DATA2020 HW4

Zhirui Li

1. Multinomial Models

(a) Give the distribution of Y by finding Pr(Y = i|X) for any i ∈ {0,1,. . . ,m}. How does this relate to the model for logistic regression? This model is similar to logistic regression and it is called multinomial logistic regression. The only difference is that for logistic regression, it is binary classification; here, the response variable has more than two classes (multiclass logistic regression). To do this, we first select a single class to serve as the baseline (without loss of generality, we select the Mth class for this role), then, we can see that the log odds between any pair of classes is linear in the features.

More specifically, E[Y = i|X] = Pr(Y = i|X) equals (for i = 0, 1, . . . , m-1) :

$$Pr(Y = i|X) = \frac{exp(\eta_i)}{1 + \sum_{j=1}^{m-1} exp(\eta_j)}$$

E[Y = M|X] = Pr(Y = M|X) equals: (M represents baseline class here)

$$Pr(Y = M|X) = \frac{1}{1 + \sum_{j=1}^{m-1} exp(\eta_j)}$$

(b) Given that Y ∈ {0,i}, let μ_i = Pr(Y = i|X,Y ∈ {0,i}) and find the link function g such that $g^{-1}(\eta_i) = \mu_i$.

Since we have:

$$g^{-1}(\eta_i) = \mu_i = Pr(Y = i|X, Y \in \{0, i\})$$

$$E[Y = i|X] = Pr(Y = i|X) = \frac{exp(\eta_i)}{1 + \sum_{j=1}^{m-1} exp(\eta_j)} = g^{-1}(\eta_i)$$

The link function for the multinomial logistic regression is:

$$g(Pr(Y = i|X, Y \in \{0, i\})) = \eta_i = log(\frac{Pr(Y = i|X)}{1 - Pr(Y = i|X)}) = log(\frac{Pr(Y = i|X)}{Pr(Y = 0|X)})$$

In summary, the link function is the logit function (log odds).

(c) Is this equivalent to fitting m separate logistic regression models? Why or why not? This is not equivalent to fitting m separate logistic regression models because multinomial logistic regression allows us to fit all classes of response together. Moreover, when we just fit m separate logistic regression models, we are essentially predicting whether this observation belongs to the baseline class or not. Moreover, for each i (each class), $Pr(Y = i|X) + Pr(Y = 0|X)$ does not equal to 1.

2. Ridge and Lasso Regression

(a) Write out the ridge regression optimization problem in this setting. Argue that in this setting, the ridge coefficient estimates satisfy $\beta^1 = \beta^2$. SEE ANOTHER UPLOADED DOCUMENT

(b) Write out the lasso optimization problem in this setting. Argue that in this setting, the lasso coefficients β^1 and β^2 are not unique. In other words, there are many possible solutions to the optimization problem. Describe these solutions. SEE ANOTHER UPLOADED DOCUMENT

3. Model Selection

We remove all the entries with missing values in the dataset.

We convert all categorical variables to factors.

For the response variable, we first convert it to dummy variable where 0 indicates voting for Obama and 0 means voting for Romney. Then, we convert it to factor to ensure the task is classification

```
train = election[split == 0, ]
test = election[split == 1, ]
dim(train)
```

```
## [1] 2560   29
```

```
dim(test)
```

```
## [1] 640  29
```

We split the whole dataset into training set and testing set. The training set contains 2560 entries and testing set contains 640 entries.

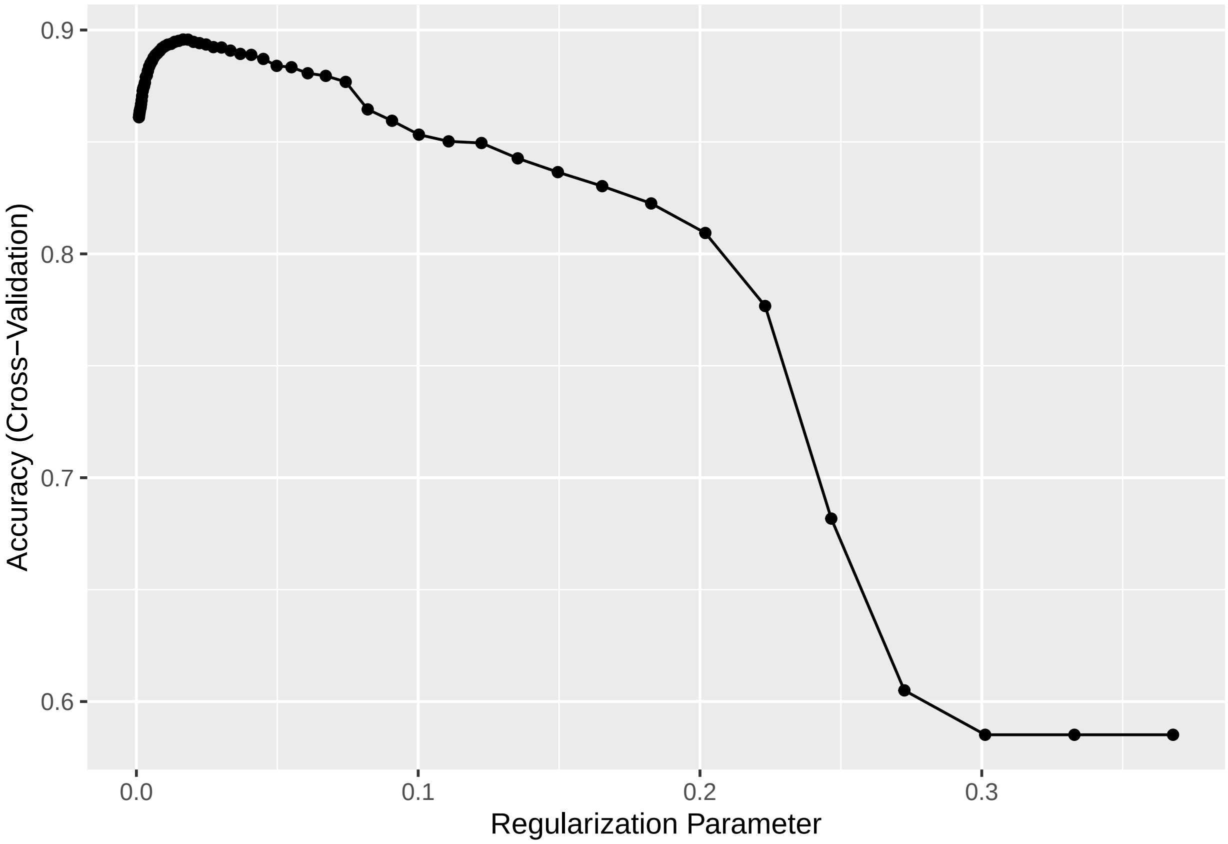
```
folds <- createFolds(train$vote12, k=10)
```

We create 10 folds for cross validation.

```
lambda_seq <- exp(seq(-7,-1,0.1))
lambda_seq
```

```
##  [1] 0.000911882 0.001007785 0.001113775 0.001230912 0.001360368 0.001503439
##  [7] 0.001661557 0.001836305 0.002029431 0.002242868 0.002478752 0.002739445
## [13] 0.003027555 0.003345965 0.003697864 0.004086771 0.004516581 0.004991594
## [19] 0.005516564 0.006096747 0.006737947 0.007446583 0.008229747 0.009095277
## [25] 0.010051836 0.011108997 0.012277340 0.013568559 0.014995577 0.016572675
## [31] 0.018315639 0.020241911 0.022370772 0.024723526 0.027323722 0.030197383
## [37] 0.033373270 0.036883167 0.040762204 0.045049202 0.049787068 0.055023220
## [43] 0.060810063 0.067205513 0.074273578 0.082084999 0.090717953 0.100258844
## [49] 0.110803158 0.122456428 0.135335283 0.149568619 0.165298888 0.182683524
## [55] 0.201896518 0.223130160 0.246596964 0.272531793 0.301194212 0.332871084
## [61] 0.367879441
```

Above values are the lambda values to test for for the lasso and ridge regularization.



From above graph, we can see that lambda value close to 0.01 yields the highest cross validation accuracy for the lasso regression.

```
lasso_mod_best <- lasso_mod_cv$finalModel$lambdaOpt
lasso_mod_best
```

```
## [1] 0.01657268
```

Above value is the best lambda value for the lasso regularization.

