# Business Analytics Practicum I

## Andreas Zaras

## Group Assignment

## Deadline: 01/04/2017

As it was said in the course outline, the group assignment accounts for 80% of the final grade (the other 20% is connected with your performance during the lectures). It should be conducted in groups of two members and both members should be either from the part time program or the full time program (full timers and part timers should not be mixed). The names of the two group members should be mailed to Andreas.Zaras@sas.com until Saturday the 25th of February. The assignment consists of three case studies, the first related to market basket analysis (association rules), the second to customer segmentation (clustering) and the third related to credit risk (predictive analytics). The first case study accounts for 15% of the assignment's grade, the second for 25% and the third for 60%.

The deliverable should be one report where you should provide answers to all the case studies. Beware that some answers are the deliverable to technical people whereas, some other, are the deliverable to business people so the writing style should be appropriate (you will be marked on this!). At the beginning of each case study, you should include an executive summary to be addressed to the management team of each organization that should contain the problem under consideration, how you tackled it, what methods you used, what decision support tool was utilized and what was the final result. Beware that an executive summary should be short (not more than half a page), clear and should not contain any technicalities.

In order to provide answers to the questions you must use SAS Enterprise Miner to explore and analyze the given data. You must create only one project named "LastNameOfFirstMember_LastNameOfSecondMember" and the data of each case study should be explored and analyzed in a different process flow named "Case_Study_1", "Case_Study_2" and "Case_Study_3" respectively. The report should be sent in word and pdf format and their names should be "LastNameOfFirstMember_LastNameOfSecondMember.docx" and "LastNameOfFirstMember_LastNameOfSecondMember.pdf" respectively.

You should send the report to Andreas.Zaras@sas.com by the 1st of April 2017. Each day that the report will be delayed, a penalty of 10% will be applied to the grade***. In the body of the mail that you will send you should also include the username and password from the SAS account that you created in order to access the software***. Since it is a group assignment only the credentials of one member of the team should be sent. The instructor will check whether the work in the software is in line with the results included in the report.

## Case Study 1 (15%)

Buy-books-on-line.com is an on line store that sells books about science and information technology. The store is very well known in the academic community so a lot of its customers are university professors and also librarians at universities buying on behalf of their institutions. A very popular category of the books that the store sells is that related to "Business Analytics". In this category the store has a list of 56 books such as "Credit Risk Analytics", "Marketing Analytics", "Analytics at Work" etc. The past year 1,896 customers have bought at least one book that belongs to the "Business Analytics" category i.e. at least one of the 56 books.

The sales department of the store wants to exploit cross selling opportunities so as to sell as many books as possible. The optimal way to achieve this, is to do wise next best offer propositions to its customers by applying associations rules. The analytics department of the store has collected a data set with 19,805 past sales transactions related to the "Business Analytics" book category. The data set is called "Final_Book_Transactions".

You are hired as an analyst by the on-line store to aid the analytics department in this market basket analysis initiative. After the data analysis **you should write a report to the analytics team of the company (technical people) to explain them what you did, which method you used, how it works and what were your results**. As already said the report should contain an executive summary in a business format. In the main body of the report you should answer the following questions:

1) What are the sales (in units) of each book? Provide a relevant chart (bar chart) using the software. Enrich the chart so as to show data labels, different color for each bar, legend at the south side of the chart, chart title, titles in both axis. This question accounts for 20% of the case study's mark.

2) Which two books should the store advertise to customers who bought/ are searching to buy only one of the following:

- Managerial Analytics
- Implementing Analytics
- Customer Analytics for Dummies
- Enterprise Analytics

In other words create the Amazon's "Customers who Bought this Item also bought" list of books. What is the biggest lift of the rules with three (3) items where each one of the above mentioned 4 books is on the left side of the rule? How is it interpreted? In order to answer the above questions you should use the association analysis node in Enterprise Miner **with the** default settings except that the maximum items in a rule should be three (3). Beware that the customers are uniquely identified by the variable PK_Customer and not by the variable Customer_Name (since two or more customers might have the same name), so be careful when you assign the role ID in the data set. Copy the results from the "Rules Table" window (200 rows) in an Excel spreadsheet and find the answers to the questions. This part accounts for 40% of the case study's mark.

