Partition

Amin Fesharaki

11/14/2021

```
churn <- read.csv(file = "/Users/datascience/Desktop/ADS 502 Data Sets/Website Data Sets/churn.csv", sep
churn = churn[, c('State', 'Account.Length','Area.Code','Phone','Intl.Plan', 'VMail.Plan', 'VMail.Messag
               'Day.Mins', 'Day.Calls', 'Day.Charge', 'Eve.Mins', 'Eve.Charge' , 'Night.Mins', 'Night.Ca
               'Night.Charge', 'Intl.Mins', 'Intl.Calls' , 'Intl.Charge', 'CustServ.Calls', 'Old.Churn'
set.seed(1)
```

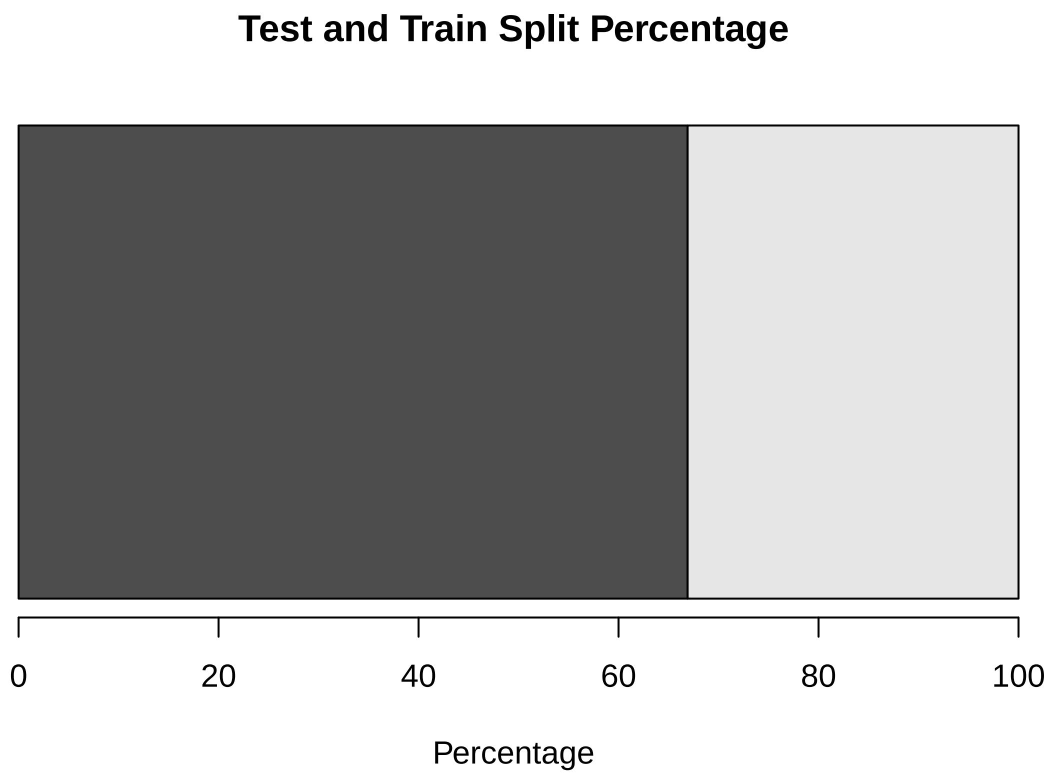
28. Partition the data set, so that 67% of the records are included in the training data set and 33% are included in the test data set. Use a bar graph to confirm your proportions

```
#  Split Train and Test Data
n <- dim(churn)[1]
train_ind <-runif(n) <0.67
churn_train <- churn[ train_ind, ]
churn_test <- churn[ !train_ind, ]

# Bar Graph
test = (nrow(churn_train)/(dim(churn)))*100
train = (nrow(churn_test)/(dim(churn))*100)
cat("Test Dataset Percentage: " , round(test), "\nTraining Dataset Percentage", round(train))
```

```
## Test Dataset Percentage:  67 10619
## Training Dataset Percentage 33 5252
```

```
Tot = rbind(test[1], train[1])
barplot(Tot, horiz = TRUE, legend = TRUE, main = "Test and Train Split Percentage", xlab = "Percentage"
```



29. Identify the total number of records in the training data set and how many records in the training data set have a churn value of true.

```
table(churn_train$Churn)
```

```
##
## False  True
##  1918   312
```

30. Use your answers from the previous exercise to calculate how many true churn records you need to resample in order to have 20% of the rebalanced data set have true churn values.

x = [0.2(2231)-330]/0.8) = 145.25 = 146

31. Perform the rebalancing described in the previous exercise and confirm that 20% of the records in the rebalanced data set have true churn values.

```
# Resample
to.resample <- which(churn_train$Churn == "True")
our.resample <- sample(x = to.resample, size = 146, replace = TRUE)
our.resample <- churn_train[our.resample,]
churn_train_bal <- rbind(churn_train, our.resample)

t.v1 <- table(churn_train_bal$Churn)
# Rebalance Percentage
t.v1/nrow(churn_train_bal)
```

```
##
##     False      True
## 0.8072391 0.1927609
```

32. Which baseline model do we use to compare our classification model performance against? To which value does this baseline model assign all predictions? What is the accuracy of this baseline model?

We can use an ‘All Positive Model’ for our data which assigns all predictions as positive. Therefore the accuracy of this model is .20 or 20%. On the other hand, we can assign our model to be an ‘All Negative Model’ which assigns all predictions as negative. In this case, the accuracy of the all negative model is .80 or 80%.

33. Validate your partition by testing for the difference in mean day minutes for the training set versus the test set.

```
#For numerical values, we use the two sample t-test for the difference in means
t.test(churn_train_bal$Day.Mins, churn_test$Day.Mins, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  churn_train_bal$Day.Mins and churn_test$Day.Mins
## t = 0.58022, df = 2207.5, p-value = 0.5618
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.767829  5.093895
## sample estimates:
## mean of x mean of y
##  181.0715  179.9084
```

The p-value for our test is 0.5618 which is greater than p-value(alpha) of 0.05. Therefore we can accept null hypothesis and say that there is sufficient evidence to support that the mean day minutes between both populations is different.

34. Validate your partition by testing for the difference in proportion of true churn records for the training set versus the test set.