# Churn in Telecom dataset

Databases and Datamining, 2009

Jonathan Vis, Rick van der Zwet
<{jvis,hvdzwet}@liacs.nl>

7 november 2009

## 1 Introduction

This report is focused towards finding association rule learning to find relations between variables in large databases. This will be done using Weka[1] and a telecom churn dataset[2].

## 2 Problem description

*Churning* -moving to a different company)- today is still a major deal within companies. Having to understand why a customer choose to go for an other company is crusial in finding flaws in the product-range or services. As more and more data about the consumer get stored, trying to find relations why he/she churned is becoming more and more interesting.

## 3 Statistics

Our dataset has 3333 entries and 21 attributes, which the charactistics shown in table 1.

We can consider ourself lucky by having an complete dataset. None of the attributes is missing at an entry. How-ever this does not mean the data is considered error-free. There might be human-errors or others of some kind inside the dataset.

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://www.dataminingconsultant.com/DKD.htm

Tabel 1: Statistical report of churn dataset

| Item | Type | Distinct | Missing | Unique | Min | Max | Mean | StdDev |
|---|---|---|---|---|---|---|---|---|
| State | Nominal | 51 | 0 | NaN | NaN | NaN | NaN | NaN |
| Account Length | Numeric | 212 | 0 | 16 | 1 | 1 | 101 | 40 |
| Area Code | Numeric | 3 | 0 | 0 | 408 | 510 | 437 | 42 |
| Phone | Nominal | 3333 | 0 | 3333 | NaN | NaN | NaN | NaN |
| Int'l Plan | Nominal | 2 | 0 | NaN | NaN | NaN | NaN | NaN |
| VMail Plan | Nominal | 2 | 0 | NaN | NaN | NaN | NaN | NaN |
| VMail Msg | Numeric | 46 | 0 | 4 | 0 | 51 | 8 | 14 |
| Day Mins | Numeric | 1667 | 0 | 770 | 0 | 351 | 180 | 54 |
| Days Calls | Numeric | 119 | 0 | 10 | 0 | 165 | 100 | 20 |
| Days Charge | Numeric | 1667 | 0 | 770 | 0 | 60 | 30 | 9 |
| Eve Mins | Numeric | 1611 | 0 | 709 | 0 | 364 | 201 | 51 |
| Eve Calls | Numeric | 123 | 0 | 17 | 0 | 170 | 100 | 20 |
| Eve Charge | Numeric | 1440 | 0 | 585 | 0 | 31 | 17 | 4 |
| Night Mins | Numeric | 1591 | 0 | 586 | 23 | 395 | 201 | 51 |
| Night Calls | Numeric | 120 | 0 | 11 | 33 | 175 | 100 | 19 |
| Night Charge | Numeric | 933 | 0 | 236 | 1 | 18 | 9 | 2 |
| Intl Mins | Numeric | 162 | 0 | 16 | 0 | 20 | 10 | 3 |
| Intl Calls | Numeric | 21 | 0 | 3 | 0 | 20 | 4.5 | 2.5 |
| Intl Charge | Numeric | 162 | 0 | 16 | 0 | 5.4 | 2.8 | 0.8 |
| CusServ Calls | Numeric | 10 | 0 | 0 | 0 | 9 | 1.5 | 1.3 |
| Churn | Nominal | 2 | 0 | 0 | NaN | NaN | NaN | NaN |

# 4 Approach

As finding association rules needs discrete values, we will discretize the attributes tagged as Numeric in table 1. At this process we will take a special look of the actual meaning of the attribute. Having 3.5 Customer Server Calls is going to be a bit impossible, so make sure to set the binaries of all *(*bins) to rounded values. We will also try *(*Weka) feature of automatic discretizion:

```
weka.filters.unsupervised.attribute.Discretize -unset-class-temporarily
-O -B 10 -M -1.0 -R first-last
```

.

Secondly table 1 shows an number of entries which are related. Like for example *Day Calls* and *Day Charge*. One could argue that both are relevant, as an consumer might churn as making many calls turns out to be problematic (bad signal, quality for example). While you could also argue the price will determine the churn. We will discard the values related to minutes and call

numbers and solely focus on the *Charge*. The *Phone* attribute also shows some interesting feature. We will make an new attributecalled *Phone-prefix* which is the first 3 numbers of the *Phone* number, to see whether this give some fine gain grouping over area code. Also the combination *Area Code-Phone-prefix* will be researched.

# 5 Implementaion

Creating phone-prefix column using standard unix tools:

```
cut -f 4 -d, churn_ooo.csv | cut -c 1-4,10 | paste -d, - churn_ooo.csv
| sed '1s/"Pho/"Phone-Prefix/' > churn_parsed.csv
```

. Using *Weka* deleted the columns `Day Mins, Day Calls, Eve Mins, Eve Calls, Night Mins, Night Calls, Intl Mins, Intl Calls` as we believe they are subsets of *Charge*. *Phone* is a unique identifier for every entry, not allowing any generalization. So it is ignored/deleted.

