**Question 10.1**

**Using the same crime data set uscrime.txt as in Questions 8.2 and 9.1, find the best model you can using (a) a regression tree model, and (b) a random forest model. In R, you can use the tree package or the rpart package, and the randomForest package. For each model, describe one or two qualitative takeaways you get**

**Answer:** Given below are the steps performed to run both the models.

**Step 1:** Import the data and install and call the required libraries.

**CODE:**

*set.seed(1)*

*uscrime <- read.table("D://MS Georgia Tech/Introduction to Analytics/HW7/uscrime.txt", header = TRUE)*

*library(randomForest)*

*install.packages("tree")*

*library(tree)*

*library(caret)*

**Building a regression Tree Model.**

**Step 1:** Building the model using R’s Tree library

**CODE:**

*TreeModelUSCrime <- tree(Crime ~ ., data = uscrime)*

*summary(TreeModelUSCrime)*

**OUTPUT:**

|  |
| --- |
| summary(TreeModelUSCrime)  Regression tree:  tree(formula = Crime ~ ., data = uscrime)  Variables actually used in tree construction:  [1] "Po1" "Pop" "LF" "NW"  Number of terminal nodes: 7  Residual mean deviance: 47390 = 1896000 / 40  Distribution of residuals:  Min. 1st Qu. Median Mean 3rd Qu. Max.  -573.900 -98.300 -1.545 0.000 110.600 490.100 |
|  |
| |  | | --- | | > | |

**Step 2:** See how the tree was slit.

**CODE:**

*# see how tree was split*

*TreeModelUSCrime$frame*

*# Ploting the tree*

*plot(TreeModelUSCrime)*

*text(TreeModelUSCrime)*

*title("USCRIME Classification Tree for Training Set")*

**OUTPUT:**

> TreeModelUSCrime$frame

var n dev yval splits.cutleft splits.cutright

1 Po1 47 6880927.66 905.0851 <7.65 >7.65

2 Pop 23 779243.48 669.6087 <22.5 >22.5

4 LF 12 243811.00 550.5000 <0.5675 >0.5675

8 <leaf> 7 48518.86 466.8571

9 <leaf> 5 77757.20 667.6000

5 <leaf> 11 179470.73 799.5455

3 NW 24 3604162.50 1130.7500 <7.65 >7.65

6 Pop 10 557574.90 886.9000 <21.5 >21.5

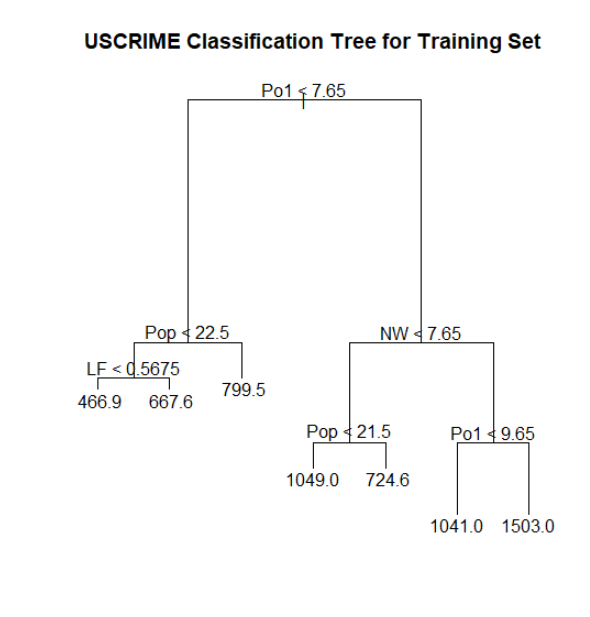
12 <leaf> 5 146390.80 1049.2000

13 <leaf> 5 147771.20 724.6000

7 Po1 14 2027224.93 1304.9286 <9.65 >9.65

14 <leaf> 6 170828.00 1041.0000

15 <leaf> 8 1124984.88 1502.8750



**Analysis**

**Step 3: Pruning leaf 5**

**Code:**

*# Prune the tree*

*termnodes <- 5*

*TreeModelUSCrime <- prune.tree(TreeModelUSCrime, best = termnodes)*

*plot(prune.TreeModelUSCrime)*

*text(prune.TreeModelUSCrime)*

*title("Pruned Tree")*

*summary(prune.crimeTreeMod)*

**OUTPUT:**

> summary(prune.crimeTreeMod)

