Research on Education and Census Data

Kathy Wu and Chelsea Lu

January 07, 2022

Introduction to the Datasets

Census Data

```
state.name <- c(state.name, "District of Columbia")
state.abb <- c(state.abb, "DC")
## read in census data
census <- read_csv("./acs2017_county_data.csv") %>%
  select(-CountyId, -ChildPoverty, -Income, -IncomeErr, -IncomePerCap, -IncomePerCapErr) %>%
  mutate(State = state.abb[match(`State`, state.name)]) %>%
  filter(State != "PR")
head(census)
```

```
## # A tibble: 6 x 31
##   State County   TotalPop   Men  Women Hispanic White Black Native Asian Pacific
##   <chr> <chr>       <dbl> <dbl>  <dbl>    <dbl> <dbl> <dbl>  <dbl> <dbl>   <dbl>
## 1 AL    Autauga~    55036 26899  28137      2.7  75.4  18.9    0.3   0.9       0
## 2 AL    Baldwin~   203360 99527 103833      4.4  83.1   9.5    0.8   0.7       0
## 3 AL    Barbour~    26201 13976  12225      4.2  45.7  47.8    0.2   0.6       0
## 4 AL    Bibb Co~    22580 12251  10329      2.4  74.6  22      0.4   0         0
## 5 AL    Blount ~    57667 28490  29177      9    87.4   1.5    0.3   0.1       0
## 6 AL    Bullock~    10478  5616   4862      0.3  21.6  75.6    1     0.7       0
## # ... with 20 more variables: VotingAgeCitizen <dbl>, Poverty <dbl>,
## #   Professional <dbl>, Service <dbl>, Office <dbl>, Construction <dbl>,
## #   Production <dbl>, Drive <dbl>, Carpool <dbl>, Transit <dbl>, Walk <dbl>,
## #   OtherTransp <dbl>, WorkAtHome <dbl>, MeanCommute <dbl>, Employed <dbl>,
## #   PrivateWork <dbl>, PublicWork <dbl>, SelfEmployed <dbl>, FamilyWork <dbl>,
## #   Unemployment <dbl>
```

Education Data

```
## read in education data
education <- read_csv("./education.csv") %>%
  filter(!is.na(`2003 Rural-urban Continuum Code`)) %>%
  filter(State != "PR") %>%
  select(-`FIPS Code`,
         -`2003 Rural-urban Continuum Code`,
         -`2003 Urban Influence Code`,
         -`2013 Rural-urban Continuum Code`,
         -`2013 Urban Influence Code`) %>%
  rename(County = `Area name`)
head(education)
```

```
## # A tibble: 6 x 42
##   State County         `Less than a high ~ `High school dipl~ `Some college (1-~
##   <chr> <chr>                        <dbl>              <dbl>              <dbl>
## 1 AL    Autauga County                6611               3757                933
## 2 AL    Baldwin County               18726               8426               2334
## 3 AL    Barbour County                8120               2242                581
## 4 AL    Bibb County                   5272               1402                238
## 5 AL    Blount County                10677               3440                626
## 6 AL    Bullock County                4245                958                305
## # ... with 37 more variables: Four years of college or higher, 1970 <dbl>,
## #   Percent of adults with less than a high school diploma, 1970 <dbl>,
## #   Percent of adults with a high school diploma only, 1970 <dbl>,
## #   Percent of adults completing some college (1-3 years), 1970 <dbl>,
## #   Percent of adults completing four years of college or higher, 1970 <dbl>,
## #   Less than a high school diploma, 1980 <dbl>,
## #   High school diploma only, 1980 <dbl>, ...
```

Preliminary data analysis

1. Census Data

```
## [1] 3142   31
```

The dimension of census data is 3142 x 31.

```
## [1] 0
```

There is no missing value in the data set.

```
## [1] 51
```

The total number of distinct values in State in Census is 51 which contains all states and a federal district.

2. Education Data

```
## [1] 3143   42
```

The dimension of education data is 3143 x 42.

```
## [1] 18
```

There are 18 distinct counties contain missing values in the data set

```
## [1] 1877
```

The total number of distinct county in education data is 1877.

```
## [1] 1877
```

The total number of distinct county in census data is 1877, which is the same as that of the education data.

Data Wrangling

3. we remove all the NA values in education data.

```
education = na.omit(education)
```

4. We then want to mutate the data set into the 6 features we want.

```
new.education = select(education, c("State",
                        "County",
                        "Less than a high school diploma, 2015-19",
                        "High school diploma only, 2015-19",
                        "Some college or associate's degree, 2015-19",
                        "Bachelor's degree or higher, 2015-19"))
new.education = mutate(new.education,
                         "Total Population of County" = rowSums(new.education[,3:6]))
```

5. We construct aggregated data sets from education data.

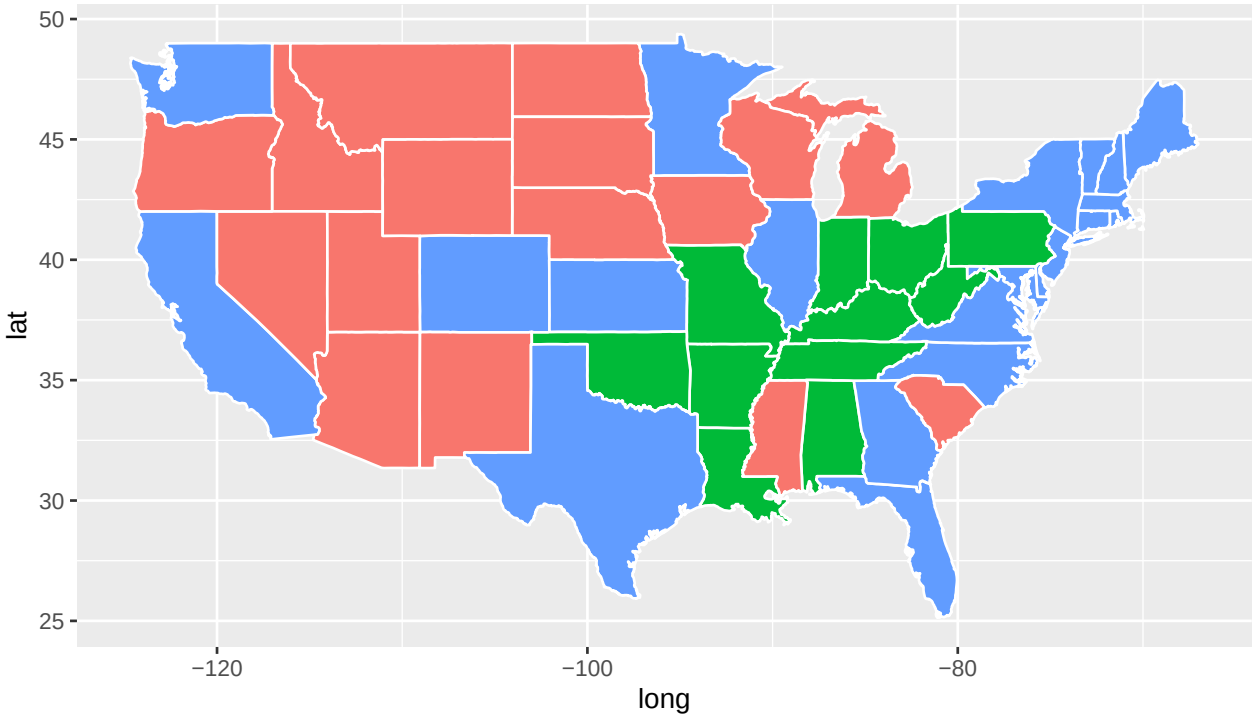
```
education.state = new.education %>%
  group_by(State)
```

6. We create a data set on the basis of education.state, where we create a new feature which is the name of the education degree level with the largest population in that state.

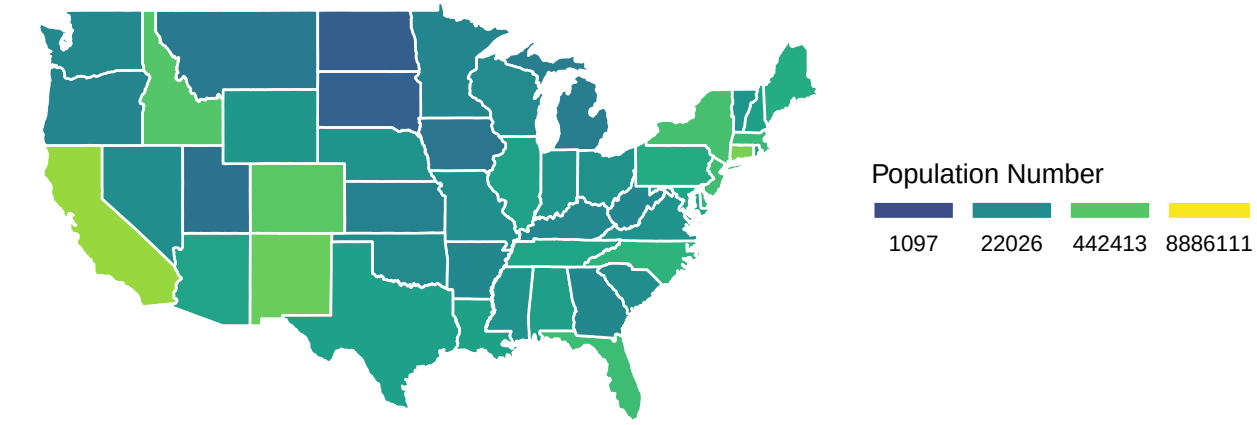
```
state.level = education.state %>%
  summarise(across(2:5, sum)) %>%
  rowwise() %>%
  mutate(edu.level = names(.)[which.max(c_across(2:5))])
```