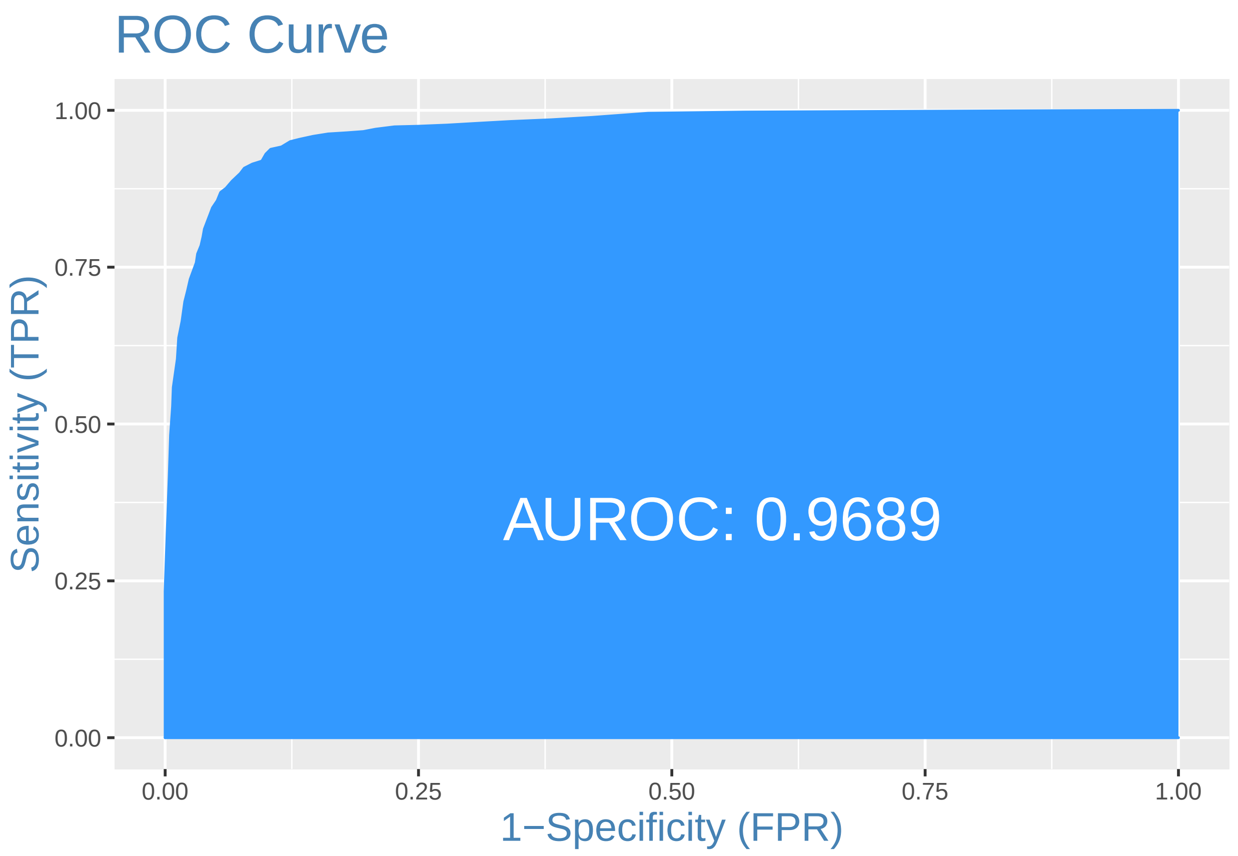
```
## 63 x 1 sparse Matrix of class "dgCMatrix"
##                                                         s1
## (Intercept)                                    -0.93568242
## ...1                                             .
## abortionMore conds                               .
## abortionNever                                    .
## abortionSome conds                              0.11053545
## congappAppWeak                                   .
## congappDisappStr                                 .
## congappDisappWk                                  .
## dem                                            -1.01960992
## deathpenAppStrng                                0.09509861
## deathpenOppose                                   .
## deathpenOppStrng                               -0.03695456
## age30-39                                         .
## age40-49                                         .
## age50-59                                         .
## age60-69                                         .
## age70-older                                      .
## educHS or less                                 .
## educSome coll                                  .
## blackYes                                      -0.48503386
## hispanicYes                                   -0.10954981
## incomeQuint2                                   .
## incomeQuint3                                   .
## incomeQuint4                                   .
## incomeQuint5                                   0.05052678
## veteran2. No                                   .
## econ_pastSame                                  0.19020470
## econ_pastWorse                                 0.61286037
## unemp_past2. About the same                    .
## unemp_past3. Worse                             0.17763583
## envir_gwarm2. Probably hasn't been happening   0.07639220
## religion2. Other Christian                     0.07337986
## religion3. Other religion                      .
## religion4. Not religious                      -0.01977507
## gay_marryYes                                  -0.37181244
## gay_adopt1                                     .
## genderMale                                     .
## gun_control2. Easier                           .
## gun_control3. Keep these rules about the same  0.15531017
## aca_app2. Favor moderately                    -0.08258346
## aca_app3. Favor a little                       .
## aca_app4. Neither favor nor oppose             0.15493691
## aca_app5. Oppose a little                      0.05397792
## aca_app6. Oppose moderately                    0.35911343
## aca_app7. Oppose a great deal                  0.88654815
## immig_citizen2. Oppose                         .
## immig_citizen3. Neither favor or oppose        .
## immig_jobs2. Very                              .
## immig_jobs3. Somewhat                          .
## immig_jobs4. Not at all                        .
## marriedYes                                     .
## owngun2. No                                   -0.03121594
## govwaste2. Waste some                          .
## govwaste3. Don't waste very much               .
## usworld_stay2. Disagree                        0.00045615
## trust_social2. Most of the time                .
## trust_social3. About half the time             .
## trust_social4. Some of the time                .
## trust_social5. Never                           .
## govcorrpt2. Most                               .
## govcorrpt3. About half                         .
## govcorrpt4. A few                              .
## govcorrpt5. None                               .
```

We can see that under lasso regression, a lot of coefficients have been eliminated.

```
lasso_probs <- predict(lasso_mod_cv, train, type="prob")
lasso_pred <- ifelse(lasso_probs[,2]>0.5, 1, 0) # 1 means vote for Romney
sum(lasso_pred == train$vote12)/nrow(train)
```

```
## [1] 0.9144531
```

The accuracy for the training data is pretty high: 91.45%.

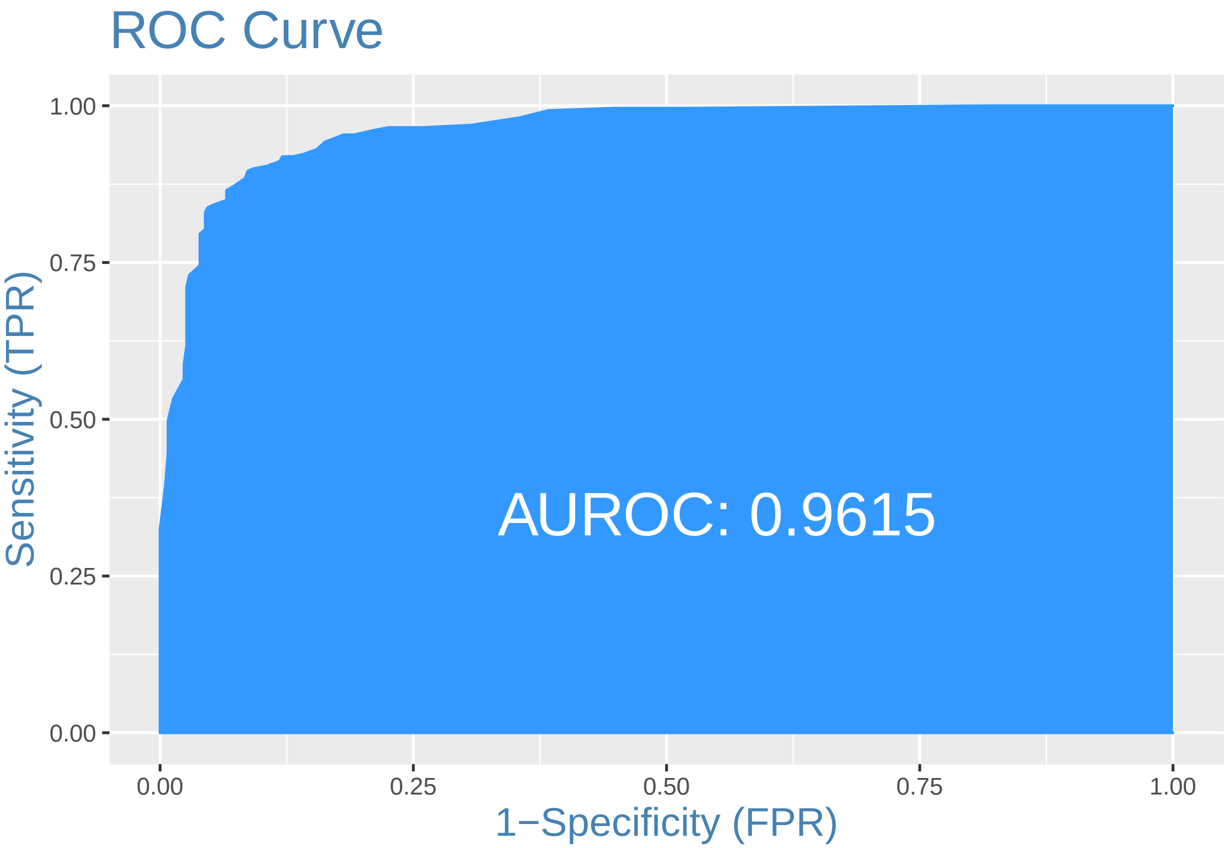


The AUROC for the training data is pretty high as well: 96.89%. Both measures indicates that the lasso regression chosen by the cross validation has done a really decent job on the training data.

```r
lasso_probs_test <- predict(lasso_mod_cv, test, type="prob")
lasso_pred_test <- ifelse(lasso_probs_test[,2]>0.5, 1, 0)  # 1 means vote for Romney
sum(lasso_pred_test == test$vote12)/nrow(test)
```

