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.

```
total_crime <- with(state_data, rmurd + rrape + rrobb + raggr +
                                  rburg + rlarc + rauto)
total_crime_quint <- as.factor(ntile(total_crime, 5))
state_data$crime_top20 <- ifelse(total_crime_quint==5, 1, 0)
summary(state_data$crime_top20)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1994  0.0000  1.0000
```

```
state_data_new <- data.frame(state_data[,-(1:7)])
```

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.

```
lpm<-lm(crime_top20 ~ poverty + gini + top_1 + urate + avwage + inc_rate,
        data=state_data_new)
summary(lpm)
```

```
##
## Call:
## lm(formula = crime_top20 ~ poverty + gini + top_1 + urate + avwage +
##     inc_rate, data = state_data_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62693 -0.22589 -0.13526  0.06016  0.92499
##
## Coefficients:
##               Estimate  Std. Error t value     Pr(>|t|)
## (Intercept)  0.50627047  0.31528974   1.606      0.10883
## poverty      0.00993580  0.00637197   1.559      0.11943
## gini        -0.16627635  0.62074900  -0.268      0.78889
## top_1        0.01883700  0.00679607   2.772      0.00574 **
## urate       -0.01661125  0.00995059  -1.669      0.09554 .
## avwage      -0.00001539  0.00000268  -5.743 0.0000000145 ***
## inc_rate     0.05493751  0.00982726   5.590 0.0000000338 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3696 on 630 degrees of freedom
## Multiple R-squared:  0.1536, Adjusted R-squared:  0.1455
## F-statistic: 19.05 on 6 and 630 DF,  p-value: < 2.2e-16
```

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.

```
lpm_pred <- ifelse(lpm$fitted.values>=0.50, 1, 0)
table(state_data_new$crime_top20, lpm_pred)
```

```
##    lpm_pred
##       0   1
##   0 499  11
##   1 118   9
```

Calculations: The classification error rate for the 0 category is

$$1 - \frac{499}{499 + 11} = 0.02$$

and this indicates that we incorrectly classify roughly two out of every hundred observations.

The classification error rate for the 1 category is

$$1 - \frac{9}{118 + 9} = 0.93$$

and this indicates that we incorrectly classify roughly ninety-three out of every hundred observations.

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.

```
logit <- glm(crime_top20 ~ poverty + gini + top_1 + urate + avwage + inc_rate,
        data=state_data_new, family=binomial(link="logit"))
summary(logit)
```

```
##
## Call:
## glm(formula = crime_top20 ~ poverty + gini + top_1 + urate +
##     avwage + inc_rate, family = binomial(link = "logit"), data = state_data_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9490  -0.6360  -0.4374  -0.1977   2.3396
##
## Coefficients:
##               Estimate  Std. Error z value     Pr(>|z|)
## (Intercept)  2.80332329  2.24100721   1.251      0.21096
## poverty      0.06904074  0.04621129   1.494      0.13517
## gini        -3.73351764  4.20416361  -0.888      0.37451
## top_1        0.14363462  0.04649321   3.089      0.00201 **
## urate       -0.11331721  0.08257394  -1.372      0.16997
## avwage      -0.00013450  0.00002361  -5.696 0.0000000122 ***
## inc_rate     0.40252395  0.07396807   5.442 0.0000000527 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 636.40  on 636  degrees of freedom
## Residual deviance: 527.27  on 630  degrees of freedom
## AIC: 541.27
##
## Number of Fisher Scoring iterations: 5
```

Interpretations:

poverty: A one percentage point increase in the poverty rate is associated with a 0.069 increase in the log odds that a state is in the top 20th percentile of the total crime rate distribution. This slope is not significant.

gini: A one point increase in the gini index of inequality is associated with a -3.73 decrease in the log odds that a state is in the top 20th percentile of the total crime rate distribution. This slope is not significant.

top_1: A one percentage point increase in the share of all personal income among the top 1% of the population is associated with a 0.14 increase in the log odds that a state is in the top 20th percentile of the total crime rate distribution. This slope is significant at $p<.01$.

urate: A one percentage point increase in the unemployment rate is associated with a 0.11 decrease in the log odds that a state is in the top 20th percentile of the total crime rate distribution. This slope is not significant.

avwage: A 1 dollar increase in the average wages in a state is associated with a 0.0001 decrease in the log odds that a state is in the top 20th percentile of the total crime rate distrubition. This slope is significant at p<.001.

inc_rate: A one per 1,000 person increase in the state incarceration rate is associated with a 0.40 increase in the log odds that a state is in the top 20th percentile of the total crime rate distribution. This slope is significant at p<.001.

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?

```
logit_pred <- ifelse(logit$fitted.values>=0.50, 1, 0)
table(state_data_new$crime_top20, logit_pred)
```

```
##    logit_pred
##       0   1
##   0 485  25
##   1 102  25
```

Calculations: The classification error rate for the 0 category is

$$1 - \frac{485}{485 + 25} = 0.05$$

and this indicates that we incorrectly classify roughly five out of every hundred observations.

