Lab 13: Receiver Operating Characteristics

SDS358: Applied Regression Analysis

Michael J. Mahometa, Ph.D.

.

> ”Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write.”
>
> *H.G. Wells*

Introduction

The basic idea of Lab is as follows: Answer a research question with the provided dataset. Each week, that research question (and data) will change depending on the topic we've covered the prior class days. Once we're done with Lab, you'll have a Lab Assignment, that will look a lot like the Lab: a research question you'll need to answer given some data. In Lab, you'll learn the procedure for answering the research question. For the Lab Assignment, you'll do that procedure for a grade (independently).

To help answer the research question, we'll follow some basic steps that we'll repeat throughout the semester:

- Reflect on the Question: Figure out the variables of interest, and the technique that's required.

- Analyze the Data: Perform the steps required for the technique.

- Draw Conclusions: Use the information that you got from the prior step to answer the research question in a concise, logical manner.

Let's get started:

Primary Research Question:

What loan applicant characteristic: Fico Score, Debt to Income Ratio, Loan Amount, or Purpose of Loan best predicts if an historic applicant will be a "Good" borrower? Use the ROC and AUC to help explain.

Step1: Reflect on the Question:

Download the syntax and data files from Canvas.

Let's load in our SDSRegressionR package so that we can use some of it's functions later:

```r
#Load our class package
library(SDSRegressionR)
```

Next, we'll load in the data. Be sure to use the basic file structure we talked about the first Lab: Put your syntax in a folder specific to this Lab. Then, make a "data" folder in that same place - use lowercase. If you do that, then all of this syntax will work like a charm.

```r
loans <- read_csv("data/historicLoans.csv")
```

Check the Data:

To make sure that we're working with the right data, and that we're all looking at it the same way, we'll answer some basic questions about the data before moving on:

1. How many observations are in the dataset for the model?

2. What was the highest Loan Amount in the first 10 observations?

3. Of the first 10 observations, were for Debt Consolodation?

These questions can be answered simply by looking at the dataset once it's loaded in:

```
View(loans)
```

Check the Variables of Interest

Let's find the variables that we need to answer the primary research question:

1. Which variable tells us the Outcome variable for this model?
   - What type of variable is this?

2. What are the variables of interest for the model?

3. Classify each of the variables of interest.

Again, these can be answered by looking at the dataframe, and with the help of the *names()* function. Also, the codebook for the data frame is our friend. You can open this in R or Excel. Remember, R is case-sensitive.

```
names(loans)
```

```
## [1] "LoanID"        "loan_amnt"     "purpose"       "fico"
## [5] "dti"           "income"        "loanIndicator"
```

Reflect on the Method

The last part of Reflect on the Question asks about the method or technique we'll use.

1. We will use Logistic Regression to answer this question. Why?

2. We'll need to run `Anova()` on our model. Why?

3. We'll need run an ROC before calculating an AUC. Why?

Step2: Analyze the Data

In this step, we'll run the provided syntax and answer some questions about the output to help us prepare for the final step.

Here's the syntax you'll need (from the .R syntax file):

```
#### Here is the R script you will use:  (remember that # indicates a comment) ####
#Lab13: ROC and AUC
library(SDSRegressionR)

#Read Data
loans <- read_csv("data/historicLoans.csv")

#Factor where needed:
table(loans$loanIndicator)
table(loans$purpose)
loans <- loans %>%
  mutate(good = factor(loanIndicator, levels=c(0,1), labels=c("Bad Borrower", "Good Borrower")),
         purpose_f = factor(purpose))
table(loans$good)

#First model:
mod1 <- glm(good ~ fico + dti+ loan_amnt + purpose_f, data = loans, family = "binomial")
summary(mod1)

#IF we wanted to remove outliers...
cooksPlot(mod1, key.variable = "LoanID", print.obs = TRUE)

#Best Predictor
library(car)
Anova(mod1, type="III")

#ROC for Fico Score
fico_mod <- glm(good ~ fico, data = loans, family = "binomial")
summary(fico_mod)

#ROC
library(pROC)
#get the data from the model.
fico_d <- modelData(fico_mod)
levels(fico_d$good)
r_fico <- roc(fico_d$good, fico_d$fico, ci = TRUE, levels=c("Bad Borrower", "Good Borrower"))
r_fico

#Plot with Youden
plot.roc(r_fico, print.thres="best", print.thres.best.method="youden")

#Let's use a TESTING data set
loan_test <- read_csv("data/historicLoans_Testing.csv")

#Original Factoring
loan_test <- loan_test %>%
  mutate(good = factor(loanIndicator, levels=c(0,1), labels=c("Bad Borrower", "Good Borrower")),
```

```r
         purpose_f = factor(purpose))

#Predict a Good Borrower
loan_test <- loan_test %>%
  mutate(pred = predict(fico_mod, loan_test, type="response")) %>%
  mutate(good_pred = case_when(fico > 709.5 ~ "Good Borrower",
                               TRUE ~ "Bad Borrower"))

#Used for Specificity and Sensitivity
t <- table(select(loan_test, good_pred, good))[2:1, 2:1]
(2158 + 452) / sum(t)
```

Question 1

The overall model was significant in the prediction of Borrower: LR chi2 (________) = ________, $p < 0.05$.

Question 2

The baseline outcome for this model is: ________.

Question 3

The best predictor of Good Borrower status is: ________, with a LR chi2 (________) = ________, $p < 0.05$

Question 4

As Fico Score increases by one whole unit, the odds of being a Good Borrower increases by ________ %.

Question 5

The AUC for Fico Score is ________.

Question 6

The Sensitivity for Fico Score predicting a Good Borrower, based on the Youden Index is ________.

Question 7

The Accuracy for Fico Score distinguishing a Good Borrower from a Bad Borrower in the NEW Testing data set, based on original Fico Score Model is ________.

Step3: Draw Conclusions

The final step is for us to Draw Conclusions. We'll take the syntax we've been given from Analyze the Question, run it, then examine the output. The questions from the prior step help set us up for the Draw Conclusions part.

We'll "fill in the blanks" in a canned paragraph for the Lab. For the Lab Assignment, you'll need to come up with a similar paragraph all on your own (please don't steal mine).

Primary Research Question
What loan applicant characteristic: Fico Score, Debt to Income Ratio, Loan Amount, or Purpose of Loan best predicts if an historic applicant will be a "Good" borrower? Use the ROC and AUC to help explain.

> Logistic Regression analysis was used to determine the best predictor of a Good Borrower, from the possible predictors of Fico Score, Debt to Income Ratio, Loan Amount, or Purpose of Loan. An initial model showed good fit with all predictors (LR chi2(10) = _______, p < 0.05). Of the predictors, the best was Fico Score, with a Likelihood Ratio chi2(1) = _______, p < 0.05. Using ROC analysis with a Youden Index, we discovered a Fico score of _______ could yeild a Sensitivity of _______ and a Specificity of _______. The corresponding Area Under the Curve was _______ (DeLong 95% CI: 0.5948 - 0.6163).
>
> Using a new testing data frame of _______ observations, and the corresponding 709.5 Fico Score as a cut of of classification, we discovered that current Fico Score model could be _______ % Accurate for new data.