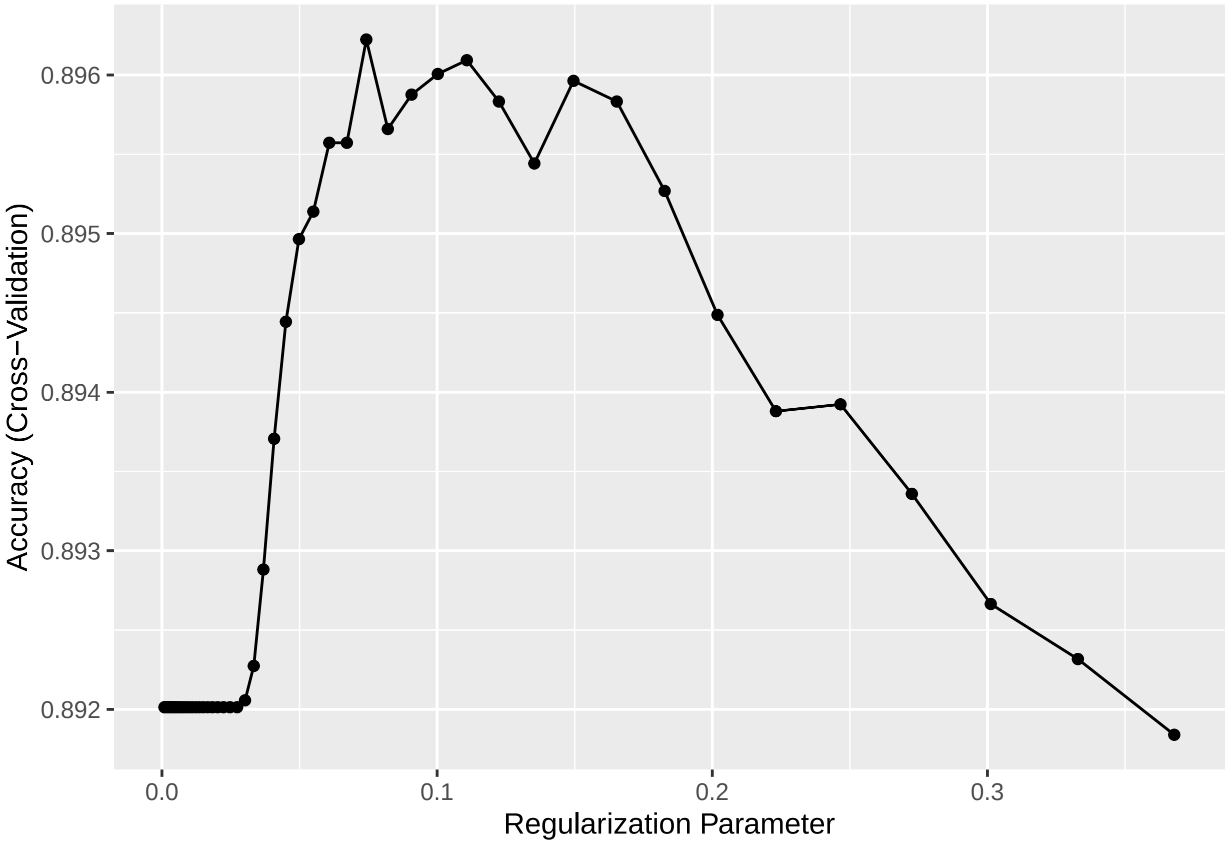
```
## [1] 0.90625
```

The accuracy for the testing data is pretty high: 90.63%.



The AUROC for the testing data is pretty high as well: 96.15%. Both measures indicates that the lasso regression chosen by the cross validation has done a really decent job on the testing data.

After we use lasso regularization to train the model, we try ridge next to compare the results.



From above graph, we can see that setting lambda to be around 0.07 seems to have the best cross-validation accuracy.

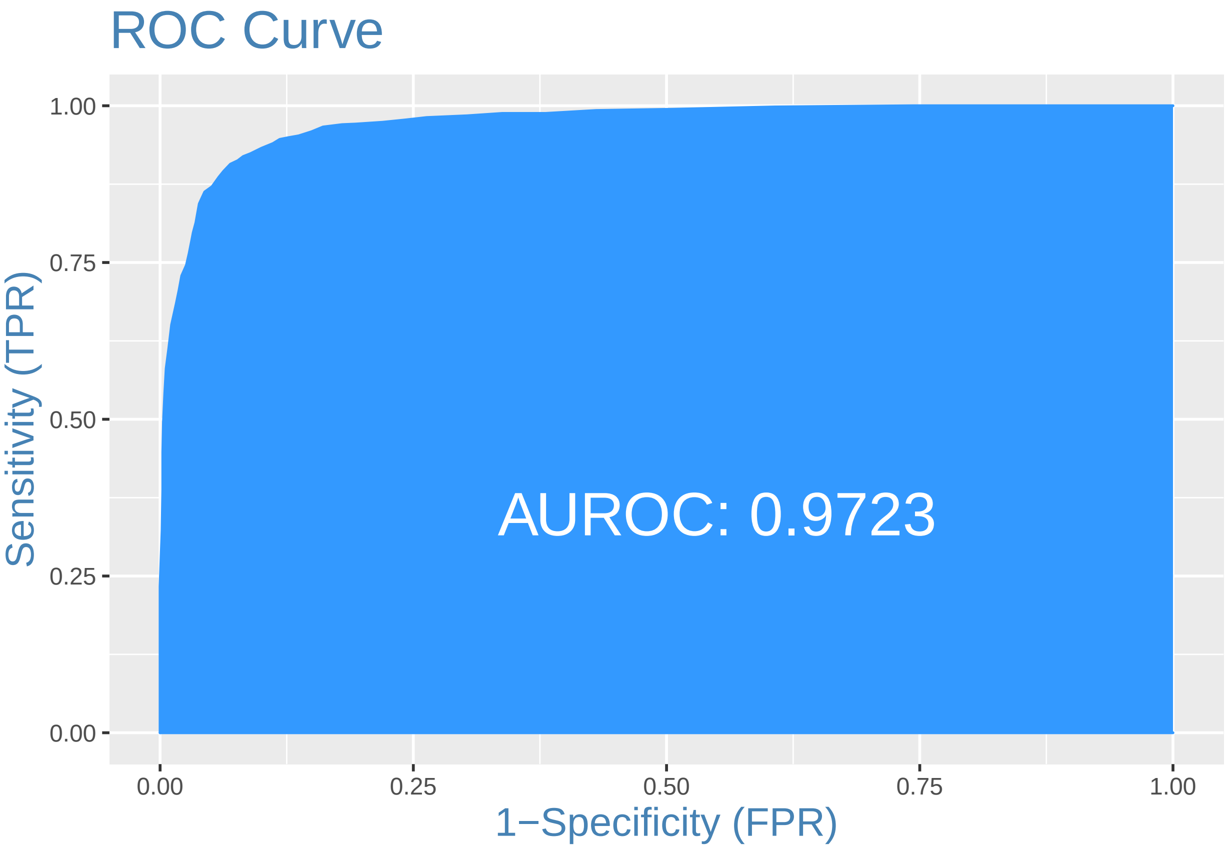
```
## 63 x 1 sparse Matrix of class "dgCMatrix"
##                                                         s1
## (Intercept)                                   -1.302116248
## ...1                                           0.044323892
## abortionMore conds                             0.081105591
## abortionNever                                  0.102736444
## abortionSome conds                             0.196979569
## congappAppWeak                                -0.004183322
## congappDisappStr                               0.007585518
## congappDisappWk                                0.067184401
## dem                                           -0.619887888
## deathpenAppStrng                               0.148790108
## deathpenOppose                                -0.051638943
## deathpenOppStrng                              -0.125023833
## age30-39                                      -0.013458834
## age40-49                                       0.002255430
## age50-59                                      -0.002958100
## age60-69                                       0.021417689
## age70-older                                    0.025630376
## educHS or less                                -0.094547938
## educSome coll                                 -0.020406639
## blackYes                                      -0.354716325
## hispanicYes                                   -0.143870530
## incomeQuint2                                   0.002688703
## incomeQuint3                                  -0.002364358
## incomeQuint4                                   0.022965907
## incomeQuint5                                   0.129165460
## veteran2. No                                   0.005468809
## econ_pastSame                                  0.129067248
## econ_pastWorse                                 0.426924371
## unemp_past2. About the same                    0.131314921
## unemp_past3. Worse                             0.305082576
## envir_gwarm2. Probably hasn't been happening   0.139803545
## religion2. Other Christian                     0.093637068
## religion3. Other religion                      0.022315875
## religion4. Not religious                      -0.082154894
## gay_marryYes                                  -0.244416124
## gay_adopt1                                    -0.116453144
## genderMale                                    -0.001058323
## gun_control2. Easier                           0.093551102
## gun_control3. Keep these rules about the same  0.199114120
## aca_app2. Favor moderately                    -0.198760112
## aca_app3. Favor a little                      -0.021917049
## aca_app4. Neither favor nor oppose             0.106495588
## aca_app5. Oppose a little                      0.086462620
## aca_app6. Oppose moderately                    0.256784721
## aca_app7. Oppose a great deal                  0.521426200
## immig_citizen2. Oppose                         0.113833878
## immig_citizen3. Neither favor or oppose        0.017846289
## immig_jobs2. Very                              0.055193811
## immig_jobs3. Somewhat                         -0.042075691
## immig_jobs4. Not at all                       -0.049706092
## marriedYes                                     0.080194188
## owngun2. No                                   -0.102019760
## govwaste2. Waste some                         -0.104210266
## govwaste3. Don't waste very much              -0.065541279
## usworld_stay2. Disagree                        0.083795086
## trust_social2. Most of the time                0.058201972
## trust_social3. About half the time             0.021856973
## trust_social4. Some of the time               -0.045651621
## trust_social5. Never                          -0.092028384
## govcorrpt2. Most                              -0.003617818
## govcorrpt3. About half                         0.037315015
## govcorrpt4. A few                             -0.025144530
## govcorrpt5. None                              -0.060937620
```

Above table shows the coefficient values under the best lambda value for the ridge regression. Unlike lasso regression, ridge regression will keep all the coefficients in the model.

```
ridge_probs <- predict(ridge_mod_cv, train, type="prob")
ridge_pred <- ifelse(ridge_probs[,2]>0.5, 1, 0)
sum(ridge_pred == train$vote12)/nrow(train)
```

```
## [1] 0.9199219
```

The accuracy for the training data is really high: 91.99%.

