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

\* 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

clear

cd "C:\Users\haniu\OneDrive\Desktop\Deepa\Deepa\Finance Core B\Business\Homework 6"

import excel "LendingClubData.xlsx", firstrow clear

\*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

sum fico dti if loan\_status==1

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

sum fico dti if loan\_status==0

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

\* Save the estimate as a new variable

\*disp normalden(720, 697.38, 32.85)

gen normal\_FICO\_good = normalden(720, 697.38, 32.85)

\* gen normal\_FICO\_good = normalden (720, 697.38, 32.85) can do this in this manner as well

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

\* Save the estimate as a new variable

\*disp normalden(720, 686.73, 24.26)

gen normal\_FICO\_bad = normalden(720, 686.73, 24.26)

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

\* Save the estimate as a new variable

\* disp normalden(25, 17.36, 8.72)

gen normal\_DTI\_good = normalden(25, 17.36, 8.72)

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

\* Save the estimate as a new variable

\*disp normalden(25, 20.41, 9.11)

gen normal\_DTI\_bad = normalden(25, 20.41, 9.11)

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

\* Save the estimate as a new variable

sum loan\_status

egen good = mean(loan\_status)

\* why mean cause its only zeros and ones which would sume up to 1

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

gen x = normal\_FICO\_good\*normal\_DTI\_good\*good

gen y = normal\_FICO\_bad\*normal\_DTI\_bad\*(1-good)

disp x/(x+y)

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

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

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

\*11. Import the LendingData excel file into Stata

clear

cd "C:\Users\haniu\OneDrive\Desktop\Deepa\Deepa\Finance Core B\Business\Homework 6"

import excel "LendingClubData.xlsx", firstrow clear

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

sum fico dti if loan\_status==1

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

sum fico dti if loan\_status==0

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

\* Save the estimate as a new variable

gen normal\_FICO\_good = normalden(fico, 697.38, 32.85)

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

\* Save the estimate as a new variable

gen normal\_FICO\_bad = normalden(fico, 686.73, 24.26)

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

\* Save the estimate as a new variable

gen normal\_DTI\_good = normalden(dti, 17.36, 8.72)

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

\* Save the estimate as a new variable

gen normal\_DTI\_bad = normalden(dti, 20.41, 9.11)

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

\* Save the estimate as a new variable

sum loan\_status

egen good = mean(loan\_status)

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

gen x = normal\_FICO\_good\*normal\_DTI\_good\*good

gen y = normal\_FICO\_bad\*normal\_DTI\_bad\*(1-good)

gen conditional\_predict = x/(x + y)

\*20. Estimate the residual of the estimate

gen residual\_conditional = conditional\_predict - loan\_status

\*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 if TestData==0

\*22. Estimate the prediction of the logit

predict predict\_logit

\*23. Estimate the residual\_

gen residual\_predict\_logit = predict\_logit - loan\_status

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

sum residual\_conditional residual\_predict\_logit if TestData==0

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

sum residual\_conditional residual\_predict\_logit if TestData==1

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

\* Topic 3: Random Forest

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

\*26. Import the LendingData excel file into Stata

clear

cd "C:\Users\haniu\OneDrive\Desktop\Deepa\Deepa\Finance Core B\Business\Homework 6"

import excel "LendingClubData.xlsx", firstrow clear

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

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

sum loan\_status if TestData==0

gen prob\_good = 0.7917

gen Initial\_Entropy = -prob\_good\*ln(prob\_good)/ln(2) - (1-prob\_good)\*ln(1-prob\_good)/ln(2)

display Initial\_Entropy

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

gen Initial\_Gini = 1 - prob\_good^2 - (1-prob\_good)^2

display Initial\_Gini

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

sum home\_ownership if Test==0

gen prob\_home =0.604

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

gen prob\_good\_home = 0.81717

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

gen prob\_good\_no\_home = 0.7529

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

display Expected\_Entropy

\*0.5[0.1log(0.1) + 0.9log(0.9)] − 0.5[0.3log(0.3) + 0.7log(0.7)]=

\*30. Calculate the Expected Information Gain

gen Expected\_Information\_Gain = Initial\_Entropy - Expected\_Entropy

display Expected\_Information\_Gain

\*31. use the code:

\*ssc install rforest

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

ssc install rforest

\* its a dotado file it is an package and it is someone else's file that we install who have all the forest thing code in it

\*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 predict\_rforest

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

\*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\_logit\_rforest

gen residual\_logit\_rforest = loan\_status - predict\_logit\_rforest

sum residual\* if TestData==0

sum residual\* if TestData==1