Visualization

7. Now we color the map (on the state level) by the education level with highest population for each state.



8. We plot the graph for the census data using the total population.



9. We clean and aggregate the information in the census data which contains county-level census information.

```
# filter out any rows with missing values
census.clean = na.omit(census)

# convert {Men, Employed, VotingAgeCitizen} attributes to percentages
census.clean = census.clean %>%
  mutate(Men = 100*census.clean$Men/census.clean$TotalPop) %>%
  mutate(Employed = 100*census.clean$Employed/census.clean$TotalPop) %>%
  mutate(VotingAgeCitizen = 100*census.clean$VotingAgeCitizen/census.clean$TotalPop)

# compute Minority attribute by combining {Hispanic, Black, Native, Asian, Pacific}
census.clean = census.clean %>%
  mutate(minority = rowSums(census[,c(6,8,9,10,11)]))

# remove these variables after creating Minority
census.clean = select(census.clean, -c(Hispanic, Black, Native, Asian, Pacific))

# remove {Walk, PublicWork, Construction, Unemployment}
census.clean = select(census.clean, -c(Walk, PublicWork, Construction, Unemployment))
```

10. Print out the cleaned census data.

```
## # A tibble: 6 x 23
##   State County TotalPop   Men  Women White VotingAgeCitizen Poverty Professional
##   <chr> <chr>     <dbl> <dbl>  <dbl> <dbl>            <dbl>   <dbl>        <dbl>
## 1 AL    Autau~    55036  48.9  28137  75.4             74.5    13.7         35.3
## 2 AL    Baldw~   203360  48.9 103833  83.1             76.4    11.8         35.7
## 3 AL    Barbo~    26201  53.3  12225  45.7             77.4    27.2         25
## 4 AL    Bibb ~    22580  54.3  10329  74.6             78.2    15.2         24.4
## 5 AL    Bloun~    57667  49.4  29177  87.4             73.7    15.6         28.5
## 6 AL    Bullo~    10478  53.6   4862  21.6             78.4    28.5         19.7
## # ... with 14 more variables: Service <dbl>, Office <dbl>, Production <dbl>,
## #   Drive <dbl>, Carpool <dbl>, Transit <dbl>, OtherTransp <dbl>,
## #   WorkAtHome <dbl>, MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>,
## #   SelfEmployed <dbl>, FamilyWork <dbl>, minority <dbl>
```

Dimensionality reduction

11. Run PCA for the cleaned county level census data (with State and County excluded).

```
pca.census = prcomp(census.clean[,colnames(census.clean)!="State" &
                                  colnames(census.clean)!="County"&
                                  colnames(census.clean)!="Women"&
                                  colnames(census.clean)!="minority"&
                                  colnames(census.clean)!="TotalPop"],
                    scale=TRUE)
pc.county = pca.census$x[,c(1,2)]
```

Here, we choose to center and scale the features before running PCA is because we want to scale the variables to have standard deviation one, and scaling makes the results less complicated.
We delete the "Women" and "TotalPop" column because it is colineared with "Men", "minority" is colineared with "White".

The three features with the largest absolute values of the first principal component are "WorkAtHome", "SelfEmployed", and "minority".

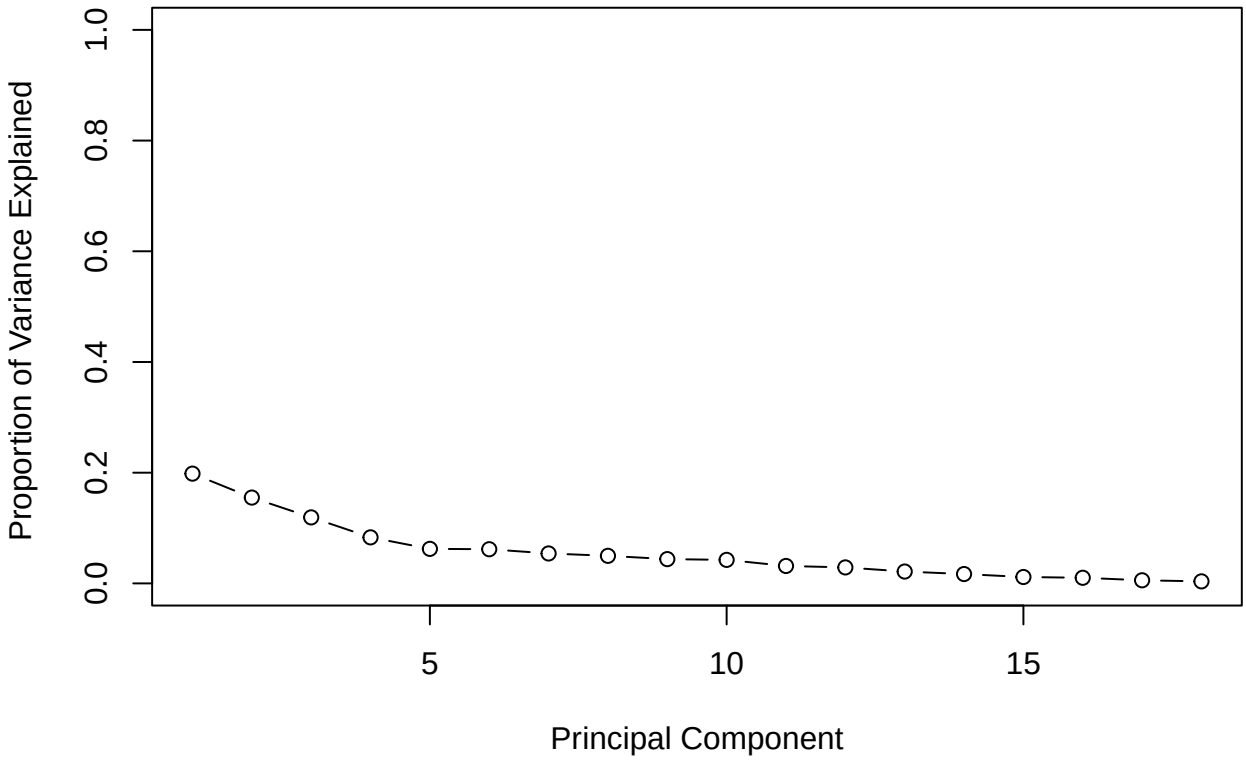
```
##   WorkAtHome SelfEmployed        Drive Professional   Production  PrivateWork
##       0.4286       0.3653       0.3589       0.3393       0.2889       0.2771
```

```
##            Drive       Production      PrivateWork          Poverty
##          -0.35886         -0.28893         -0.27714         -0.23543
##      MeanCommute           Office          Service          Carpool
##          -0.18175         -0.15340         -0.08843         -0.06468
## VotingAgeCitizen              Men
##           0.02701          0.07154
```

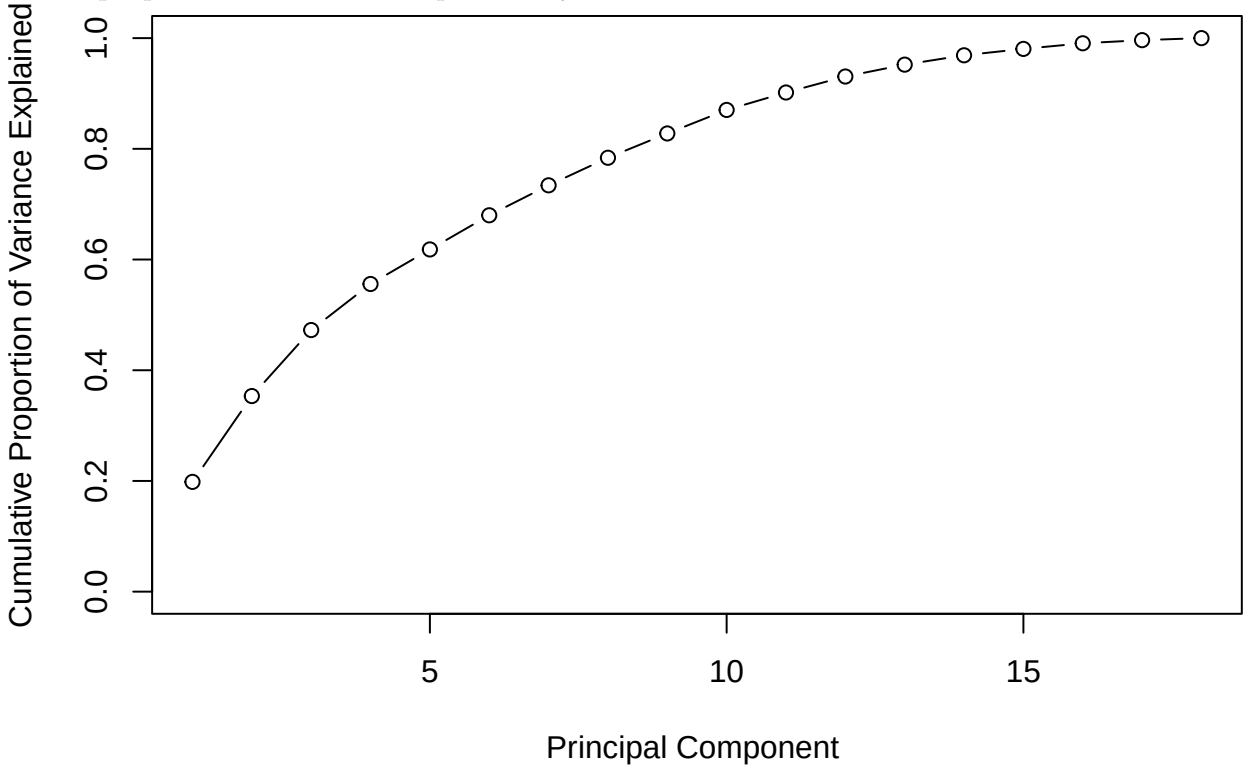
The features that have opposite signs are "Drive", "Production", "Privatework", "Poverty", "MeanCommute", "Office", "Service", and "Carpool".
This means that there is a negative correlation of the variables in the first PC. For example, with an increasing of one of the negative-sign features, there is a decrease in the response.

