PROG8430 – Data Analysis, Modeling and Algorithms

Assignment 5

Classification

DUE BEFORE 10 pm, APRIL 13, 2023

1. Submission Guidelines

All assignments must be submitted via the econestoga course website before the due date into the assignment folder.
You may make multiple submissions, but only the most current submission will be graded.

SUBMISSIONS

In the Assignment 4 Folder submit:

1. Your *.Rmd file. This file must have all output already run *and* your comments and answers to the questions. If you do not include your output, I will *not* be running your code to generate it. I may, however run the code to verify the results.
2. The *.pdf or *.doc file that is produced from your code.

DO NOT PUT THE DOCUMENTS IN TO A ZIP FILE!

PLEASE NOTE: The marks on the assignment are generally awarded 50% for the actual R code and calculations and 50% for interpretation and demonstration that you understand what you have done.

EXAMPLES: The example output provided is simply to demonstrate what a typical submission might look like. You can use it as a basis, but your submission must be in your own words. Submissions that simply "cut and paste" my example commentary will be marked 0.

All variables in your code must abide by the naming convention [variable_name]_[intials]. For example, my variable for State would be State_DM.

THIS IS AN INDIVIDUAL ASSIGNMENT. UNAUTHORIZED COLLABORATION IS AN ACADEMIC OFFENSE AS IS DIRECT 'CUTTING AND PASTING' FROM OTHER SOURCES. Please see the Conestoga College Academic Integrity Policy for details.

Remember the discussion forums on eConestoga are a great place to ask questions.

2. Grading

This assignment will be marked out of 30 and is worth 12.5% of your total grade in the course.

Assignments submitted after 10pm will be reduced 20%. Assignments received after 8:00am the morning after the due date will receive a mark of 0%.

Assignments which do not follow the submission instructions may have marks deducted.

3. Data

Each student will be using the study dataset:

STUDY DATASET:

PROG8430-Assign04-23W.txt

Appendix one contains a data dictionary for the study file.

4. Background

A major mail-order company tracks the time (in days) it takes for customers to receive their orders (each row in the dataset represents one order).

The company has a goal of making all deliveries in at least 8.5 days. Any parcel delivered in at least 8.5 days is considered 'On Time' and anything after that is considered late.

Your task will be to use a variety of classifiers to determine the factors which contribute to, and help predict, an 'On Time' delivery. As always, imagine this analysis will be generating a report that will be presented to a client.

All of the tasks have been demonstrated in class. A careful review of your notes from the lectures should give you everything you need to complete these tasks.

5. Assignment Tasks

NOTE – For each step/task, include sufficient commentary in your submission to demonstrate understanding of the steps you have taken.

6. Assignment Tasks

Nbr	Description	Marks
1	Preliminary Data Preparation 1. Rename all variables with your initials appended (just as was done in previous assignments). Remember that any variables you subsequently create need to have your initials appended. 2. Since this is the same dataset as used in Assignment 4, you do not need to do the regular descriptive analysis and outlier detection. As in	3

	Assignment 4, remember to delete the observation with PG <0 using the following code: ClssAssign <- ClssAssign[!ClssAssign$PG < 0,] NOTE – Your variable names will be different of course! 3. Create a new variable in the dataset called OT_**[Intials]** which will have a value of 1 if DL ≤ 8.5 and 0 otherwise. If you have forgotten how to do this, the code to accomplish it is included in the appendix. Remember to also delete the DL variable after this.	
2	Exploratory Analysis 1. Correlations: Create correlations (as demonstrated) and comment on what you see. Are there co-linear variables?	1
3	Model Development As demonstrated in class, create two logistic regression models. 1. A full model using all of the variables. 2. An additional model using backward selection. For each model, interpret and comment on the main measures we discussed in class: (1) Fisher iterations (2) AIC (3) Residual Deviance (4) Residual symmetry (5) z-values (6) Parameter Co-Efficients 3. As demonstrated in class, analyze the output for any significantly influential datapoints. Are there any? 4. Based on your preceding analysis, recommend which model should be selected and explain why.	 1 1 3 1 1
PART B		
In this section, all classifiers should be built using OT_**[Intials]** as the dependant variable and the remaining variables as the independent variables.		
1	Logistic Regression – stepwise 1. As above, use the step option in the glm function to fit the model (using stepwise selection). 2. Summarize the results in Confusion Matrices for both train and test sets. 3. As demonstrated in class, calculate the time (in seconds) it took to fit the model and include this in your summary.	 1 2 1
2	Naïve-Bayes Classification 1. Use all the variables in the dataset to fit a Naïve-Bayesian classification model. 2. Summarize the results in Confusion Matrices for both train and test sets. 3. As demonstrated in class, calculate the time (in seconds) it took to fit the model and include this in your summary.	 1 2 1

3	Recursive Partitioning Analysis 1. Use all the variables in the dataset to fit an recursive partitioning classification model. 2. Summarize the results in Confusion Matrices for both train and test sets. 3. As demonstrated in class, calculate the time (in seconds) it took to fit the model and include this in your summary.	1 2 1
3a	**BONUS SECTION** This section is bonus marks. There is no need to complete this, but successful completion of this section will be worth 3 bonus marks. Neural Network 1. Use all the variables in the dataset to fit a Neural Network classification model. Set the seed to 8430, the size to 3, rang=0.1 and maxit=1200. 2. Summarize the results in Confusion Matrices for both train and test sets. 3. As demonstrated in class, calculate the time (in seconds) it took to fit the model and include this in your summary.	
4	Compare All Classifiers For all questions below please provide evidence. 1. Which classifier is most accurate? 2. Which classifier seems most consistent (think train and test)? 3. Which classifier is most suitable when processing speed is most important? 4. Which classifier minimizes false positives? 5. In your opinion, classifier is best overall?	5
5	Professionalism and Clarity	2

APPENDIX ONE: STUDY FILE DATA DICTIONARY

Variable	Description
DL	Time for delivery (in days, rounded to nearest 10th)
VN	Vintage of product (i.e. how long it has been in the warehouse).
PG	How many packages of product have been ordered
CS	How many orders the customer has made in the past
ML	Distance the order needs to be delivered (in km)
DM	Indicator for if the product is manufactured in Canada (C) or elsewhere (I)
HZ	Indicator for if the product is designated as Hazardous (H) or not (N).
CR	Indicator for which Carrier delivered the item (Def Post, or Sup Del)
WT	Weight of the shipment (in decagrams)

Example Code for outcome

```
df$OT_DM <- as.factor(ifelse(df$DL_DM  <= 8.5, 1,0))
```