B.Sc. in IT Management
B.Sc in Computing
Enterprise Database Technologies
CA 1 - INDIVIDUAL
This CA is worth 30%
Released: 24th February 2017
Upload by 24th March 2017

This CA is designed to assess your skills and abilities to pre-process the data from a data understanding and exploratory data analysis perspective in the context of the data mining process. Additionally, it will also assess your skills in the data mining tasks assigned. There are two sections to the report.

Section 1 - Data Understanding and Data Exploration (50%). Use the .csv file provided on Moodle.

Section 2 – Data Mining (50%)

This CA is on an individual basis. A brief viva/presentation of findings may be required as part of this work. Any work not directly your own must be referenced. NB: Institute plagiarism rules apply to this and every CA. These rules will be strictly enforced.

The assignment will be checked through the Turnitin Plagiarism Prevention system, for identifying unoriginal material, copied (without reference to the source) from an electronic source on the Internet, electronic libraries, other assignments.

It is critical that all terms are properly explained and that the significance of your observations and comments are properly communicated to the reader.

It is expected that authors will seek to tie in appropriate material from the lectures, and readings into this report.

Marks will be awarded for completeness and correctness of the analysis, synthesis of ideas, conciseness and clarity of thought and argument.

NB The report should be no longer than 6 pages (excluding appendices). Evidence of the use of R should be correctly referenced in the appendix. Significant graphs and statistics to justify your findings and argument should only appear in the main body of the report

You are required to upload one PDF file containing your report.

Scenario 1

EuroCom provides new fixed line, mobile phone and broadband services to customers in the Euro-zone. It uses the Euro-zone telecommunications network and competes for customers against other major telecommunications companies in the market. Currently the company has a 50,000 customer base and there are roughly 250,000 calls per day.

One of the biggest problems facing EuroCom is customer retention. Customers are free to move between telecommunications service providers and some regularly change service providers, a phenomenon known as churn in the telecommunications industry. Some companies have churn rates as high as 20% of their customers changing service providers per annum. EuroCom has placed a high emphasis on churn and are always trying to find new ways of reducing it. The directors of EuroCom are looking for the answer to the main question of how we detect customers who are going to churn

Overall, the directors are looking for answers to these types of questions.

1. What is it that makes a customer churn?
2. Are some customers more likely to churn than others?
3. How can we identify these customers before they churn?

A representative sample dataset is provided. It is expected that you **will use R to explore and prepare the data**. Your approach to the problem should follow the CRISP-DM process: data understanding (also called exploratory data analysis), data pre-processing stages followed by the Data Mining task.

Section 1 (50 marks)

Before starting your analysis, please familiarise yourself with the churn dataset provided. You can assume the sample was randomly selected from a population that is normally distributed. Please refer to Appendix 1 for description of the predictor variables and response (target) variable.

As part of an appendix to your report provide clearly commented supporting R code you used to carry out your tasks).

1. Perform data processing on the data set, as required. Give evidence whether there are problems with data quality, duplicate data or missing data. **Note:** With respect to predictor variables, missing categorical values should be replaced by the mode value by gender for that predictor variable. e.g. if the voice_mail flag is missing for a record that has gender of male, the replacement value should be the most popular voice_mail flag value for males. Missing numeric values should be replaced by the median value for the variable. Comment on your findings as well as the actions you carried out.
2. Discretise the Income predictor variable as follows;
   Income >= 88,000 -> **High Income**
   Income < 88,000 && Income >= 38,000 -> **Medium Income**
   Income < 38,000 -> **Low Income**
3. For each predictor variable, where appropriate, find the following information:
   a. The attribute type, e.g. nominal, ordinal, numeric.
   b. Percentage of missing values in the data.
   c. Max, min, mean, mode, median standard deviation.
   d. The type of distribution that the numeric attribute seems to follow (e.g. normal).
   e. Study the histogram of the attribute and note how it seems to influence the risk of churning.
   f. Whether the numeric data is skewed and the type of skewness.
   g. Using graphical methods, are there any outliers for the attribute under consideration?

   Overall, provide a short commentary on overall findings\observations. Use graphical methods as appropriate to support your answer.
4. Choose a numeric predictor variable that has possible outliers based on your analysis in 2 above. Use the IQR method and Z-Score Standardisation method to identify outliers. Discuss your findings.
5. Choose a numeric predictor variable you have determined as skewed. Transform the data using the following transformation methods in an attempt to achieve normality:
   a. Z-Score Standardisation
   b. Natural Log Transformation
   c. Square Root Transformation

   You may use graphical methods to assist you. Comment on your results with regard to the normality of the data.
6. Construct a histogram of each numerical variable, with an overlay of the response variable.
   a. Discuss the relationship, if any, each of the variables has with an overlay of the response variable.
   b. Which variable would you expect to make a significant appearance in any data mining classification model. Justify your answer.
7. Investigate whether there are any correlated variables. Using R visualize 2D-scatter plots for each pair of numeric attributes:
   a. Does any pair of variable look to be correlated?
   b. Verify your assertions using statistical methods.
   c. Which attributes seem to be the most/least linked to the risk of churning? Summarise in a table your findings concerning the predictive value of each attribute with respect to the churn attribute.
   d. Are there any variables that can be eliminated? Justify your answer and express the possible benefits of doing so (if any).