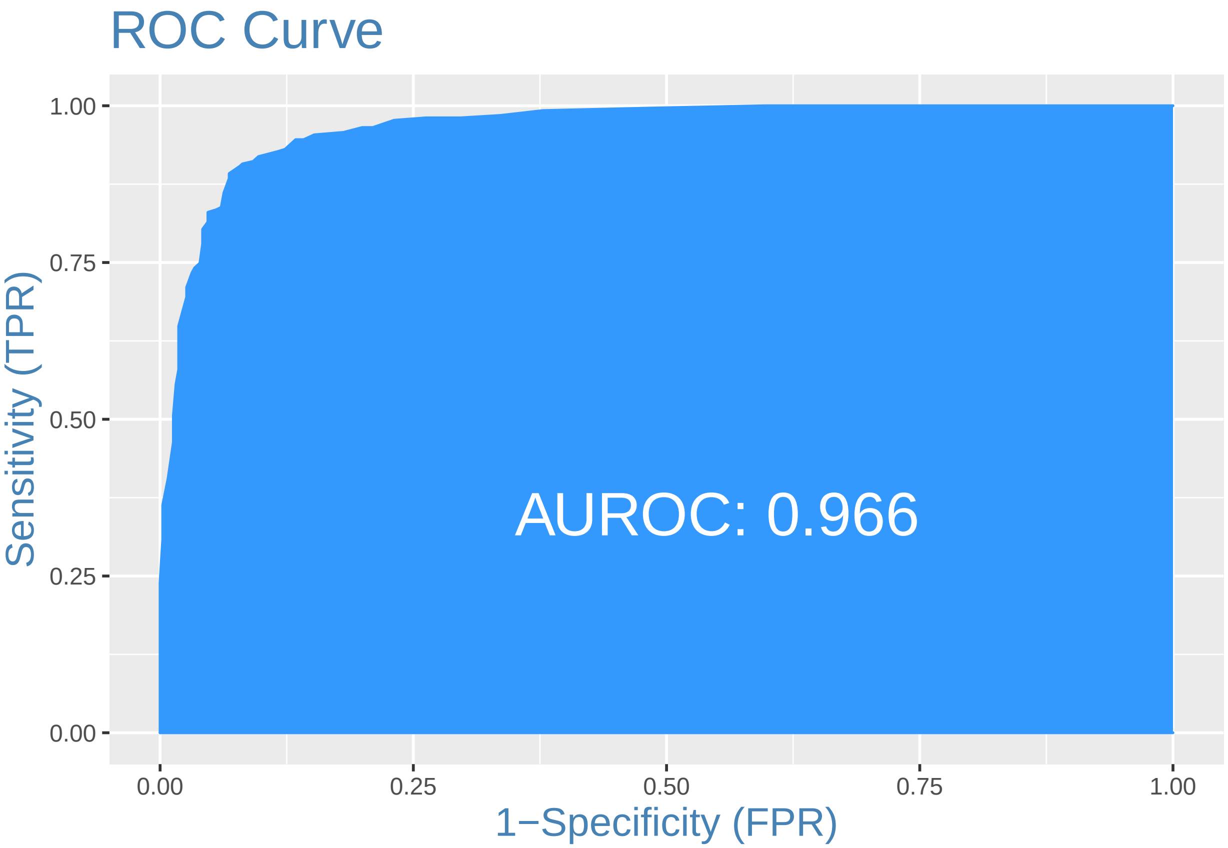


The AUROC for the training data is pretty high as well: 97.23%. Both measures indicates that the ridge regression chosen by the cross validation has done a really decent job on the training data.

```
ridge_probs_test <- predict(ridge_mod_cv, test, type="prob")
ridge_pred_test <- ifelse(ridge_probs_test[,2]>0.5, 1, 0)
sum(ridge_pred_test == test$vote12)/nrow(test)
```

```
## [1] 0.90625
```

The accuracy for the testing data is really high: 90.63%.



The AUROC for the testing data is pretty high as well: 96.6%. Both measures indicates that the ridge regression chosen by the cross validation has done a really decent job on the testing data. Compared to the lasso, ridge has the exact same accuracy on the testing data, but a slight increase in the AUROC.