3) If you set the maximum items in a rule to 3, which are the 3 books most bought together by customers? How many occurrences of this set of 3 books are found together? What does this number mean? What is the support metric of this set of 3 books and how is it calculated? This part accounts for 40% of the case study's mark.

## Case Study 2 (25%)

Sports-OnLine.com is an on line retailer that sells sports clothes and shoes and it is operating in the market since October 2001. On January 2007, after six year of operation, the management team of the store wants to exploit the electronic data captured the previous years to better understand the market. After a meeting with the marketing department, it was decided that a customer segmentation analysis should be performed and, based on the available data, a Recency Frequency Monetary (RFM) analysis would be the most suitable technique for the desired objective.

During the period Oct 2001 – Dec 2006, 995 customers have done 4906 sales transactions that have been recorded by the on line store and have been stored in the following data set:

4,906
transactions

| Customer_ID | Date_OF_Transaction | Amount_Of_Transaction |
|---|---|---|
| Cust345 | 05/03/2005 | 123 |
| Cust120 | 10/01/2004 | 34 |
| Cust657 | 23/02/2006 | 53 |
| ……. | ……. | ……. |
| Cust213 | 03/03/2003 | 12 |
| Cust086 | 29/07/2002 | 65 |

The IT department in cooperation with the Business Analytics department have transformed the above data set into RFM format, and have produced the SAS data set named RFM_Assignemnt_Final.sas7bdat that is presented below. Since the 4,906 transactions of the previous data set have been produced by 995 customers the RFM data set has 995 rows, each one corresponding to a single customer.

995
Customers

| Customer_ID | R | F | M |
|---|---|---|---|
| Cust001 | 4 | 5 | 485 |
| Cust002 | 14 | 4 | 350 |
| Cust003 | 19 | 2 | 233 |
| ……. | ……. | ……. | ……. |
| Cust994 | 24 | 1 | 185 |
| Cust995 | 6 | 2 | 187 |

You are hired as the Marketing Analytics consultant to perform the RFM segmentation with the data mining software SAS Enterprise Miner. Do the clustering of the customers and the profiling of the segments created. Name the segments (e.g. churners, good customers, bad customers, first time customers etc) and describe briefly what marketing actions are appropriate for each segment (e.g. customer reactivation program, contact customers for feedback, cross sell activities, special promotions etc) and why.

## Case Study 3 (60%)

This case study refers to a fictitious bank - XYZ - that does business in the retail banking sector. Until recently the bank granted consumer loans to its customers by using a generic scorecard developed by a credit agency using external data, coupled with the intuition and experience of its officers. The past year the regulator of the banking system has obliged the credit institutions to adopt objective methods for credit operations and more specifically to develop in house statistically based credit risk systems using their own data.

The credit risk department of the bank, in cooperation with the IT department, has collected the data set named “Applicants”, which contains data about past loan applications. The data set contains customer’s application characteristics (for example age, amount of loan, purpose of loan and so on) and a target variable that indicates whether the applicant proved to be a good client or a bad client i.e. whether the client returned to the bank the money that he borrowed or not. You can find the data dictionary of this data set at the end of this document.

You are hired as a data mining analyst to aid the credit risk department develop a statistical model that will predict whether a customer applying for a load will prove to be a good or a bad client. After the model is developed by using the historical data (“Applicants” data set), new customers asking for a loan should be scored so as to predict whether they will return their loan or not and hence whether they will be granted the loan or not. The new customers along with their characteristics are stored in the “Score” data set. Please follow the following steps and answer the relevant questions:

Open SAS Enterprise Miner.

Create a new project.

Create a new diagram and name it Credit_Risk.

Create a new data source (Applicants) by consulting the relevant data dictionary.

Drag and drop the Applicants data set from the project pane to the diagram workspace.

Go to options -- > preferences and change the sample option to Random and the fetch option to Max.

1) Are there any missing values in the variables of the data set? Provide a screenshot of the software to prove this. What is the proportion of good and bad clients in the data set? Provide a screenshot from the software to prove it (pie chart). This part accounts for 2.5% of the case study's mark.

2) It should be noted that the data set, as it is provided, it is adjusted for separate sampling. The original proportion of goods and bads was **90% good - 10% bad** and according to the predictive analytics rule of thumb the data set was sampled in a way that the proportion became 70% - 30%. Explain the reasons for making this adjustment. How many applicants were in the original data set? How many of them were good and how many were bad? How was practically the sample created? This part accounts for 2.5% of the case study's mark.

