

Inngangur að gagnanámi – Project 2

Þorsteinn Sigurðsson



Content

Objectives.....	2
Data set description.....	2
Attribute information.....	2
Preprocessing.....	3
Deleting attributes.....	3
Discretizing.....	4
Aggregation.....	6
Results.....	7
Rule mining process.....	8
Resulting rules.....	9
Recomendations.....	11

Objectives

Introduction:

The data set i choose for this project is the Bank marketing Data Set. It is related with direct marketing campaigns of a Portuguese banking institution and the marketing was based on phone calls, often more than one contact to the same client was required in order to access if the product which is a bank term deposit would be(Yes) or not(No) subscribed.

The goal is to predict if the client will subscribe a term deposit and we need to find the best association rules to reccomend the best strategies for the next campaign.

Data set description

Attribute description:

The data includes 17 attributes

1. age, numeric.
2. job, nominal (admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown).
3. marital, nominal (divorced, married, single).
4. education, nominal (primary, secondary, tertiary, unknown).
5. default(if the client has credit in default), categorical (no,yes).
6. balance, numeric.
7. housing (if the client has a housing loan), nominal (yes,no).
8. Loan (if the client has a loan), nominal (yes,no).
9. contact (contact communication type) nominal (cellular, telephone, unknown).
10. day (last contact day of month), numeric (1-31).
11. month (last contact month of year), nominal(jan,feb,mar, apr, may, jun, jul, aug, sep, okt, nov, dec)
12. duration (last contact duration in seconds), numeric.
13. campaign (number of contacts performed before this campaign and for this client), numeric.
14. pdays (number of days that passed by after the client was last contacted), numeric.
15. previous, (number of contacts performed before this campaign and for this client), numeric, note that -1 means that the client was not contacted before.
16. poutcome (outcome of the previous marketing campaign), nominal (failure, other,success, unknown)

Output variable.

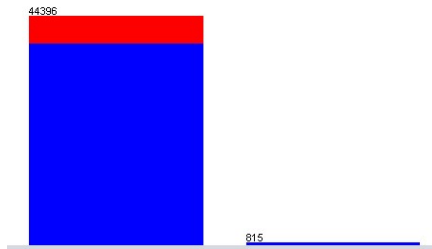
17. y (has the client subscribed a term deposit), binary (yes,no)

Preprocessing.

Deleting attributes.

First we need to find attribute candidates for deletions, the attributes i choose for deletion either had almost zero to no support, redundant or attributes that had no relative information for our goal.

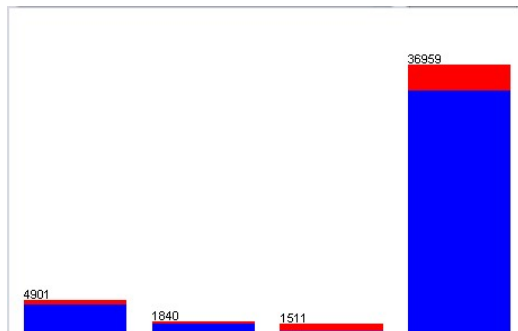
The first attribute is default.



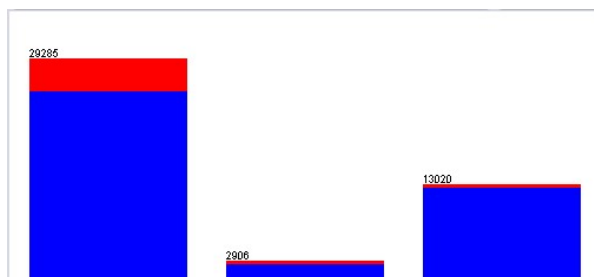
We have almost no support if the client has credit in default(yes).

The attributes day and month hold no valuable information for our goal because there is no correlation if the client was contacted in some day of the month or month of the year. For example we don't want a association rule that says the client will subscribe a term deposit in some particular day or month.

