The classes are now balanced, i.e. weight is standardized, but distinct count is left the same.

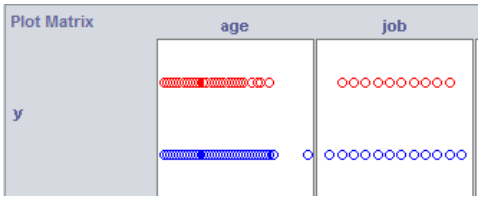
Weight of 1640 for each class result

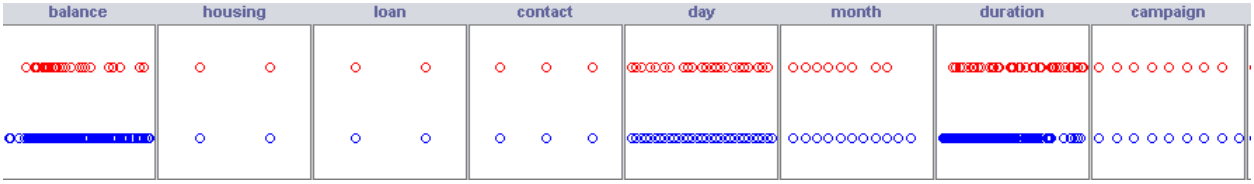
## Visualization of Data

Histograms: A histogram is used to display the distribution of data values along the real number line. It competes with the probability plot as a method of assessing normality [7]. The histogram here in this case shows that data is non-normal, except to some degree for age attribute.




According to the data: age, balance, duration, pdays and poutcome are most closely related to the class.




## Attributes to be Included

### Best First + CfsSubsetEval

Attribute Subset Evaluator (supervised, Class (nominal): 17 y):

CFS Subset Evaluator

Including locally predictive attributes

Selected attributes: 1,6,8,10,12 : 5

Attributes: Age, Balance, Loan, Day, Duration

**PT I:**

**Select Attributes tab >**
**Nom(y) = classWeka Attribute Evaluator: cfsSubseteval-P1-E1**
**Search Method: BestFirst – D1 – N5**

| Attributes | Cross-validation Fold:10 Seed:1 | Cross-validation Folds: 5 Seed:1 | Cross-validation Folds: 3 Seed: 1 | Use Full Training Set, (Nom) Y |
|---|---|---|---|---|
| Age | 9 (90%) | 5 (100%) | 3 (100%) | X |
| Job | 0 (0%) | 0 (0%) | 0 (0%) | |
| Marital | 0 (0%) | 0 (0%) | 0 (0%) | |
| Education | 0 (0%) | 0 (0%) | 0 (0%) | |
| Default | 0 (0%) | 0 (0%) | 0 (0%) | |
| Balance | 6 (60%) | 5 (100%) | 2 (67%) | X |
| Housing | 0 (0%) | 0 (0%) | 0 (0%) | |
| Loan | 8 (80%) | 3 (60%) | 3 (100%) | X |
| Contact | 2 (20%) | 2 (40%) | 2 (67%) | |

| | | | | |
|---|---|---|---|---|
| Day | 6 (60%) | 3 (60%) | 1 (33%) | X |
| Month | 2 (20%) | 0 (0%) | 0 (0%) | |
| Duration | 10 (100%) | 5 (100%) | 3 (100%) | X |
| Campaign | 0 (0%) | 0 (0%) | 0 (0%) | |
| Pdays | 0 (0%) | 0 (0%) | 0 (0%) | |
| Previous | 0 (0%) | 0 (0%) | 0 (0%) | |
| Poutcome | 0 (0%) | 0 (0%) | 0 (0%) | |

## Ranker + GainRatioAttributeEval

**PT II:**

**Select Attributes tab >**

**Nom(y) = class**

**Weka Attribute Evaluator: GainRatioAttributeEval**
**Search Method: Ranker – T – 1.79… - N – 1**

Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.

GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute)

| Attributes | Use Full Training Set - Ranking Attributes |
|---|---|
| Age | 0.03274 **(#4)** |
| Job | 0.00887 |
| Marital | 0.00635 |
| Education | 0.00475 |
| Default | 0.00128 |
| Balance | 0.0189 |
| Housing | 0.01864 |
| Loan | 0.02973 |
| Contact | 0.04042 **(#3)** |
| Day | 0.05437**(#2)** |
| Month | 0.02475 |
| Duration | 0.11813 **(#1)** |
| Campaign | 0.01027 |
| Pdays | |
| Previous | |
| Poutcome | |

**PT III:**

**Nom(y) = class**

**Weka Attribute Evaluator: InfoGainAttributeEval**
**Search Method: Ranker – T – 1.79… - N – 1**

InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute)

| Attributes | 10 fold, 1 seed- Avg Merit | Use Full Training Set - Ranking Attributes |
|---|---|---|
| Age | 0.042 +- 0.004 **(#4)** | 0.041698 **(#4)** |
| Job | 0.028 +- 0.003 | 0.02761 |
| Marital | 0.009 +- 0.002 | 0.008658 |
| Education | 0.008 +- 0.002 | 0.00763 |
| Default | 0 | 0.000205 |
| Balance | 0.026 +- 0.012 | 0.029794 |
| Housing | 0.019 +- 0.003 | 0.018631 |
| Loan | 0.016 +- 0.002 | 0.015951 |
| Contact | 0.047 +- 0.005 **(#3)** | 0.047144 **(#3)** |
| Day | 0.012 +- 0.01 | 0.005435 |
| Month | 0.076 +- 0.002 **(#2)** | 0.075347 **(#2)** |
| Duration | 0.296 +- 0.013 **(#1)** | 0.284187 **(#1)** |
| Campaign | 0.006 +- 0.002 | 0.004764 |
| Pdays | 0 | 0.03553361 |
| Previous | 0 | 0.01622639 |
| Poutcome | 0 | 0.03758116 |

Rationale:

Data type and analytical problem outcome: The data used for analysis is based on historical outcomes. The instances indicate a potential consumer who was contacted and the outcomes of the attributes and class for said customer. By analyzing the attributes and class outcome, correlations can be determined and a more effective marketing strategy can be made to target those who are more inclined to subscribing to a long-term deposit account. Considering there is prior knowledge on classes, and classes are pre-defined, the best course of action for analysis is Classification (supervised learning).

