Programming Exercise 9 - Answer Key

Crime Analytics (CJUS 6106)

Instructions

For this exercise, you will be tasked with estimating a linear probability models, a logistic regression model, and a random forest. You will be using the state_data file I have used in prior lectures. Note that this file must be knitted as a PDF. I strongly encourage you to use the .rmd file I provide for the exercise to begin your assignment.

Question 1

Create a total crime rate variable called **total_crime** outside of the **state_data** data frame.

Then, create a dummy variable named **crime_top20** within the **state_data** data frame indicating whether an observation is in the top 20th percentile of the total crime rate distribution (**Hint** - you'll need to use the **ntile** function from the last exercise and an **ifelse** function to accomplish this). Provide a summary of this dummy variable to confirm that its mean is 0.20 (within rounding error).

Finally, create a new data frame that removes the individual crime rate variables but keeps all other variables (including the dummy variable you just made). There should be 22 variables in this reduced data frame.

Question 2

Estimate a **linear probability model** predicting **crime_top20** using the following predictor variables: poverty, gini, top_1, urate, avwage, and inc_rate. Refer to prior lectures/assignments for assistance in understanding what each variable measures. You do not need to interpret these coefficients. Be sure to use the reduced data frame for this.

After this, convert the predicted values (**fitted.values**) from this model into 1s and 0s based upon if they meet or exceed a value of 0.50. Create a table that compares the predicted values for **crime_top20** to the actual values of **crime_top20** and compute classification error rates for both categories. Interpret both values.

Question 3

Now, estimate a logistic regression model using those same variables. Provide a summary for this model and interpret each coefficient (remember that these are now measured in changes to the log odds of belonging to the 1 category!). Provide an indication for each coefficient interpretation about whether it is statistically significant and at what level.

Next, convert the predicted values (**fitted.values**) from this model into 1s and 0s based upon if they meet or exceed a value of 0.50. Create a table that compares the predicted values for **crime_top20** to the actual values of **crime_top20** and compute classification error rates for both categories. Interpret both values. Are these similar to those obtained from the linear probability model?

Question 4

Create training and test data sets from the reduced **state_data** data frame you created at the end of Question 1. The training data should be a 75% sample of this data frame while the remaining 25% of observations go into the test data set.

Question 5

Using the training data set you just created in Question 4, estimate a random forest model predicting **crime_top20** using all other independent variables in the data frame. Use the following options to estimate this model: **importance=TRUE**, **proximity=FALSE**, and **ntree=250**. These options should make the random forest easier for your computer to estimate (though we usually leave **ntree** at the default value of 500).

After estimating this model, plot the importance of the predictors and provide an interpretation for the pattern of results. Be sure to mention if the lists of most important variables differ when using the **mean decease in accuracy** or the **mean decrease in Gini impurity** measures.

Question 6

Using the random forest model you estimated in Question 5, predict outcomes for the test data set created in Question 4. As you did with prior questions, compute the classification error rates for each category of the outcome variable and interpret them.

Finally, which model (LPM, logistic, or random forest) seems to perform best for predicting outcome values?