Statistics Learning, Module 2, 2021

Syndicate Problem Set # 1

All responses are to be neatly typed up. This includes any equations that you choose to present. The page setup should have at least 2cm margins on all sides, with all texts formatted to 12pt font size and at least 1.5 spacing. Present your numerical results neatly in tables and plots. R screenshots are not acceptable unless explicitly requested.

The submitted report should not exceed 10 pages. Graphs and tables may be presented in an Appendix, with the length of the appendix not exceeding 5 pages.

The assignment is due on Monday May 31st at 9PM. The assignment is to be submitted as a PDF file via Canvas. You are expected to work through these problems over the weeks prior, as indicated below.

Credit card balances (To be completed during Week 1)

The data file "credit.csv" collects observations on customer level credit card. The key variable is "Balance", which is the most recent balance on each customer's credit card account held at the bank. The bank would like to explore the factors that contribute to the credit balance. This type of study will help the bank in identifying the more active credit card users. The model can be used in an active reach-out initiative, where customers are offered a higher limit to meet their level of credit card usage; and also have potential in identifying suspicious balances due to fraud.

1. What relationship do you expect between the Balance and the potential predictors? Write a short paragraph discussing the expected relationships (max. 150 words).

2. Run a multiple linear regression with all possible predictors given. Does this regression model confirm your suspected relationship discussed in 1.) above? Discuss in a short paragraph (max. 150 words).

   Note that there are several categorical variables involved here. Make sure that you are aware of what the "reference groups" are for accurate interpretation.

3. Conduct a model selection exercise. Your process should include consideration of any potential nonlinear and interaction relationships that may be present. Briefly describe the process that you have taken to arrive at the final model. Present your choice of model and provide an analysis of the driver(s) of credit account balances. Your response to this should not exceed 1 page.

4. Discuss the limitation(s) of your analysis (max. 150 words)

Bank's Marketing Success

The data file "bankTD.csv" contains data on a certain European Bank's marketing campaign to promote term deposit products. The description of the data is given in the text file "bankDescriptions.txt" in the assignment pack. The goal of the analytics of this data set is to model what makes the campaign successful, with "success" being measured by whether the customer who partakes in the campaign took up the offer of the term deposit product (denoted by "y" in the data table). In exploring the analytics problem, explore the following tasks.

(To be completed during Week 2)

1. Suppose that you are only interested in linking the duration of the phone call to the "y" variable that indicate whether the customer took up the term deposit product. Write down the logistic model for this purpose.

2. Write down the log-likelihood of the model you presented in 1).

3. Using the mle() function in R, estimate the logistic model in 1). Estimate the same model using the glm() function. Provide an extract of the R code and output.

4. Provide an analysis of the relationship between the duration of the phone call and the likelihood that a customer takes up the term deposit product offer. Your analysis should include a discussion of the nature of the relationship, odds ratio and/or marginal effects. Supportive plots may be provided. Your response to this problem should not exceed 1 page.

(To be completed during Week 3)

5. Now consider all given predictor variables in the logistic model. Estimate the logistic regression with all predictors. Present the estimated model and discuss whether this model improves upon the model you estimated in 3.) Your response to this should not exceed 1 page.

   Note that there are several categorical variables involved here. Make sure that you convert all categorical variables into "factor" variables in R before fitting the model. You also have to be aware of what the "reference groups" for accurate interpretation. Also note that part of this question can be done during week 2).

6. Conduct a model selection exercise. As part of this, consider any potential nonlinear terms in your model. Briefly describe the process that you have taken to arrive at the final model. Present your choice of model and provide an analysis of the indicator for the take up of the term deposit product. Does the phone call duration still play a role? How does your result differ from the models constructed in 3) and 5)? Your response to this should not exceed 1 page.

7. Construct the hit and miss table for your chosen model, given that the cut-off probability is 50%. How accurate is your model at predicting a successful marketing campaign, given the customer's characteristics and the campaign related variables? What is the error rate associated with a false negative? What is the error rate associated with a false positive?

   The use of a 50% cut-off rate is one that is often done by default. Investigate how the choice of cut-off rate impacts the predictive power of the model. Consider using a grid of regular intervals, e.g. a grid with 5% incremental interval = [5%, 10%,15%,…,90%,95%]. How does the accuracy/error rate change over the varying cut-off points? Provide an appropriate recommendation about the cut-off point based on your analysis.

   Your response to this should not exceed 1 page.