Determining attributes to use:

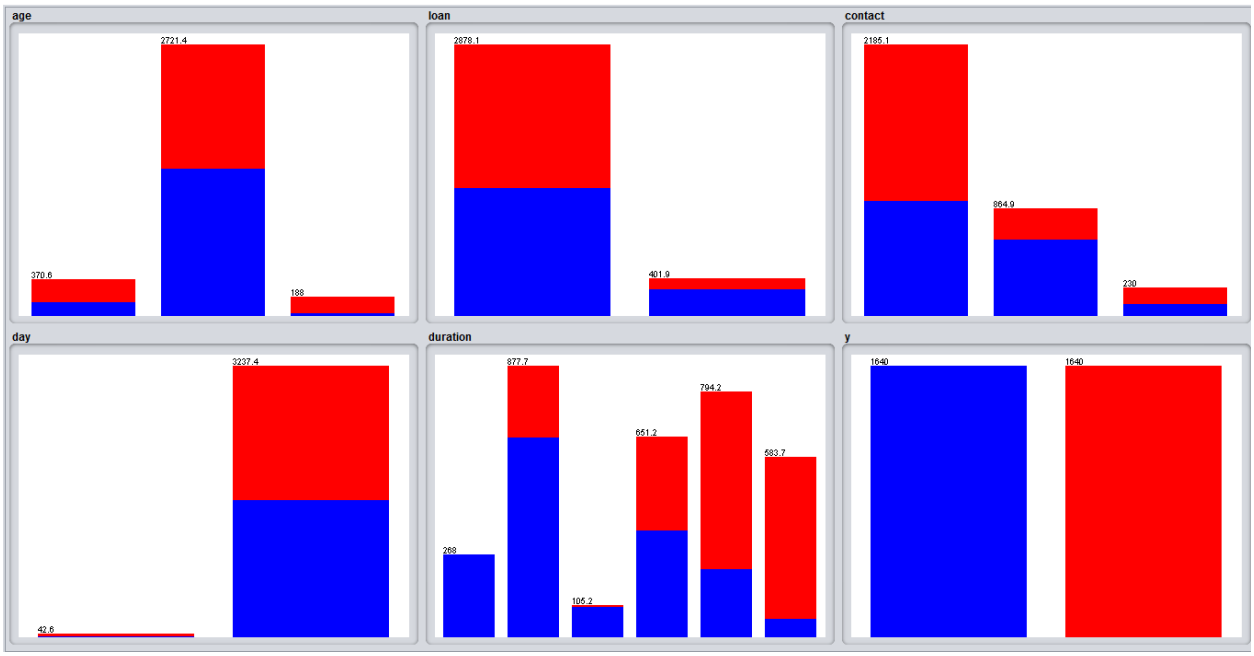
Discretization was applied to the numeric attributes in order to narrow the large differences between the maximum values and minimum values and essentially put the data into bins. When the data is discretized there is always some amount of discretization error.

Preprocessing > Filter > Supervised > AttributeSelection

Attributes used:

| No. | | Name |
|---|---|---|
| 1 | ☐ | age |
| 2 | ☐ | loan |
| 3 | ☐ | contact |
| 4 | ☐ | day |
| 5 | ☐ | duration |
| 6 | ☐ | y |

**Attributes vs. Class Correlation**




Blue = no, Red = yes

Included attributes:

Using different evaluators, cfsesubsetval, infogain and gainratio, all showed Age and Contact, Loan, Day, and Duration most highly correlated to the class in a supervised learning environment.

**|AGE, LOAN, CONTACT, DAY, DURATION, Y|**

Excluded attributes:

All the rest of the attributes were not considered in the predictive modeling.

# Predictive Modeling (Classification)

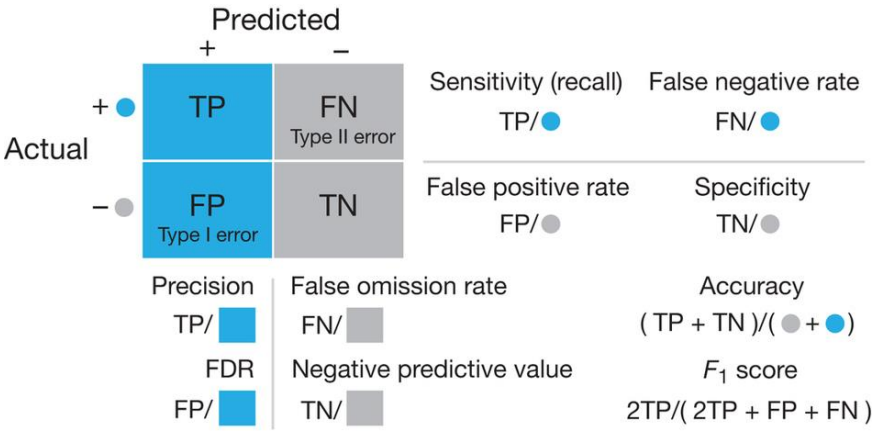



*Figure 1:http://www.nature.com/nmeth/journal/v13/n8/full/nmeth.3945.html*

## Decision Tree - Supervised Learning

### J-48

The WEKA tool provides a number of options associated with tree pruning. In case of potential overfitting pruning can be used as a tool for précising. In other algorithms, the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy [Ref-8].

Confidence factor: 0.25

| | Cross-validation Fold:10 Seed:1 | Cross-validation Folds: 5 Seed:1 | Use Training Set | Percentage Split 60%-40% |
|---|---|---|---|---|
| Correctly Classified Instances | 2574.3785 | 2578.1912 | 2641.6705 | 1067.5223 |
| Incorrectly Classified Instances | 705.6228 | 701.8101 | 638.3308 | 241.147 |
| Precision /Recall - N | 0.818/0.733 | 0.818/ 0.736 | 0.848/0.745 | 0.895/ 0.716 |
| TP/FP Rate - N | 0.733/0.164 | 0.736/0.164 | 0.745/0.134 | 0.716/0.084 |
| Precision /Recall - Y | 0.758/0.836 | 0.760/0.836 | 0.772/0.866 | 0.762/0.916 |
| TP/FP Rate - Y | 0.836/0.267 | 0.836/0.264 | 0.866/ 0.255 | 0.916/0.284 |

a 1202.63, 437.37| 1206.44 , 433.56| 470.05, 186.28 |1221.15, 418.85

b 268.25, 1371.75| 268.25, 1371.75|54.87, 597.47| 219.48, 1420.5|
**a = no**
**b = yes**

## Naive Bayes

The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naive yet the algorithm tends to perform well and learn rapidly in various supervised classification problems [9].

