B3T1102 Data Modelling & Analytics

Assignment 2: Classification Problems with Discrete Outcomes

Answer Form

Instructions

1. This is an individual assignment. You are to submit your attempts individually via Canvas.
2. You are required to submit two documents to Canvas, namely, this Answer Form, which you shall name as “<your student ID>\_AF2.docx” , and an Excel file, which you shall name as “<your student ID>\_NA2.xlsx” . Errors in your file name may potentially lead to a loss of marks, so please ensure that the file name is correct.
3. In this assignment, you will be required to write all numerical answers in the provided Excel answer file. The Excel answer file contains two columns. The first column lists the question numbers to which an answer is expected in the Excel answer file. Do not edit the contents of the first column. In the second column, you are to key in your answer to the corresponding question there.
4. For any numerical answers, when keying in to the Excel answer file, please round off your answers to 3 significant figures. If they are whole numbers, leave them as such.
5. For any numerical answers, if the value of the answer is less than 10-5, you may leave your answer as 0.
6. For text-based answers, you can type your answers into this answer form. Where a text-based answer is expected of you, you will find a box. You may type you answers in the boxes provided.
7. This assignment counts towards 25% of the final grade. There is a maximum of 50 points awarded for this assignment. Any student who scores lower than 28 points for this assignment can opt for the improvement option. However, if a student gives a nil-return for this assignment, we will not grant the improvement option, as the student is assumed to have forfeited the assignment.
8. Late submissions will not be entertained, unless special approval has been obtained.

Deadline for handing in this assignment: Friday May 13, 2022, 23:59.

**Initialization**

On the Canvas assignment page, you will find a link. Go to the link. You are required to key in your 6-digit student ID, and you will be provided with two data sets, one that is for the training data set and one that is for the testing data set. Download them. You will be required to answer all of the questions in this Assignment using the two datasets.

It is highly recommended for you to store copies of the data sets in case you mess them up in your working. The link will not be active for long.

In both of the training and testing data frames, you will find 6 columns. The column ‘y’ is to be assumed as the outcome / dependent variable, which is binary. The other columns labelled from ‘x1’ to ‘x5’ are to be assumed as the predictors / independent variables.

**Question 1**

**1 point**

Find the proportion of observations of y in the **training** data set that are labelled as ‘1’. Key your answer into the Excel answer file.

**Question 2**

**1 point**

Find the proportion of observations of y in the **testing** data set that are labelled as ‘1’. Key your answer into the Excel answer file.

**Question 3**

**2 points**

Explain the significance of the answer you computed in Question 2 when assessing the out-of-sample accuracy of any model.

The difference in the proportion of y observations is due to variables being randomly selected and categorized into Training data set (Q1) Testing data set (Q2).

We use the Training data set (Q1) to build our model to make predictions. At the time the model was built, we do not consider the Testing data set or so-called *Out-of-sample* data (Q2). Assessing "Out-of-sample accuracy" means assess the percentage of correct predictions that the model makes on *Out-of-sample* data, which is the data that the model has not been trained on.

**Question 4**

**3 points**

Build a logistic regression model with the dependent variable ‘y’ and all variables ‘x1’ to ‘x5’ as the independent variables. Call this model ‘mod1’. Do not yet do any transformations on the variables. How many variables are significant under the significance level of 0.05? Key your answer into the Excel answer file.

**Question 5**

**2 points**

Run a test to verify if there is multi-collinearity in the model. In your test, you would compute one statistic for each of the predictors. Key in the **largest** values of these 5 statistics into the Excel answer file.

**Question 6**

**1 point**

Given your answers in Question 5, conclude whether there is multi-collinearity in ‘mod1’. Key in ‘Yes’ or ‘No’ into the Excel answer file.

**Question 7**

**3 points**

Build predictions of the model ‘mod1’ on the outcome variable ‘y’, **using the training data set**. Use a cut-off probability of 50%. Form a confusion matrix and report the accuracy in the Excel answer file.

**Question 8**

**1 point**

Also compute the recall and input this into the Excel answer file.

**Question 9**

**3 points**

Build predictions of the model ‘mod1’ on the outcome variable ‘y’, **using the testing data set** now. Use a cut-off probability of 50%. Form a confusion matrix and report the accuracy in the Excel answer file.

**Question 10**

**1 point**

Also compute the recall and input this into the Excel answer file.

**Question 11**

**2 points**

Examine if your answers for Questions 7 and 8 differ from Questions 9 and 10. Explain why this occurs.

The reason follows from question 3:

The confusion matrix was built based on variables and predictions from the Training dataset (Q7&Q8). At the time the matrix was built, the testing data set (Q9&Q10) was unseen. Afterward, the matrix is then used to test the Testing dataset. Thus, differences in Prediction accuracy between two datasets occur.