12. Determine the number of minimum number of PCs needed to capture 90% of the variance for the analysis.

Plot of proportion of variance explained by each component:



Plot of proportion of variance explained by cumulative PVE:



We need 11 PCs in order to explain 90% of the total variation in the data.

Clustering

13. With census.clean (with State and County excluded), perform hierarchical clustering with complete linkage.
Hierarchical Clustering for 10 clusters:

```
## clust1
##    1    2    3    4    5    6    7    8    9   10
## 2789  245   73    2   12    1    2    5   12    1
```

```
## [1] 2
```

For hierarchical clustering with 10 cluster, we observe that "Santa Barbara County" is in cluster 2.

First 2 Principal Componets
Then we use the first 2 principal components from pc.county as inputs instead of the original features to run hierarchical clustering algorithm again.

```
## clust2
##    1    2    3    4    5    6    7    8    9   10
## 1490  432   99  533  461   18   83    1   11   14
```

```
## [1] 5
```

For clustering the first 2 principal components with 10 cluster, we observe that "Santa Barbara County" is in cluster 5.

Apparently, using the first 2 principal components include more information related to "Santa Barbara County". As a result, the second method seems to put "Santa Barbara County" in a more appropriate cluster.

Modeling

Question we want to answer: *Can we use census information as well as the education information in a county to predict the level of poverty in that county?*

```
# we join the two data sets
all = census.clean %>%
  left_join(education, by = c("State"="State", "County"="County")) %>%
  na.omit
```

14. Transform the variable Poverty into a binary categorical variable with two levels: 1 if Poverty is greater than 20, and 0 if Poverty is smaller than or equal to 20. Remove features that you think are uninformative in classification tasks.

```
all = all %>%
  mutate(Poverty = factor(as.integer(Poverty > 20), level=c(0,1)))
all = all %>%
  select(c("Men","Poverty","Professional","Employed","PrivateWork","SelfEmployed","FamilyWork","minorit
colnames(all) = make.names(colnames(all))
```

We then partition the dataset into 80 training and 20 test data.

The following code to define 10 cross-validation folds:

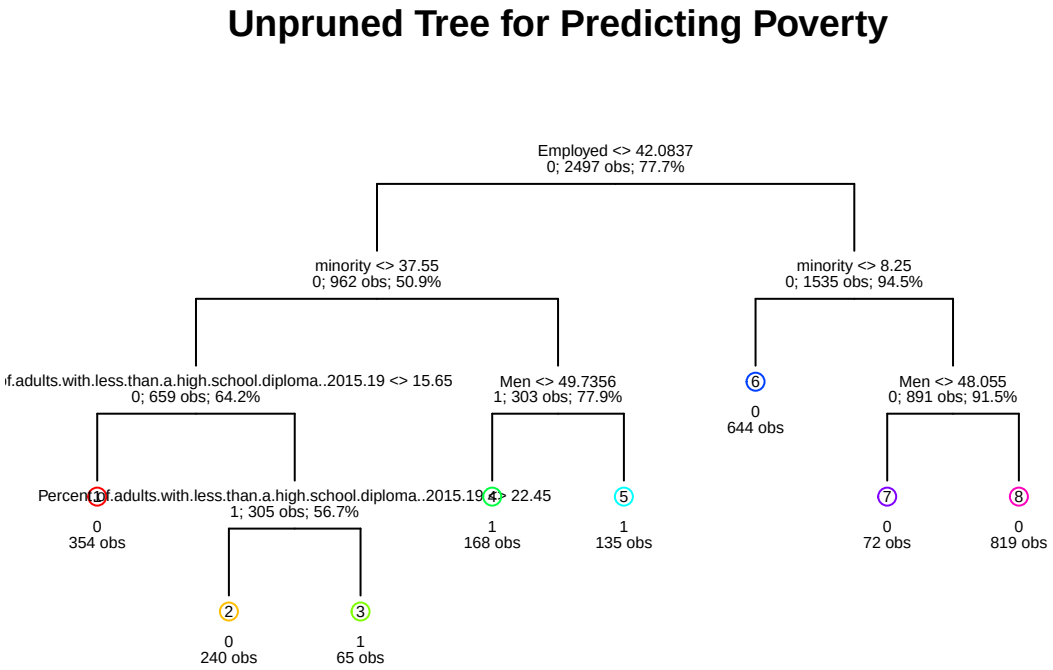
```
set.seed(123)
nfold <- 10
folds <- sample(cut(1:nrow(all.tr), breaks=nfold, labels=FALSE))
```

The following is the error rate function. And the object records is used to record the classification performance of each method in the subsequent problems.

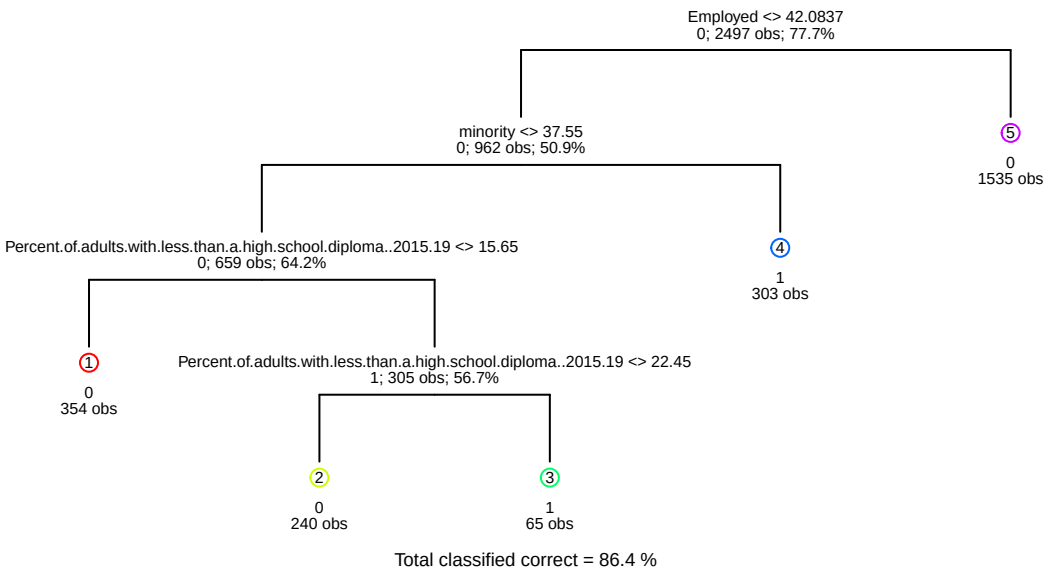
```
calc_error_rate = function(predicted.value, true.value){
  return(mean(true.value!=predicted.value))
}
records = matrix(NA, nrow=3, ncol=2)
colnames(records) = c("train.error","test.error")
rownames(records) = c("tree","logistic","lasso")
```

Classification

15. Decision Tree



Tree before pruning:



Tree After Pruning:

The *Training Error Rate* is 0.1358.

The *test Error Rate* is 0.1472.

Both Unpruned and Pruned Tree were first splitted by the "Employed" feature which implies that it could be the most influential variable in predicting Poverty.
Overall, the Pruned Tree tells us that unemployed minorities with lower education level may result in Poverty.

16. Logistic Regression

We first fit a logistic regression model.

```
##
## Call:
## glm(formula = Poverty ~ ., family = "binomial", data = all.tr)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.340  -0.458  -0.196  -0.045   3.484
##
## Coefficients:
##                                                                 Estimate
## (Intercept)                                                   144.28969
## Men                                                            -0.34884
## Professional                                                   -0.02790
## Employed                                                       -0.26022
## PrivateWork                                                    -0.10797
## SelfEmployed                                                   -0.13657
## FamilyWork                                                     -0.16249
## minority                                                        0.01905
## Percent.of.adults.with.less.than.a.high.school.diploma..2015.19 -0.98173
```

```
## Percent.of.adults.with.a.high.school.diploma.only..2015.19                -1.06762
## Percent.of.adults.completing.some.college.or.associate.s.degree..2015.19  -1.17431
## Percent.of.adults.with.a.bachelor.s.degree.or.higher..2015.19             -1.02735
##                                                                          Std. Error
## (Intercept)                                                               116.01155
## Men                                                                         0.02965
## Professional                                                                0.02104
## Employed                                                                    0.01914
## PrivateWork                                                                 0.01425
## SelfEmployed                                                                0.02789
## FamilyWork                                                                  0.17422
## minority                                                                    0.00421
## Percent.of.adults.with.less.than.a.high.school.diploma..2015.19             1.16024
## Percent.of.adults.with.a.high.school.diploma.only..2015.19                  1.15975
## Percent.of.adults.completing.some.college.or.associate.s.degree..2015.19    1.16027
## Percent.of.adults.with.a.bachelor.s.degree.or.higher..2015.19               1.15929
##                                                                          z value
## (Intercept)                                                                 1.24
## Men                                                                       -11.77
## Professional                                                               -1.33
## Employed                                                                  -13.59
## PrivateWork                                                                -7.58
## SelfEmployed                                                               -4.90
## FamilyWork                                                                 -0.93
## minority                                                                    4.53
## Percent.of.adults.with.less.than.a.high.school.diploma..2015.19            -0.85
## Percent.of.adults.with.a.high.school.diploma.only..2015.19                 -0.92
## Percent.of.adults.completing.some.college.or.associate.s.degree..2015.19   -1.01
## Percent.of.adults.with.a.bachelor.s.degree.or.higher..2015.19              -0.89
##                                                                          Pr(>|z|)
## (Intercept)                                                                 0.21
## Men                                                                      < 2e-16
## Professional                                                                0.18
## Employed                                                                 < 2e-16
## PrivateWork                                                              3.5e-14
## SelfEmployed                                                             9.7e-07
## FamilyWork                                                                  0.35
## minority                                                                 6.0e-06
## Percent.of.adults.with.less.than.a.high.school.diploma..2015.19             0.40
## Percent.of.adults.with.a.high.school.diploma.only..2015.19                  0.36
## Percent.of.adults.completing.some.college.or.associate.s.degree..2015.19    0.31
## Percent.of.adults.with.a.bachelor.s.degree.or.higher..2015.19               0.38
##
## (Intercept)
## Men                                                                      ***
## Professional
## Employed                                                                 ***
## PrivateWork                                                              ***
## SelfEmployed                                                             ***
## FamilyWork
## minority                                                                 ***
## Percent.of.adults.with.less.than.a.high.school.diploma..2015.19
## Percent.of.adults.with.a.high.school.diploma.only..2015.19
## Percent.of.adults.completing.some.college.or.associate.s.degree..2015.19
```

```
## Percent.of.adults.with.a.bachelor.s.degree.or.higher..2015.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2650.6  on 2496  degrees of freedom
## Residual deviance: 1457.0  on 2485  degrees of freedom
## AIC: 1481
##
## Number of Fisher Scoring iterations: 6
```

The significant variables are "Men", "Employed", "PrivateWork", "SelfEmployed", "minority".
"Employed" and "minority" seem to have consistency comparing to the influential variables in the Pruned Tree.
Increasing "Employed" by 1-unit implies multiply the odds by $e^{coefficient-estimate-of-Employed}$.
Increasing "minority" by 1-unit implies multiply the odds by $e^{coefficient-estimate-of-minority}$.