The poutcome holds no value for our goal because we are not looking if the prevoius campaign was a success or not. Also it has the unknown value for over 80% of the instances.



The contact variable holds no value for our goal because the marketing campaign was base on phonecalls so it does not matter if the contact was from cellular phone or a regular phone, the telephone attribute also has little to to no support.

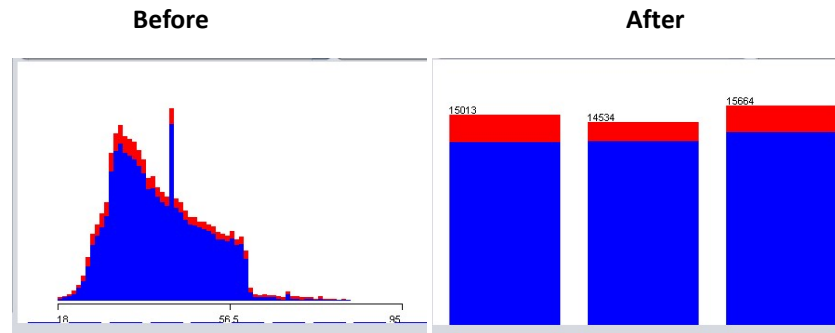


Discretizing

We need to discretize all the attribute with numeric values and with the right amount of bins for our graph.

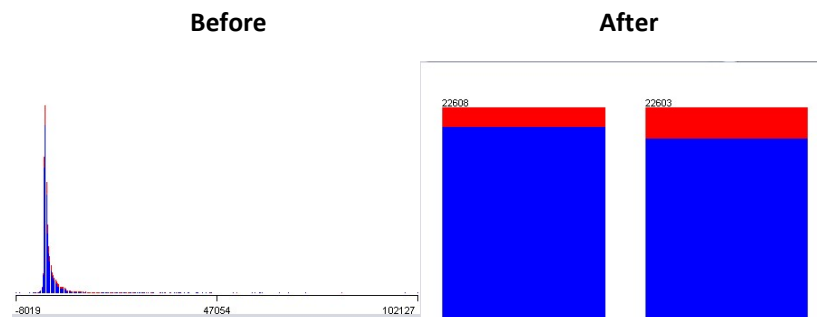
Age

The first attribute is age. I chose 3 bins with equal frequency for ages between 18 - 34.5, 34.5-44.5 and 44.5-95.



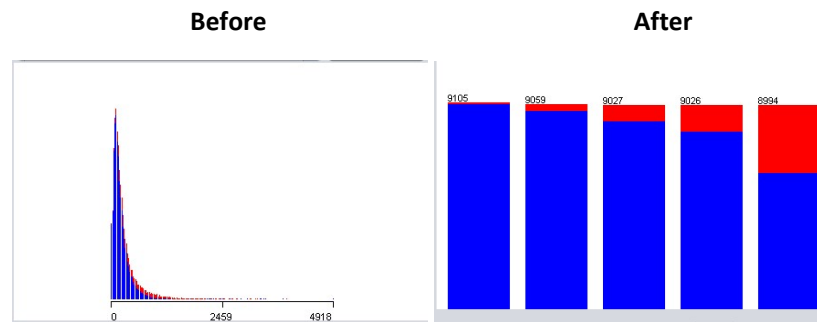
Balance

The attribute balance is numeric so we need to discretise it, we can see from the graph that it is not clear correlation between clients so i chose 2 bins with equal frequency. The outcome then becomes more clear with two labels (-8019-448.5) and (448.5-102127)



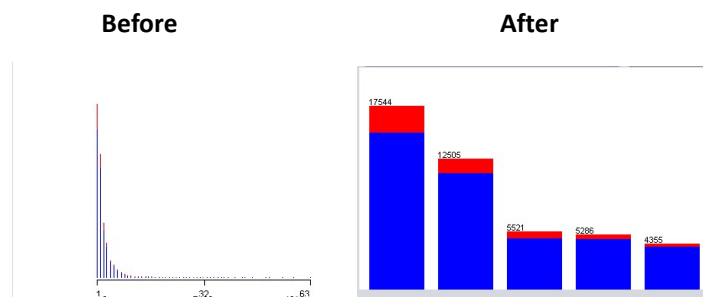
Duration

I chose the amount of bins that showed the increase of the „Yes“ instance in the output class.



