STA30004: Data Mining Assignment 2021 Part 1

Total marks 50 (15% over all)

Due date: 11th April Midnight

This assignment relates to motor insurance claim data. In every tutorial class we will spend some time working on this assignment and at the end of week 5 you will be expected to submit your part 1 assignment. There are two data sets used for this assignment.

Description of the first dataset

There is data for more than 73000 policies in the data file **motor20pct.csv** that are associated with claims in a particular year. The variables for each of the policies in this data set are explained below:-

CAR_AGE measures the age of the insured car in years
DRIVERS measures the number of people who are specified as designated drivers
EXPOSURE measures the fraction of the year for which the policy was active
MILEAGE measures the expected mileage travelled in a single year
PRIMAGE gives the age of the primary driver in years
TOTAL gives the total amount claimed on the policy in the year
EXCESS = 0, 75 or 100 indicating the excess claim amount associated with each policy. The insurance company will not pay out claims below this excess amount.
USAGE specifies how the car is used (S=only social, SB=strictly business, SC=social and business, ST=social and taxi)
CLAIM=1 if there was at least one claim during the year, 0 otherwise.

Create a variable called CatClaim set equal to “Yes” when CLAIM=1 and “No” for CLAIM=0.

You will be working with a random sample of 10000 of these policies. Instructions for generating this sample are provided in Tutorial 1.

Question 1 (5 marks)
Suggest a list of questions that could be answered using this data in your assignment. Consider the CLAIM variable as a possible TARGET variable in your models and the total claim amount as a possible RISK variable, reflecting the risk associated with any claim.

Question 2 (20 marks)
Instructions for this question are provided in Tutorials 2 and 3.

Summarise your data using descriptive statistics and graphs. Some suggestions are provided below. All tables and graphs must be numbered/labelled and discussed/interpreted.

i) Produce summary statistics for your data
i) Boxplots for numeric input variables for claim categories
ii) Pairs plot for all numeric input variables
iii) Correlation Plot for all numeric input variables
iv) Hierarchical Correlation Plot for all numeric input variables
v) Bar charts for the categorical variables Usage and the claim variable
vi) Other exciting plots

Question 3: (15 marks)

Tutorial 5 provides guidelines for this question

Partition your data with 70% for training, 15% for validation and 15% for testing. Number and label all your tables and graphs and discuss/interpret the results.

a) Produce a Tree to predict CatClaim. Then Draw your tree and ask for the Rules.

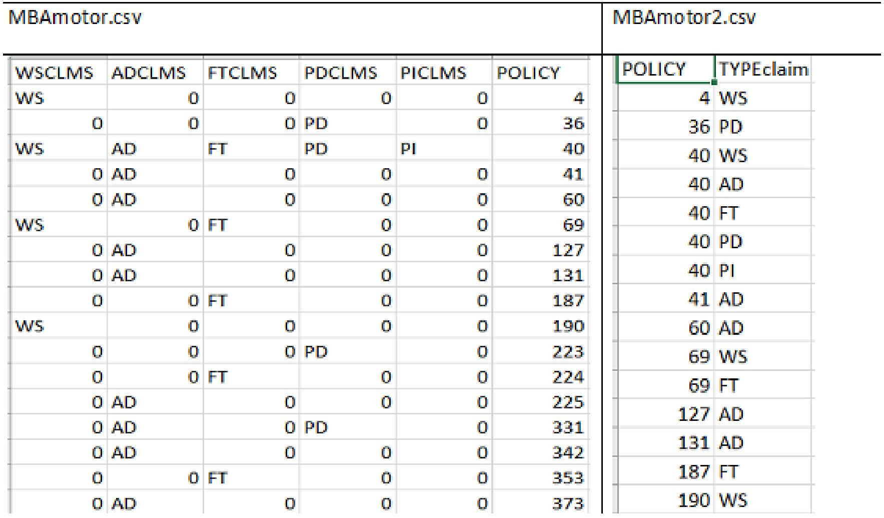
b) How do the results change when you re-run your tree assuming a loss matrix with losses half as big for a false positive (CatClaim=”Yes”) than a false negative (CatClaim=”No”).

c) How do your results change when you re-run your tree assuming priors of 20% for CatClaim = Yes and 80% for CatClaim = No. These were the percentages for the original data file “motor20pct”.

Description for the second data set

For this question consider the data set MBAmotor2.csv which was created using MBAmotor.csv. This file tells us what type of claim was posted by each of the policy holders during the year. There is at least one type of claim for all these policies.

WSCLMS=WS for windshield claims
ADCLMS=AD for accidental damage
FTCLMS =FT for fire or theft
PDCLMS = PD for personal damage claims
PICLMS = PI for personal injury claims



MBAmotor.csv

WSCLMS	ADCLMS	FTCLMS	PDCLMS	PICLMS	POLICY
WS	0	0	0	0	4
0	0	0	PD	0	36
WS	AD	FT	PD	PI	40
0	AD	0	0	0	41
0	AD	0	0	0	60
WS	0	FT	0	0	69
0	AD	0	0	0	127
0	AD	0	0	0	131
0	0	FT	0	0	187
WS	0	0	0	0	190
0	0	0	PD	0	223
0	0	FT	0	0	224
0	AD	0	0	0	225
0	AD	0	PD	0	331
0	AD	0	0	0	342
0	0	FT	0	0	353
0	AD	0	0	0	373

MBAmotor2.csv

POLICY	TYPEclaim
4	WS
36	PD
40	WS
40	AD
40	FT
40	PD
40	PI
41	AD
60	AD
69	WS
69	FT
127	AD
131	AD
187	FT
190	WS

Question 4 (10 marks)

Tutorial 4 provides guidelines for this question.

Conduct an association analysis using these data and discuss your results. In particular, you should define the terms support and confidence and determine the strongest and the most common associations between the above types of claim. Number and label all tables and figures and discuss/interpret the results.