Converting CDC Data to Binary Data in R

Noah Ferrel

12/13/2020

Introduction

In the real world, data is collected from a variety of different sources over a wide range of time. The data is not always formatted in the most optimal manner and over time the data formatting may change. This project aims to combine and clean ten years of CDC data in a such a way that it is ripe for analysis. The cleaned data can be analyzed in a multitude of different ways, but logistic regressions were preformed to find the important predictors of Diabetes and Alcoholism.

Data

- The combine_clean_data.r file takes in 16 data files from the 'data' folder
  - CDC 10 Years of Data: patient data from 2006 to 2015
  - BI_DATA_sample.Rda: Sample file to test what the final data should look like
  - new_names.Rda: A file containing a list of drug ids and diagnosis
  - ICD9_codes.RDA: A list of ICD9 codes
  - RFV_codes.RDA: A list of RFV codes
  - testdf.Rda: A file containing four Data Frames with different class types, used to test the bicols() function
  - OpioidCodesOnly.csv: A list of opioid codes
  - Parameter_OP_codes.csv: User inputed list of opioid codes

Functions in combine_clean_data.r

- combine_columns(df, list) : The function takes in a df and a column to grab the selected columns from each dataframe
- whitespace2NA(column) :The function takes in a column and changes the white space to NA
- dash_to_nothing(column) : The function takes in a column and gets rid of the dashes
- bicols(data, list): The function takes in a data frame and a list of codes to return binary column containing if the code was in the list

Results

The project output was two new RDA files. The first file contains the codes of different diagnosis or ailments. The columns are the different Identifications and the rows are the patients. Bellow is a sample of what the outputted data looks like.

The second data file is the binary data frame that shows if an individual had the diagnosis or ailment.

```
head(newDF[,1:9])
```

```
##   DRUGID1 DRUGID2 DRUGID3 DRUGID4 DRUGID5 DRUGID6 DRUGID7 DRUGID8 DRUGID9
## 1  d03423    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 2  d00046  d03393    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 3  a70674    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
```

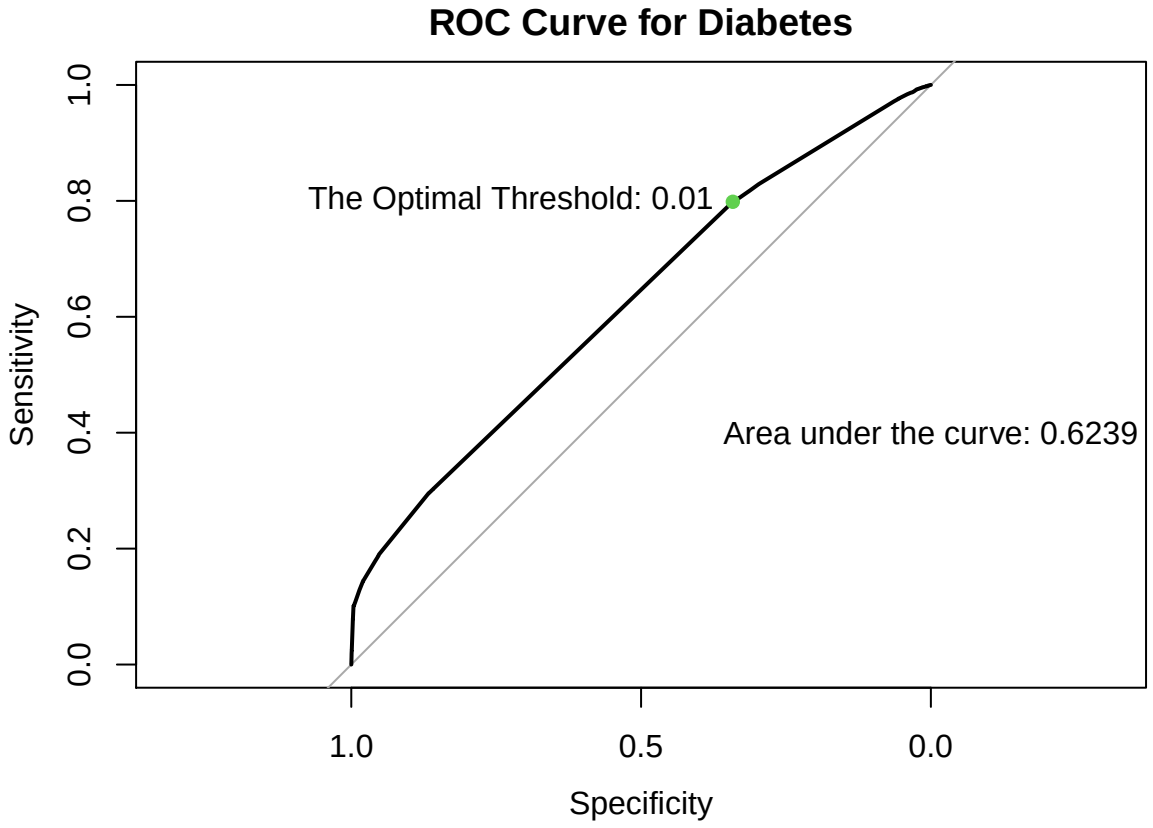
```
## 4  d00021  d00787  d00212    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 5  d00321  d00170    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 6  d00096  d00059    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
```

```
head(BI_DATA[,1:9])
```

```
##   p.ab p.arth p.back p.cancer p.chest p.chol p.dent p.fibro p.frac
## 1    0      0      0        0       0      0      0       0      0
## 2    0      0      0        0       0      0      0       0      0
## 3    0      0      0        0       0      0      0       0      0
## 4    0      0      0        0       0      0      0       0      0
## 5    0      0      0        0       1      0      0       0      0
## 6    0      0      0        0       0      0      0       0      0
```