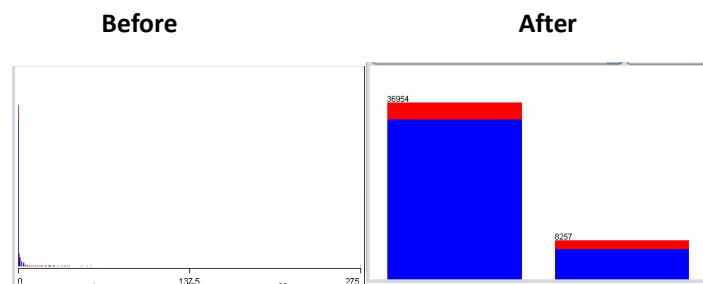
Campaign

I chose five bins here because to see if there is a correlation if the client was contacted 0-1 times and to see if the „Yes“ instance would decrease or increase by the amount of contacts that was performed for the client.



Previous

Here i chose two bins for if the client was not contacted in the previous campaign or if he was contacted.



After discretizing, i saw that the attributes pday and previous were redundant so i removed the attribute pdays.

Aggregation:

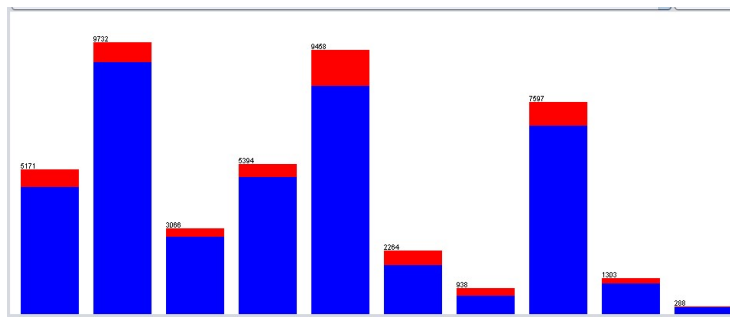
By looking at the values in the job attribute i saw some values that could be aggregated.

Name: V2		Distinct: 12		Type: Nominal	
Missing: 0 (0%)				Unique: 0 (0%)	
No.	Label	Count		Weight	
1	admin.	5171		5171.0	
2	blue-collar	9732		9732.0	
3	entrepreneur	1487		1487.0	
4	housemaid	1240		1240.0	
5	management	9458		9458.0	
6	retired	2264		2264.0	
7	self-employed	1579		1579.0	
8	services	4154		4154.0	
9	student	938		938.0	
10	technician	7597		7597.0	
11	unemployed	1303		1303.0	
12	unknown	288		288.0	

Entrepreneur and self-employed are more then less the same thing and also the variables housemaid and services because houskeeping falls into the same category as a service.

No.	Label	Count	Weight
1	admin.	5171	5171.0
2	blue-collar	9732	9732.0
3	entrepreneur_self-employed	3066	3066.0
4	housemaid_services	5394	5394.0
5	management	9458	9458.0
6	retired	2264	2264.0
7	student	938	938.0
8	technician	7597	7597.0
9	unemployed	1303	1303.0
10	unknown	288	288.0

