B.Sc in Computing

# Enterprise Database Technologies CA 1

Daniel Mateus Pires, x00132886

https://github.com/dmateusp/R\_CA1

Released: 24th February 2017

Table of contents

[Enterprise Database Technologies CA 1 0](#_Toc478218916)

[Section 1 - Data Understanding and Data Exploration 1](#_Toc478218917)

[1. Data pre-processing 1](#_Toc478218918)

[2. Discretizing income 1](#_Toc478218919)

[3. Finding information 1](#_Toc478218920)

[4. Finding outliers mathematically 3](#_Toc478218921)

[5. Skewness in TOT\_MINUTES\_USAGE 3](#_Toc478218922)

[6. Relationship between variables and response 4](#_Toc478218923)

[7. Correlated variables 5](#_Toc478218924)

[Section 2 – Data Mining 6](#_Toc478218925)

[Results 6](#_Toc478218926)

[Attributes kept 6](#_Toc478218927)

[What the algorithms do 6](#_Toc478218928)

[Attributes used in the predictions 7](#_Toc478218929)

[Interpreting the models 7](#_Toc478218930)

[Key predictors of churning 8](#_Toc478218931)

[Significant decisions paths 8](#_Toc478218932)

[Overall assessment 8](#_Toc478218933)

Appendix

## Section 1 - Data Understanding and Data Exploration

### Data pre-processing

First operation carried out was getting the number of null or empty string values per column (appendix 1.a).

We can see that only very few values are empty (single digit); and some columns have no value missing which indicates that our data is of good quality in terms of completeness.

The next step was to replace null numeric values with the median of the respective columns (appendix 1.b).

Then I created a function to get the mode by gender of a given column and replaced missing values in categorical columns by their respective modes (appendix 1.c).

The mode of PHONE\_PLAN is International for both Males and Females, there are also more Males churners than Females churners (plotting the influence of gender could be interesting).

### Discretizing income

One should be careful with the inclusion / exclusion of lower and upper ranges, the Low-Income category as an example end before 38,000 (37,999 is the last value), this is taken in account in the code.

### Finding information

For 3.c (appendix 3.c), the get\_mode function (created earlier), was used along with the summary function.

See appendix 3c, 3.d, 3.e, 3.f, 3.g for this question

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Predictor | A | B | C | D | E | F | G |
| AREA\_CODE | Nominal | 0 | **Mode**: 10040 | X | X | X | X |
| CUST\_MOS | Numeric | 3/2071 | **Min**: 1,  **Median**: 11,  **Mean**: 16.05,  **Max**:  50 | Most customers seem to stay during 5 to 15 months | It seems that in the first months, the customer has more chances to Churn, Around 10 months the customer will churn as well (end of one year contract?) | Skewness: 1.131244  Positively skewed (skewed to the right) | One outlier found in the box plot |
| LONGDIST\_FLAG | Nominal | 0 | **Mode**: 1 | X | X | X | X |
| CALLWAITING\_FLAG | Nominal | 0 | **Mode**: 0 | X | X | X | X |
| NUM\_LINES | Numeric | 0 | **Min**: 1,  **Median**: 1,  **Mean**: 1.391,  **Max**:  3 | Most users seem to have 1 number only and only very few have 3 numbers | The number of lines do not seem to bring much insight | Skewness: 2.057503  Strongly positively skewed (skewed to the right) | X |
| VOICEMAIL\_FLAG | Nominal | 0 | **Mode**: 1 | X | X | X | X |
| MOBILE\_PLAN | Nominal | 0 | **Mode**: 0 | X | X | X | X |
| CONVERGENT\_BILLING | Nominal | 0 | **Mode**: No | X | X | X | X |
| GENDER | Nominal | 0 | **Mode**:  M | X | Does not bring insight | X | X |
| INCOME | Ordinal | 0 | **Mode**:  Medium Income | Most users have medium income, twice as many users have high income compared to low income users | Users with low income or high income tend not to churn while medium incomes tend to churn more | X | X |
| PHONE\_PLAN | Ordinal | 4/2071 | Mode: **International** | Only few users choose the Euro-zone, most of the users opt for the International and National plans | Users having the Euro-Zone or the International phone plan tend to churn while users with a National or Promo\_plan tend to churn less | X | X |
| EDUCATION | Nominal | 8/2071 | Mode: **Post Primary** | Post-Primary is the dominant group, the number of High school and Primary school are very low | Primary are churners, Masters tend to churn, PhD tend not to churn | X | X |
| TOT\_MINUTES\_USAGE | Numeric | 4/2071 | **Min**: 0,  **Median**: 264,  **Mean**: 2036,  **Max**:  36237 | Clear majority of users use less than 2500 mins | Do not seem to bring insight | Skewness: 1.088757  Positively skewed (skewed to the right) | Graphically, from the boxplot we can see that the data contains a lot of outliers that will need to be cleaned out |