Using *Weka* embedded discretize function on all Numeric Columns of table 1. Any charge value was taken to be full integer values e.g rounded currency:

```
weka.filters.unsupervised.attribute.NumericCleaner -min -1.7E308
-min-default -1.708 -max 1.7E308 -max-default 1.7E308 -closeto
0.0 -closeto-default 0.0 -closeto-tolerance 1.0E-6 -R 8-11
-decimals 0
```

And made discrete:

```
weka.filters.unsupervised.attribute.NumericToNominal -R 8-11
```

Account length was set to be a bin of 'weight' 1, assuming a 1 months[3] contract:

```
weka.filters.unsupervised.attribute.NumericToNominal -R 3
```

*Phone-Prefix* is set to be a set on every number unique:

```
weka.filters.unsupervised.attribute.NumericToNominal -R 1
```

*CustServ Calls* it set to be rounded values, as one cannot make half calls:

```
weka.filters.unsupervised.attribute.NumericToNominal -R 12
```

**Result 1** association algoritm *Apriori* - run 1

```
1. Area Code=415 Int'l Plan=no VMail Plan=yes 423 ==> Churn?=False. 405 conf :(0.96)
2. Int'l Plan=no VMail Plan=yes 830 ==> Churn?=False. 786    conf:(0.95)
...
```

Then running the association algoritm *Apriori*, with the *Churn* value as the result of the equations:

Mostly negative results e.g. proving when a consumer is not going to churn. Secondly data seems to specialise, rule 1 for example is a specialisation of rule 2. Try running without the requirement that churn needs to be on the right end of the rule.

**Result 2** association algoritm *Apriori* - run 2

```
1. VMail Message=0 2411 ==> VMail Plan=no 2411    conf:(1)
2. VMail Plan=no 2411 ==> VMail Message=0 2411    conf:(1)
...
```

Seems like *VMail Plan=no* seems to equal the *VMail Message=0*. So *VMail Plan* can be safely deleted from the attribute list. *VMail Message* does not seems to be a very clear description for it's claimed purpose. Running without *VMail Plan* did not show improvement. Hence we decided to go for an normalisation on the Churn number. By taking a random sample of $Churn = False$ values such that it equals the number of $Churn = True$ values[4] we re-run the experiments, but found no improvement in the experiments.

# 6 Conclusions

The churn dataset does not classify itself properly associations rules. Mainly due to the fact that the so called 'hidden factors' for churning, like 'if calling more than X minutes at rate Y I will churn'. cannot be mined using this current dataset. Further research could include this relations by means of formula's, but it require domain specific knowledge to include for example relations between *Day Min* and Day Calls. Alternative methods like scat and plat analysis[5] seems to lead to more promensing results. This could (of course) also be done in Weka. Take for a brief preview in Appendix 1.

[3]This might as well be days, years of some other value, but assuming fixed phone contracts, given the range (0-244) months seems to be make most sense

[4]`sort -r -t, -k 22,22 churn_parsed.csv | sed '484,2366d' > churn_equal.csv`

[5]Like done at http://meru.cecs.missouri.edu/courses/cecs401_data_mining/projects/group2/finproject1.htm

# 7 Appendix 1

Result using *Weka* classifier:

```
weka.classifiers.trees.J48 -C 0.25 -M 2
```

shows interesting details, like

```
'Day Mins' > 254.4 and 'VMail Plan' = no and 'Eve Mins > 187.7 => True.
```

**Result 3** J48 pruned tree of raw churn dataset

```
=== Classifier model (full training set) ===
J48 pruned tree
------------------
Day Mins <= 264.4
|   CustServ Calls <= 3
|   |   Int'l Plan = no
|   |   |   Day Mins <= 223.2: False. (2221.0/60.0)
|   |   |   Day Mins > 223.2
|   |   |   |   Eve Mins <= 242.3: False. (296.0/22.0)
|   |   |   |   Eve Mins > 242.3
|   |   |   |   |   VMail Plan = yes: False. (20.0)
|   |   |   |   |   VMail Plan = no
|   |   |   |   |   |   Night Mins <= 174.2
|   |   |   |   |   |   |   Day Mins <= 246.8: False. (12.0)
|   |   |   |   |   |   |   Day Mins > 246.8: True. (5.0/1.0)
|   |   |   |   |   |   Night Mins > 174.2: True. (50.0/8.0)
|   |   Int'l Plan = yes
|   |   |   Intl Calls <= 2: True. (51.0)
|   |   |   Intl Calls > 2
|   |   |   |   Intl Mins <= 13.1: False. (173.0/7.0)
|   |   |   |   Intl Mins > 13.1: True. (43.0)
|   CustServ Calls > 3
|   |   Day Mins <= 160.2
|   |   |   Eve Charge <= 19.83: True. (79.0/3.0)
|   |   |   Eve Charge > 19.83
|   |   |   |   Day Mins <= 120.5: True. (10.0)
|   |   |   |   Day Mins > 120.5: False. (13.0/3.0)
|   |   Day Mins > 160.2
|   |   |   Eve Charge <= 12.05
|   |   |   |   Eve Calls <= 125: True. (16.0/2.0)
|   |   |   |   Eve Calls > 125: False. (3.0)
|   |   |   Eve Charge > 12.05: False. (130.0/24.0)
Day Mins > 264.4
|   VMail Plan = yes: False. (53.0/6.0)
|   VMail Plan = no
|   |   Eve Mins <= 187.7
|   |   |   Day Mins <= 280.4: False. (30.0/7.0)
|   |   |   Day Mins > 280.4: True. (27.0/9.0)
|   |   Eve Mins > 187.7: True. (101.0/5.0)
```