**Question 12**

**3 points**

Plot the residuals of ‘mod1’ against all of the 5 independent variables. One of them will result in a plot that is of a different pattern from the others. Key in which variable it is into the Excel answer sheet. Your answer should be in the format: ‘x<a number>’.

**Question 13**

**3 points**

For that variable, paste the residuals plot into the box provided below. Explain the significance of the plot.

Chart, histogram

Description automatically generated

The residuals do not evenly spread around the horizontal line (not randomly distributed but skewed right). Therefore, assumption A1. Linearity is violated. As a result, when the functional form is incorrect, both the estimated coefficients and standard errors are not reliable. We get biased predictions related to x5, and the p-values are not reliable anymore.

**Question 14**

**3 points**

Split the training data set into two, one which contains all of the observations where the variable you identified in Question 12 is **less than** 0 (called data set “df\_train\_under”) and another which contains all remaining observations (called data set “df\_train\_over”). On the first data set “df\_train\_under”, build a logistic regression model called ‘mod1a’ using all of the variables. Key in the p-value associated with the variable you identified in Question 12 into your Excel answer file.

**Question 15**

**2 points**

Form predictions of both ‘mod1’ and ‘mod1a’ on ‘df\_train\_under’. Still using a cut-off probability of 0.5, create confusion matrices for these predictions using ‘df\_train\_under’. Report the accuracy of the predictions formed by ‘mod1’ in your Excel answer file.

**Question 16**

**1 point**

Report the accuracy of the predictions formed by ‘mod1a’ in your Excel answer file.

**Question 17**

**1 point**

Comment on why you would or would not see a difference.

There is a difference in prediction accuracy because the two models were built from different datasets. The confusion matrix produced by model ‘mod1’ was based on variables and predictions from the whole training dataset. Meanwhile, the confusion matrix produced by model ‘mod1a’ was based on variables and predictions from the filtered training dataset. This filtered dataset only selected rows where x5 is below zero.

**Question 18**

**1 point**

On the other half of the training data set, “df\_train\_over”, build a logistic regression model named ‘mod1b’. Explain how you can use the models ‘mod1a’ and ‘mod1b’ to create a model for predicting the outcome of ‘y’.

As described from question 13, when plotting the residuals vs. x5, we can see that we get biased predictions related to x5 (when x5 < 0 the model is more likely to predict 0, and x5 > 0 the model is more likely to predict 1).

The act of splitting the dataset fixed this issue. Below is the figure of the two models’ residuals against x5:

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

The residuals are now randomly dispersed around the horizontal axis. Therefore, we can split the datasets into to subset (x5<0, x5>=0). Then we form predictions separately on the two subsets by two models, ‘mod1a’ and ‘mod1b’, respectively. Afterward, we can join the two resulting predictions and build the confusion matrix.

**Question 19**

**2 points**

Implement what you described in Question 18 and build predictions for your model on the testing data set. Key in the accuracy of your model in the Excel answer file.

**Question 20**

**1 point**

Discuss the implications of the result you observed in Question 19.

Graphical user interface

Description automatically generated with low confidence

The model predicted more True Positives and True Negatives. The prediction accuracy increased to 0.845. It can be implied that accuracy increases when there are no biased predictions.

**Question 21**

**3 points**

Build a classification tree model on the original training data set using all of the predictors, up till a maximum depth of 3. Report what is the variable used in the first split of your classification tree in the Excel answer file. Your answer should be in the format: ‘x<a number>’.

**Question 22**

**1 point**

Discuss the implications of the result you observed in Question 21.

Diagram

Description automatically generated

The classification tree starts off by splitting x5. In the first leaf (x5 < -0.014), the proportion of ‘yes’ in the datasets is 0.5, the largest proportion that can be split.

**Question 23**

**2 points**

Report the **out-of-sample** prediction accuracy of your classification tree model in your Excel answer file.

**Question 24**

**3 points**

Build another classification tree model to maximum depth 6 this time. Report the **out-of-sample** prediction accuracy of your classification tree model in your Excel answer file.

**Question 25**

**2 points**

Build another classification tree model to maximum depth 10 this time. Report the **out-of-sample** prediction accuracy of your classification tree model in your Excel answer file.

**Question 26**

**2 points**

Discuss the implications of your results in Questions 24 and 25. Back up your claims with numerical or graphical evidence.

After increasing the number of edges or max depth, the out-of-sample prediction accuracy of the model has improved to 0.85. The phenomenon is the so-called bias-variance tradeoff

Bias-variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.