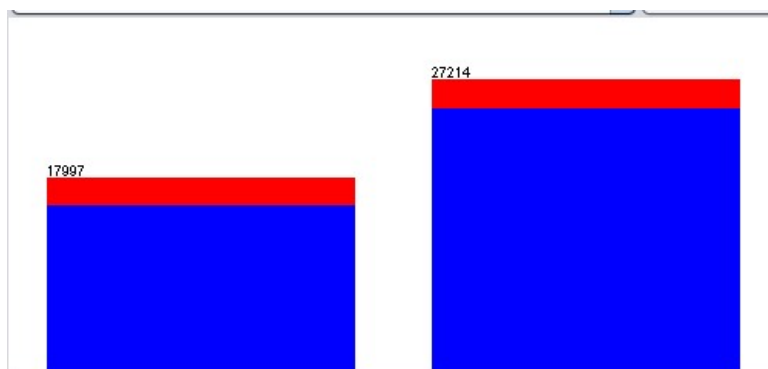
Graph



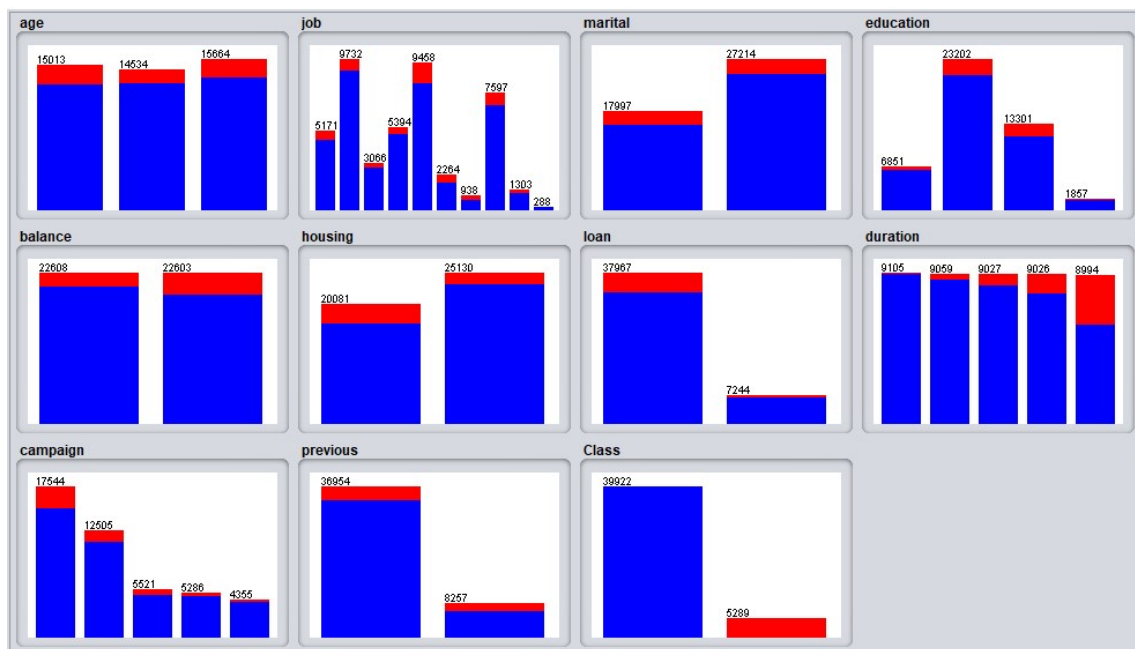
For the attribute marital i wanted to have two values either you're married or not so i merged single and the divorced variables.

Name: marital		Distinct: 2		Type: Nominal	
Missing: 0 (0%)				Unique: 0 (0%)	
No.	Label	Count		Weight	
1	divorced_single	17997		17997.0	
2	married	27214		27214.0	

Graph

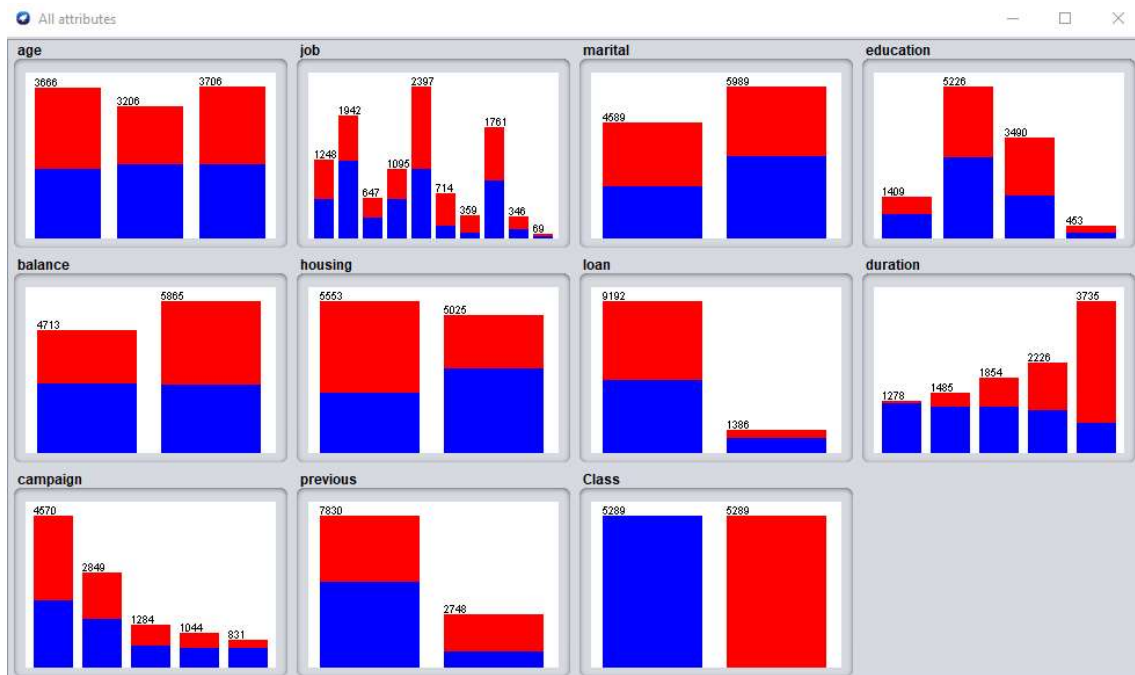


Result after preprocessing.



Because we only care about the outcome for the instance „yes“ in the output class, we need to balance the data so there is equal amount of „yes“ and „no“ instances.

After balancing



Rule mining process

For the parameter setting in apriori i chose the lowerbound minimum support as 0.1 and the delta to 0.05 to try to catch all possible rules that are close in support. I also chose confidence as the metric type for our rankings of the best rules.

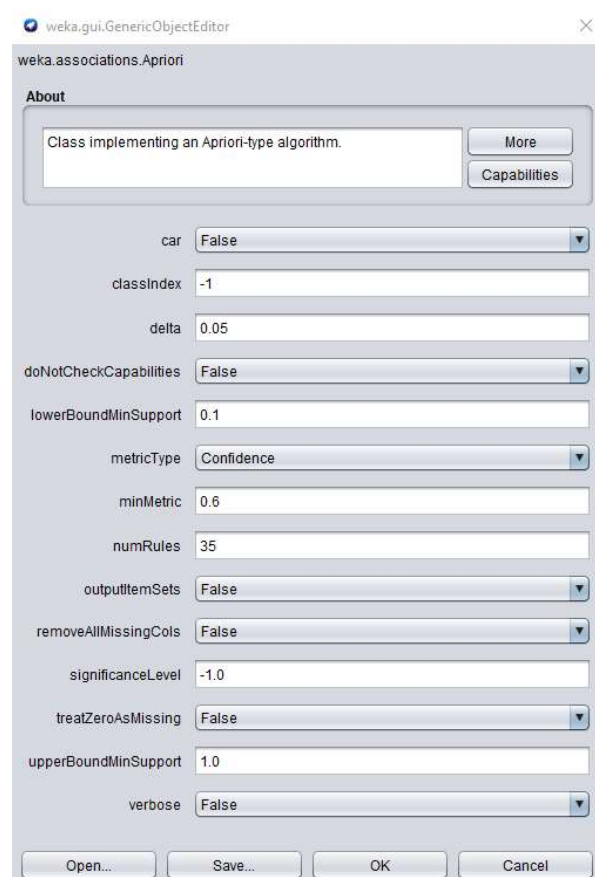
After the first run for the Apriori associator for the best 10 rules we can see that the attributes loan and previous are skewed so almost all the rules are inflated by these attributes so i decided to exclude them