### Finding outliers mathematically

I chose TOT\_MINUTES\_USAGE since its box graph seems to indicate a lot of outliers.

I found 176 outliers using the IQR method while the Z-standardisation method found 69 outliers (appendix 4).

### Skewness in TOT\_MINUTES\_USAGE

My approach was, to first get the skewness value of TOT\_MINUTES\_USAGE before transformation: 1.088757, (appendix 5.), this positive skewness indicates that the data is skewed on the right (graphically we can see a long right tail). Most of the records will be on the left of the graph.

Z-score standardisation obtained the same skewness so no value was added, my observation is that Z-score uses mean and standard deviation which both are influenced by outliers (which are very present in TOT\_MINUTES\_USAGE). (appendix 5.a)

Natural log reduced skewness and made it a left-skewness (as opposed to the previous right skewness), it added value (appendix 5.b): -0.7042918

Square root increased the skewness, so it is not appropriate to use it with this data (appendix 5.c): 1.288432

### Relationship between variables and response

**a.**

To study the relationships, I used the same graphs plotted in appendix 3.e.

My approach was to plot histograms for each variable, colour encoded by the response variable.

Histograms where there are no disproportions between churners and non-churners on at least one of the values or range, might not be of any value for the prediction.

The question mentioned using only numeric variables, but, exploring the data showed more interesting results for overlaid graphs in some ordinal / nominal variables (included in the summary).

#### Variables that seem to influence churning (from this graphical method):

Income, where from the graph we can infer that Low and High Incomes are more frequent churners than Medium Incomes.

Phone plan, where we see that Euro-zone users are almost only churners, International users also have big churning rates while National and Promo-plan have low churning rates.

Education, where Primary have big churning rates and disproportions can be observed in all other categories besides Post Primary which seems to be balanced (might not help inferring rules).

Area codes where there are clear disproportions on each area.

#### Variables that seem to have no influence on churning (from this graphical method):

Number of lines and Gender seem to be both almost perfectly balanced, so they might be irrelevant to infer rules.

#### Variables for which the graph is not explicit:

Customer loyalty and Total minutes usage seem to show balance on some values while some other values show disproportions.

**b.**

I would expect Income or/and Area code to show up in the classification models as they seem to influence churning rate.

### 7. Correlated variables

**a.**

The sum of MINUTES\_CURR\_MONTH, MINUTES\_PREV\_MONTH, MINUTES\_3MONTHS\_AGO correlate with TOT\_MINUTES\_USAGE, the other variables are not correlated (appendix 7)

**b.**

The high correlation coefficient confirms the correlation between TOT\_MINUTES\_USAGE and the other usage metrics (0.9916) confirms the conclusions from the graphical analysis.

The 3 other correlation coefficients confirm that no other pair of numerical variables are correlated.

**c.**

As demonstrated in 3.c, the attributes that seem to have an important influence on churning are:

* Income
* Phone plan
* Education
* Area code

The attributes that seem to have some influence are:

* Customer loyalty (CUS\_MOS)
* Convergent billing

The attributes that seem to be of no value to find churners are:

* Number of lines
* Gender
* Total minutes usage

**d.**

The variables MINUTES\_CURR\_MONTH, MINUTES\_PREV\_MONTH and MINUTES\_3MONTHS\_AGO should be eliminated because they correlate with Total minutes usage.

However, we showed that Total minutes usage does not seem to bring value in finding churners, so the usage times could be dropped altogether.

Gender and Number of lines could be dropped as well as it seems that they are not bringing any value, the benefit is to keep the training from trying to use meaningless variables and it will also simplify our Decision Trees (less splits).