After calculating the error rates, the Records Matrix is:

```
##          train.error test.error
## tree          0.1358     0.1472
## logistic      0.1237     0.1376
## lasso             NA         NA
```

17. Now Consider a Lasso Regression Model
We first fit the model to select the best tuning parameter λ, and

the optimal value tuning parameter λ is 2×10^{-4}.

The coefficients of lasso regression are:

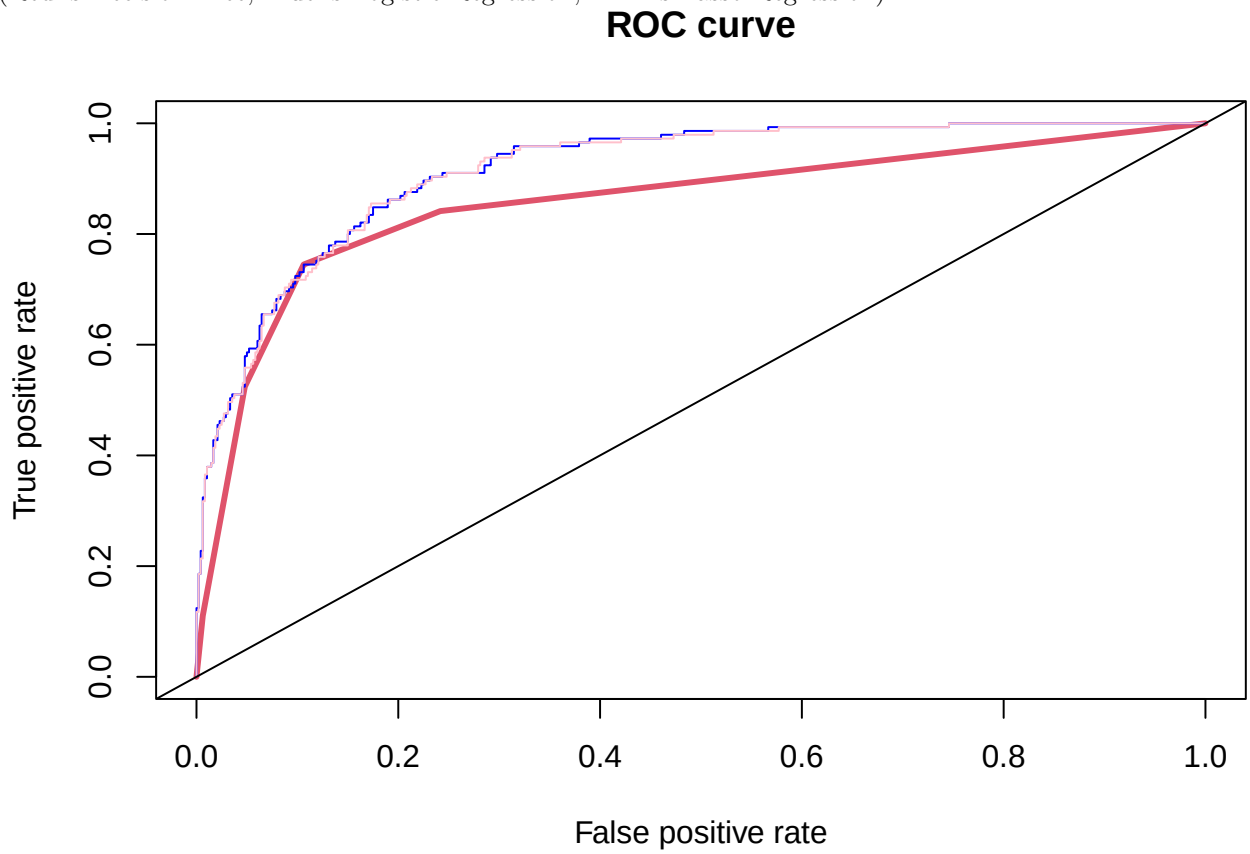
```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                                                       s1
## (Intercept)                                                    40.91527
## Men                                                            -0.34644
## Professional                                                   -0.02548
## Employed                                                       -0.25754
## PrivateWork                                                    -0.10669
## SelfEmployed                                                   -0.13537
## FamilyWork                                                     -0.15155
## minority                                                        0.01909
## Percent.of.adults.with.less.than.a.high.school.diploma..2015.19 0.04750
## Percent.of.adults.with.a.high.school.diploma.only..2015.19     -0.03747
## Percent.of.adults.completing.some.college.or.associate.s.degree..2015.19 -0.14345
## Percent.of.adults.with.a.bachelor.s.degree.or.higher..2015.19   .
```

The we find out the non-zero coefficients are "Men", "Poverty", "Professional", "Employed", "PrivateWork", "SelfEmployed", "FamilyWork", "minority", "Percent of adults with less than a high school diploma, 2015-19", "Percent of adults with a high school diploma only, 2015-19", "Percent of adults completing some college or associate's degree, 2015-19".
And the only zero coefficient feature is "Percent of adults with a bachelor's degree or higher, 2015-19".

The penalized logistic regression does model selection and takes less predicting variables than the unpenalized

logistic regression.
Higher education has a less significant negative influence on Poverty than the postive influence of the lower education has on Poverty.

18. ROC curve for the decision tree, logistic regression and LASSO logistic regression using predictions on the test data.
(Red is Decision Tree, Blue is Logistic Regression, Pink is Lasso Regression)



The current error rates for Decision Tree, Logistic Regression, and Lasso Regression:

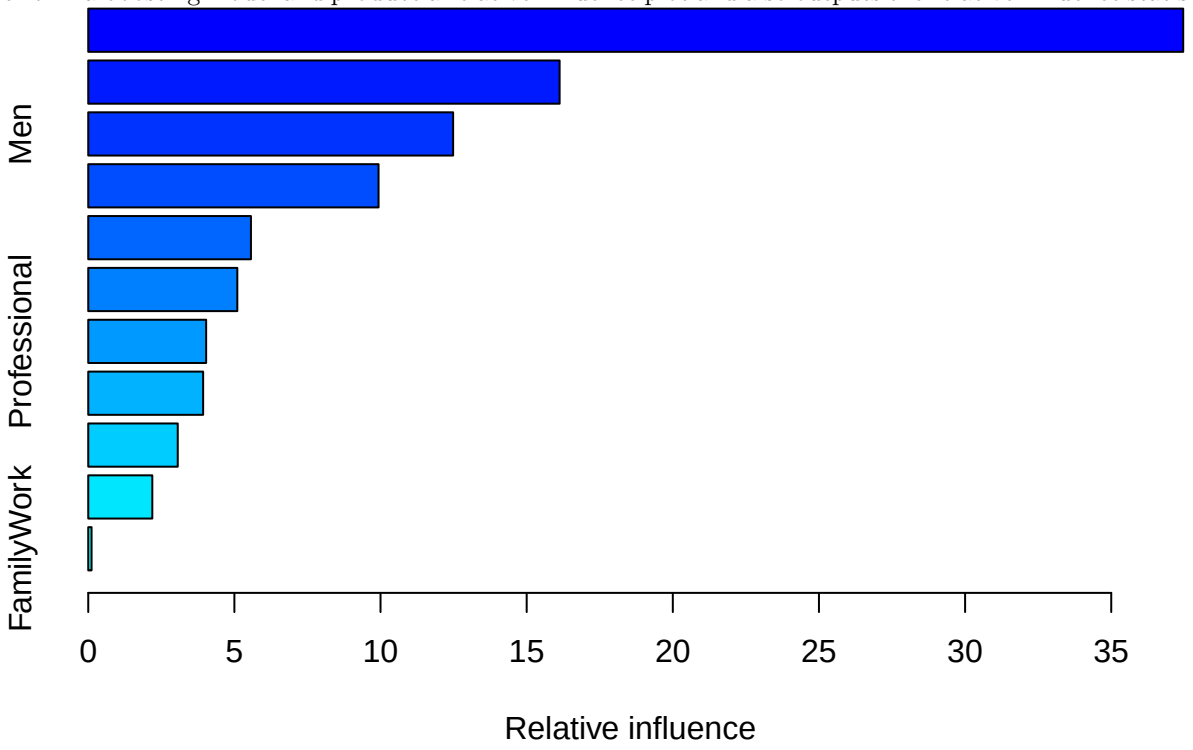
```
##          train.error test.error
## tree          0.1358     0.1472
## logistic      0.1237     0.1376
## lasso         0.1241     0.1408
```

According to the error rate records matrix, we can see that logistic regression with the lowest test error rate, has a better performance than the other two methods.
Decision Tree and Lasso have higher test error rate.
If we are more interested in predicting the Poverty using census data, logistic regression is a good method to use because in the model we fit previously, all the significant variables are from census data.
Decision Tree seems to be less appropriate for predicting Poverty in this situation, while Logistic and Lasso Regression turns out to provide a relatively better AUC.