Best rules found:

```
1. Class=2 5289 ==> loan=no 4805    <conf:(0.91)> lift:(1.05) lev:(0.02) [208] conv:(1.43)
2. balance='(448.5-inf)' 5865 ==> loan=no 5298    <conf:(0.9)> lift:(1.04) lev:(0.02) [201] conv:(1.35)
3. housing=no 5553 ==> loan=no 4990    <conf:(0.9)> lift:(1.03) lev:(0.02) [164] conv:(1.29)
4. previous='(-inf-0.5]' 7830 ==> loan=no 6715    <conf:(0.86)> lift:(0.99) lev:(-0.01) [-89] conv:(0.92)
5. marital=married 5989 ==> loan=no 5129    <conf:(0.86)> lift:(0.99) lev:(-0.01) [-75] conv:(0.91)
6. Class=1 5289 ==> previous='(-inf-0.5]' 4446    <conf:(0.84)> lift:(1.14) lev:(0.05) [531] conv:(1.63)
7. education=secondary 5226 ==> loan=no 4390    <conf:(0.84)> lift:(0.97) lev:(-0.01) [-151] conv:(0.82)
8. Class=1 5289 ==> loan=no 4387    <conf:(0.83)> lift:(0.95) lev:(-0.02) [-209] conv:(0.77)
9. marital=married 5989 ==> previous='(-inf-0.5]' 4482    <conf:(0.75)> lift:(1.01) lev:(0) [48] conv:(1.03)
10. loan=no 9192 ==> previous='(-inf-0.5]' 6715    <conf:(0.73)> lift:(0.99) lev:(-0.01) [-89] conv:(0.96)
```

For the next apriori run i increased the amount of rules because we only care about the rules that include the „yes“ instance for the output class and have a confidence above 0.6

Parameter settings.



The screenshot shows the 'weka.associations.Apriori' dialog box in the Weka GUI. The 'About' section at the top indicates it's a class implementing an Apriori-type algorithm. Below this, various parameters are configured:

- car:** False
- classIndex:** -1
- delta:** 0.05
- doNotCheckCapabilities:** False
- lowerBoundMinSupport:** 0.1
- metricType:** Confidence
- minMetric:** 0.6
- numRules:** 35
- outputItemSets:** False
- removeAllMissingCols:** False
- significanceLevel:** -1.0
- treatZeroAsMissing:** False
- upperBoundMinSupport:** 1.0
- verbose:** False

At the bottom, there are buttons for 'Open...', 'Save...', 'OK', and 'Cancel'.

Resulting rules

=== Run information ===

Scheme: weka.associations.Apriori -N 35 -T 0 -C 0.6 -D 0.05 -U 1.0 -M 0.1 -S
-1.0 -c -1

Instances: 10578

Attributes: 9

age

job

marital

education

balance

housing

duration

campaign

Class

=== Associator model (full training set) ===

Apriori

=====

Minimum support: 0.15 (1587 instances)

Minimum metric <confidence>: 0.6

Number of cycles performed: 17

enerated sets of large itemsets:

Size of set of large itemsets L(1): 21

Size of set of large itemsets L(2): 76

Size of set of large itemsets L(3): 19

Best rules found:

1. job=management 2397 ==> education=tertiary 2030 <conf:(0.85)> lift:(2.57)
lev:(0.12) [1239] conv:(4.36)

2. housing=no duration='(369.5-inf)' 1903 ==> Class=2 1607 <conf:(0.84)>
lift:(1.69) lev:(0.06) [655] conv:(3.2)

3. balance='(448.5-inf)' duration='(369.5-inf)' 2141 ==> Class=2 1736
<conf:(0.81)> lift:(1.62) lev:(0.06) [665] conv:(2.64)

4. duration='(369.5-inf)' 3735 ==> Class=2 2989 <conf:(0.8)> lift:(1.6)
lev:(0.11) [1121] conv:(2.5)

5. age='(44.5-inf)' housing=no 2401 ==> marital=married 1809 <conf:(0.75)>
lift:(1.33) lev:(0.04) [449] conv:(1.76)

6. age='(44.5-inf)' balance='(448.5-inf)' 2274 ==> marital=married 1713
<conf:(0.75)> lift:(1.33) lev:(0.04) [425] conv:(1.76)

7. age='(44.5-inf)' 3706 ==> marital=married 2729 <conf:(0.74)> lift:(1.3)
lev:(0.06) [630] conv:(1.64)

8. housing=no campaign='(-inf-1.5)' 2409 ==> Class=2 1653 <conf:(0.69)>
lift:(1.37) lev:(0.04) [448] conv:(1.59)

9. balance='(448.5-inf)' Class=2 3240 ==> housing=no 2162 <conf:(0.67)>
lift:(1.27) lev:(0.04) [461] conv:(1.43)

10. age='(44.5-inf)' marital=married 2729 ==> housing=no 1809 <conf:(0.66)>
lift:(1.26) lev:(0.04) [376] conv:(1.41)

11. balance='(-inf-448.5)' housing=yes 2483 ==> Class=1 1626 <conf:(0.65)>
lift:(1.31) lev:(0.04) [384] conv:(1.45)

12. marital=married housing=yes 2919 ==> Class=1 1910 <conf:(0.65)> lift:(1.31)
lev:(0.04) [450] conv:(1.45)

13. balance='(448.5-inf)' housing=no 3323 ==> Class=2 2162 <conf:(0.65)>
lift:(1.3) lev:(0.05) [500] conv:(1.43)

14. age='(44.5-inf)' 3706 ==> housing=no 2401 <conf:(0.65)> lift:(1.23)
lev:(0.04) [455] conv:(1.35)

15. marital=divorced_single housing=no 2483 ==> Class=2 1608 <conf:(0.65)>
lift:(1.3) lev:(0.03) [366] conv:(1.42)

16. age='(-inf-34.5)' 3666 ==> marital=divorced_single 2374 <conf:(0.65)>
lift:(1.49) lev:(0.07) [783] conv:(1.61)

17. campaign='(-inf-1.5)' Class=2 2561 ==> housing=no 1653 <conf:(0.65)>
lift:(1.23) lev:(0.03) [308] conv:(1.34)

18. housing=no Class=2 3354 ==> balance='(448.5-inf)' 2162 <conf:(0.64)>
lift:(1.16) lev:(0.03) [302] conv:(1.25)

19. marital=married Class=2 2755 ==> balance='(448.5-inf)' 1758 <conf:(0.64)>
lift:(1.15) lev:(0.02) [230] conv:(1.23)

20. marital=divorced_single Class=2 2534 ==> housing=no 1608 <conf:(0.63)>
lift:(1.21) lev:(0.03) [277] conv:(1.3)

21. Class=2 5289 ==> housing=no 3354 <conf:(0.63)> lift:(1.21) lev:(0.05) [577]
conv:(1.3)

22. marital=married Class=2 2755 ==> housing=no 1746 <conf:(0.63)> lift:(1.21)
lev:(0.03) [299] conv:(1.3)