As a result, we will choose lasso because it has significantly less predictors.

```
## Start:  AIC=3476.29
## vote12 ~ 1
##
##                   Df Deviance    AIC
## + aca_app          6   1844.0 1858.0
## + dem              1   2327.5 2331.5
## + econ_past        2   2415.0 2421.0
## + unemp_past       2   2615.6 2621.6
## + gun_control      2   3060.3 3066.3
## + gay_marry        1   3062.9 3066.9
## + black            1   3094.9 3098.9
## + abortion         3   3151.2 3159.2
## + envir_gwarm      1   3193.7 3197.7
## + deathpen         3   3208.8 3216.8
## + immig_citizen    2   3220.7 3226.7
## + gay_adopt        1   3240.2 3244.2
## + govwaste         2   3242.3 3248.3
## + owngun           1   3326.5 3330.5
## + immig_jobs       3   3337.3 3345.3
## + married          1   3387.3 3391.3
## + govcorrpt        4   3393.0 3403.0
## + religion         3   3402.5 3410.5
## + hispanic       1   3422.0 3426.0
## + income         4   3419.8 3429.8
## + congapp        3   3425.4 3433.4
## + trust_social   4   3439.4 3449.4
## + age            5   3440.8 3452.8
## + veteran        1   3456.8 3460.8
## + gender         1   3465.9 3469.9
## + ...1           1   3470.3 3474.3
## + educ           2   3469.6 3475.6
## <none>               3474.3 3476.3
## + usworld_stay   1   3473.6 3477.6
##
## Step:  AIC=1857.99
## vote12 ~ aca_app
##
##                  Df Deviance    AIC
## + dem             1   1450.0 1466.0
## + econ_past       2   1633.2 1651.2
## + unemp_past      2   1682.5 1700.5
## + black           1   1714.4 1730.4
## + gay_marry       1   1741.1 1757.1
## + abortion        3   1765.2 1785.2
## + religion        3   1787.0 1807.0
## + deathpen        3   1794.0 1814.0
## + gun_control     2   1798.1 1816.1
## + owngun          1   1804.6 1820.6
## + income          4   1799.2 1821.2
## + gay_adopt       1   1805.3 1821.3
## + envir_gwarm     1   1811.3 1827.3
## + hispanic        1   1812.3 1828.3
## + govwaste        2   1811.2 1829.2
## + married         1   1815.3 1831.3
## + trust_social    4   1815.7 1837.7
## + immig_citizen   2   1820.0 1838.0
## + usworld_stay    1   1831.0 1847.0
## + congapp         3   1827.7 1847.7
## + age             5   1824.6 1848.6
## + immig_jobs      3   1830.9 1850.9
## + educ            2   1834.5 1852.5
## + govcorrpt       4   1831.5 1853.5
## + veteran         1   1841.2 1857.2
## <none>                1844.0 1858.0
## + ...1            1   1842.5 1858.5
## + gender          1   1844.0 1860.0
##
## Step:  AIC=1466.03
## vote12 ~ aca_app + dem
##
##                  Df Deviance    AIC
## + econ_past       2   1281.9 1301.9
## + unemp_past      2   1342.0 1362.0
## + gay_marry       1   1372.3 1390.3
## + black           1   1380.3 1398.3
## + abortion        3   1377.4 1399.4
## + religion       3   1402.9 1424.9
## + gay_adopt      1   1411.6 1429.6
## + deathpen       3   1411.0 1433.0
## + owngun         1   1427.7 1445.7
## + gun_control    2   1426.4 1446.4
## + envir_gwarm    1   1431.8 1449.8
## + immig_jobs     3   1428.9 1450.9
## + married        1   1434.5 1452.5
## + immig_citizen  2   1433.5 1453.5
## + hispanic       1   1436.0 1454.0
## + income         4   1433.2 1457.2
## + govwaste       2   1437.5 1457.5
## + age            5   1435.5 1461.5
## + usworld_stay   1   1443.7 1461.7
## + trust_social   4   1439.6 1463.6
## + ...1           1   1447.8 1465.8
## <none>               1450.0 1466.0
## + gender         1   1448.3 1466.3
## + veteran        1   1449.1 1467.1
## + congapp        3   1445.9 1467.9
## + govcorrpt      4   1444.2 1468.2
## + educ           2   1449.3 1469.3
##
## Step:  AIC=1301.92
## vote12 ~ aca_app + dem + econ_past
##
##                 Df Deviance    AIC
## + black          1   1217.2 1239.2
## + gay_marry      1   1227.4 1249.4
## + abortion       3   1236.8 1262.8
## + religion       3   1249.7 1275.7
## + income         4   1248.0 1276.0
## + owngun         1   1257.2 1279.2
## + gay_adopt      1   1260.5 1282.5
## + gun_control    2   1261.7 1285.7
## + married        1   1264.5 1286.5
## + deathpen       3   1262.4 1288.4
## + usworld_stay   1   1266.7 1288.7
## + hispanic       1   1268.5 1290.5
## + unemp_past     2   1266.6 1290.6
## + trust_social   4   1262.8 1290.8
## + envir_gwarm    1   1272.3 1294.3
## + educ           2   1272.7 1296.7
## + immig_citizen  2   1277.1 1301.1
## + govwaste       2   1277.5 1301.5
## + age            5   1271.9 1301.9
## <none>               1281.9 1301.9
## + ...1           1   1280.0 1302.0
## + govcorrpt      4   1274.3 1302.3
## + immig_jobs     3   1276.8 1302.8
## + veteran        1   1281.3 1303.3
## + gender         1   1281.9 1303.9
## + congapp        3   1278.3 1304.3
##
```

```
## Step:  AIC=1239.18
## vote12 ~ aca_app + dem + econ_past + black
## 
##                 Df Deviance    AIC
## + gay_marry      1   1158.6 1182.6
## + abortion       3   1176.2 1204.2
## + religion       3   1180.5 1208.5
## + hispanic       1   1193.9 1217.9
## + gay_adopt      1   1193.9 1217.9
## + usworld_stay   1   1197.5 1221.5
## + owngun         1   1197.9 1221.9
## + gun_control    2   1196.5 1222.5
## + income         4   1193.1 1223.1
## + deathpen       3   1201.7 1229.7
## + unemp_past     2   1203.8 1229.8
## + married        1   1207.1 1231.1
## + envir_gwarm    1   1207.3 1231.3
## + educ           2   1207.7 1233.7
## + trust_social   4   1204.6 1234.6
## <none>               1217.2 1239.2
## + immig_citizen  2   1213.3 1239.3
## + immig_jobs     3   1211.4 1239.4
## + ...1           1   1215.5 1239.5
## + govcorrpt      4   1209.8 1239.8
## + veteran        1   1215.9 1239.9
## + govwaste       2   1214.8 1240.8
## + gender         1   1217.2 1241.2
## + age            5   1211.2 1243.2
## + congapp        3   1215.5 1243.5
## 
## Step:  AIC=1182.6
## vote12 ~ aca_app + dem + econ_past + black + gay_marry
## 
##                 Df Deviance    AIC
## + hispanic       1   1130.8 1156.8
## + religion       3   1136.0 1166.0
## + gun_control    2   1139.8 1167.8
## + usworld_stay   1   1142.4 1168.4
## + income         4   1136.7 1168.7
## + abortion       3   1139.2 1169.2
## + owngun         1   1146.5 1172.5
## + deathpen       3   1143.1 1173.1
## + educ           2   1146.2 1174.2
## + unemp_past     2   1149.9 1177.9
## + trust_social   4   1146.2 1178.2
## + envir_gwarm    1   1152.6 1178.6
## + married        1   1153.3 1179.3
## + ...1           1   1155.4 1181.4
## <none>               1158.6 1182.6
## + gay_adopt      1   1156.7 1182.7
## + govcorrpt      4   1150.7 1182.7
## + immig_jobs     3   1153.3 1183.3
## + immig_citizen  2   1156.3 1184.3
## + veteran        1   1158.4 1184.4
## + gender         1   1158.5 1184.5
## + govwaste       2   1157.0 1185.0
## + congapp        3   1157.6 1187.6
## + age            5   1157.2 1191.2
##
## Step:  AIC=1156.85
## vote12 ~ aca_app + dem + econ_past + black + gay_marry + hispanic
##
##                 Df Deviance    AIC
## + abortion       3   1108.8 1140.8
## + usworld_stay   1   1114.8 1142.8
## + gun_control    2   1114.9 1144.9
## + religion       3   1114.9 1146.9
## + deathpen       3   1115.7 1147.7
## + income         4   1114.2 1148.2
## + educ           2   1121.2 1151.2
## + owngun         1   1123.5 1151.5
## + unemp_past     2   1123.1 1153.1
## + gay_adopt      1   1126.1 1154.1
## + married        1   1126.2 1154.2
## + envir_gwarm    1   1126.4 1154.4
## + trust_social   4   1122.0 1156.0
## + govcorrpt      4   1122.5 1156.5
## + ...1           1   1128.6 1156.6
## <none>               1130.8 1156.8
## + immig_jobs     3   1126.4 1158.4
## + veteran        1   1130.8 1158.8
## + gender         1   1130.8 1158.8
## + immig_citizen  2   1129.6 1159.6
## + govwaste       2   1129.8 1159.8
## + congapp        3   1130.2 1162.2
## + age            5   1130.4 1166.4
##
## Step:  AIC=1140.76
## vote12 ~ aca_app + dem + econ_past + black + gay_marry + hispanic +
##     abortion
##
##                 Df Deviance    AIC
## + deathpen       3   1089.3 1127.3
## + usworld_stay   1   1093.4 1127.4
## + gun_control    2   1092.9 1128.9
## + income         4   1089.5 1129.5
## + educ           2   1096.7 1132.7
## + religion       3   1096.6 1134.6
## + owngun         1   1100.8 1134.8
## + unemp_past     2   1101.0 1137.0
## + trust_social   4   1099.1 1139.1
## + married        1   1105.2 1139.2
## + envir_gwarm    1   1105.3 1139.3
## + govcorrpt      4   1099.6 1139.6
## + ...1           1   1106.0 1140.0
## <none>               1108.8 1140.8
## + gay_adopt      1   1107.2 1141.2
## + immig_jobs     3   1104.7 1142.7
## + gender         1   1108.7 1142.7
## + veteran        1   1108.7 1142.7
## + govwaste       2   1107.1 1143.1
## + immig_citizen  2   1107.5 1143.5
## + congapp        3   1108.4 1146.4
## + age            5   1107.8 1149.8
##
## Step:  AIC=1127.34
## vote12 ~ aca_app + dem + econ_past + black + gay_marry + hispanic +
##     abortion + deathpen
##
##                 Df Deviance    AIC
## + usworld_stay   1   1071.8 1111.8
## + income         4   1069.1 1115.1
## + educ           2   1074.3 1116.3
## + gun_control    2   1077.5 1119.5
## + religion       3   1076.9 1120.9
## + trust_social   4   1077.3 1123.3
## + owngun         1   1083.9 1123.9
## + unemp_past     2   1082.0 1124.0
## + govcorrpt      4   1078.5 1124.5
## + ...1           1   1085.2 1125.2
## + envir_gwarm    1   1086.5 1126.5
## + married        1   1086.7 1126.7
## <none>               1089.3 1127.3
## + gay_adopt      1   1088.2 1128.2
## + immig_jobs     3   1085.1 1129.1
## + gender         1   1089.2 1129.2
## + veteran        1   1089.3 1129.3
## + immig_citizen  2   1088.2 1130.2
## + govwaste       2   1088.6 1130.6
## + congapp        3   1088.9 1132.9
## + age            5   1088.2 1136.2
##
## Step:  AIC=1111.81
## vote12 ~ aca_app + dem + econ_past + black + gay_marry + hispanic +
##     abortion + deathpen + usworld_stay
##
##                 Df Deviance    AIC
## + gun_control    2   1059.1 1103.1
## + educ           2   1059.3 1103.3
## + income         4   1056.3 1104.3
## + religion       3   1060.9 1106.9
## + owngun         1   1065.6 1107.6
## + unemp_past     2   1065.2 1109.2
## + govcorrpt      4   1061.6 1109.6
## + envir_gwarm    1   1068.5 1110.5
## + ...1           1   1068.7 1110.7
## + trust_social   4   1063.2 1111.2
## + married        1   1069.5 1111.5
## <none>               1071.8 1111.8
## + gay_adopt      1   1070.5 1112.5
## + immig_jobs     3   1067.7 1113.7
## + gender         1   1071.8 1113.8
## + veteran        1   1071.8 1113.8
## + govwaste       2   1070.6 1114.6
## + immig_citizen  2   1070.9 1114.9
## + congapp        3   1071.5 1117.5
## + age            5   1071.0 1121.0
##
## Step:  AIC=1103.09
## vote12 ~ aca_app + dem + econ_past + black + gay_marry + hispanic +
##     abortion + deathpen + usworld_stay + gun_control
##
##                 Df Deviance    AIC
## + educ           2   1046.1 1094.1
## + income         4   1043.8 1095.8
## + religion       3   1048.3 1098.3
## + ...1           1   1053.6 1099.6
## + govcorrpt      4   1048.6 1100.6
## + unemp_past     2   1053.0 1101.0
## + owngun         1   1056.3 1102.3
## + envir_gwarm    1   1056.4 1102.4
## + trust_social   4   1051.0 1103.0
## <none>               1059.1 1103.1
## + married        1   1057.2 1103.2
## + gay_adopt      1   1057.9 1103.9
## + gender         1   1058.5 1104.5
## + veteran        1   1059.0 1105.0
## + immig_jobs     3   1055.6 1105.6
## + govwaste       2   1057.9 1105.9
## + immig_citizen  2   1058.2 1106.2
## + congapp        3   1058.8 1108.8
## + age            5   1057.9 1111.9
##
## Step:  AIC=1094.08
## vote12 ~ aca_app + dem + econ_past + black + gay_marry + hispanic +
##     abortion + deathpen + usworld_stay + gun_control + educ
##
##                 Df Deviance    AIC
## + religion       3   1035.8 1089.8
## + unemp_past     2   1038.1 1090.1
## + owngun         1   1042.6 1092.6
## + govcorrpt      4   1036.9 1092.9
## + envir_gwarm    1   1043.0 1093.0
## + income         4   1037.3 1093.3
## <none>               1046.1 1094.1
## + married        1   1044.2 1094.2
## + gay_adopt      1   1044.5 1094.5
## + gender         1   1044.9 1094.9
## + ...1           1   1045.0 1095.0
## + immig_jobs     3   1041.7 1095.7
## + trust_social   4   1039.8 1095.8
## + veteran        1   1045.9 1095.9
## + govwaste       2   1044.0 1096.0
## + immig_citizen  2   1045.8 1097.8
## + congapp        3   1045.9 1099.9
## + age            5   1043.7 1101.7
```

```
##
## Step:  AIC=1089.76
## vote12 ~ aca_app + dem + econ_past + black + gay_marry + hispanic +
##     abortion + deathpen + usworld_stay + gun_control + educ +
##     religion
##
##                 Df Deviance    AIC
## + unemp_past     2   1027.6 1085.6
## + income         4   1026.0 1088.0
## + envir_gwarm    1   1032.6 1088.6
## + owngun         1   1032.6 1088.6
## + govcorrpt      4   1027.2 1089.2
## <none>               1035.8 1089.8
## + married        1   1034.4 1090.4
## + gay_adopt      1   1034.8 1090.8
## + trust_social   4   1029.3 1091.3
## + immig_jobs     3   1031.3 1091.3
## + gender         1   1035.3 1091.3
## + ...1           1   1035.4 1091.4
## + veteran        1   1035.6 1091.6
## + govwaste       2   1034.2 1092.2
## + immig_citizen  2   1035.5 1093.5
## + congapp        3   1035.3 1095.3
## + age            5   1033.5 1097.5
##
## Step:  AIC=1085.63
## vote12 ~ aca_app + dem + econ_past + black + gay_marry + hispanic +
##     abortion + deathpen + usworld_stay + gun_control + educ +
##     religion + unemp_past
##
##                 Df Deviance    AIC
## + income         4   1016.8 1082.8
## + envir_gwarm    1   1024.6 1084.6
## + owngun         1   1024.8 1084.8
## <none>               1027.6 1085.6
## + married        1   1025.9 1085.9
## + govcorrpt      4   1020.5 1086.5
## + immig_jobs     3   1022.9 1086.8
## + trust_social   4   1020.9 1086.9
## + gay_adopt      1   1027.0 1087.0
## + gender         1   1027.2 1087.2
## + ...1           1   1027.3 1087.3
## + veteran        1   1027.5 1087.5
## + govwaste       2   1025.9 1087.9
## + immig_citizen  2   1027.3 1089.3
## + congapp        3   1027.2 1091.2
## + age            5   1025.7 1093.7
##
## Step:  AIC=1082.82
## vote12 ~ aca_app + dem + econ_past + black + gay_marry + hispanic +
##     abortion + deathpen + usworld_stay + gun_control + educ +
##     religion + unemp_past + income
##
##                 Df Deviance    AIC
## + envir_gwarm    1   1013.7 1081.7
## <none>               1016.8 1082.8
## + owngun         1   1015.0 1083.0
## + immig_jobs     3   1011.9 1083.9
## + gay_adopt      1   1016.0 1084.0
## + govcorrpt      4   1010.2 1084.2
## + gender         1   1016.3 1084.3
## + ...1           1   1016.4 1084.4
## + veteran        1   1016.6 1084.6
## + married        1   1016.7 1084.7
## + trust_social   4   1011.0 1085.0
## + govwaste       2   1015.5 1085.5
## + immig_citizen  2   1016.7 1086.7
## + congapp        3   1016.4 1088.4
## + age            5   1015.4 1091.4
##
## Step:  AIC=1081.69
## vote12 ~ aca_app + dem + econ_past + black + gay_marry + hispanic +
##     abortion + deathpen + usworld_stay + gun_control + educ +
##     religion + unemp_past + income + envir_gwarm
##
##                 Df Deviance    AIC
## <none>               1013.7 1081.7
## + owngun         1   1011.8 1081.8
## + gay_adopt      1   1013.0 1083.0
## + gender         1   1013.0 1083.0
## + ...1           1   1013.1 1083.2
## + immig_jobs     3   1009.3 1083.3
## + govcorrpt      4   1007.4 1083.3
## + veteran        1   1013.4 1083.4
## + married        1   1013.5 1083.5
## + trust_social   4   1008.2 1084.2
## + govwaste       2   1012.5 1084.5
## + immig_citizen  2   1013.4 1085.4
## + congapp        3   1013.3 1087.3
## + age            5   1012.3 1090.3
```