3) Provide a graph (pie chart) that shows the proportion of good and bad clients in ages over 60. What do you observe? This question accounts for 2.5% of the case study's mark.

4) What is the average age of a) bad clients and b) good clients (run a stat explore node and check the output in the results window). What does this mean with respect to the target variable? This question accounts for 2.5% of the case study's mark.

5) Add a data partition node to the diagram and connect it with the data source. Assign 70% of the data for training and 30% for validation. The sampling in the data partition node is stratified. What does this mean? Provide output from the software to justify that. This question accounts for 2.5% of the case study's mark.

As it was previously said, the data set is adjusted for separate sampling. In order to get accurate results we must tell the software what was the original proportion of goods and bads. Provide the necessary information to the software by setting prior probabilities (Select the "Applicants" data source and in the properties panel press Decisions in the Columns tab. Go to the Prior Probabilities tab, select Yes in the question and input the prior probabilities).

Add a decision tree node to the workspace and connect it to the data partition node. Create a decision tree interactively.

6) What is the variable used for the first split? What is its logworth? Which cases are directed to the left node and which to the right node? Where are the missing values directed to? This part accounts for 2.5% of the case study's mark.

**Within the interactive tree facility, create the maximal tree (Select the root node -- > right click -- > Train Node).**

Add a second decision tree node to the workspace and connect it to the data partition node. Name the new decision as largest tree. In the subtree tab of the properties window change the method to largest. Run the tree node.

7) How many terminal leaves does the tree have? How is this tree called? Check the performance of the training and validation data set when the misclassification rate is used as the assessment criterion. Provide the relevant graph (subtree assessment plot) in your report. How is the phenomenon presented in line for the training data set (blue line) called? Explain it briefly in a couple of sentences. Describe what is the solution to the phenomenon. Provide a screenshot of the largest tree in your report. This part accounts for 7.5% of the case study's mark.

8) Run the first decision tree node by using the average squared error as the performance criterion (Properties panel -- > Subtree Tab -- > Assessment Measure = Average Squared Error). How many terminal leaves does the optimal tree have? Provide a screenshot of the optimal tree. Provide a screenshot of the subtree assessment plot when average square error is selected as the performance criterion and comment on it (in a couple of sentences). This part accounts for 5% of the case study's mark.

9) Beware that the decision tree and the decision tree model are two different concepts. In the previous part you provided a screenshot of the decision tree. In this part provide a description of the decision tree **model**. This part accounts for 5% of the case study's mark.

10) Assume that the cut - off point for separating goods from bads is 90% or 0.9 (if an applicant has a probability of being a good applicant >= 90% s/he is classified as good. Write a paragraph to interpret the decision tree as you would explain it to the management team of the bank i.e. to non - technical people. What are the most important variables that separate good applicants from bad applicants? This part accounts for 7.5% of the case study's mark.

From the model tab, add a regression node to the diagram workspace and connect it to the data partition node. Accept the default settings and run the regression node. Notice that, because the target is binary, the regression node estimates a logistic regression model.

From the model tab, add a neural network node to the diagram workspace and connect it to the data partition node. Accept the default settings and run the neural network node.

From the assess tab, add a model comparison node and connect it to the decision tree node, the neural network node and the regression node.

The management team of the credit risk department has come up with the following profit matrix to be used for the evaluation of the models. The numbers represent monetary units e.g. dollars, euro, pounds etc.

| | | **Prediction** | |
|---|---|---|---|
| | | **Good -- > Accept** | **Bad -- > Reject** |
| **Actual** | **Good** | 2000 | -2000 |
| | **Bad** | -12000 | 0 |

11) Using any assumptions you like, give an interpretation of the profit matrix presented below and input the profits matrix into your process flow. This part accounts for 7.5% of the case study's mark.

12) Based on the above profit matrix what minimum probability (cut - off point) should an applicant have so as to be considered a good applicant and hence to be granted the loan? Provide the mathematical calculations. This part accounts for 7.5% of the case study's mark.