19. Other ways to Fit a Model

(a) Fitting the model using Boosting

We perform a boosting model and produce a relative influence plot and also outputs the relative influence statis-



tics.

```
##
## Employed
## Percent.of.adults.with.less.than.a.high.school.diploma..2015.19                  Percent.of.adults.
## Men
## minority
## Percent.of.adults.with.a.bachelor.s.degree.or.higher..2015.19                      Percent.of.adults
## Percent.of.adults.completing.some.college.or.associate.s.degree..2015.19 Percent.of.adults.completing
## Professional
## PrivateWork
## SelfEmployed
## Percent.of.adults.with.a.high.school.diploma.only..2015.19                             Percent.of.adu
## FamilyWork
##                                                                          rel.inf
## Employed                                                                 37.4622
## Percent.of.adults.with.less.than.a.high.school.diploma..2015.19          16.1266
## Men                                                                      12.4835
## minority                                                                  9.9318
## Percent.of.adults.with.a.bachelor.s.degree.or.higher..2015.19             5.5663
## Percent.of.adults.completing.some.college.or.associate.s.degree..2015.19  5.1002
## Professional                                                              4.0331
## PrivateWork                                                               3.9303
## SelfEmployed                                                              3.0617
## Percent.of.adults.with.a.high.school.diploma.only..2015.19                2.1913
## FamilyWork                                                                0.1129
```

From the above graph, we can see that "Employed" is is by far the most important variable. Comparing to the other variables, "Employed" has a relative relative influence to the response value "Poverty".

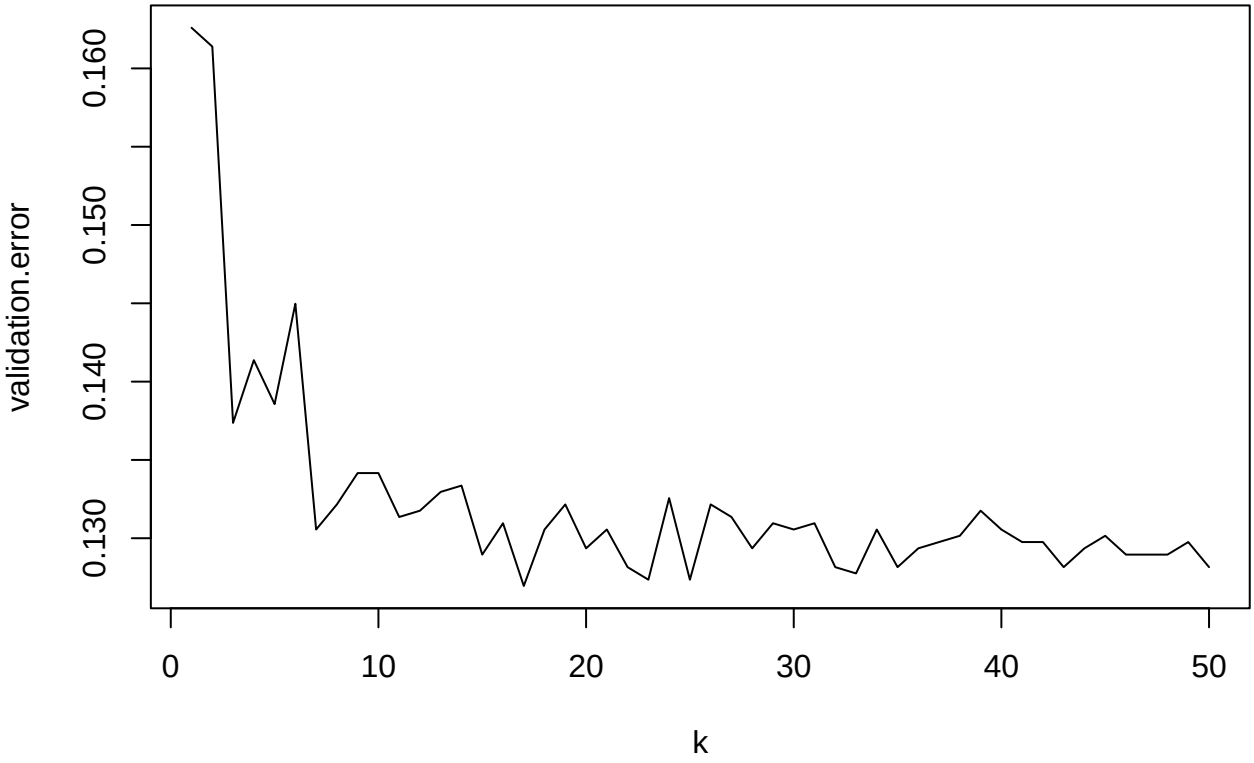
Then we would like to know how does this boosting model perform by calculating its error rate using training

and testing data.

```
##          train.error test.error
## tree          0.1358     0.1472
## logistic      0.1237     0.1376
## lasso         0.1241     0.1408
## boosting      0.1005     0.1376
```

(b) Fitting the model using KNN
We perform a boosting model and produce a relative influence plot and also outputs the relative influence statistics.

LOOCV to find the best K:



Then we choose the number of k to be:

```
## [1] 17
```

Confusion Matrix of Training Data:

```
##           true
## predicted    0    1
##         0 1861  208
##         1   79  349
```

Confusion Matrix of Test Data:

```
##           true
## predicted   0   1
##         0 451  63
##         1  29  82
```

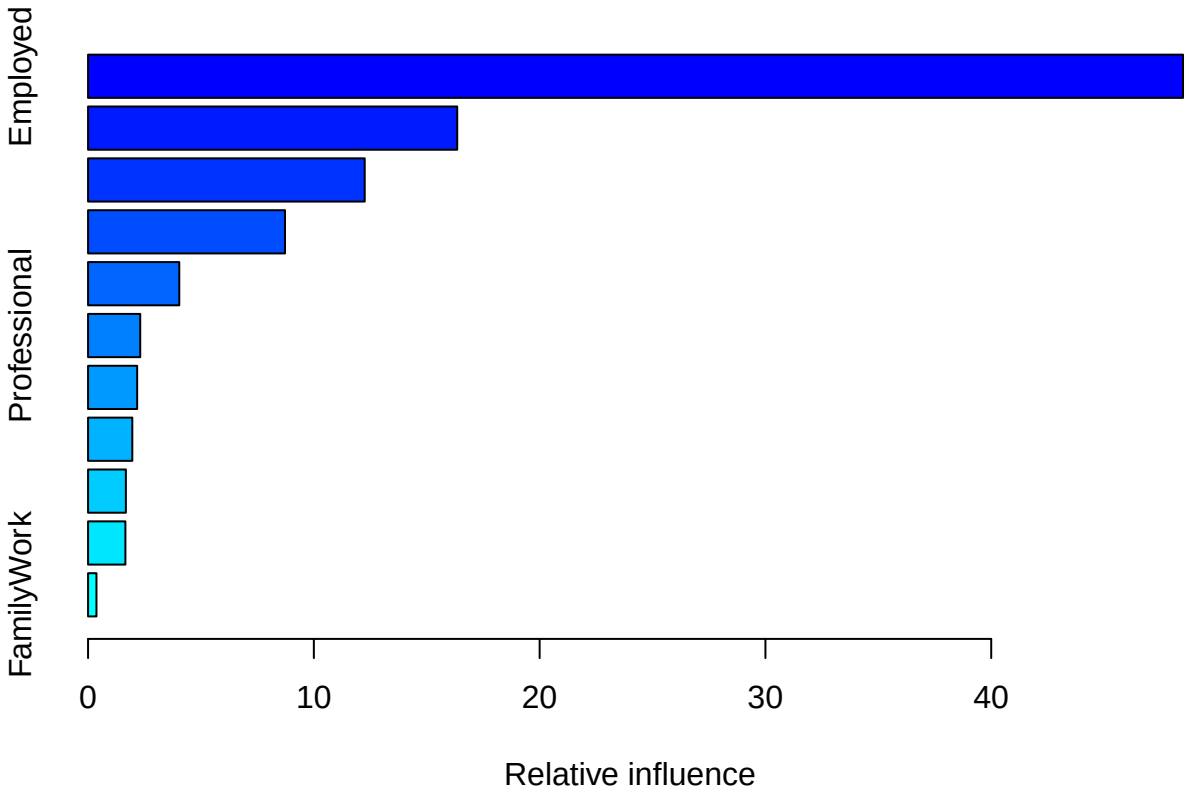
Now we can see the error rate for all the methods we used is:

```
##          train.error test.error
## tree          0.1358     0.1472
## logistic      0.1237     0.1376
## lasso         0.1241     0.1408
## boosting      0.1005     0.1376
## knn           0.1149     0.1472
```

Comparison
According to the new error rate record matrix, while logistic regression and boosting have the same test error rate, boosting seems to have a lower training error rate which may implies overfitting in the boosting model compare to the logistic regression model.
And interestingly, Decision tree and KNN have the same test error as well, and similarly, KNN seems to have a lower training error rate which may also implies overfitting in the KNN model compare to the decision Tree model.
Overall, Logistic Regression gives us a better performance on this data set.