Next, we will compare forward stepwise selection with lasso and ridge.

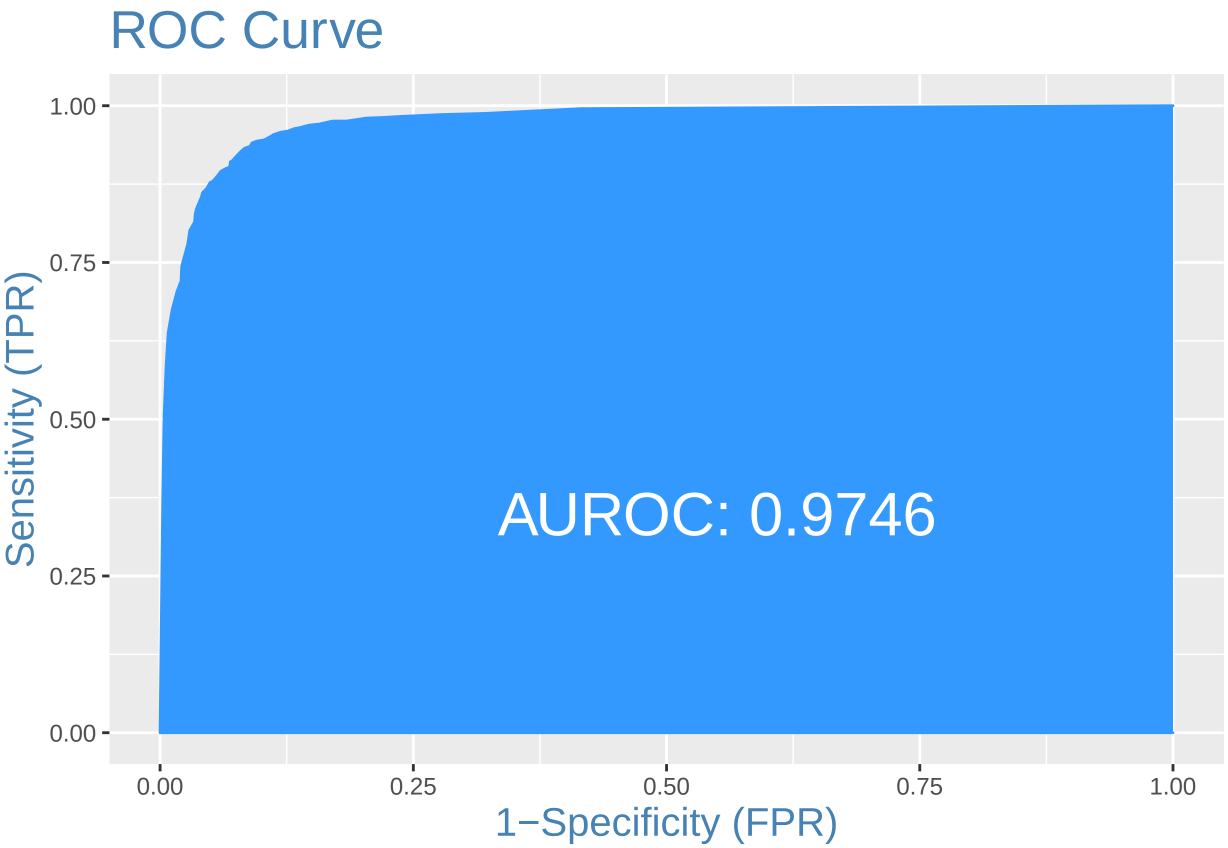
```
##                                    (Intercept)
##                                     -4.1437740
##                     aca_app2. Favor moderately
##                                      0.4903494
##                       aca_app3. Favor a little
##                                      1.2384568
##             aca_app4. Neither favor nor oppose
##                                      1.9129731
##                      aca_app5. Oppose a little
##                                      2.1394910
##                    aca_app6. Oppose moderately
##                                      2.5488809
##                  aca_app7. Oppose a great deal
##                                      3.6239675
##                                            dem
##                                     -2.4026797
##                                 econ_pastSame 
##                                     0.9621199 
##                                econ_pastWorse 
##                                     2.0727225 
##                                      blackYes 
##                                    -2.8688248 
##                                  gay_marryYes 
##                                    -0.8481541 
##                                   hispanicYes 
##                                    -0.8938859 
##                            abortionMore conds 
##                                     0.5098186 
##                                 abortionNever 
##                                     0.9457779 
##                            abortionSome conds 
##                                     0.9735224 
##                              deathpenAppStrng 
##                                     0.6388472 
##                                deathpenOppose 
##                                     0.2286725 
##                              deathpenOppStrng 
##                                    -0.4151118 
##                       usworld_stay2. Disagree 
##                                     0.5512666 
##                          gun_control2. Easier 
##                                     0.5062330 
## gun_control3. Keep these rules about the same 
##                                     0.5703571 
##                                educHS or less 
##                                    -0.5973549 
##                                 educSome coll 
##                                    -0.1705721 
##                    religion2. Other Christian 
##                                     0.4870508 
##                     religion3. Other religion 
##                                     0.4474086 
##                      religion4. Not religious 
##                                    -0.1812039 
##                   unemp_past2. About the same 
##                                     0.5548666 
##                            unemp_past3. Worse 
##                                     0.8297521 
##                                  incomeQuint2 
##                                     0.3533962 
##                                  incomeQuint3 
##                                     0.3009105 
##                                  incomeQuint4 
##                                     0.4503326 
##                                  incomeQuint5 
##                                     0.9361708 
##  envir_gwarm2. Probably hasn't been happening 
##                                     0.4027137
```