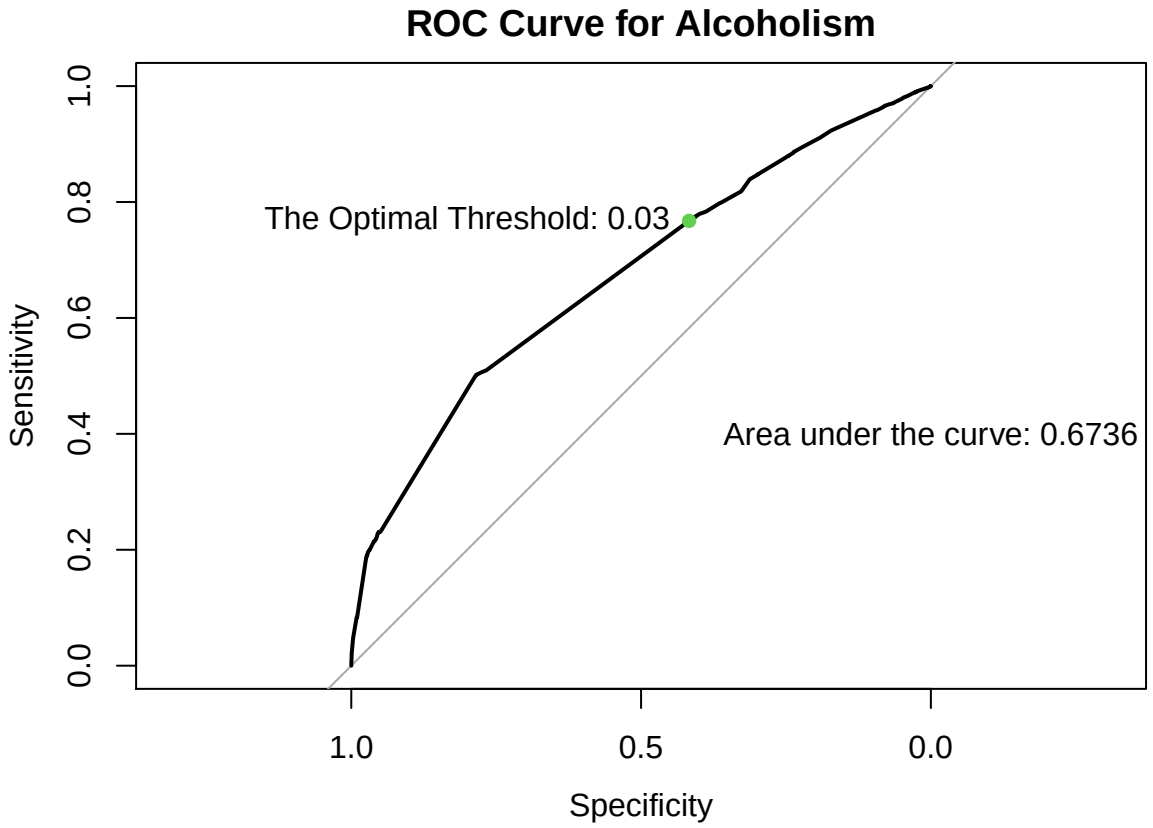
Using logistic regression, it was found that arthritis pain, back pain, chest pain, chol pain, headache, pelvic pain, kidney pain,the reason for visit was endo related, schedule 2 drugs, and schedule 6 drugs were important predictors of diabetes. The area under the curve was 0.62 which is poor. With more time I would try to combine this data set with another that contained more information on the individual directly related to the known factors of diabetes .

```
## Diabetes ROC curve
roc(BI_DATA$Diabetes,pred_dibep,plot = T,main = "ROC Curve for Diabetes") # Area under the curve: 0.623
points(Diab_curve$specificities[Diab_ind],Diab_curve$sensitivities[Diab_ind],pch=16,col=3)
text(x = Diab_curve$specificities[Diab_ind], y = Diab_curve$sensitivities[Diab_ind], labels = "The Optir
text(0,0.4,labels = "Area under the curve: 0.6239")
```



The important predictors of Alcoholism are abdominal pain, dent pain, headache, nonfrac pain, the reason for visit was circ, skin, mental, took opioids and schedule 2,3,4, and 6 drugs. This model performed beter than the diabetes model, but It still can be improved. I would want data on how often they drink and family history to improve the model.

```r
## Alcoholism ROC curve
roc(BI_DATA$Alcohol,pred_ALC,plot = T,main = "ROC Curve for Alcoholism")
points(ALC_curve$specificities[ALC_ind],ALC_curve$sensitivities[ALC_ind],pch=16,col=3)
text(x = ALC_curve$specificities[ALC_ind], y = ALC_curve$sensitivities[ALC_ind], labels = "The Optimal T
text(0,0.4,labels = "Area under the curve: 0.6736")
```



Changes for final

- Removed an extra variable 'e' that had unnecessary memory cost
- Removed a function that did not have direct applications to the project
- Tested the mclapply from the parallel package to read in the data for line 17 (was line 13), but It did not improve run-time. When ran in parallel the run time increased from 20s to 1450s.
- Added additional documentations to the README and to the Combine_clean_data.R especially lines 11-60 (was line 11-36)
- Added an input file to select opioid codes that are not wanted on line 73,75 (was line 73)