20. Interesting Questions

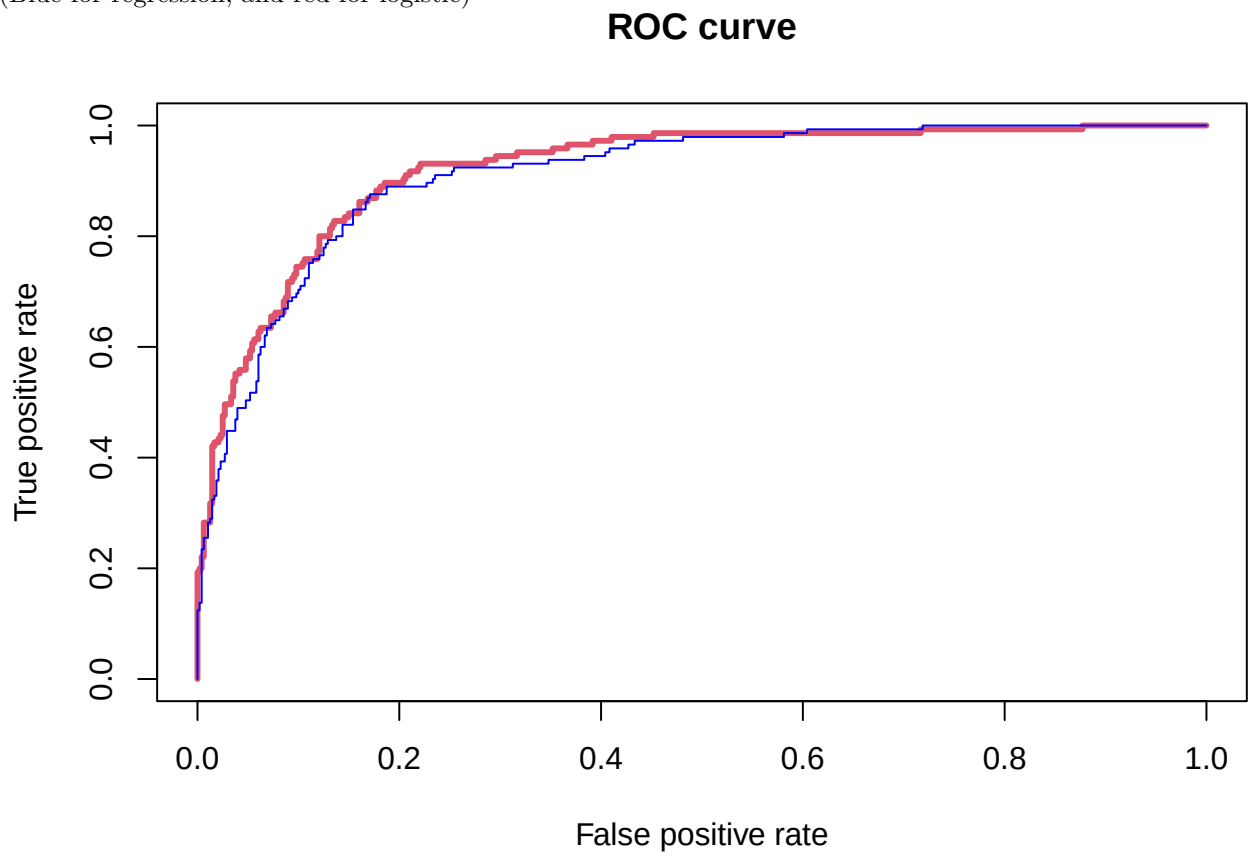
(a) First, we want to use Boosting model to predict the actual value of Poverty by County, and compare it with the Boosting model in the classification setting.




```
##
## Employed
## `Percent of adults with less than a high school diploma, 2015-19`                  `Percent of adul
## minority
```

```
## Men
## PrivateWork
## Professional
## `Percent of adults with a bachelor's degree or higher, 2015-19`                      `Percent of adu
## SelfEmployed
## `Percent of adults completing some college or associate's degree, 2015-19` `Percent of adults comple
## `Percent of adults with a high school diploma only, 2015-19`                           `Percent of
## FamilyWork
##                                                                                  rel.inf
## Employed                                                                         48.4946
## `Percent of adults with less than a high school diploma, 2015-19`                16.3486
## minority                                                                         12.2522
## Men                                                                               8.7223
## PrivateWork                                                                       4.0392
## Professional                                                                      2.3097
## `Percent of adults with a bachelor's degree or higher, 2015-19`                   2.1741
## SelfEmployed                                                                      1.9592
## `Percent of adults completing some college or associate's degree, 2015-19`        1.6746
## `Percent of adults with a high school diploma only, 2015-19`                      1.6539
## FamilyWork                                                                        0.3716
```

The ROC curve for Logistic Boosting and Regression Boosting:
(Blue for regression, and red for logistic)



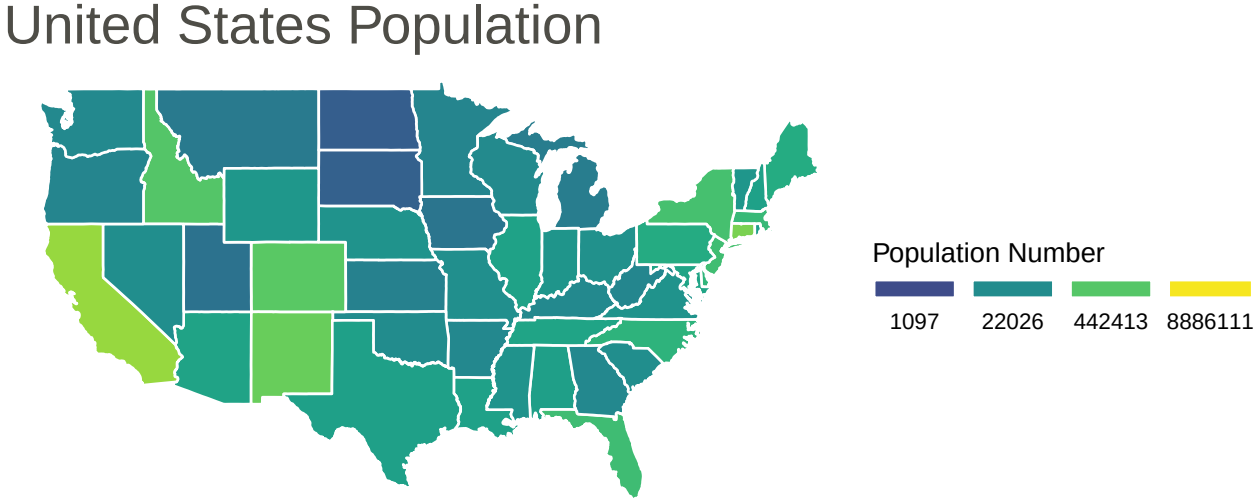
The AUC for Logistic Boosting is 0.920158045977012.
The AUC for Regression Boosting is 0.91008620689655.

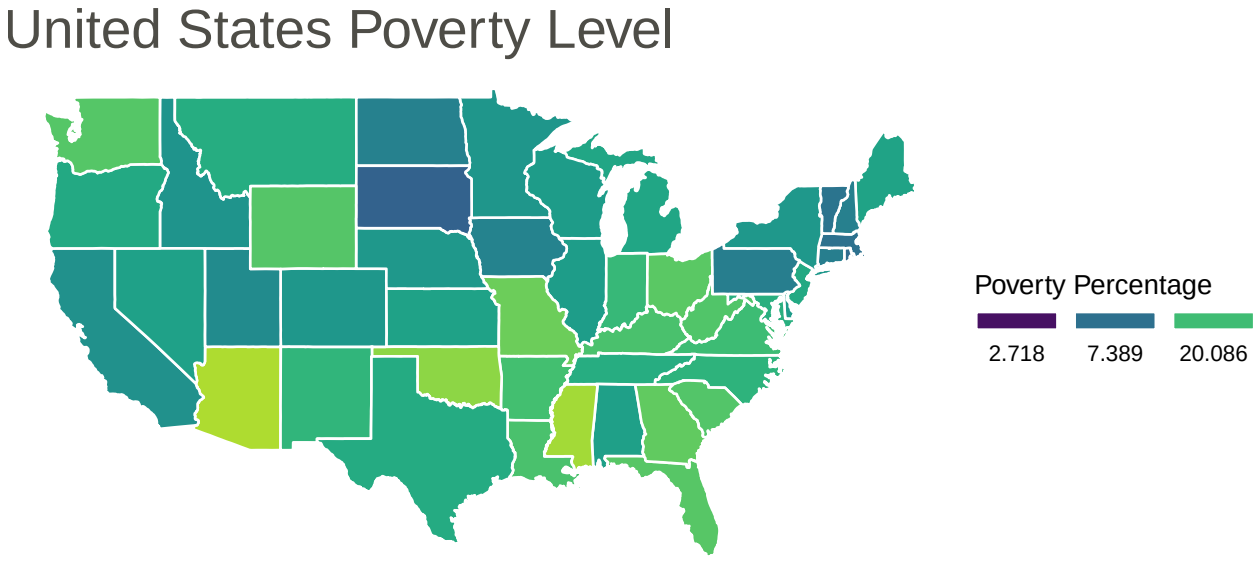
Apparently, we obtain a higher AUC value for the binary response variable than the actual response while

using the same modeling method, which is quite different from our previous assumption of the error rate for the result. One of possible reason could be this problem is more of a identification problem, which may works better with Classification Algorithms.
Specifically for the current dataset we are using, we would prefer classification models because this dataset we use is basically identification problem where we want to predict if there is Poverty exists.
On the other hand, if we consider problems (not in the datasets we currently used) that predict house pricing might perform better using regression method, because house price prediction is a continuous data and the output variable is continuous nature or real value.
As a result, choosing Regression or Classification method is usually based on what kind of output variables we want to predict.

After calculate all the training and test rate of the various models, we want to find out most related variables in order to improve our models and make a better prediction.
Since we did not use "TotalPop" as one of of predicting variables in fitting the models to predict "Poverty", we first want to find out how the Total Population related to the percentage of Poverty in different States.

(b) Display the Poverty level of each State using the ggplot and compare it with the Population Level graph. Will more population result in a higher Poverty level?
In this case, we will use the census.clean without make the Poverty column into level 0 and 1 (using the actual vales), then plot the graph.
The graph is shown below with comparison to the previous graph of Population:





From the above graphs, we can see that there might exists some correlation between the Total Population of each states and percentage of Poverty. For example, California has a really high population and a relatively high poverty percentage. Also for South Dakota, it has a lower population level and a relatively low poverty percentage. As a result, we want to discuss if there exists correlation between Total Population and Poverty in each states using the bootstrapping method below.