| | Cross-validation Fold:10 Seed:1 | Cross-validation Folds: 5 Seed:1 | Use Training Set | Percentage Split 60%-40% |
|---|---|---|---|---|
| Correctly Classified Instances | 2603.4566 | 2584.0955 | 2629.4955 | 1114.2404 |
| Incorrectly Classified Instances | 676.5447 | 695.9058 | 650.5058 | 194.4289 |
| Precision /Recall - N | 0.826/0.744 | 0.804/ 0.762 | 0.827/0.745 | 0.942/0.750 |
| TP/FP Rate - N | 0.744/0.156 | 0.762/0.186 | 0.763/0.160 | 0.750/0.047 |
| Precision /Recall - Y | 0.773/0.844 | 0.760/0.814 | 0.780/0.840 | 0.791/0.953 |
| TP/FP Rate - Y | 0.844/0.256 | 0.814/0.238 | 0.840/ 0.237 | 0.953/0.250 |
| | | a b | a b | a b |
| | | 1219.52 420.49 \|1248.93, 391.07 \|1251.65, 388.35\|492.38, 163.95\| | | |
| | | 256.06 1383.94 \| 304.83, 1335.17 \| 262.16, 1377.84\|30.48, 621.86 \| | | |

## Random Forest

Random forest is an ensemble method for classification, regression and other task, that operated by constructing a multitude of decision trees at training time and outputting the class that is the mode of classes (classification) or mean prediction (regression) of the individual tree. Random decision forest corrects for decision tree's habit of overfitting to their training set [10].

| | Cross-validation Fold:10 Seed:1 | Cross-validation Folds: 5 Seed:1 | Use Training Set | Percentage Split 60%-40% |
|---|---|---|---|---|
| Correctly Classified Instances | 2556.7568 | 2552.3994 | 2653.3374 | 1073.5137 |
| Incorrectly Classified Instances | 723.2445 | 727.6019 | 626.6639 | 235.1557 |
| Precision /Recall - N | 0.798/0.749 | 0.797/ 0.746 | 0.840/0.763 | 0.897/0.725 |
| TP/FP Rate - N | 0.749/0.190 | 0.746/0.190 | 0.763/0.145 | 0.725/0.084 |
| Precision /Recall - Y | 0.763/0.810 | 0.761/0.810 | 0.783/0.855 | 0.768/0.916 |
| TP/FP Rate - Y | 0.810/0.251 | 0.810/ 0.254 | 0.855/ 0.237 | 0.916/0.275 |

A 1227.69, 412.32 | 1223.33, 416.67 |1251.11, 388.89 |476.04, 180.29|
b 310.93, 1329.07 |310.93, 1329.07 |237.77, 1402.23 |54.87, 597.47|

For training, cross-validation and testing data is resampled by using the Resampling filter. Then undo to get all instances back.

Resample - noreplacement option = true, invert selection option = true. this gives the Training set. Then resample again - invert selection = false, sample size percent to 50%, to get 50% of the training set for cross-validation.
Test set instances: resample, invert selection to true, 50% of the cross-validation instances which are not included in the cv set. save each of these arff files.

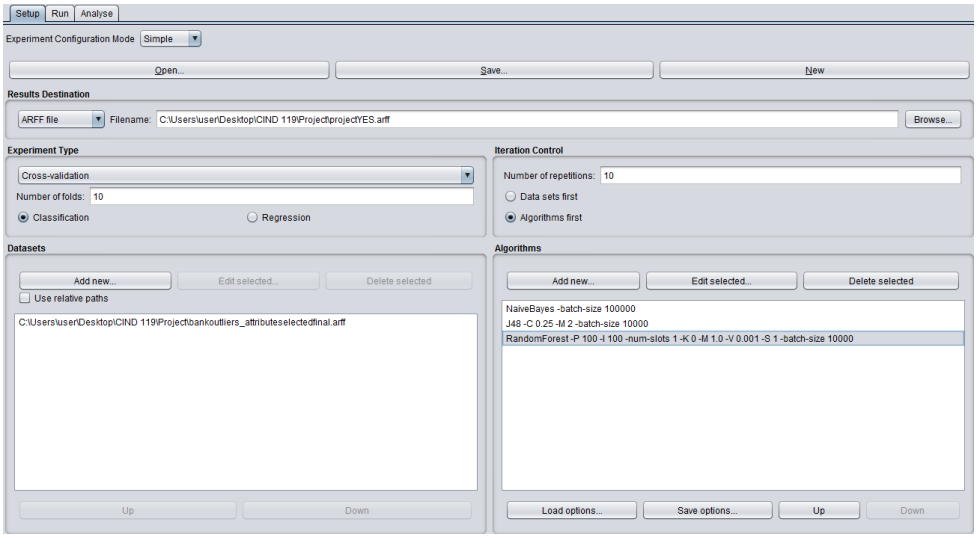
**Results**

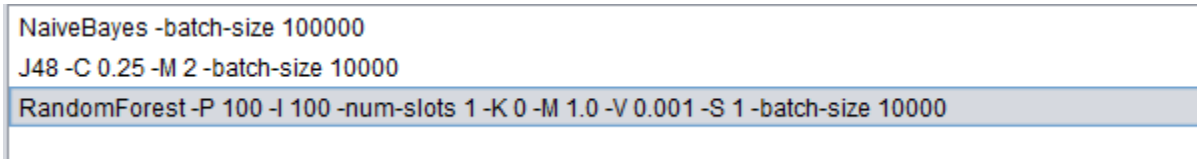
The data was split into a training and testing set. It was then run through different parameter changes in a J48 Decision Tree classifier, Naive Bayes and Random Forest. For each classifier, a 10, 5 - fold cross validation, 60-40% splitting and a full run was done on the data.

## EXPERIMENTER

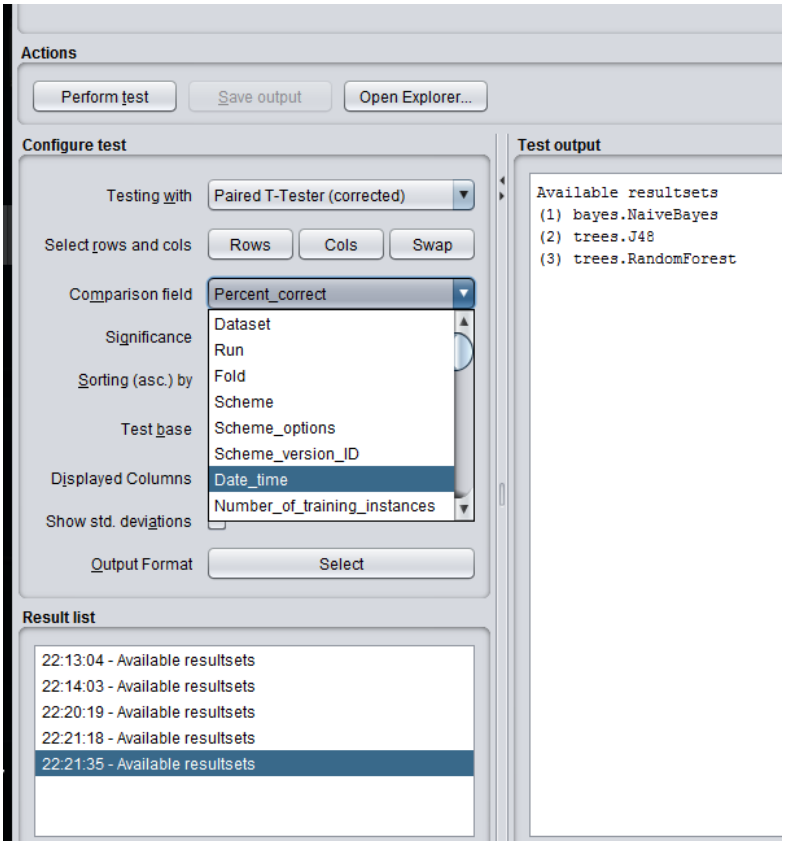
In WEKA to test multiple data sets against each other