Section 2 Learning Algorithms (50 marks)

It has been agreed by the domain expert, that the variables CUST_ID, MINUTES_CURR_MONTH, MINUTES_PREV_MONTH, MINUTES_3MONTHS_AGO should not be part of the data mining task

Before you attempt this task, make sure to normalise any numerical data and deal with any perfectly correlated variables if any. Justify your decisions.

Use the Weka Data mining tool for this task. For learning schemes you have not seen in lectures, explain in a few sentences the main principle of the learning algorithm. **For training\testing use the Hold-out method with a 66%\34% split.**

Use the following learning schemes to analyse the churns data:

weka.classifiers.rules.PART
weka.classifiers.rules.JRip
weka.classifiers.trees.J48

Use ZeroR as your baseline classifier

For each classifier answer the following questions:

Describe briefly each classifier and indicate which parameters you have used to run it. In interpreting the model, how does the classifier determine whether a customer is a churner or not? Do the decisions made by the classifiers make sense to you? What are the key predictors of churn? What are the significant rules\decision tree paths? Provide evidence for your assertions.

What can you say about the accuracy\ error rate of these classifiers when classifying a churner? As part of your answer, you should comment on the following measures on the test dataset:

- **i** Proportion of false positives
- **ii** Proportion of false negatives
- **iii** Overall error rate and overall model accuracy
- **iv** Precision
- **v** Sensitivity(Recall) - True Positive Rate
- **vi** Specificity - False Positive Rate
- **vii** ROC

Overall assessment

Do the classifiers agree with each other? Are there significant differences? As part of your analysis, comment on the models produced by the classifiers. Justify your reasoning. Comment on the models produced by the classifiers. Provide an overall assessment of the classifiers. Is there preferred model? Explain your reasoning. From your analysis, what is your advice with regard to the original questions posed.

1. What is it that makes a customer churn?
2. Are some customers more likely to churn than others?
3. How can we identify these customers before they churn?

Appendix 1

The dataset provided is representative sample of EuroCom customer accounts on at a point in time. It also shows the account holders who EuroCom lost and the ones they retained.	
CUST_ID	Unique Account Number that identifies an account holder
AREA_CODE	Geographical area account holder resides; Should be treated as a categorical variable
MINUTES_CURR_MONTH	Phone Minutes currently for current months (to the time the data was extracted)
MINUTES_PREVIOUS_MONTH	Phone Minutes used in the previous month
MINUTES_3MONTHS_AGO	Phone Minutes used in the 3 month PREVIOUS
CUST_MOS	The number of continuous months the Customer is with the provider
LONGDIST_FLAG	Whether has signed up for the off-peak long distance call package
CALLWAITING_FLAG	Whether the customer has call waiting
NUM_LINES	The number of fixed lines the customer has leased.
VOICEMAIL_FLAG	Whether the customer has voice mail
MOBILE_PLAN	Has the customer signed up to the mobile phone plan.
CONVERGENT_BILLING	All service charges consolidated onto one bill
GENDER	Account holder's gender
INCOME	Account holder's annual income (€)
PHONE_PLAN	The phone plan the customer has signed up for national, euro-zone, international (outside Euro-zone) and promo_plan (signed up to the promotional plan)
EDUCATION	Highest Level of education attainment the account holder has achieved
TOT_MINUTES_USAGE	The total number of minutes used to date
CHURNER (response)	Attrition: Whether the Customer left the Telco or not

Notes

1. To save your dataframe in R to a csv datafile, use write.csv() function

   write.csv(churn, file = "c:/mydir/churn.csv")

2. To read a CSV file into WEKA, ensure you choose the .csv file format when you click on Open File in the PreProcess tab in Weka Explorer
3. Variables that are read in to Weka as numeric variable but need to be treated as nominal can be converted as follows:
   Step 1: Select your filter in the preprocess tab by clicking “Choose”
   Step 2: Navigate to unsupervised and search for "NumerictoNominal".
   Step 3: In the attributeIndices box enter your custom ranges for attributes you wish to convert. Click okay and press Apply