\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Homework 6

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Instructions:

\* To create this document, first copy and paste the full text here into a .Do document (a STATA Do-File).

\* Below each question, write the code you used to answer the question

\* Next, write your actual answer to the question by commenting out your writing (by starting the line with a \*)

\* Next, copy and paste the entire document (my writing and yours) into a Word document. This will allow me to see your code on Canvas without downloading every homework.

\* The goal is that I should be able to copy and paste your entire text into a .Do File and run the code without any errors.

\* Finally, submit file as Homework 6 on Canvas

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Topic 1: Naive Bayes Classifier (Single Case)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*1. Import the LendingData excel file into Stata

\*DONE\*

\*2. For simplicity drop the Testing Data

drop if TestData == 1

\*3. Estimate the Mean/SD of FICO scores and Debt-to-Income for good loans

codebook fico if loan\_status == 1

\*mean = 697.384\*

\*SD = 32.8479\*

codebook dti if loan\_status == 1

\* mean = 17.3633 \*

\* SD = 8.7204 \*

\*4. Estimate the Mean/SD of FICO scores and Debt-to-Income for bad loans

codebook fico if loan\_status == 0

\*mean = 686.732\*

\*SD = 26.2648\*

codebook dti if loan\_status == 0

\* mean = 20.4104 \*

\* SD = 9.11308 \*

\*5. Estimate the Probability density for FICO of 720 conditional on the loan is good

disp normalden(720, 697.38, 32.85)

\* Save the estimate as a new variable

gen probDensFico720\_1 = .00958108

\*6. Estimate the Probability density for FICO of 720 conditional on the loan is bad

\* Save the estimate as a new variable

gen probDensFico720\_0 = normalden(720,686.732,26.2648)

\*7. Estimate the Probability density for DTI of 25 conditional on the loan is good

\* Save the estimate as a new variable

gen probDensDTI25\_1 = normalden(25,17.3633,8.7204)

\*8. Estimate the Probability density for DTI of 25 conditional on the loan is bad

\* Save the estimate as a new variable

gen probDensDTI25\_0 = normalden(25, 20.4104,9.11308)

\*9. Estimate the probability a loan is good using the Training Data:

\* Save the estimate as a new variable

gen probLoanGood = 5542/(5542+1458)

\*you can take an average because it's just ones and zeros\*

\*10. Estimate the Conditional probability of a good loan given a FICO=720 and dti=25

gen probLoanBad = 1 - probLoanGood

gen good\_prob = probDensFico720\_1 \* probDensDTI25\_1 \* probLoanGood

gen bad\_prob = probDensFico720\_0 \* probDensDTI25\_0 \* probLoanBad

gen conditional\_prob = good\_prob/(good\_prob + bad\_prob)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Topic 2: Naive Bayes Classifier (Entire Sample)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

clear

\*11. Import the LendingData excel file into Stata

import excel "/Users/henryvelasquez/Documents/MBA/MBA S3/2nd 7/Machine Learning in Finance K579/HW6/LendingClubData.xlsx", sheet("Sheet1") firstrow

\*do only for training data\*

drop if TestData == 0

\*12. Estimate the Mean/SD of FICO scores and Debt-to-Income for good loans

codebook fico if loan\_status == 1

/\*

mean = 697.384

std dev = 32.8479

\*/

codebook dti if loan\_status == 1

/\*

mean = 17.3633

std dev = 8.7204

\*/

\*13. Estimate the Mean/SD of FICO scores and Debt-to-Income for bad loans

codebook fico if loan\_status == 0

/\*

mean = 686.732

std dev = 24.2648

\*/

codebook dti if loan\_status == 0

/\*

mean = 20.4104

std dev = 9.11308

\*/

\*14. Estimate the Probability density for each FICO conditional on the loan is good

\* Save the estimate as a new variable

gen probDensFico\_1 = normalden(fico,697.384, 32.8479)

\*15. Estimate the Probability density for each FICO conditional on the loan is bad

\* Save the estimate as a new variable

gen probDensFico\_0 = normalden(fico, 686.732, 24.2648)

\*16. Estimate the Probability density for each DTI conditional on the loan is good

\* Save the estimate as a new variable

gen probDensDTI\_1 = normalden(dti, 17.3633, 8.7204)