Regression tree:

snip.tree(tree = crimeTreeMod, nodes = c(4L, 6L))

Variables actually used in tree construction:

[1] "Po1" "Pop" "NW"

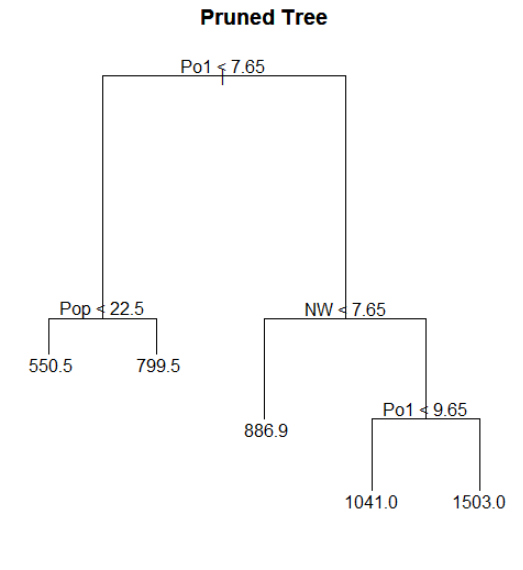
Number of terminal nodes: 5

Residual mean deviance: 54210 = 2277000 / 42

Distribution of residuals:

Min. 1st Qu. Median Mean 3rd Qu. Max.

-573.9 -107.5 15.5 0.0 122.8 490.1



**Question 10.2**

**Describe a situation or problem from your job, everyday life, current events, etc., for which a logistic regression model would be appropriate. List some (up to 5) predictors that you might use.**

**Answer:**

I once worked in a financial firm based in Dallas Texas as an Analytics Consultant providing Analytics support to a line of business called as Post-Closing (also known as Post-Funding). To understand why a logistic regression model will be appropriate here, I would first give an understanding of what Post-Closing means.

**Post-Closing Overview:** When a borrower want to purchase a house and he/she applies for a home loan with a bank or a money lending firm. The bank approves the loan and funds the borrower. Once the loan is funded by the bank, the bank decides to sell this loan to another firm (known as Investor) with small profit margin so that the bank selling the loan receives more money to fund more loans. This process of bank selling a funded loan to an investor is called Post-Closing.

When a bank tries to sell this loan, it has to provide all the documents to the investor and the document must be in the correct format with the correct data as expected by the investor. If there are issues with errors in the documents then these issues are called as Defects Whenever a Defect is detected, it needs to be fixed before an investor finalizes the purchase. However, there are certain defects which cannot be fixed and as a result those loans either cannot be sold or has to be sold at a heavy loss. Such loans which cannot be sold are called as ‘Scratch & Dent’ loans.

A logistic regression model would be appropriate for the scenario mentioned above predict if a loan would be ‘Scratch & Dent’ Loan. To check if a loan would become a Scratch & Dent loan, it is dependent on the following predictors.

1. **Channel**: A loan can come through 4 channels. They are

* Wholesale – Where a bank gets bulk of loans from a wholesale realtor
* Retail – Where loan originates from a retail medium.
* Correspondent – Where loan originates from another local bank later purchased.
* Consumer Direct – Where loan originates from the borrower directly applying loan online.

For all the above channels, we get different form of documentation leading to different defects

1. **Loan Type**: A borrower can be eligible for different types of loans based on different parameters like income, job status etc. Different types of loans are USDA loans (for farmers or people in agriculture), VA loans (for veterans), FHA loans etc.
2. **Number of Defects:** As the number of defects increases, the chances of it being not fixed also increases which makes a loan fall into scratch and dent category.
3. **State:** Different states in USA have different laws for granting of loans. Most of the laws per state may be same but there may be only few laws which change causing a loan to fall in the scratch and dent category.