Naives Bayes vs. J48 vs. Random forest




After running the three tests for 10 iterations, the analysis tab is used.



**Compare output:**

**Analysing: True_positive_rate**

Datasets: 1

Resultsets: 3

Confidence: 0.05 (two tailed)

Dataset (1) bayes.N | (2) tree (3) tree

---

'bank-weka.filters.unsupe(100) 0.75 | 0.74 | 0.76

---

(v/ /*) | (0/1/0) (0/1/0)

**Analysing: F_measure**

Datasets: 1

Resultsets: 3

Confidence: 0.05 (two tailed)

Dataset (1) bayes.N | (2) tree (3) tree

---

'bank-weka.filters.unsupe(100) 0.79 | 0.77 0.78

---

(v/ /*) | (0/1/0) (0/1/0)

No significant difference between tests.
v = victory, * = loss, blank = cannot tell if statistically significant or not

**Analysing: SF_mean_entropy_gain**

Datasets: 1

Resultsets: 3

Confidence: 0.05 (two tailed)

Dataset (1) bayes.N | (2) trees (3) trees.

---

'bank-weka.filters.unsupe(100) 0.36 | -9.90 * -10.02 *

---

(v/ /*) | (0/0/1) (0/0/1)

**Analysing: Percent_correct**

Datasets: 1

Resultsets: 3

Confidence: 0.05 (two tailed)

Dataset (1) trees.J4 | (2) bayes (3) trees

---

'bank-weka.filters.unsupe(100) 78.20 | 79.34 78.21

---

(v/ /*) | (0/1/0) (0/1/0)

Discussion:

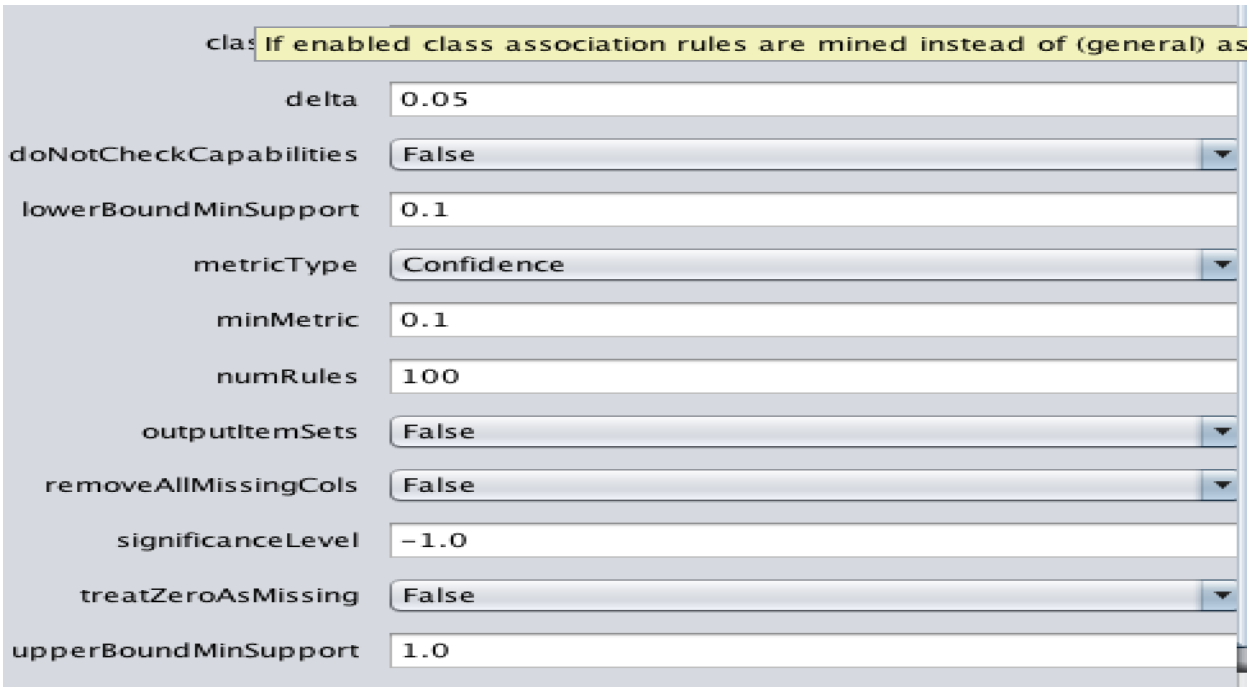
Naive Bayes is the best algorithm for the dataset, when comparing true/false positive, confusion matrices, accuracy, recall, and precision. As well when comparing the 3 classification algorithms in Experimenter, Bayes consistently came out on top as the best algorithm (albeit not by a truly large measure in percent correct compared to the other two algorithms but came out much better in the mean entropy gain test. Naive Bayes assumes the attributes are independent The Decision Tree algorithm tended to overfit.

# Post-Predictive Analysis

Association rules were used due to categorical variables. The dataset was filtered in favor of business's interest: "yes" for subscription. Because it was an association rule, numeric attributes in the dataset has been discretized to categorical variable from pre-defined filter. Association rule was performed within Weka based on three specific cases. First case was designed, what are the outcomes if the lowerBoundMinSupport value and confidence level was set low? Second case was designed, what are the outcomes if the lowerBoundMinSupport value and confidence level was set high? Third case was designed, what are the outcomes if lowerBoundMinSupport and confidence level was set like we did in the class? and here are the following results:

⇒ default upperbound and delta is 1 and 0.05 respectively.