Above are the coefficients chosen by the forward stepwise selection with AIC.

```
best_probs <- predict(step_best, train, type="response")
best_pred <- ifelse(best_probs>0.5, 1, 0)    # 1 means vote for Romney
sum(best_pred == train$vote12)/nrow(train)
```

```
## [1] 0.9222656
```

The accuracy score on the training set for the forward stepwise selection with AIC is 92.23%, which indicates this model has done a great job on training set.

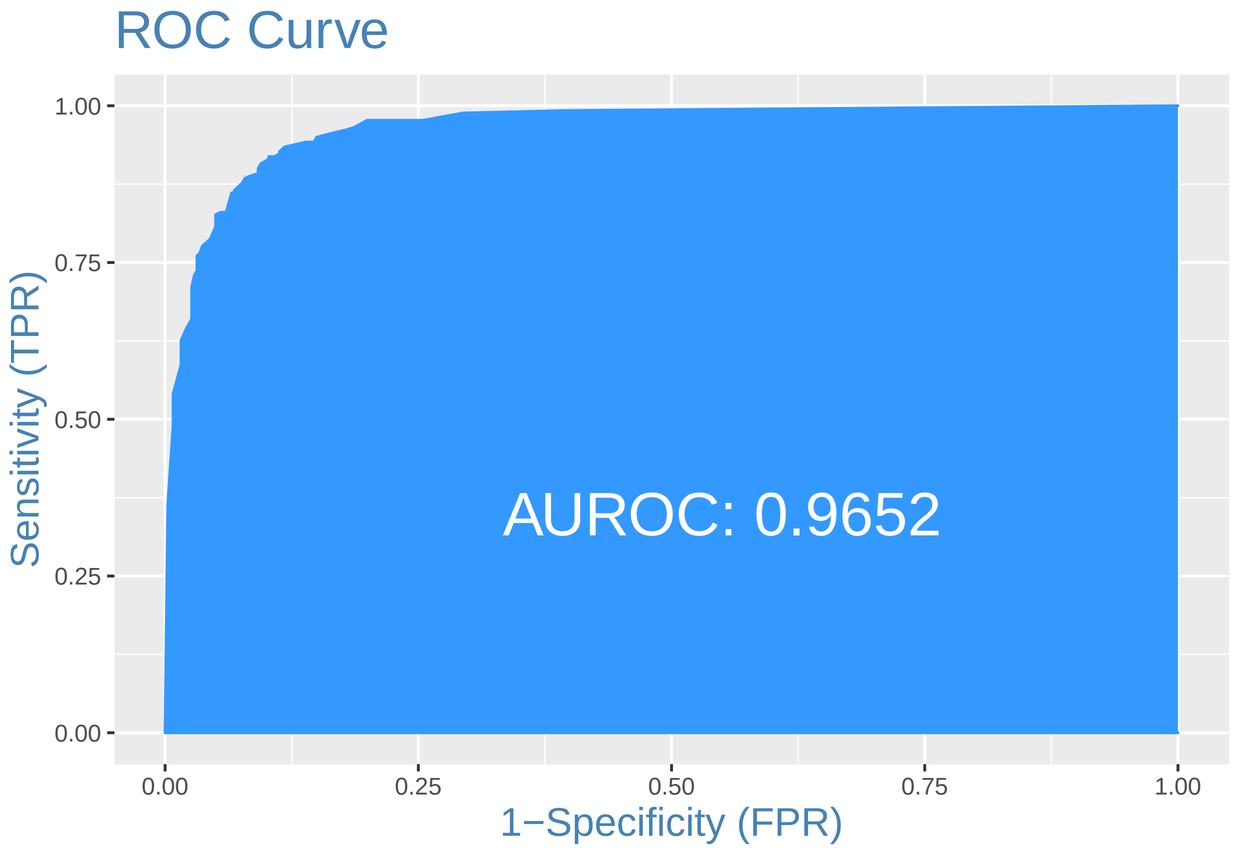


The AUROC on the training set is pretty good as well: 97.46%. This model is pretty good at distinguishing between the two classes (either voting for Obama or voting for Romney).

```
best_probs_test <- predict(step_best, test, type="response")
best_pred_test <- ifelse(best_probs_test>0.5, 1, 0)
sum(best_pred_test == test$vote12)/nrow(test)
```

```
## [1] 0.9046875
```

The accuracy score on the testing set for the forward stepwise selection with AIC is 90.47%, which is slightly lower compared to the lasso and ridge.