(c) Then, we want to use Bootstrap Method to discuss the correlation between Poverty and three variables we are interested in.

1. Poverty & Total Population

As we discussed in the previous section, we want to see if there is correlation between Poverty and Total Population.

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = all1, statistic = fc1, R = 500)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1* -0.06665 -0.002727     0.01647
```

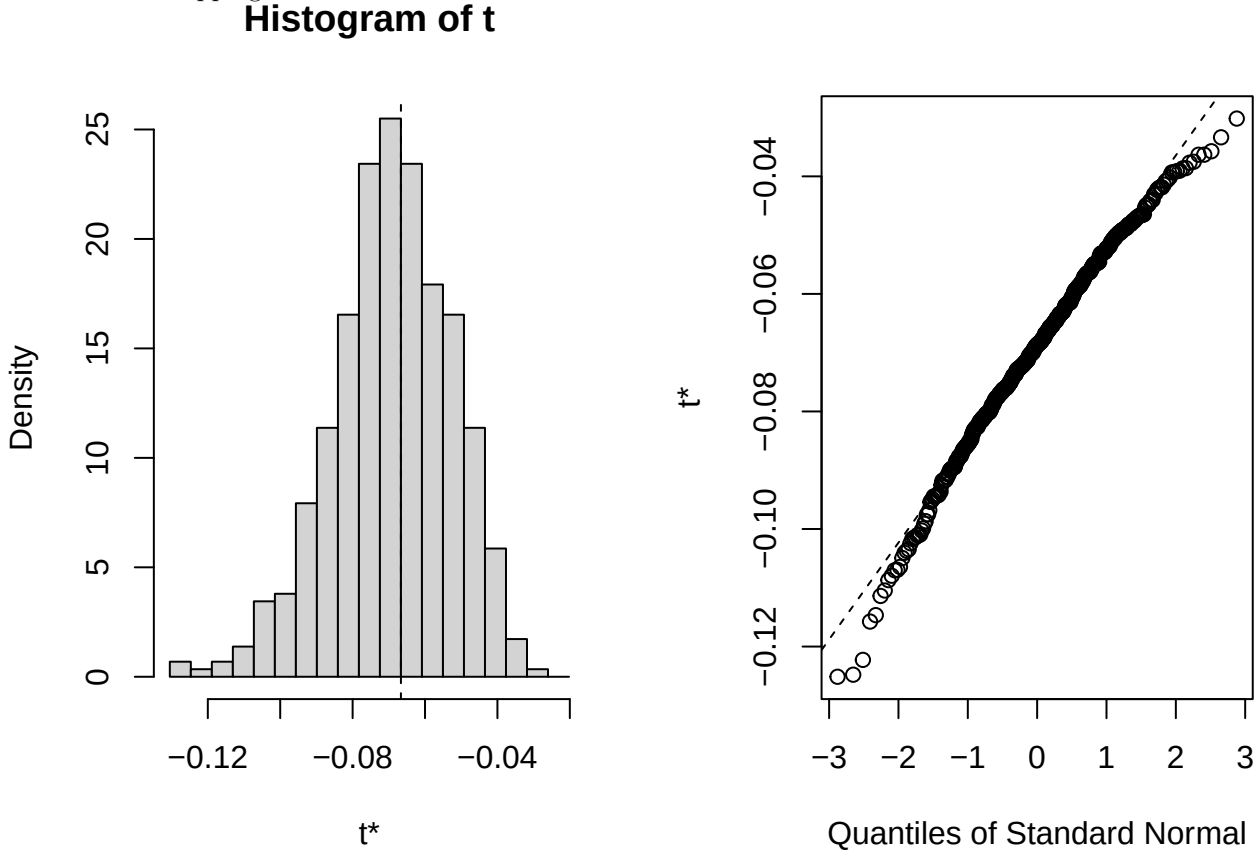
The Characteristic of the Coefficient:

```
## [1] -0.12514 -0.03017
```

```
## [1] -0.06938
```

```
## [1] 0.01647
```

Plot of Bootstrapping:



Confidence Interval of Coefficient:

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 500 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootcorr_pop, type = c("norm"))
##
## Intervals :
## Level      Normal
## 95%   (-0.0962, -0.0316 )
## Calculations and Intervals on Original Scale
```

Range of the correlation coefficient: [-0.12514 -0.03017]
Mean: -0.06938
Standard deviation: 0.01647
95% confidence interval: [-0.0962, -0.0316]

As we can see, the range of the correlation coefficient is from -0.13 to -0.03, and the 95% CI is from -0.096 to -0.032.
The statistics suggest that these two variables are from sightly to moderately negative correlated. From the plot we have in problem 8, and the plot in problem We thought poverty would have a more strong correlation with the total population, but it turns out not like that.

2. Poverty & Percent of adults with less than a high school diploma, 2015-19
Next, we want to see how correlated are Poverty with "Percent of adults with less than a high school diploma, 2015-19", which was shown to have strong influence in boosting method above.

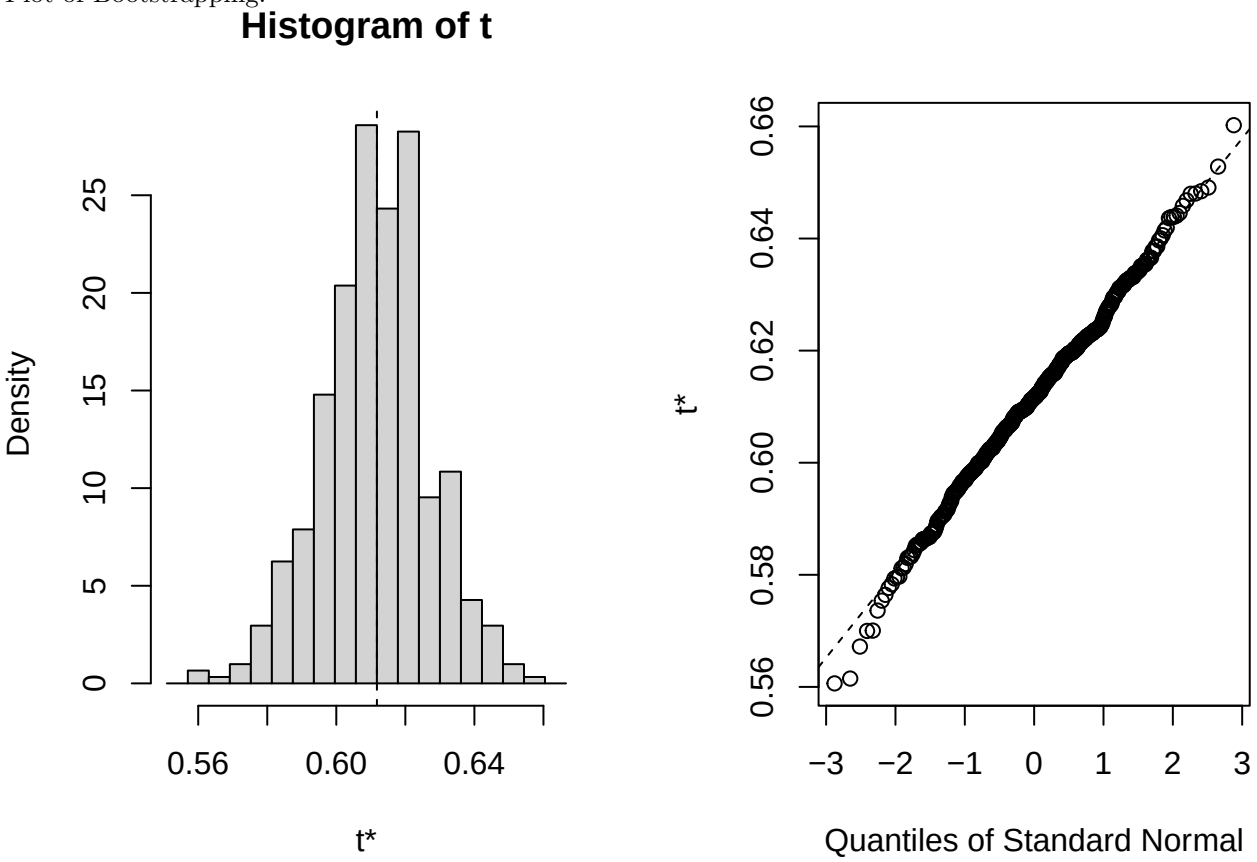
Characteristic of Coefficient:

```
## [1] 0.5606 0.6602
```

```
## [1] 0.6115
```

```
## [1] 0.01544
```

Plot of Bootstrapping:



Confidence Interval of Coefficient:

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 500 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootcorr_ed, type = c("norm"))
##
## Intervals :
## Level      Normal
## 95%   ( 0.5817,  0.6423 )
## Calculations and Intervals on Original Scale
```

As we can see, the range of the correlation coefficient is from 0.5606 to 0.6602, and the 95% CI is from 0.5817

to 0.6423.
The statistics suggest that these two variables have some strong positive correlation, which implies that a low education level has a positive influence on poverty.

3. Poverty & Employed
Employed also shown to be a strong factor in boosting method, so we performed boostrap on it as well:

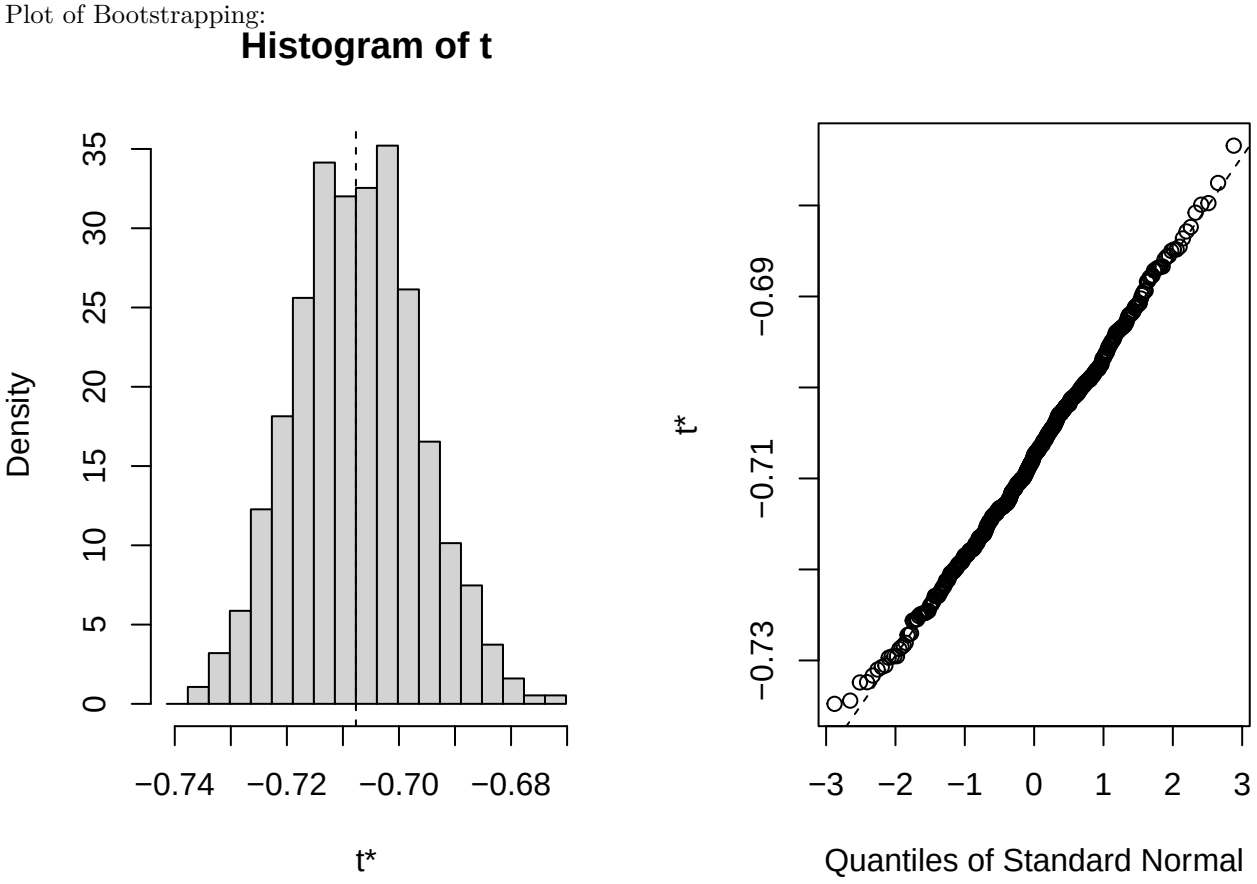
Characteristic of Coefficient:

```
## [1] -0.7348 -0.6734
```

```
## [1] -0.7075
```

```
## [1] 0.01097
```

Plot of Bootstrapping:



Confidence Interval of Coefficient:

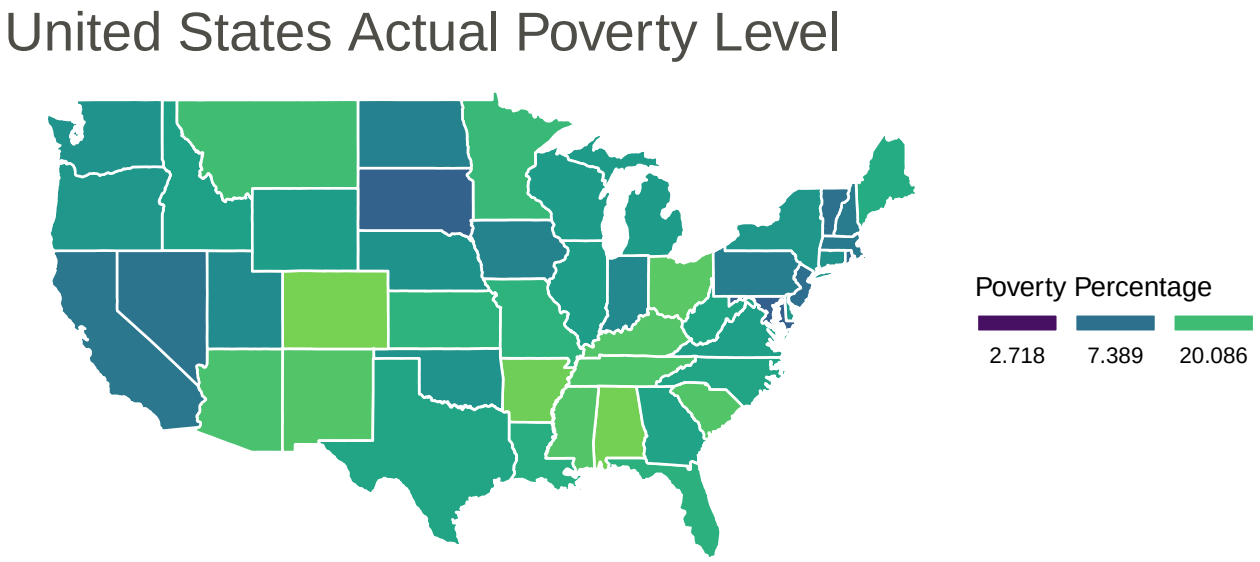
```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 500 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootcorr_em, type = c("norm"))
##
## Intervals :
```

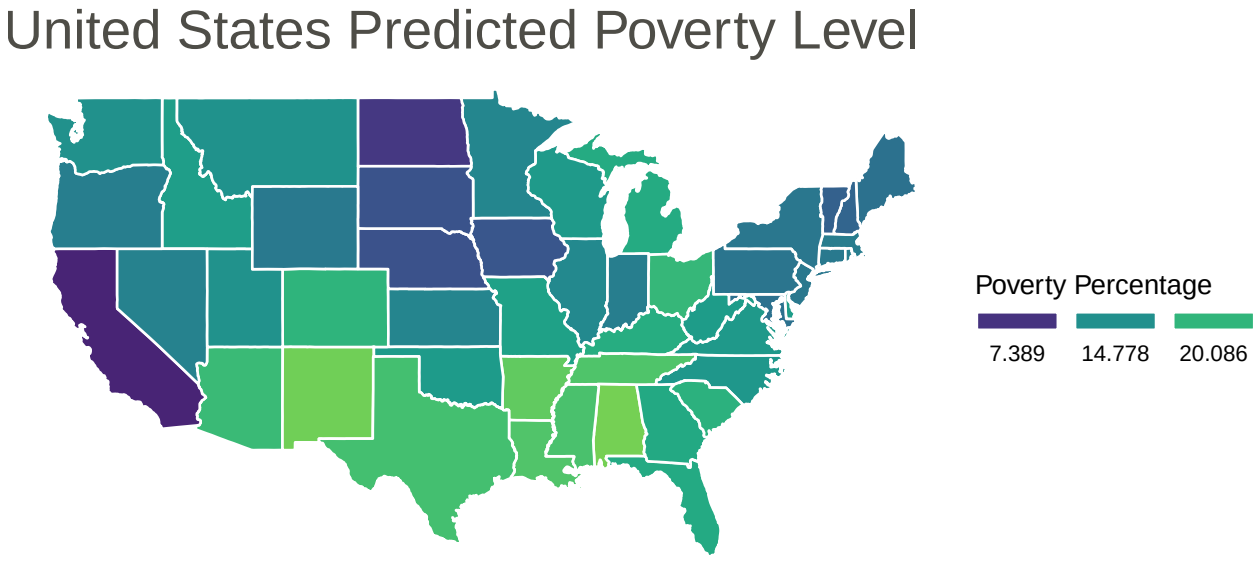
```
## Level      Normal
## 95%   (-0.7293, -0.6863 )
## Calculations and Intervals on Original Scale
```

As we can see, the range of the correlation coefficient is from -0.7348 to -0.6734, and the 95% CI is from -0.7293 to -0.6863.
The statistics suggest that these two variables have some strong negative correlation, which implies that employed groups are less possible to have poverty.

21. Interpret and discuss any overall insights gained in this analysis and possible explanations.

a. We use the Regression Boosting Model to predict the percentage of the Poverty and compare with the actual percentage of Poverty.





As we can see in the above graphs, California has a relatively low predicted Poverty percentage compare to

the actual high percentage in Poverty, which means the prediction using regression might cause some false prediction. One of possible reason could be since this problem is more of a identification problem, so it might not work as good as we assumed for boosting method.

b. One result that we did not expected was the influence of higher education level has a lighter influence on poverty than the lower education level. Before reading the result of LASSO regression, the only thing we knew was LASSO can help to reduce coefficient amount, but now we learned that it could also show some relative significance between coefficients with similar measure.

c. Before we use bootstrap to find the correlation between Poverty and Total Population of each States, we thought a larger population might result in a higher percentage of Poverty because more people means higher possibility of identify as "Yes" in Poverty. However, after we use bootstrap to calculate the confidence interval of the coefficients, we realize that Total Population has a really low correlation compare to the other variables, such as "Unemployed" and "Percent of adults with less than a high school diploma, 2015-19" that we are interested in.

d. We might be able to compare some other aspects that have relations with poverty from some industry data. In this data set, we complied several variables from census and education, which mainly focused on the lifestyle and identification of the population (race, transportation, jobs, gender, and education background). In addition to that, for example, the population in the US who are in the Medicaid insurance group, which is a public program designed for the low-income population. The industry data tends to have a cleaner structure and has a more direct relation with the financial situation. Other industry data examples could be government financial aid and support programs.