23. age='(44.5-inf)' marital=married 2729 ==> balance='(448.5-inf)' 1713
<conf:(0.63)> lift:(1.13) lev:(0.02) [199] conv:(1.2)

24. education=secondary housing=yes 2743 ==> Class=1 1706 <conf:(0.62)>
lift:(1.24) lev:(0.03) [334] conv:(1.32)

```

25. campaign='(-inf-1.5]' Class=2 2561 ==> balance='(448.5-inf)' 1591
<conf:(0.62)> lift:(1.12) lev:(0.02) [171] conv:(1.18)

26. balance='(448.5-inf)' Class=1 2625 ==> marital=married 1626 <conf:(0.62)>
lift:(1.09) lev:(0.01) [139] conv:(1.14)

27. housing=yes Class=1 3090 ==> marital=married 1910 <conf:(0.62)> lift:(1.09)
lev:(0.02) [160] conv:(1.14)

28. marital=married balance='(-inf-448.5]' 2605 ==> Class=1 1608 <conf:(0.62)>
lift:(1.23) lev:(0.03) [305] conv:(1.31)

29. marital=married housing=no 3070 ==> balance='(448.5-inf)' 1894 <conf:(0.62)>
lift:(1.11) lev:(0.02) [191] conv:(1.16)

30. housing=yes 5025 ==> Class=1 3090 <conf:(0.61)> lift:(1.23) lev:(0.05) [577]
conv:(1.3)

31. education=secondary Class=1 2776 ==> housing=yes 1706 <conf:(0.61)>
lift:(1.29) lev:(0.04) [387] conv:(1.36)

32. age='(34.5-44.5]' 3206 ==> marital=married 1968 <conf:(0.61)> lift:(1.08)
lev:(0.01) [152] conv:(1.12)

33. age='(44.5-inf)' 3706 ==> balance='(448.5-inf)' 2274 <conf:(0.61)>
lift:(1.11) lev:(0.02) [219] conv:(1.15)

34. Class=2 5289 ==> balance='(448.5-inf)' 3240 <conf:(0.61)> lift:(1.1)
lev:(0.03) [307] conv:(1.15)

35. Class=1 5289 ==> marital=married 3234 <conf:(0.61)> lift:(1.08) lev:(0.02)
[239] conv:(1.12)

```

Remember we only care about the rules with Class=2 which is the „yes“ instance of the output class so we have the rules 2,4,8,13 and 15 we can show to the bank.

```

2. housing=no duration='(369.5-inf)' 1903 ==> Class=2 1607 <conf:(0.84)>
lift:(1.69) lev:(0.06) [655] conv:(3.2)

4. duration='(369.5-inf)' 3735 ==> Class=2 2989 <conf:(0.8)> lift:(1.6)
lev:(0.11) [1121] conv:(2.5)

8. housing=no campaign='(-inf-1.5]' 2409 ==> Class=2 1653 <conf:(0.69)>
lift:(1.37) lev:(0.04) [448] conv:(1.59)

13. balance='(448.5-inf)' housing=no 3323 ==> Class=2 2162 <conf:(0.65)>
lift:(1.3) lev:(0.05) [500] conv:(1.43)

15. marital=divorced_single housing=no 2483 ==> Class=2 1608 <conf:(0.65)>
lift:(1.3) lev:(0.03) [366] conv:(1.42)

34. Class=2 5289 ==> balance='(448.5-inf)' 3240 <conf:(0.61)> lift:(1.1)
lev:(0.03) [307] conv:(1.15)

```

However there are some rules that are not interesting for example we don't care for the duration of the phonecall, because we can't recommend to have the phonecall as long as possible so we ignore rule 2 and 4, but it is note worthy to mention them.

Recommendations

The client that will subscribe a bank term deposit will most likely not have a housing loan and has a balance over 448.5 so i would recomend the bank to target those clients in the next campaign