Exploratory Data Analysis (EDA)
by Melvin Ott, PhD
September, 2017

Introduction

The Masters in Predictive Analytics program at Northwestern University offers graduate courses that cover predictive modeling using several software products such as SAS, R and Python. The Predict 410 course is one of the core courses and this section focuses on using Python.

Predict 410 will follow a sequence in the assignments. The first assignment will ask you to perform an EDA(See Ratner[1] Chapters 1&2) for the Ames Housing Data dataset to determine the best single variable model. It will be followed by an assignment to expand to a multivariable model. Python software for boxplots, scatterplots and more will help you identify the single variable. However, it is easy to get lost in the programming and lose sight of the objective. Namely, which of the variable choices best explain the variability in the response variable?

(You will need to be familiar with the data types and level of measurement. This will be critical in determining the choice of when to use a dummy variable for model building. If this topic is new to you review the definitions at Types of Data before reading further.)

This report will help you become familiar with some of the tools for EDA and allow you to interact with the data by using links to a software product, Shiny, that will demonstrate and interact with you to produce various plots of the data. Shiny is located on a cloud server and will allow you to make choices in looking at the plots for the data. Study the plots carefully. This is your initial EDA tool and leads to your model building and your overall understanding of predictive analytics.

Single Variable Linear Regression EDA

1. Become Familiar With the Data

Identify the variables that are categorical and the variables that are quantitative. For the Ames Housing Data, you should review the Ames Data Description pdf file.

2. Look at Plots of the Data

For the variables that are quantitative, you should look at scatter plots vs the response variable saleprice. For the categorical variables, look at boxplots vs saleprice. You have sample Python code to help with the EDA and below are some links that will demonstrate the relationships for the a different building_prices dataset.

For the boxplots with Shiny:

Click here

For the scatterplots with Shiny:

Click here

3. Begin Writing Python Code

Start with the shell code and improve on the model provided.

Single Variable Logistic Regression EDA

1. *Become Familiar With the Data*

In 411 you will have an introduction to logistic regression and again will ask you to perform an EDA. See the file credit data for more info. Make sure you recognize which variables are quantitative and which are categorical. And, for several of these variables, what is the level of measurement?

2. *Look at Plots of the Data*

For logistic regression, the response variable is of the type yes/no. In this dataset it is coded as good/bad. So, the EDA may include histograms for quantitative variables with a separate histogram for each of the response values. For numeric coded explanatory categorical variables, if the response good/bad is recoded as 0/1 then the mean for the response variable for each of the categories will indicate if there is a relationship.

For the histograms with Shiny:

Click here

For the means with Shiny:

Click here

3. *Begin Writing Python Code*

OK. You have looked at the plots, which variable do you think will be most useful for predicting or explaining bad credit? After you answer this question, begin writing Python code to see if you can replicate these plots.

The data set **CREDIT** contains information on 1000 customers. There are 21 variables in the data set:

Name	Model Role	Measurement Level	Description
AGE	Input	Interval	Age in years
AMOUNT	Input	Interval	Amount of credit requested
CHECKING	Input	Nominal or Ordinal	Balance in existing checking account: 1 = less than 0 DM 2 = more than 0 but less than 200 DM 3 = at least 200 DM 4 = no checking account
COAPP	Input	Nominal	Other debtors or guarantors: 1 = none 2 = co-applicant 3 = guarantor
DEPENDS	Input	Interval	Number of dependents
DURATION	Input	Interval	Length of loan in months
EMPLOYED	Input	Ordinal	Time at present employment: 1 = unemployed 2 = less than 1 year 3 = at least 1, but less than 4 years 4 = at least 4, but less than 7 years 5 = at least 7 years
EXISTCR	Input	Interval	Number of existing accounts at this bank
FOREIGN	Input	Binary	Foreign worker: 1 = Yes 2 = No
GOOD_BAD	Target	Binary	Credit Rating Status (good or bad)

HISTORY	Input	Ordinal	Credit History: 0 = no loans taken / all loans paid back in full and on time 1 = all loans at this bank paid back in full and on time 2 = all loans paid back on time until now 3 = late payments on previous loans 4 = critical account / loans in arrears at other banks
HOUSING	Input	Nominal	Rent/Own: 1 = rent 2 = own 3 = free housing
INSTALLP	Input	Interval	Debt as a percent of disposable income
JOB	Input	Ordinal	Employment status: 1 = unemployed / unskilled non-resident 2 = unskilled resident 3 = skilled employee / official 4 = management / self-employed / highly skilled employee / officer
MARITAL	Input	Nominal	Marital status and gender 1 = male – divorced/separated 2 = female – divorced/separated/married 3 = male – single 4 = male – married/widowed 5 = female – single
OTHER	Input	Nominal or Ordinal	Other installment loans: 1 = bank 2 = stores 3 = none
PROPERTY	Input	Nominal or Ordinal	Collateral property for loan: 1 – real estate 2 = if not 1, building society savings agreement / life insurance 3 = if not 1 or 2, car or others 4 = unknown / no property

PURPOSE	Input	Nominal	Reason for loan request: 0 = new car 1 = used car 2 = furniture/equipment 3 = radio / television 4 = domestic appliances 5 = repairs 6 = education 7 = vacation 8 = retraining 9 = business x = other
RESIDENT	Input	Interval	Years at current address
SAVINGS	Input	Nominal or Ordinal	Savings account balance: 1 = less than 100 DM 2 = at least 100, but less than 500 DM 3 = at least 500, but less than 1000 DM 4 = at least 1000 DM 5 = unknown / no savings account
TELEPHON	Input	Binary	Telephone: 1 = none 2 = yes, registered under the customer's name

Exploratory Data Analysis (EDA)

Ratner[1] describes 'data mining' "as any process that finds unexpected structures in data and uses the EDA framework to ensure that the process explores the data, not exploits it." Unexpected suggests that the word exploratory is very appropriate to this process.

Tukey[2] in his book and in many presentations gave structure to EDA. Others have extended it to include 'big' data. Big data has occurred due to our ability to capture huge datasets, store it on servers cost effectively, and analyze it with software that will handle it.

Shiny Apps

To learn more about Shiny applications with RStudio click on the link below:

http://rstudio.github.io/shiny/tutorial/

Types of Data

Quantitative data are numeric and represent counts or measurements.

Categorical data are names or labels such as a,b,c but can often be shown as 1,2,3. They do not suggest counts or measurements.

Discrete data are finite or countable numeric data.

Continuous data are values that represent a continuous scale of measurement.

A **nominal** level of measurement suggests names or categories. There is no apparent order suggested.

Ordinal level data suggest a sequential ordering but mathematical calculations should not be performed on this data.

Interval level data are ordinal plus the difference between two data values is meaningful. And, there is no zero level.

Ratio level data are interval and have a zero level plus differences and ratios may be calculated.

References:

1. Ratner, B. (2012). *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data* (2nd ed.). New York: CRC Press [ISBN-13: 9781439860915]

2. Tukey, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley.