UNSW Business School

COMM3501 Quantitative Business Analytics

A4 Individual Assignment (40%)

Due date: Monday 5th August 2024, 12:00 PM (noon) week 11

1. Assignment overview

In this assessment, you will analyse a dataset with an emphasis on practical business analytics and develop authentic outputs. The task aims to enhance your problem-solving skills in real-world scenarios. It is also intended to develop your skills in research, critical thinking and problem solving, your data analysis and programming skills, and your ability to communicate your ideas and solutions concisely and coherently.

2. Assignment scenario

You are an analyst at a data analytics consulting firm. Your manager has tasked you with providing a report to an American client. The client is a major U.S. wireless telecommunications company which provides cellular telephone service. They require assistance in developing a statistical model to predict customer churn, establish a target customer profile for implementing a proactive churn-management program, and rolling the solution out to their customer-facing call centres.

These days, the telecommunications industry faces fierce competition in satisfying its customers. Churn is a marketing term, referring to a current customer deciding to take their business elsewhere—in the current context, switching from one mobile service provider to another. As with many other sectors, churn is an important issue for the wireless telecommunications industry. For this client, the role of the desired churn model is not only to accurately predict customer churn, but also to understand customer behaviours.

3. Assignment details

3.1. Task details

Your main tasks will involve: data manipulation and cleaning; statistical modelling; writing a technical report. Your client also wants a non-technical description of the characteristics of customers that churned, to assist in the development of a risk-management strategy, i.e., a proactive churn-management program.

In your report, your manager wants you to include: some details on your data manipulation, cleaning, and descriptive analysis; a brief summary and comparison of the models you fitted; a detailed description of your selected model/s and interpretation of the results; your main findings, recommendations and conclusions.

The client is familiar with machine learning. All your modelling results should be included, mostly in an appendix to the report.

In addition, among the 10,000 customers in the `eval_data.csv` evaluation dataset, you must identify 3000 customers which you believe are most likely to churn.

See the submission details section and marking criteria section for more information.

3.2. Data Description

The data provides details of 30,000 customers in the training dataset, and 10,000 customers in the evaluation dataset:

1. `training_data.csv`
2. `eval_data.csv`

The datasets can be downloaded from the Moodle website in the *A4 Individual Project – A4 Datasets* section.

For each of the observations in the training dataset, there is information on 44 attributes describing the customer care service details, customer demography and personal details, etc. These are described below.

Similar, but not identical, datasets are provided here. You may also wish to have a look at the following analysis based on the Kaggle datasets to give you an idea: Churn Prediction (weblink). This analysis is just a brief example and is not based on your datasets. Different and more variables may be of interest for your analysis. Extra readings are given in the Resources section.

3.2.1. `training_data.csv` (Training dataset)

This dataset provides insights about the customers and whether they are churned customers.

Variable Name	**Description**
CustomerID	A unique ID assigned to each customer/subscriber
Churn	Is churned? (categorical: “no”,“yes”)
MonthlyRevenue	Mean monthly revenue for the company
MonthlyMinutes	Mean monthly minutes of use
TotalRecurringCharge	Mean total recurring charges (recurring billing)
OverageMinutes	Mean overage minutes of use
RoamingCalls	Mean number of roaming calls
DroppedCalls	Mean number of dropped voice calls

BlockedCalls	Mean number of blocked voice calls
UnansweredCalls	Mean number of unanswered voice calls
CustomerCareCalls	Mean number of customer care calls
ThreewayCalls	Mean number of three-way calls
OutboundCalls	Mean number of outbound voice calls
InboundCalls	Mean number of inbound voice calls
DroppedBlockedCalls	Mean number of dropped or blocked calls
CallForwardingCalls	Mean number of call forwarding calls
CallWaitingCalls	Mean number of call waiting calls
MonthsInService	Months in Service
ActiveSubs	Number of Active Subscriptions
ServiceArea	Communications Service Area
Handsets	Number of Handsets Issued
CurrentEquipmentDays	Number of days of the current equipment
AgeHH1	Age of first Household member
AgeHH2	Age of second Household member
ChildrenInHH	Presence of children in Household (yes or no)
HandsetRefurbished	Handset is refurbished (yes or no)
HandsetWebCapable	Handset is web capable (yes or no)
TruckOwner	Subscriber owns a truck (yes or no)
RVOwner	Subscriber owns a recreational vehicle (yes or no)
BuysViaMailOrder	Subscriber Buys via mail order (yes or no)
RespondsToMailOffers	Subscriber responds to mail offers (yes or no)
OptOutMailings	Subscriber opted out mailings option (yes or no)
OwnsComputer	Subscriber owns a computer (yes or no)
HasCreditCard	Subscriber has a credit card (yes or no)
RetentionCalls	Number of calls previously made to retention team
RetentionOffersAccepted	Number of previous retention offers accepted
ReferralsMadeBySubscriber	Number of referrals made by subscriber
IncomeGroup	Income group
OwnsMotorcycle	Subscriber owns a motorcycle (yes or no)
MadeCallToRetentionTeam	Customer has made call to retention team (yes or no)
CreditRating	Credit rating category
PrizmCode	Living area
Occupation	Occupation category
MaritalStatus	Married (yes or no or unknown)

3.2.2. `eval_data.csv` (Evaluation dataset)

The evaluation dataset comprises 10,000 current customers. From these 10,000 customers, select 3000 which you believe are most likely to churn. This evaluation dataset has the same format as the training dataset but doesn't include the column Churn. The true values for the column Churn will be released after the due date of the assignment.