\*17. Estimate the Probability density for each DTI conditional on the loan is bad

\* Save the estimate as a new variable

gen probDensDTI\_0 = normalden(dti, 20.4104, 9.11308)

\*18. Estimate the probability a loan is good using the Training Data:

\* Save the estimate as a new variable

egen probGood = mean(loan\_status)

\*19. Estimate the Conditional probability of a good loan

gen probBad = 1 - probGood

gen mixed\_good\_prob = probDensFico\_1 \* probDensDTI\_1 \* probGood

gen mixed\_bad\_prob = probDensFico\_0 \* probDensDTI\_0 \* probBad

gen conditional\_prob = mixed\_good\_prob/(mixed\_good\_prob + mixed\_bad\_prob)

\*20. Estimate the residual of the estimate

gen manualResidualCalc = loan\_status - conditional\_prob

\*21. Estimate a logit model on the training data with loan\_status as the y-variable and fico and dti as x-variables

logit loan\_status fico dti

\*22. Estimate the prediction of the logit

predict loanStatusPrediction

\*23. Estimate the residual

gen loanStatusPredResiduals = loan\_status - loanStatusPrediction

\*24. Compare the MSEs from the Naive Bayes Classifier and Logit Model for the Training Data

\*drop if TestData==1

summarize loanStatusPredResiduals if TestData==0, detail

summarize manualResidualCalc if TestData==0, detail

\*25. Compare the MSEs from the Naive Bayes Classifier and Logit Model for the Testing Data

\* rerun previous code with drop if TestData == 0

summarize loanStatusPredResiduals if TestData==1, detail

summarize manualResidualCalc if TestData==1, detail

/\*

Variance lower in logit

Many assumptions in Naive model. Normal distribution, and assuming different variables dont have relationships to eachother.

\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Topic 3: Random Forest

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*26. Import the LendingData excel file into Stata

clear

import excel "/Users/henryvelasquez/Documents/MBA/MBA S3/2nd 7/Machine Learning in Finance K579/HW6/LendingClubData.xlsx", sheet("Sheet1") firstrow

\*27. Estimate the Initial Entropy in the training data

\* Hint: log base 2 = ln(x)/ln(2)

sum loan\_status if Test == 0

egen probGood = mean(loan\_status)

gen initialEntropy = -probGood\*ln(probGood)/ln(2) - (1-probGood)\*ln(1-probGood)/ln(2)

\*28. Estimate the Initial Gini Uncertainty in the training data

gen initialGini = 1 - probGood^2 - (1 - probGood)^2

\*29. Estimate the Expected Entropy in the trainng data based on knowing the home ownership status of the applicant

sum home\_ownership if TestData == 0

gen prob\_home = .604

sum loan\_status if TestData == 0 & home\_ownership == 1

gen prob\_good\_home = .81717

sum loan\_status if TestData == 0 & home\_ownership == 0

gen prob\_good\_no\_home = .7529

gen entropy\_home = -prob\_good\_home\*ln(prob\_good\_home)/ln(2) - (1-prob\_good\_home) \* ln(1-prob\_good\_home)/ln(2)

gen entropy\_no\_home = -prob\_good\_no\_home\*ln(prob\_good\_no\_home)/ln(2) - (1 - prob\_good\_no\_home) \* ln(1-prob\_good\_no\_home)/ln(2)

gen expected\_entropy = prob\_home\*entropy\_home + (1-prob\_home) \* entropy\_no\_home

\*30. Calculate the Expected Information Gain

gen expectedGain = initialEntropy - expected\_entropy

\*31. use the code:

\*ssc install rforest

\*( This installs a new new command: rforest)

/\*done\*/

\*32. run a random forest on the training data with loan\_status as the y-variable and Home Ownership, Income, Debt-to-Income, and FICO score as x-variables

rforest loan\_status home\_ownership income dti fico if TestData==0, type(class)

\*33. Predict Loan Status from the random forest and estimate the residual

predict rForestPrediction

\*34. Compare the MSEs from the random forest to the logistic regression

logit loan\_status home\_ownership income dti fico if TestData==0

predict predict\_logit4

gen residual\_logit4 = loan\_status-predict\_logit4

sum residual\* if Test == 0

sum residual \* if Test ==1