IFI 8420

Assignment 3: Logistic Regression

(Group Submission)

Note: *Quiz 3B in iCollege will be based on this Assignment. Please have your R program available in running condition when you take the quiz. You will need solutions of your program to take the quiz. Quiz 3B is not under lock down browser.*

Part 1 **(100 points)**

Analyze the data in the **CreditCard** dataset in AER package. (Note that you have to install AER package and any other additional package that are required by AER)

The following variables are included in the dataset:

1. card: was the application for a card accepted? (Binary: 1/0) Response Variable
2. reports: Number of major derogatory reports
3. income: Yearly income (in USD 10,000)
4. Age: Age in years plus 12ths of a year
5. Owner: Does the individual own his/her home?
6. dependents: number of dependents
7. months: Months living at current address
8. share: ratio of monthly credit card expenditure to yearly income
9. selfemp: Is the individual self-employed?
10. majorcards: number of major credit cards held
11. active: number of active credit accounts
12. expenditure: average monthly credit card expenditure

Use variables 2 to 8 to determine which of the predictors influence the probability that an application is accepted. Online **Quiz 3B** will be based on your analysis below from A – H:

A. Provide summary stat of the predictors. (5 points)

B. There are some values of variable age under one year. Consider data with age>18 for your analysis for the rest of the questions. (5 points)

C. Plot of income vs. reports (Number of major derogatory reports): mark individuals with card application accepted as blue, and not accepted as red. (5 points)

D. Boxplots of income as a function of card acceptance status. Boxplots of reports as a function of card acceptance status (mark card application accepted as blue, and not accepted as red). (Display two boxplots in same page). (5 points)

E. Construct the histogram for the predictors. (5 points)
Note that share is highly right-skewed, so log(share) will be used in the analysis. reports is also extremely right skewed (most values of reports are 0 or 1, but the maximum value is 14. To reduce the skewness, log(reports+1) will be used for your analysis. Highly skewed predictors have high leverage points and are less likely to be linearly related to the response.

F. Using Logistic Regression with variables 2 to 8 determine which of the predictors influence the probability that an application is accepted. Use the summary function to print the results. (5 points)

G. To predict whether the application will be accepted or not, convert the predicted probabilities into class labels yes with the following condition: probs >.5="yes". Compute the confusion matrix and overall fraction of correct predictions. (15 points)

H. Now fit the logistic regression model using a training data for observations 1 to 1000. Compute the confusion matrix and the overall fraction of correct predictions for the test data (that is, the data for observations 1001 to end of data.) (20 points)

I. Apply Discriminant Analysis (LDA), Nearest-Neighbors, Naïve Bayes in the training and test data created in Part H and provide the performance of the models that analyze credit card application acceptance. (20 points)

J. Comparing H and I - Which model performs the best and why? Provide your reasoning in the final model selection and validation. You need to provide a summary of your reasoning in the Python Program (15 points)

Deliverables:

1. Submit Python program electronically in iCollege in the corresponding Assignment tab. (individual submission)

2. Please submit **one Python program (one file)** containing all the parts of the assignment (mark/comment so that each part is separated clearly in the program). Python code should provide comments on each sections of the assignment the code is intended for.

3. The assignment submission grade will be based on whether you have completed each part of the analysis and whether your code runs through all the parts of the analysis. Your

grade will be based not only on the correctness of the program but also how efficiently the program executes the tasks.

4. Note that **you do not have to write your response** to the above questions related to the interpretation of the model results in the code, **except for Part J**. If you do write the responses in the program - it will not be part of your grade.

5. **The quiz 3B will have questions (parts A – H) that will test your conceptual understanding of the output/results of the model and the data. Make sure you understand the relevant concepts of the analysis in each part before you take the online quiz.**

   Do not submit a separate word document explaining your results