1) LowerBoundMinSupp: 0.1
   Confidence: 0.1

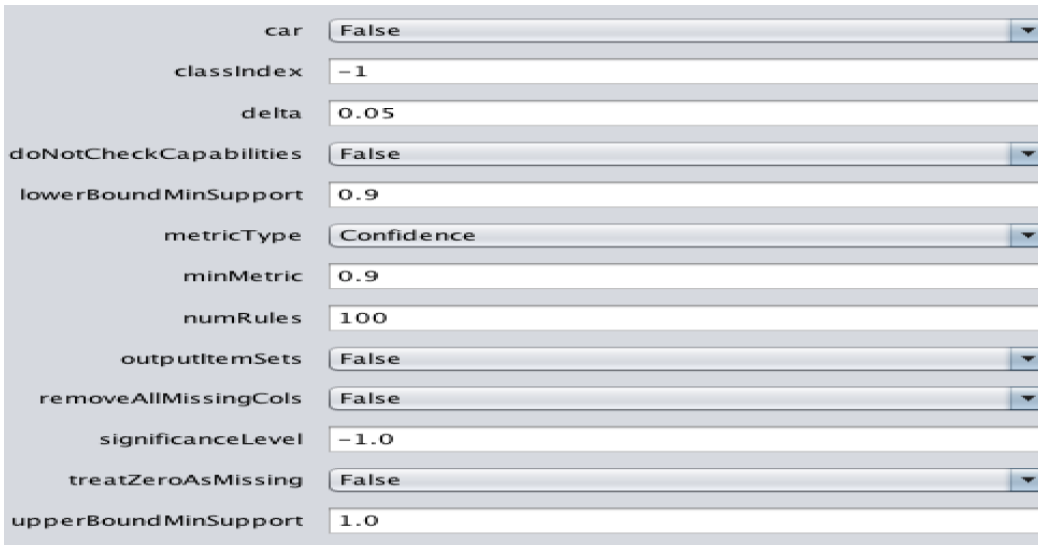



Best rules found:

1. default=no 512 ==> y=yes 512 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2. loan=no 478 ==> y=yes 478 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. default=no loan=no 471 ==> y=yes 471 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4. contact=cellular 416 ==> y=yes 416 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5. default=no contact=cellular 409 ==> y=yes 409 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
6. loan=no contact=cellular 383 ==> y=yes 383 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
7. default=no loan=no contact=cellular 378 ==> y=yes 378 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
8. previous=0 337 ==> pdays=-1 337 <conf:(1)> lift:(1.55) lev:(0.23) [119] conv:(119.02)
9. pdays=-1 337 ==> previous=0 337 <conf:(1)> lift:(1.55) lev:(0.23) [119] conv:(119.02)
10. poutcome=unknown 337 ==> pdays=-1 337 <conf:(1)> lift:(1.55) lev:(0.23) [119] conv:(119.02)

2) LowerBoundMinSupp: 0.9

Confidence: 0.9



```
Best rules found:

 1. default=no 512 ==> y=yes 512    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. loan=no 478 ==> y=yes 478    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. default=no loan=no 471 ==> y=yes 471    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. loan=no 478 ==> default=no 471    <conf:(0.99)> lift:(1) lev:(0) [1] conv:(1.03)
 5. loan=no y=yes 478 ==> default=no 471    <conf:(0.99)> lift:(1) lev:(0) [1] conv:(1.03)
 6. loan=no 478 ==> default=no y=yes 471    <conf:(0.99)> lift:(1) lev:(0) [1] conv:(1.03)
 7. y=yes 521 ==> default=no 512    <conf:(0.98)> lift:(1) lev:(0) [0] conv:(0.9)
 8. default=no 512 ==> loan=no 471    <conf:(0.92)> lift:(1) lev:(0) [1] conv:(1.01)
 9. default=no y=yes 512 ==> loan=no 471    <conf:(0.92)> lift:(1) lev:(0) [1] conv:(1.01)
10. default=no 512 ==> loan=no y=yes 471    <conf:(0.92)> lift:(1) lev:(0) [1] conv:(1.01)
11. y=yes 521 ==> loan=no 478    <conf:(0.92)> lift:(1) lev:(0) [0] conv:(0.98)
12. y=yes 521 ==> default=no loan=no 471    <conf:(0.9)> lift:(1) lev:(0) [0] conv:(0.98)
```

3) Class-defined:

LowerBoundMinSupp: 0.5

Confidence: 0.8

```
Best rules found:

 1. default=no 512 ==> y=yes 512    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. loan=no 478 ==> y=yes 478    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. default=no loan=no 471 ==> y=yes 471    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. contact=cellular 416 ==> y=yes 416    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. default=no contact=cellular 409 ==> y=yes 409    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. loan=no contact=cellular 383 ==> y=yes 383    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. default=no loan=no contact=cellular 378 ==> y=yes 378    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. previous=0 337 ==> pdays=-1 337    <conf:(1)> lift:(1.55) lev:(0.23) [119] conv:(119.02)
 9. pdays=-1 337 ==> previous=0 337    <conf:(1)> lift:(1.55) lev:(0.23) [119] conv:(119.02)
10. poutcome=unknown 337 ==> pdays=-1 337    <conf:(1)> lift:(1.55) lev:(0.23) [119] conv:(119.02)
```

There were total of 416 rules; by looking at the top 10 best rule figured by Apriori algorithm with respect to different confidence level and LowerBoundMinSupport; cases combined reveals the following suggestions:

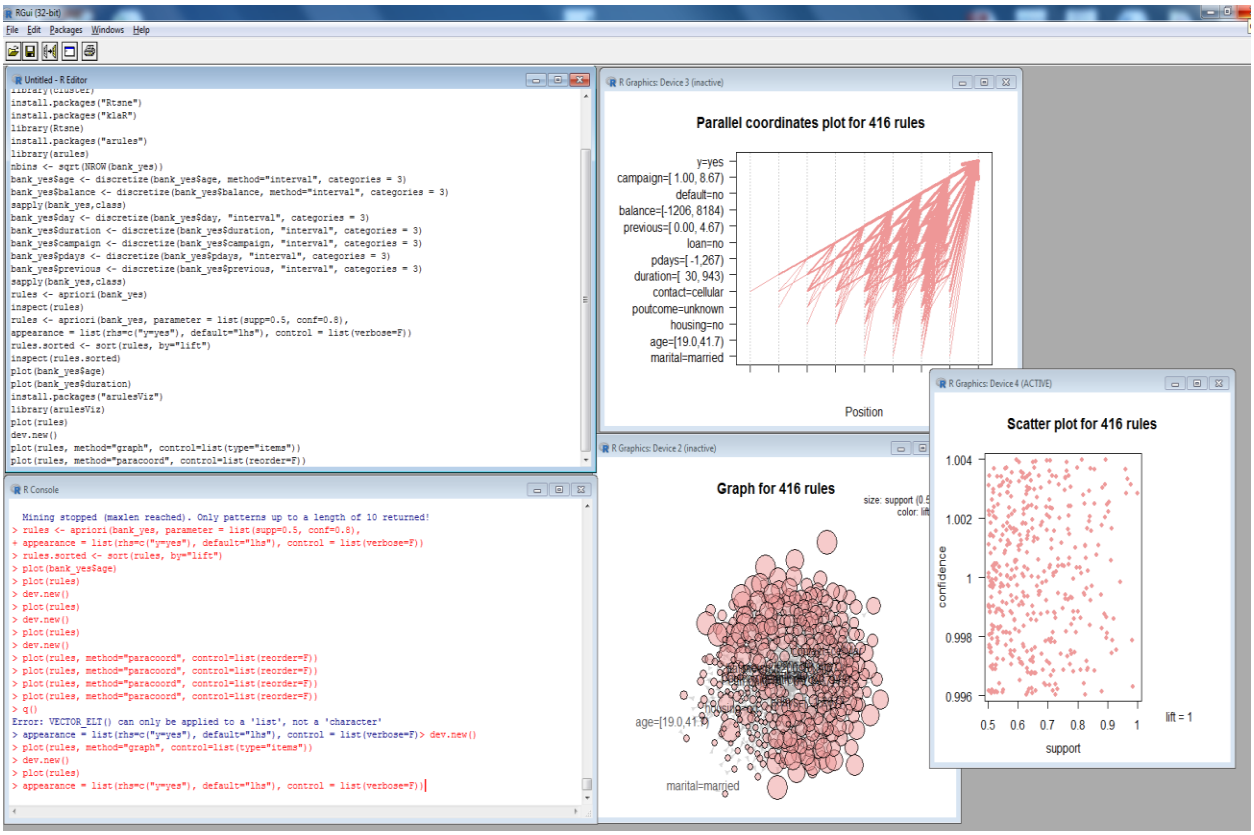
default = no 512 ⇒ y=yes 512

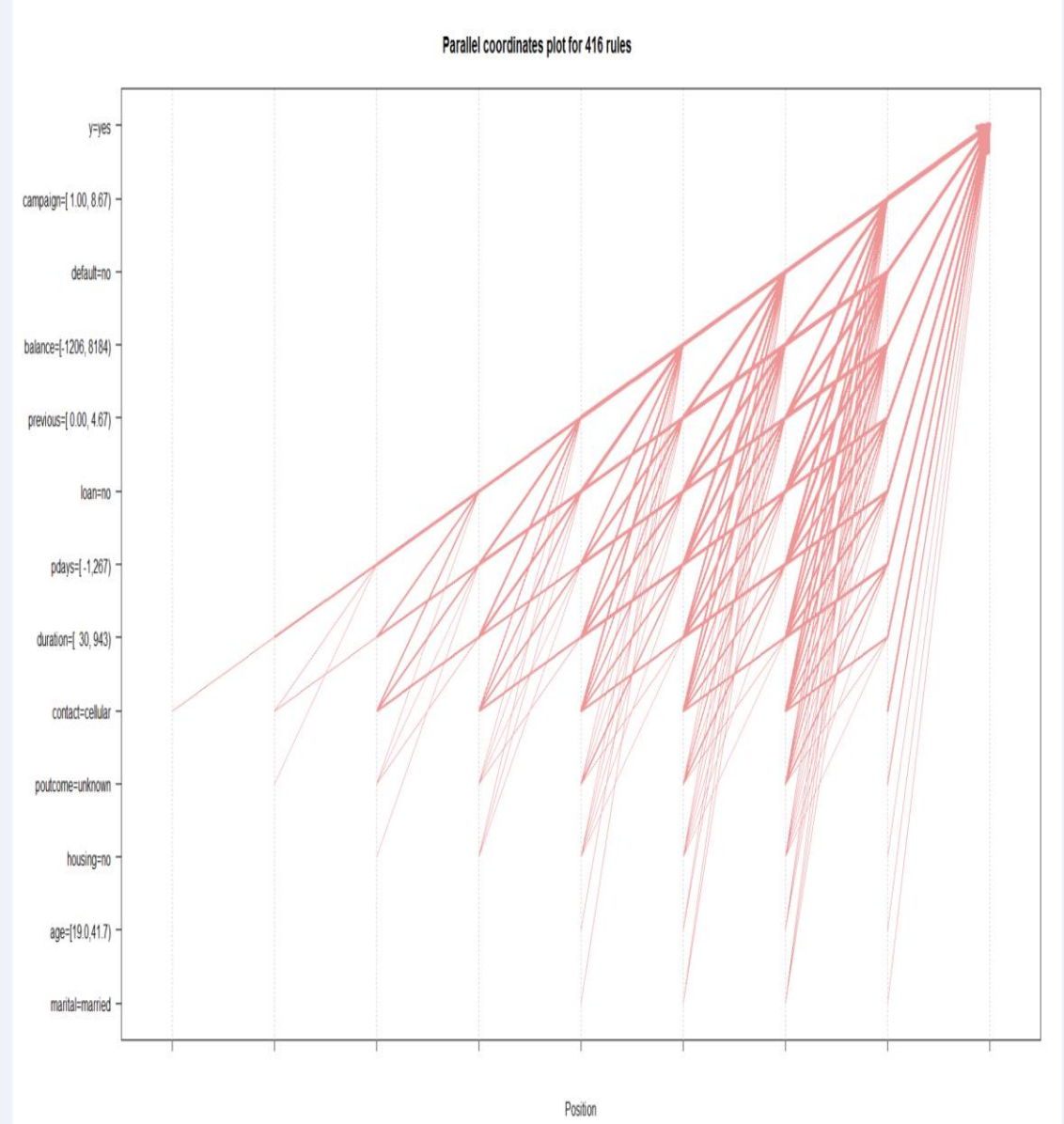
loan = no 478 ⇒ y=yes 478

