

Business Analytics Practicum I

Andreas Zaras

Group Assignment

Deadline: 02/06/2024

This group 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 churn prediction. 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 Visual Data Mining and Machine Learning on SAS Viya to explore and analyze the given data. The report should be sent in word and pdf format and their titles should be "LastName_FirstName.docx" and "LastName_FirstName.pdf" respectively.

You should hand in the report by the 2nd of June 2024 through the turnitin facilities of moodle. 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 SAS Viya for Learners software.*** Since this is a group assignment the credentials of one member of the group should be provided. 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 “On_Line_Book_Store”.

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. Use as minimum support level the 0.05 and as minimum confidence level the 0.1. Also set the maximum number of items in a rule equal to 3 (three) (in the interface this option is referred erroneously as minimum number of items in a rule). Save the rules table in the CASUSER library with the title MBA_Results. In the main body of the report you should answer the following questions:

- 1) Write the Executive Summary. This part accounts for 20% of the case study’s mark.
- 2) What are the sales (in units) of each book? Provide a relevant chart (bar chart) using the SAS Visual Analytics software. Enrich the chart so as to show data labels, chart title, titles in both axis. This question accounts for 20% of the case study’s mark.
- 3) 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? This part accounts for 30% of the case study's mark.

4) 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 30% 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

	Cust219	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 sports_on_line_rfm.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.

					Customer_ID	R	F	M
995 Customers					Cust001	4	5	485
					Cust002	14	4	350
					Cust003	13	2	233
				
					Cust994	24	1	185
					Cust995	6	2	187

The company also has in its repositories the following data about its customers:

					Customer_ID	Gender	Age Range	Most Frequent Payment Method
995 Customers					Cust001	M	18-25	Pay Pal
					Cust002	F	18-25	Debit Card
					Cust003	F	25-35	Debit Card
				
					Cust994	M	25-35	Credit Card
					Cust995	M	18-25	Cash

The IT department in cooperation with the Business Analytics department have merged the two previously mention data sets and have produced the SAS data set named Sports_On_Line_RFM-CD.sas7bdat that is presented below.

995
Customers

Customer_ID	R	F	M	Gender	Age Range	Most Frequent Payment Method
Cust001	4	5	485	M	18-25	Pay Pal
Cust002	14	4	350	F	18-25	Debit Card
Cust003	13	2	233	F	25-35	Debit Card
.....
Cust994	24	1	185	M	25-35	Credit Card
Cust995	6	2	187	M	18-25	Cash

You are hired as the Marketing Analytics consultant to perform the RFM segmentation with the software SAS Viya. Do the clustering of the customers and the profiling of the segments created and provide answers to the below questions. **Be as detailed as possible and justify your answers with the relevant output from the software.** Do not include instructions and screenshots on how you used the software to derive the output (e.g. We pressed this button on so as to import the data set etc). You will be graded based on your business propositions, methodology, and interpretation of graphs and tables derived from the software.

[Q1]: Write an Executive Summary to be addressed to the management team of the 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 results. Beware that an executive summary should be short (not more than half a page), clear and should not contain any technicalities. This question counts for 20% of the case study mark.

[Q2]: Describe whether any variable transformations, removal of outliers or/ and replacement of missing values was needed before running the clustering analysis. This question counts for 10% of the case study mark.

[Q3]: Name the segments created (e.g. churners, high value customers, weak customers, first time customers etc) and justify how you selected the names based on the R, F, M variables. This question counts for 20% of the case study mark.

[Q4]: Do a further profiling of the segments created by using the three extra variables i.e. Age, Most Frequent Payment Method and Age Range by creating an interactive dashboard. This question counts for 10% of the case study mark.

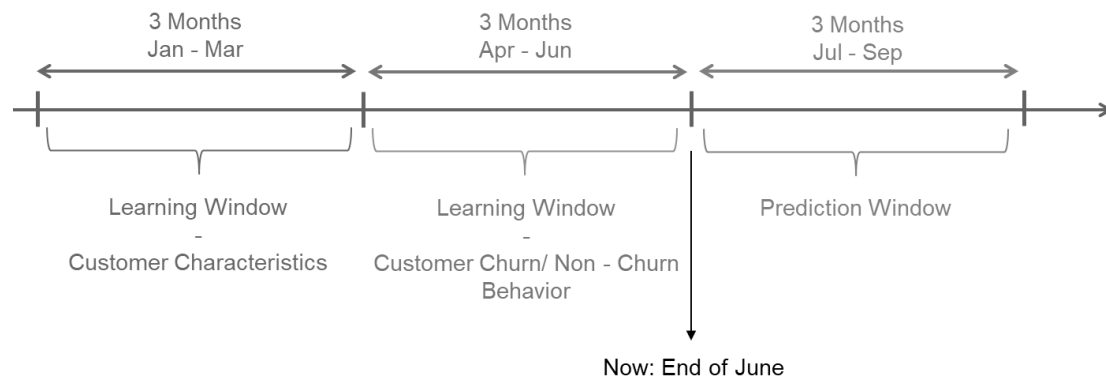
[Q5]: Describe what marketing actions are appropriate for each segment (e.g. customer reactivation program, contact customers for feedback, cross sell activities, special promotions etc) and why. This question counts for 40% of the case study mark