## Section 2 – Data Mining

### Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ZeroR | PART | JRip | J48 |
| FP | 0.479 | 0.114 | 0.084 lower than the other algorithms | 0.114 |
| FN | 0 | 0.288 | 0.306 higher than the other algorithms | 0.294 |
| Model accuracy | 47.8% accuracy | 80.25% | 80.96% | 79.97% |
| Precision | 0.229 | 0.809 | 0.822 | 0.806 |
| True Positive Rate | 0.479 | 0.803 | 0.810 | 0.8 |
| False Positive Rate | 0 | 0.205 | 0.2 | 0.208 |
| ROC | 0.5 | 0.859 | 0.846 | 0.845 |

### Attributes kept

The attributes kept for the training are: Income, Phone plan, Education, Area code, Customer loyalty (CUS\_MOS), Convergent billing

### What the algorithms do

ZeroR trains on the proportions of the Response Classes and replicates this proportion on the test set by assigning randomly, that is how we get a result close to a random guessing (50% precision).

PART (appendix 8.a) and JRip (appendix 8.b) both create rules by combining prediction variables.

PART builds a partial C4.5 decision tree in each iteration and makes the “best” leaf into a rule.

J48 (appendix 8.c) creates a decision tree.

### Attributes used in the predictions

PART seem to use AREA\_CODE greatly along with EDUCATION and INCOME while JRip rely a lot on AREA\_CODE and less on EDUCATION / INCOME.

J48 uses every variable almost equally, besides INCOME which is only used for small part of the decision and CONVERGENT\_BILLING that was completely dropped off.

### Interpreting the models

The PART inferred rules are to be read in the following manner:

If the phone plan is Euro-Zone and area code is 36785 then that person will churn (39 were classified correctly using this rule),

Else, if income is high and no convergent billing then that person will not churn (510 were classified correctly using this rule, 137 were wrongly classified)

Etc...

The JRip inferred rules are to be read in the following manner:

If income is high and no convergent billing then that person will not churn (510 were classified correctly using this rule, 137 were wrongly classified) -> same rule as in PART

Else, if area code is 10040 then that person will not churn (310 were classified correctly using this rule, 125 were wrongly classified)

Etc…

The J48 tree is to be read in the following manner:

If phone plan is international and education is master then that person will churn (180 were correctly classified using this rule)

If phone plan is euro zone then that person will churn (59 correctly classified using this rule, 9 incorrectly classified)

Etc…

The decisions align with the conclusions drawn by the graphical analysis.

### Key predictors of churning

INCOME and AREA\_CODE seem to be the two main predictors of churning across all algorithms.

### Significant decisions paths

From the PART and the JRip output, “Income high and convergent billing no” is a rule that classified 510 records correctly and 137 incorrectly (a very used rule), this rule is significant and the result shows that it is meaningful.

From the J48 tree, the rule “Phone plan international and education post-primary” is a rule that classified 351 records correctly and 71 incorrectly, making it also a significant rule.

### Overall assessment

1. From the conclusions above we can see that some areas a more prone to churning than others, it might be low coverage or low quality of the service in those areas.

Education is also a factor, that is combined with phone plans or area codes to give multiple rules. The education might have a relation with the age of the customer and therefore its needs and an inadequate offer or phone plan could make these customers churn.

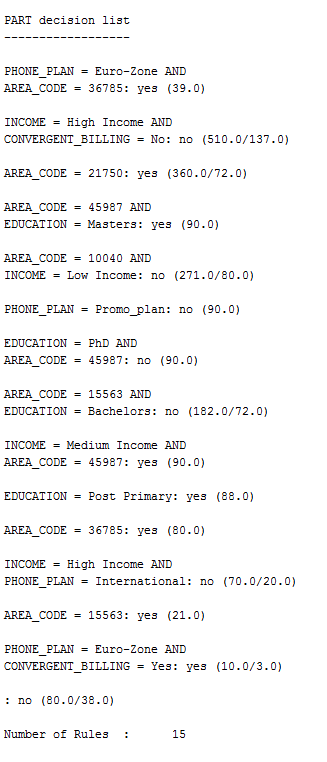
1. The persons with an international phone plan and that have post-primary education or Masters seem to be more likely to churn (J48). The area code 21750 seems to be prone to churning as well.

These are the clearest rules that we can infer from the algorithms output.

1. The customers that enter in these categories can be monitored more closely, marketing measures can be taken to try and improve customer retention in these areas as well as getting feedback from these customers could be insightful for the company.

# Appendix WRITTEN REPORT

## 8.a



## 8.b

## 

## 8.c