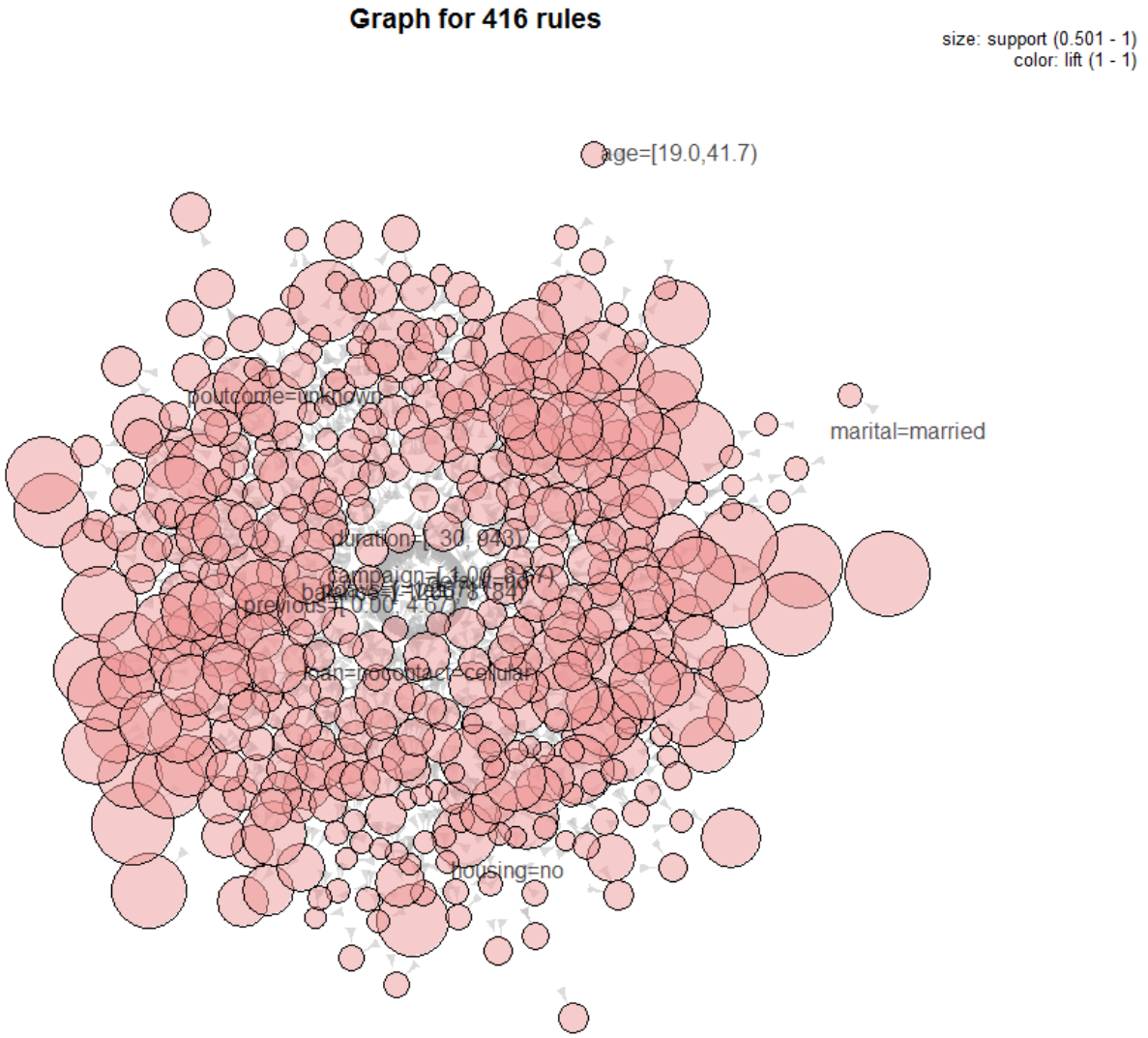
default = no, loan = no 471 ⇒ y = yes 471

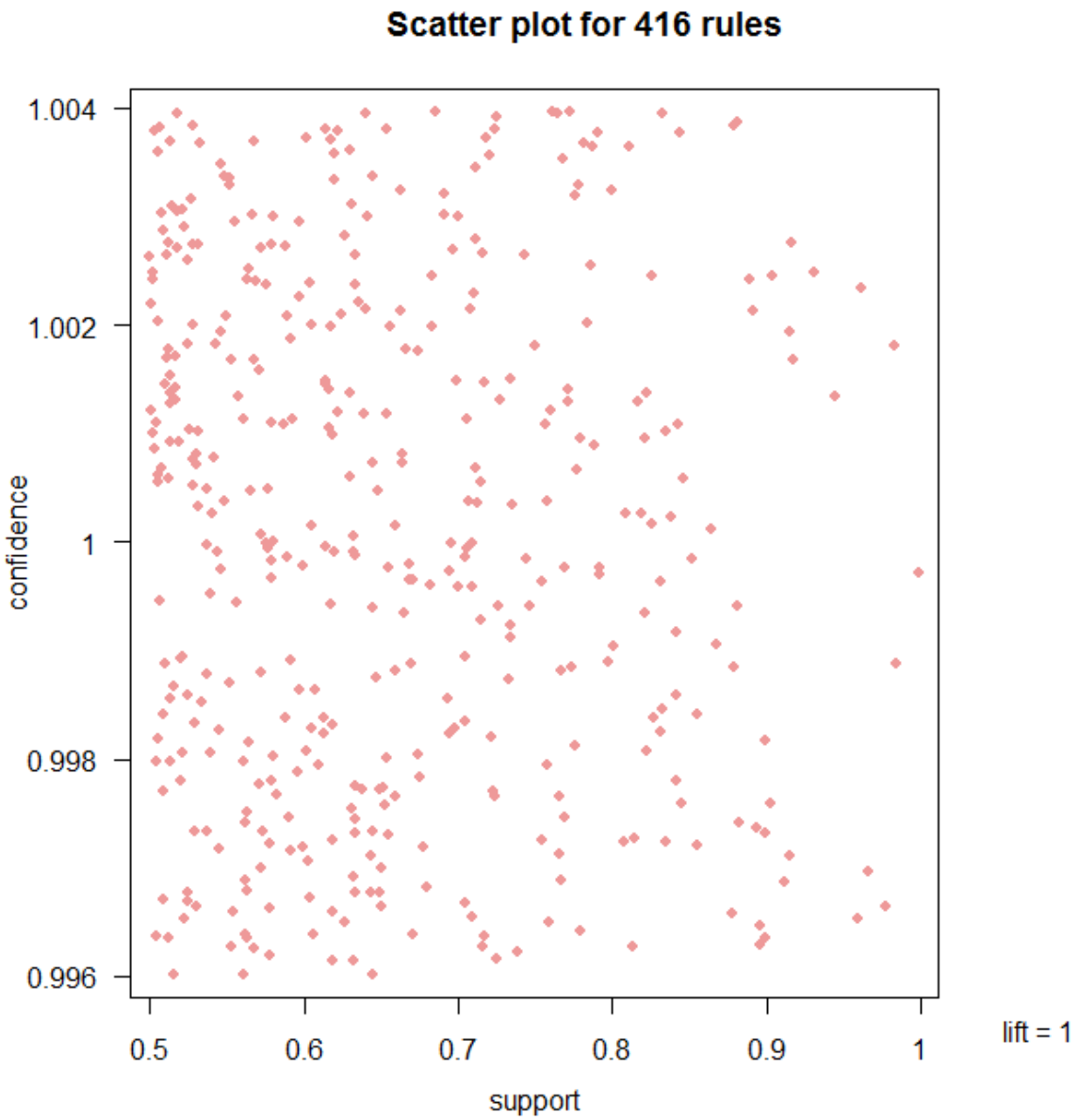
These are top 3 that is common to all the cases.