Case Study 3 (60%)

This case study refers to a fictitious mobile phone company - XYZ - that is operating for the past three years. The mobile telecommunications market in the country is becoming more and more competitive because of deregulation and slowly the industry will reach a maturity stage. The increasing competition and the maturity of the market makes the company focus on its existing customers and try to retain as many of them as it can so as to maintain its market share, since acquiring a new customer is more expensive than retaining an existing one. In order for this to be achieved, the management of the company has decided to invest in analytical CRM technology and more specifically to develop a machine learning based churn prediction process. If the above system predicts that a customer is likely to churn (or else attrite), the marketing department can prevent him/ her from leaving by offering relevant incentives.

The customer retention department of the company, in cooperation with the IT department, has collected the data set named “telco_data_jan_jun_partition.sas7bdat”, which contains data about customers from the period 1st of January 2017 to 30th of June 2017. The data set contains customer’s characteristics (input variables) such as the number of calls the customer has done during the day, the number of time that s/he called the customer service, the number of minutes s/he used the service during the night etc. You can find the relevant data dictionary of this data set at the end of this document. The data set consists also of a target variable that is coded as 1/ 0 and that it indicates whether the customer has left the company during the period April – June 2017 or not (1=churner, 0=non-churner).

You are hired as a machine learning engineer to aid the customer retention department develop a statistical model that will predict whether a customer will prove to be a churner or a non-churner. The model should be built from the above mentioned data set derived from the period January - June 2017 according to the image 1 below. After the model is developed by using the historical data, the input data set will be rolled ahead three months (so the data will be taken from the period 1st of April to the 30th of September 2017) and will be scored for the three month period July – September 2017 i.e. the prediction will be which customers are likely to churn (attrite) during this three month period. The customers that are more likely to churn according to the model, should be motivated to stay by giving them incentives. The characteristics of the customers to be scored for the July – September 2017 are stored in the “telco_data_apr_sep.sas7bdat” data set and are related to the period April – June 2017.



Please follow the following steps and answer the relevant questions:

Open SAS VDMML on SAS Viya (You will also need to open SAS Visual Analytics on SAS Viya to explore the data).

Create a new project.

Create a new data source ("telco_data_jan_jun_partition") by consulting the relevant data dictionary at the end of this document.

- 1) Write the Executive Summary. This part accounts for 10% of the case study's mark. The management team of the marketing department has come up with the following profit matrix to be used for the evaluation of the models to be created. The numbers represent monetary units e.g. dollars, euro, pounds etc.

		Prediction	
		Churner -- > Motivate Stay	Non-Churner -- > Do Nothing
Actual	Churner	1000	-1500
	Non-Churner	-500	0

- 2) Using any assumptions you like, give an interpretation of the profit matrix presented above. This part accounts for 7.5% of the case study's mark.
- 3) Based on the above profit matrix what minimum probability (cut - off point) should a customer have so as to be considered a churner and hence to be considered for a contact to prevent him/ her from attrition (churn)? Provide the mathematical calculations. This part accounts for 5.5% of the case study's mark.

- 4) Use the project settings for this question (the gear on the upper right corner of the screen). Partition the historical data set to training and validation using the 70% - 30% rule of thumb? Why this process must be done? The sampling in the data partition is stratified. What does this mean? Also use the Misclassification Rate (Event) as the performance criterion and input the previously calculated cut-off point into the software. This question accounts for 2.5% of the case study's mark.
- 5) Are there any missing values in the variables of the data set? Provide a screenshot of the SAS Visual Analytics software to prove this. What is the proportion of churners and non-churners in the data set? Provide a screenshot from the SAS Visual Analytics software to prove it (pie chart). This part accounts for 2.5% of the case study's mark.
- 6) The proportion of churners to non – churners in the historical data set is 14.14% - 85.86%. What would you do if this proportion was 3% - 97%? This part accounts for 5% of the case study's mark.
- 7) Provide a graph (pie chart) using SAS Visual Analytics on SAS Viya that shows the proportion of churners and non- churners for those customers that have contacted more than 6 (six) times the customer service. What do you observe? This question accounts for 2.5% of the case study's mark.
- 8) Use SAS Visual Analytics on SAS Viya to show the average amount of minutes that the customers speak on their phone during the day (Day_Mins) for a) churners and b) non-churners. What does this mean with respect to the target variable? This question accounts for 2.5% of the case study's mark.

Add a decision tree node to the workspace and connect it to the data source node. Use the following options: CHAID, No Bonferroni, Reduced Error

- 9) What is the variable used for the first split? Explain briefly why this variable is selected (hint: 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.

Add a second decision tree node to the workspace. Name the new decision as Maximal tree. In the properties window of the tree change the method to largest (maximal) i.e. Pruning options --> Selection Method --> Largest. Run the tree node.

10) 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.

11) Run the first decision tree node. 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 Misclassification Rate is selected as the performance criterion and comment on it (in a couple of sentences). This part accounts for 7,5% of the case study's mark.

12) 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 7,5% of the case study's mark. (Please interpret the model by using only 5 of the terminal leaves).

13) Write a paragraph to interpret the decision tree as you would explain it to the management team of the telco organization i.e. to non - technical people. What are the most important variables that separate churners from non – churners? This part accounts for 7.5% of the case study's mark. (Please interpret the tree by using only 5 of the terminal leaves).

Add a logistic regression node to the pipeline. Accept the default settings and run the regression node.

Add a neural network node to the pipeline. Accept the default settings and run the neural network node.

14) 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 customers to be churners). This part accounts for 5.5% of the case study's mark.

15) 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 25% point in the x axis. This part accounts for 5.5% of the case study's mark.

16) Check the cumulative lift chart for the validation data set. Explain what this graph shows by using the 20% point in the x axis. This part accounts for 5.5% of the case study's mark.

17) Check the cumulative % captured response graph for the validation data set. Explain what this graph shows by using the 40% point in the x axis. This part accounts for 5.5% 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 "telco_data_apr_sep" i.e. to predict which customers are more likely to be churners in the next scoring period. Insert the necessary node (Score Data) to do that, run it and provide a screenshot with the completed process flow (In the score data node attach the telco_data_apr_sep data set and for the output library select the CASUSER). You should also notice that because this data set needs to be scored it does not contain a target variable. Name the new scored table as Scored_Telco.

In order to answer the final three questions, do the following: Select the Score data node and go to the results. Select the Output data tab and View Output. Press the Explore and Visualize button, select the CASUSER library and name the table as Scored_Telco_Visualize. You will be transferred to the SAS Visual Analytics environment.

18) How many customers are there in the "telco_data_apr_sep" data set? How many of them are predicted as churners and how many as non-churners? Provide a relevant bar chart using SAS Visual Analytics. This part accounts for 2.5% of the case study's mark.

19) What is the biggest probability of being a churning assigned to a customer? What is the smallest one? This part accounts for 2.5% of the case study's mark.

20) Select one customer from the score data set (whoever you like). Based on which column of the score data set and why the software assigns 1 / 0 to the customer you selected i.e. predicts that s/he will be churning / non - churning? This part accounts for 2.5% of the case study's mark.

Data Dictionary for “telco data jan jun partition” Data Set

Data Dictionary for CHURN Data Set

Variable	Role	Level	Meaning
Account_Length	Input	Interval	Indicating how long the account is active
Area_code2	Rejected	Nominal	Used to identify specific areas
Churn	Target	Binary	Indicates whether a customer churned or not
CustServ_Calls	Input	Interval	Indicates how many calls the customer made to customer service, during the period in examination
Day_Calls	Input	Interval	Indicates the number of calls the customer made during the day
Day_Charge	Input	Interval	Indicates how much the customer paid for the day calls
Day_Mins	Input	Interval	Indicates the number of minutes the customer used the service during the day
Eve_Calls	Input	Interval	Indicates the number of calls the customer made during the evening
Eve_Charge	Input	Interval	Indicates how much the customer paid for the evening calls
Eve_Mins	Input	Interval	Indicates the number of minutes the customer used the service during the evening
Intl_Calls	Input	Interval	Indicates the number of international calls the customer made
Intl_Charge	Input	Interval	Indicates how much the customer paid for the international calls
Intl_Mins	Input	Interval	Indicates the number of minutes the customer used in international calls
Intl_Plan	Input	Binary	Indicates whether the customer has an international plan
Night_Calls	Input	Interval	Indicates the number of calls the customer made during the night
Night_Charge	Input	Interval	Indicates how much the customer paid for the night calls
Night_Mins	Input	Interval	Indicates the number of minutes the customer used the service during the night
Phone	Rejected	Nominal	Phone number, discrete, having unique values; can be used as ID
State	Rejected	Nominal	Used to identify the 50 states in the USA and the District of Columbia
Vmail_Message	Input	Interval	Number of voice mail messages sent
Vmail_Plan	Input	Binary	Indicates whether the customer has a voice mail plan
Partition Indicator	Partition	Binary	This is the partition indicator to reassure the same training and validation to data sets any time we run the project