Input the profit matrix into the software to incorporate the decision analysis you did into your model by maximizing the expected profit (Select the "Applicants" data source and in the properties panel press Decisions in the Columns tab). Name Decision 1 as Accept and Decision 2 as Reject in the Decisions tab. Don't forget to select YES to the question "Do you want to use the decisions?" in the Decisions tab. Run the model comparison node.

13) Go to the results window of the model comparison node and focus on the score rankings overlay plots. Check the cumulative % response chart for the validation data set. Explain what this graph shows by using the 20% and 100% points in the x axis (the 20% and 100% most highly ranked applicants). This part accounts for 7.5% of the case study's mark.

14) Check the % response chart for the validation data set. How is this graph constructed and what do the values of the x axis represent? Explain what this graph shows by using the 50% point in the x axis. This part accounts for 7.5% of the case study's mark.

15) Check the cumulative lift chart for the validation data set. Explain what his graph shows by using the 20% point in the x axis. This part accounts for 7.5% of the case study's mark.

16) Check the cumulative % captured response graph for he validation data set. Explain what this graph shows by using the 40% point in the x axis. This part accounts for 7.5% of the case study's mark.

17) Check the cumulative total expected profit chart for the validation data set. To which customers would you grant a loan in the validation data set and what model would you choose as the optimal one for scoring new applicants? This part accounts for 7.5% of the case study's mark.

By now you must have selected the optimal model, so it is time to put it into production and score the data set named "Score" that contains the new applicants. After you insert the necessary nodes to do that **and run them** provide a screenshot with the completed process flow. Don't forget to set the role of the score data set to score. You should also notice that because this data set needs to be scored it does not contain a target variable.

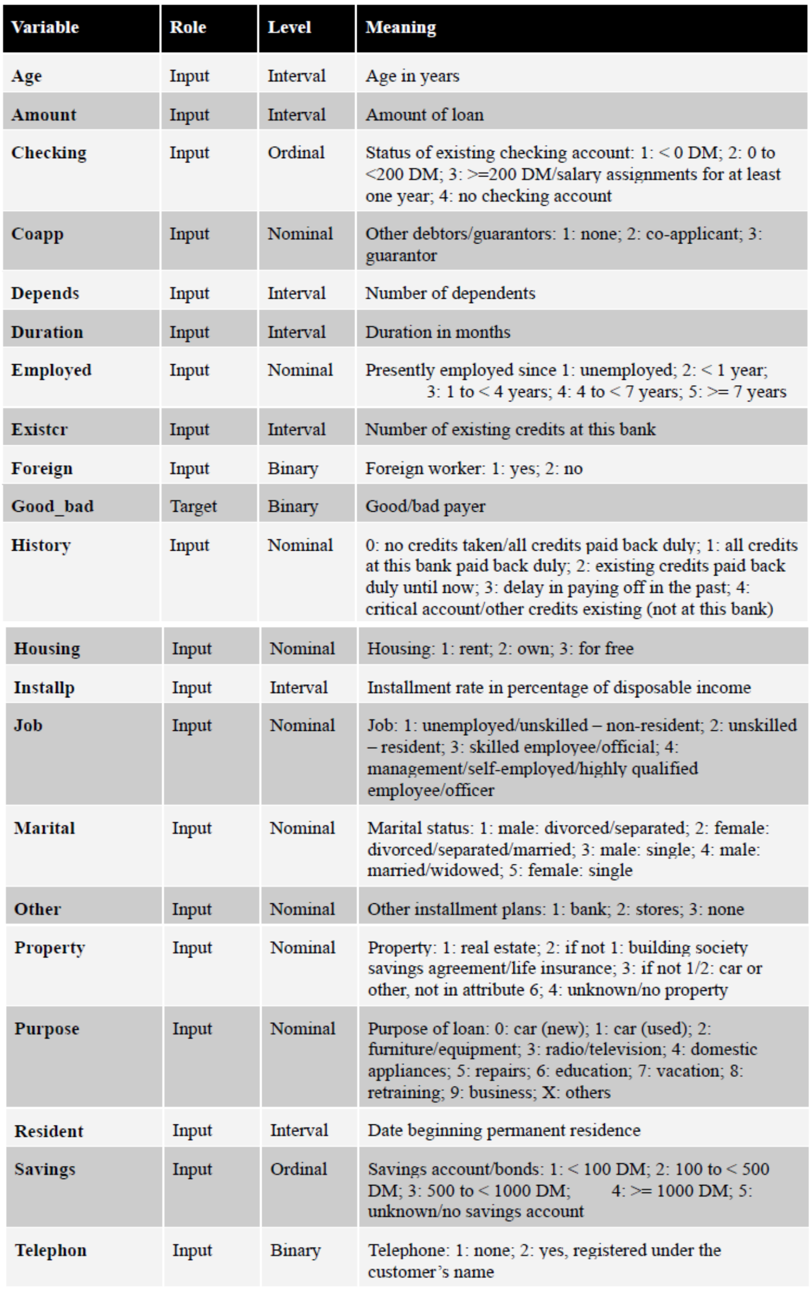
In order to answer the final three questions do the following: Select the Score node. Go to the properties panel. At the General tab select the exported data property. Select the score data set and press Browse. Right click on the table and select Export to Excel.