3.3. Software

You may choose which software package or program to use, e.g., R or python. The code enabling you to perform most of the computing can be found in the course learning activities.

3.4. Resources

- Extra information on the original dataset and on the context can be found here: link 1 and link 2
- Data manipulation with R with the 'dplyr' package (weblink)
- Tidy data in R (weblink)
- Exploratory Data Analysis with R (weblink)
- Data visualisation in R with ggplot2 for fancy plots (weblink)
- He and Garcia (2009), for strategies for dealing with imbalanced data in classification problems
- Yadav and Roychoudhury (2018), for some strategies to deal with missing attribute values in R (available on Moodle)
- If you are interested in using R Markdown, here is a guide for creating PDF documents (weblink)
- For any code-related questions, google.com or stackoverflow.com are pretty helpful!

3.5. Marking criteria

You will be assessed against the following criteria:

1. Data manipulation, cleaning, and descriptive analysis
2. Modelling
3. Recommendations and discussion
4. Report writing
5. Predictive accuracy

The mark allocation and details for each marking criteria are given below and in the rubric. The materials you submit should be your own. Familiarise yourself with the UNSW policies for plagiarism before submitting.

3.5.1. Criteria 1-3

There are potentially multiple valid approaches to this task, so you must choose an approach that is both justifiable and justified.

You may also wish to engage in extra research beyond the course content. Please feel free to do so. Although the marks for each component of the assignment are capped, innovations are encouraged.

Any assumptions must be clearly identified and justified, if used. Sufficient details, e.g., calculations and results, must be provided. Include an appendix to the report for non-essential but useful results; however, the appendix will not be directly assessed. Ensure that the body of your report is self-contained and addresses all marking criteria.

3.5.2. Criteria 4

Communication of quantitative results in a concise and easy-to-understand manner is a skill that is vital in practice. As such, marks will be given for report writing. To maximize your marks for this component, you may wish to consider issues such as: table size/readability, figure axes/formatting, text readability, grammar/spelling, page layout, and referencing of external sources.
Include a brief introduction section in your report.

A maximum page limit of 8 pages is applicable to the main body of the report. This limit includes tables and graphs, but excludes the cover page, table of contents, references, and any appendices. There is no limit to the length of the appendix. Exceeding the page limit will attract a proportional penalty to the overall assignment mark. Your report must be a self-contained document (i.e., not multiple files), with all pages in portrait format.

Consider how the overall look, feel and readability of your document is affected by choices like margin size, line and paragraph spacing, typeface/font, and text size. If in doubt, don't stray too far from the defaults in your word processor / typesetting program, or use something like the following settings: margins of 2.54cm for each edge, 1.15 line spacing, Calibri size 11 text.

3.5.3. Criteria 5

Provide a comma-separated values (CSV) file following the format in the sample file provided on Moodle (`selected_customers_example_for_submission.csv`), predicting the 3000 (out of 10,000) customers in the evaluation dataset which you believe are the most likely to churn. See the submission section for details.

The accuracy of your predictions on the evaluation data will have a (minor) impact on your mark. The marks you get for the accuracy criterion will be given by the following formula.

$$\text{Marks} = \begin{cases} \dfrac{5}{Y} \times \text{No. churned customers identified,} & \text{if No. churned customers identified} < Y \\ 5 + \dfrac{5}{C - Y}(\text{No. churned customers identified} - Y), & \text{if No. churned customers identified} \geq Y, \end{cases}$$

where we will take C as the maximum number of churned customers correctly identified by a student in the class, and Y as the number of churned customers you would correctly identify on average if your prediction algorithm were to just return a pure random sample of the 10,000 customers in the evaluation dataset. Therefore, if your prediction accuracy is below that expected by random sampling, your mark for this component will scale from 0 to 5 based on how many predictions were correct. If your prediction accuracy is above that expected by random sampling, then your mark is scaled from 5 to 10 based on the accuracy.

4. Assignment submissions

Your final submission should include:

1) A technical report in .docx or .pdf format
2) Your sample of predicted churn customers in a CSV file named `selected_customers_`**`yourStudentzID`**`.csv` *
3) Reproducible codes with brief instructions on how to use them, e.g., R script/s with comments (this item will not be assessed).

Upload your final submission using the submission links on Moodle. Check your report displays properly on-screen once it is submitted.

* If your zID were z1234567, you would call the file `selected_customers_z1234567.csv`

5. References

He, Haibo, and Edwardo A. Garcia. 2009. “Learning from imbalanced data.” IEEE Transactions on Knowledge and Data Engineering 21 (9): 1263–84. https://doi.org/10.1109/TKDE.2008.239.

Yadav, Madan Lal, and Basav Roychoudhury. 2018. “Handling missing values: A study of popular imputation packages in R.” Knowledge-Based Systems 160 (April): 104–18. https://doi.org/10.1016/j.knosys. 2018.06.012.