# Visualizing & Comparing Association Rules with R




Parallel coordinates plot for 416 rules
y=yes
campaign=[1.00, 8.67)
default=no
balance=[-1206, 8184)
previous=[0.00, 4.67)
loan=no
pdays=[-1,267)
duration=[ 30, 943)
contact=cellular
poutcome=unknown
housing=no
age=[19.0,41.7)
marital=married
Position


Graph for 416 rules
size: support (0.501 - 1)
color: lift (1 - 1)
age=[19.0,41.7)
poutcome=unknown
marital=married
duration=[ 30, 643)
housing=no




As readers may notice from diagrams above, parallel coordinates, “bubble” graph, and scatterplot, customers with characteristics such as default, loan, and (contact = cellular) are considered important customer. As a note, “contact = cellular” relation was discovered by weka within top 4 when apriori algorithm’s  was set like in-class value of lowerBoundMinSupp and Confidence level; figures above proves this case.

# Conclusion

The banking industry is one of the most competitive industries in every nation. Within the banking industry, optimizing targeting for telemarketing is a key issue, under a growing pressure to increase profits and reduce costs. Under this context, the use of a data based model to predict the result of a telemarketing phone call to sell long term deposits is a valuable tool to support client selection decisions of bank campaign managers [11].

In this study, we proposed to use client's data with an effective data mining approach for the selection of telemarketing clients. We analyzed more than 4500 client's historical data and initially considering 16 attributes. The data had some outliers and were treated with proper tools including box plot and WEKA tool. Discretization was applied to the numeric attributes in order to narrow the large differences between the maximum values and minimum values.

Realizing the fact that all attributes may or may not be effective attribute to run model, we used various tools to screen out attributes. For this purpose, we used CFS Subset Evaluator, gainratioattributeeval, and infogainattibuteeval; using these tools we came with the selection of only 5 attributes including, age, loan, contact, day, and duration. A particular emphasis was given on considering both social and economic indicators.

During the modelling phase, we used a classification method to ensure high degree of accuracy. We used WEKA for the tools, including Decision Tree - Supervised Learning J-48, Naïve Bayes, and random forest classification model to see how the selected attributes could provide the best answer with a high degree of accuracy. Naïve Bayes consistently came out on top as the best algorithm, not by a truly large measure in percent correct compared to the other two algorithms.

Recommendations (Strategies)

All socioeconomic variables may have an impact on buying decision of certain banking product. The buying behavior of an individual to other environmental factors combined or solely plays role in investment decision in general. Keeping these views into consideration, we came with recommendations related to individual variable and then their combine use.

- Age: the data set shows the average age of most clients called in the past are about 41 years. The investment decision is influenced by age [12] so, the same kind of script may not be more appealing to all age groups and prospects may not subscribe the product.

- Contact: Most people were called on their cellular phones in the past to maintain relationship. It is better to see the behavioral pattern of clients with their success rate to speak with a bank representative frankly and in no rush. The home phone could be another option. Further, calling them in different time of the day can also influence the impact of the message; options could be day, evening, weekend etc. [13].

- Loan/No Default: The investment decision is determined by saving and possibility of borrowing for investment [14]. Most respondents in the data set had no personal loan; perhaps a good strategy could be providing them financing strategy together while selling long term investment account.

- Day: Most people were called on and between 15th and 16th day of the month. Marketing is an

ongoing activity, having said that, people may like to think about investing their money more enthusiastically during pay day or time or week, so, screening clients with their payday (weekly, biweekly and monthly) and then calling accordingly could work effectively [15].

· Duration-Most clients were called in an average of 4 and half minute, the question is whether people were getting enough time to interact and get their answers about the investment alternatives or not? Educating is one of the key factors in long term investment. Natural and normal script may not help often, so, increasing time of call may provide sufficient information to prospects and needed interactions [16].

The recommendations made above are based on the assumption that other 11 variables have no or little impact in the decision making. Things and environments change over a period of time, so, continuous data analysis and model validation could improve bank's marketing strategies and increase sales.

# References

[0] Lan McGugan, Aggressive selling by banks? The Globe and Mail, 2017.

[1] Philip Kotler, Kevin Lane Keller, Framework for Marketing Management, 5th edition Pearson, 2012.

[2] Hyperlink: https://en.oxforddictionaries.com/definition/stakeholder.

[3] NC Cary, JMP Statistics and Graphics Guide, SAS Institute Inc., 2003.

[4] Foster Provost, Machine Learning from Imbalanced Data Sets 101, New York University, 2000.

[5] Marie Špiláčková , Historical Research in Social Work – Theory and Practice, ERIS Web Journal, 2012.

[6] Svetlana S. Aksenova, Machine Learning with WEKA, School of Engineering and Computer Science Department of Computer Science California State University, Sacramento California, 95819, 2004.

[7] NCSS Statistical Software, Histograms, 2006

[8] Gaganjot Kaur and Amit Chhabra, Improved J48 Classification Algorithm for the Prediction of Diabetes, International Journal of Computer Applications (0975 – 8887) Volume 98-2014.

[9] Tina R. Patil, Mrs. S. S. Sherekar , Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, International Journal Of Computer Science And Applications Vol. 6, No.2, 2013.

[10] Trevor Hastie, Robert Tibshirani, The Elements of Statistical Learning, 2nd Ed, Springer, ISBN 0-387-95284-5.

[11] Moro Sérgio, Cortez Paulo, Rita Paulo, A data-driven approach to predict the success of bank telemarketing, ISCTE-IUL, Business Research Unit (BRU-IUL), Lisboa, Portugal, 2014.

[12] Zipporah Nyaboke Onsomu, Effect of Age on Investor Decisions, International Journal of Innovative Research and Development, 2015.

[13] Chris Fill and Barbara Jamieson, Marketing Communications, Heriot-Watt University, 2006.

[14] Okumu Argan Wekesa, Mwalili Samuel , and Mwita Peter, Modelling Credit Risk for Personal Loans Using Product-Limit Estimator, International Journal of Financial Research, Vol. 3, No. 1; January 2012.

[15] Deaune E. Sharp, Organizing and Managing the Call Center, Digital Press, 2003.

[16] Software Co-marketing: Developing Skills Series, Telemarketing, IBM Corporation 2003.