18) How many applicants are there in the "Score" data set? How many of the applicants were predicted as bad and how many as good? Provide a relevant bar chart. This part accounts for 2.5% of the case study's mark.

19) What was the biggest probability of being a good client assigned to an applicant? What was the smallest one? What is the expected profit to be realized by the bank if it applies the model scoring results? This part accounts for 2.5% of the case study's mark.

20) What are the custid's of the 5% most highly ranked clients (1st bucket) with respect to the probability of repaying the loan to be granted? What are the custid's of the 5% lowest ranked clients (20th bucket) with respect to the probability of repaying the loan to be granted? This part accounts for 2.5% of the case study's mark.

## Data Dictionary for the Applicants Data Set



| Variable | Role | Level | Meaning |
|---|---|---|---|
| Age | Input | Interval | Age in years |
| Amount | Input | Interval | Amount of loan |
| Checking | Input | Ordinal | Status of existing checking account: 1: < 0 DM; 2: 0 to <200 DM; 3: >=200 DM/salary assignments for at least one year; 4: no checking account |
| Coapp | Input | Nominal | Other debtors/guarantors: 1: none; 2: co-applicant; 3: guarantor |
| Depends | Input | Interval | Number of dependents |
| Duration | Input | Interval | Duration in months |
| Employed | Input | Nominal | Presently employed since 1: unemployed; 2: < 1 year; 3: 1 to < 4 years; 4: 4 to < 7 years; 5: >= 7 years |
| Existcr | Input | Interval | Number of existing credits at this bank |
| Foreign | Input | Binary | Foreign worker: 1: yes; 2: no |
| Good_bad | Target | Binary | Good/bad payer |
| History | Input | Nominal | 0: no credits taken/all credits paid back duly; 1: all credits at this bank paid back duly; 2: existing credits paid back duly until now; 3: delay in paying off in the past; 4: critical account/other credits existing (not at this bank) |
| Housing | Input | Nominal | Housing: 1: rent; 2: own; 3: for free |
| Installp | Input | Interval | Installment rate in percentage of disposable income |
| Job | Input | Nominal | Job: 1: unemployed/unskilled – non-resident; 2: unskilled – resident; 3: skilled employee/official; 4: management/self-employed/highly qualified employee/officer |
| Marital | Input | Nominal | Marital status: 1: male: divorced/separated; 2: female: divorced/separated/married; 3: male: single; 4: male: married/widowed; 5: female: single |
| Other | Input | Nominal | Other installment plans: 1: bank; 2: stores; 3: none |
| Property | Input | Nominal | Property: 1: real estate; 2: if not 1: building society savings agreement/life insurance; 3: if not 1/2: car or other, not in attribute 6; 4: unknown/no property |
| Purpose | Input | Nominal | Purpose of loan: 0: car (new); 1: car (used); 2: furniture/equipment; 3: radio/television; 4: domestic appliances; 5: repairs; 6: education; 7: vacation; 8: retraining; 9: business; X: others |
| Resident | Input | Interval | Date beginning permanent residence |
| Savings | Input | Ordinal | Savings account/bonds: 1: < 100 DM; 2: 100 to < 500 DM; 3: 500 to < 1000 DM; 4: >= 1000 DM; 5: unknown/no savings account |
| Telephon | Input | Binary | Telephone: 1: none; 2: yes, registered under the customer's name |