The classification error rate for the 1 category is

$$1 - \frac{25}{102 + 25} = 0.8$$

and this indicates that we incorrectly classify roughly eighty out of every hundred observations.

These error rates are a bit different than those from the linear probability model. The logistic regression model performs slightly worse when predicting 0s but is considerably better at predicting 1s.

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.

```
samp_size<-(0.75*nrow(state_data_new))
train_obs<-sample(seq_len(nrow(state_data_new)), size=samp_size)

train<-state_data_new[train_obs,]
test<-state_data_new[-train_obs,]
```

Question 5

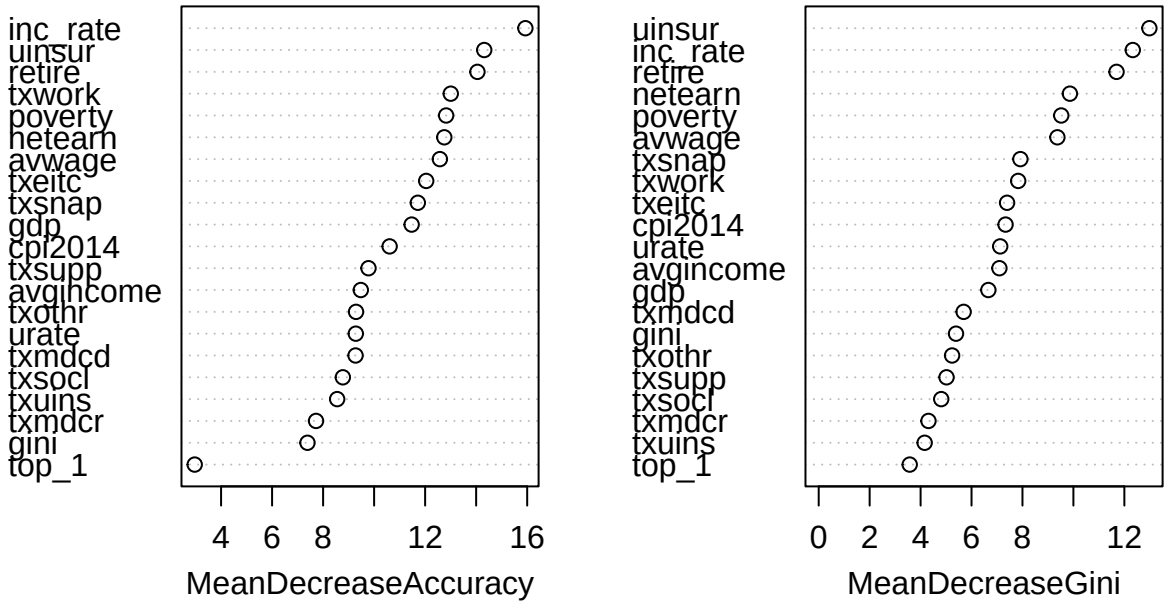
Using the training data set you just created in Question 4, estimate a random forest model predicting **crime_top20** using all other independent variables in the data frame. Use the following options to estimate this model: **importance=TRUE**, **proximity=FALSE**, and **ntree=250**. These options should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

```
rf<-randomForest(as.factor(crime_top20)~., data=train, importance=TRUE,
                         proximity=FALSE, ntree=250)
rf
```

```
##
## Call:
##  randomForest(formula = as.factor(crime_top20) ~ ., data = train,      importance = TRUE, proximity =
##                Type of random forest: classification
##                      Number of trees: 250
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 9.64%
## Confusion matrix:
##     0  1 class.error
## 0 370  9   0.0237467
## 1  37 61   0.3775510
```

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.

```
varImpPlot(rf, main="Variable Importance Plot")
```



Interpretations:

The top five variables for both measures are the same. Whether we use the mean decrease in accuracy or the mean decrease in gini impurity measure, the variables uinsur (state dollars spent on unemployment insurance), inc_rate (incarceration rate per 1,000 population), retire (state dollars spent on retiree wages), netearn (average net earnings of a state resident), and avwage (average gross wages among state residents). Following this, the lists differ in order but essentially have the same variables, usually just two or there places different from one another. This suggests that either measure would be appropriate to use in determining the importance of the variables in this classification model.

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.

```
rf_test <- predict(rf, test)
table(test$crime_top20, rf_test)
```

```
##    rf_test
##       0   1
##   0 130   1
##   1  12  17
```

Calculations: The classification error rate for the 0 category is

$$1 - \frac{128}{128 + 1} = 0.01$$

and this indicates that we incorrectly classify roughly one out of every hundred observations.

The classification error rate for the 1 category is

$$1 - \frac{21}{10 + 21} = 0.32$$

and this indicates that we incorrectly classify roughly thirty-two out of every hundred observations.

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

The classification error rates for the random forest model are the lowest for both categories. Therefore, the random forest model performs the best in predicting outcome values.