The AUROC on the testing set is pretty good as well, and it is slightly greater than the AUROC for lasso but slightly less than the AUROC for ridge.

```
##                                                  lasso        ridge
## (Intercept)                                -0.93568242 -1.302116248
## ...1                                        0.00000000  0.044323892
## abortionMore conds                          0.00000000  0.081105591
## abortionNever                               0.00000000  0.102736444
## abortionSome conds                          0.11053545  0.196979569
## congappAppWeak                              0.00000000 -0.004183322
## congappDisappStr                            0.00000000  0.007585518
## congappDisappWk                             0.00000000  0.067184401
## dem                                        -1.01960992 -0.619887888
## deathpenAppStrng                            0.09509861  0.148790108
## deathpenOppose                              0.00000000 -0.051638943
## deathpenOppStrng                           -0.03695456 -0.125023833
## age30-39                                    0.00000000 -0.013458834
## age40-49                                    0.00000000  0.002255430
## age50-59                                    0.00000000 -0.002958100
## age60-69                                    0.00000000  0.021417689
## age70-older                                 0.00000000  0.025630376
## educHS or less                              0.00000000 -0.094547938
## educSome coll                               0.00000000 -0.020406639
## blackYes                                   -0.48503386 -0.354716325
## hispanicYes                                -0.10954981 -0.143870530
## incomeQuint2                                0.00000000  0.002688703
## incomeQuint3                                0.00000000 -0.002364358
## incomeQuint4                                   0.00000000  0.022965907
## incomeQuint5                                   0.05052678  0.129165460
## veteran2. No                                   0.00000000  0.005468809
## econ_pastSame                                  0.19020470  0.129067248
## econ_pastWorse                                 0.61286037  0.426924371
## unemp_past2. About the same                    0.00000000  0.131314921
## unemp_past3. Worse                             0.17763583  0.305082576
## envir_gwarm2. Probably hasn't been happening   0.07639220  0.139803545
## religion2. Other Christian                     0.07337986  0.093637068
## religion3. Other religion                      0.00000000  0.022315875
## religion4. Not religious                      -0.01977507 -0.082154894
## gay_marryYes                                  -0.37181244 -0.244416124
## gay_adopt1                                     0.00000000 -0.116453144
## genderMale                                     0.00000000 -0.001058323
## gun_control2. Easier                           0.00000000  0.093551102
## gun_control3. Keep these rules about the same  0.15531017  0.199114120
## aca_app2. Favor moderately                    -0.08258346 -0.198760112
## aca_app3. Favor a little                       0.00000000 -0.021917049
## aca_app4. Neither favor nor oppose             0.15493691  0.106495588
## aca_app5. Oppose a little                      0.05397792  0.086462620
## aca_app6. Oppose moderately                    0.35911343  0.256784721
## aca_app7. Oppose a great deal                  0.88654815  0.521426200
## immig_citizen2. Oppose                         0.00000000  0.113833878
## immig_citizen3. Neither favor or oppose        0.00000000  0.017846289
## immig_jobs2. Very                              0.00000000  0.055193811
## immig_jobs3. Somewhat                          0.00000000 -0.042075691
## immig_jobs4. Not at all                        0.00000000 -0.049706092
## marriedYes                                     0.00000000  0.080194188
## owngun2. No                                   -0.03121594 -0.102019760
## govwaste2. Waste some                          0.00000000 -0.104210266
## govwaste3. Don't waste very much               0.00000000 -0.065541279
## usworld_stay2. Disagree                        0.00045615  0.083795086
## trust_social2. Most of the time                0.00000000  0.058201972
## trust_social3. About half the time             0.00000000  0.021856973
## trust_social4. Some of the time                0.00000000 -0.045651621
## trust_social5. Never                           0.00000000 -0.092028384
## govcorrpt2. Most                               0.00000000 -0.003617818
## govcorrpt3. About half                         0.00000000  0.037315015
## govcorrpt4. A few                              0.00000000 -0.025144530
## govcorrpt5. None                               0.00000000 -0.060937620

##                                                     lasso        ridge
## (Intercept)                                   -0.93568242 -1.302116248
## ...1                                                   NA  0.044323892
## abortionMore conds                                     NA  0.081105591
## abortionNever                                          NA  0.102736444
## abortionSome conds                             0.11053545  0.196979569
## congappAppWeak                                         NA -0.004183322
## congappDisappStr                                       NA  0.007585518
## congappDisappWk                                        NA  0.067184401
## dem                                           -1.01960992 -0.619887888
## deathpenAppStrng                               0.09509861  0.148790108
## deathpenOppose                                         NA -0.051638943
## deathpenOppStrng                              -0.03695456 -0.125023833
## age30-39                                               NA -0.013458834
## age40-49                                               NA  0.002255430
## age50-59                                               NA -0.002958100
## age60-69                                               NA  0.021417689
## age70-older                                            NA  0.025630376
## educHS or less                                         NA -0.094547938
## educSome coll                                          NA -0.020406639
## blackYes                                      -0.48503386 -0.354716325
## hispanicYes                                   -0.10954981 -0.143870530
## incomeQuint2                                           NA  0.002688703
## incomeQuint3                                           NA -0.002364358
## incomeQuint4                                           NA  0.022965907
## incomeQuint5                                   0.05052678  0.129165460
## veteran2. No                                           NA  0.005468809
## econ_pastSame                                  0.19020470  0.129067248
## econ_pastWorse                                 0.61286037  0.426924371
## unemp_past2. About the same                            NA  0.131314921
## unemp_past3. Worse                             0.17763583  0.305082576
## envir_gwarm2. Probably hasn't been happening   0.07639220  0.139803545
## religion2. Other Christian                     0.07337986  0.093637068
## religion3. Other religion                              NA  0.022315875
## religion4. Not religious                      -0.01977507 -0.082154894
## gay_marryYes                                  -0.37181244 -0.244416124
## gay_adopt1                                             NA -0.116453144
## genderMale                                             NA -0.001058323
## gun_control2. Easier                                   NA  0.093551102
## gun_control3. Keep these rules about the same  0.15531017  0.199114120
## aca_app2. Favor moderately                    -0.08258346 -0.198760112
## aca_app3. Favor a little                               NA -0.021917049
## aca_app4. Neither favor nor oppose             0.15493691  0.106495588
## aca_app5. Oppose a little                      0.05397792  0.086462620
## aca_app6. Oppose moderately                    0.35911343  0.256784721
## aca_app7. Oppose a great deal                  0.88654815  0.521426200
## immig_citizen2. Oppose                                 NA  0.113833878
## immig_citizen3. Neither favor or oppose                NA  0.017846289
## immig_jobs2. Very                                      NA  0.055193811
## immig_jobs3. Somewhat                                  NA -0.042075691
## immig_jobs4. Not at all                                NA -0.049706092
## marriedYes                                             NA  0.080194188
## owngun2. No                                   -0.03121594 -0.102019760
## govwaste2. Waste some                                  NA -0.104210266
## govwaste3. Don't waste very much                       NA -0.065541279
## usworld_stay2. Disagree                        0.00045615  0.083795086
## trust_social2. Most of the time                        NA  0.058201972
## trust_social3. About half the time                     NA  0.021856973
## trust_social4. Some of the time                        NA -0.045651621
## trust_social5. Never                                   NA -0.092028384
## govcorrpt2. Most                                       NA -0.003617818
## govcorrpt3. About half                                 NA  0.037315015
## govcorrpt4. A few                                      NA -0.025144530
## govcorrpt5. None                                       NA -0.060937620
##                                                     best
## (Intercept)                                   -4.1437740
## ...1                                                  NA
## abortionMore conds                             0.5098186
## abortionNever                                  0.9457779
## abortionSome conds                             0.9735224
## congappAppWeak                                        NA
## congappDisappStr                                      NA
## congappDisappWk                                       NA
## dem                                           -2.4026797
## deathpenAppStrng                               0.6388472
## deathpenOppose                                 0.2286725
## deathpenOppStrng                              -0.4151118
## age30-39                                              NA
## age40-49                                              NA
## age50-59                                              NA
## age60-69                                              NA
## age70-older                                           NA
## educHS or less                                -0.5973549
## educSome coll                                 -0.1705721
## blackYes                                      -2.8688248
## hispanicYes                                   -0.8938859
## incomeQuint2                                   0.3533962
## incomeQuint3                                   0.3009105
## incomeQuint4                                   0.4503326
## incomeQuint5                                   0.9361708
## veteran2. No                                          NA
## econ_pastSame                                  0.9621199
## econ_pastWorse                                 2.0727225
## unemp_past2. About the same                    0.5548666
## unemp_past3. Worse                             0.8297521
## envir_gwarm2. Probably hasn't been happening   0.4027137
## religion2. Other Christian                     0.4870508
## religion3. Other religion                      0.4474086
## religion4. Not religious                      -0.1812039
## gay_marryYes                                  -0.8481541
## gay_adopt1                                            NA
## genderMale                                            NA
## gun_control2. Easier                           0.5062330
## gun_control3. Keep these rules about the same  0.5703571
## aca_app2. Favor moderately                     0.4903494
## aca_app3. Favor a little                       1.2384568
## aca_app4. Neither favor nor oppose             1.9129731
## aca_app5. Oppose a little                      2.1394910
## aca_app6. Oppose moderately                    2.5488809
## aca_app7. Oppose a great deal                  3.6239675
## immig_citizen2. Oppose                                NA
## immig_citizen3. Neither favor or oppose               NA
## immig_jobs2. Very                                     NA
## immig_jobs3. Somewhat                                 NA
## immig_jobs4. Not at all                               NA
## marriedYes                                            NA
## owngun2. No                                           NA
## govwaste2. Waste some                                 NA
## govwaste3. Don't waste very much                      NA
## usworld_stay2. Disagree                        0.5512666
## trust_social2. Most of the time                       NA
## trust_social3. About half the time                    NA
## trust_social4. Some of the time                       NA
## trust_social5. Never                                  NA
## govcorrpt2. Most                                      NA
## govcorrpt3. About half                                NA
## govcorrpt4. A few                                     NA
## govcorrpt5. None                                      NA
```

Then, we create a table for the coefficient values for lasso, ridge, and forward stepwise selection. We can see that only lasso and best perform variable selection. Moreover, lasso has fewer coefficients than forward stepwise selection. Forward stepwise selection doesn't perform regularization, so that it has greater magnitude for the coefficients.

Finally, we perform 10-folds cross validation on the maximum number of steps to take for the forward stepwise selection (we put no penalty on the number of predictors included in the model).

The optimal number of steps under cross validation is 16.

```
##                                   (Intercept)
##                                   -4.36325273
##                      aca_app2. Favor moderately
##                                    0.67384810
##                        aca_app3. Favor a little
##                                    1.36229282
##                aca_app4. Neither favor nor oppose
##                                    1.93809500
##                       aca_app5. Oppose a little
##                                    2.58084730
##                     aca_app6. Oppose moderately
##                                    2.71386602
##                   aca_app7. Oppose a great deal
##                                    3.78973674
##                                           dem
##                                   -2.31034129
##                                  econ_pastSame
##                                    0.90685611
##                                 econ_pastWorse
##                                    1.94027210
##                                      blackYes
##                                   -2.67465971
##                                   gay_marryYes
##                                   -0.86527619
##                                  incomeQuint2
##                                    0.59215159
##                                  incomeQuint3
##                                    0.55007328
##                                  incomeQuint4
##                                    0.75525008
##                                  incomeQuint5
##                                    1.11660923
##                             abortionMore conds
##                                    0.61736638
##                                  abortionNever
##                                    1.10884991
##                             abortionSome conds
##                                    0.96145134
##                    religion2. Other Christian
##                                    0.36332303
##                     religion3. Other religion
##                                    0.37632153
##                      religion4. Not religious
##                                   -0.39382061
##                  unemp_past2. About the same
##                                    0.56085618
##                           unemp_past3. Worse
##                                    0.98953028
##                             deathpenAppStrng
##                                    0.35136874
##                               deathpenOppose
##                                    0.05962663
##                             deathpenOppStrng
##                                   -0.59735434
##              trust_social2. Most of the time
##                                    0.49605271
##           trust_social3. About half the time
##                                    0.45131568
##              trust_social4. Some of the time
##                                    0.26880571
##                         trust_social5. Never
##                                   -0.82550816
##                                  hispanicYes
##                                   -0.64259333
##                       usworld_stay2. Disagree
##                                    0.49931134
##                         gun_control2. Easier
##                                    0.37436058
## gun_control3. Keep these rules about the same
##                                    0.46466090
##                               educHS or less
##                                   -0.48599336
##                                educSome coll
##                                   -0.02348075
##                            immig_jobs2. Very
##                                    0.12423849
##                        immig_jobs3. Somewhat
##                                   -0.32257201
##                      immig_jobs4. Not at all
##                                   -0.08289446
```

Then, we retrain the forward stepwise selection model using the number of steps chosen by the cross validation on the whole dataset.

```
length(coef(step_best_final))
```

```
## [1] 40
```

The number of predictors in the forward stepwise selection model with CV is 40.

```
length(coef(step_best))
```

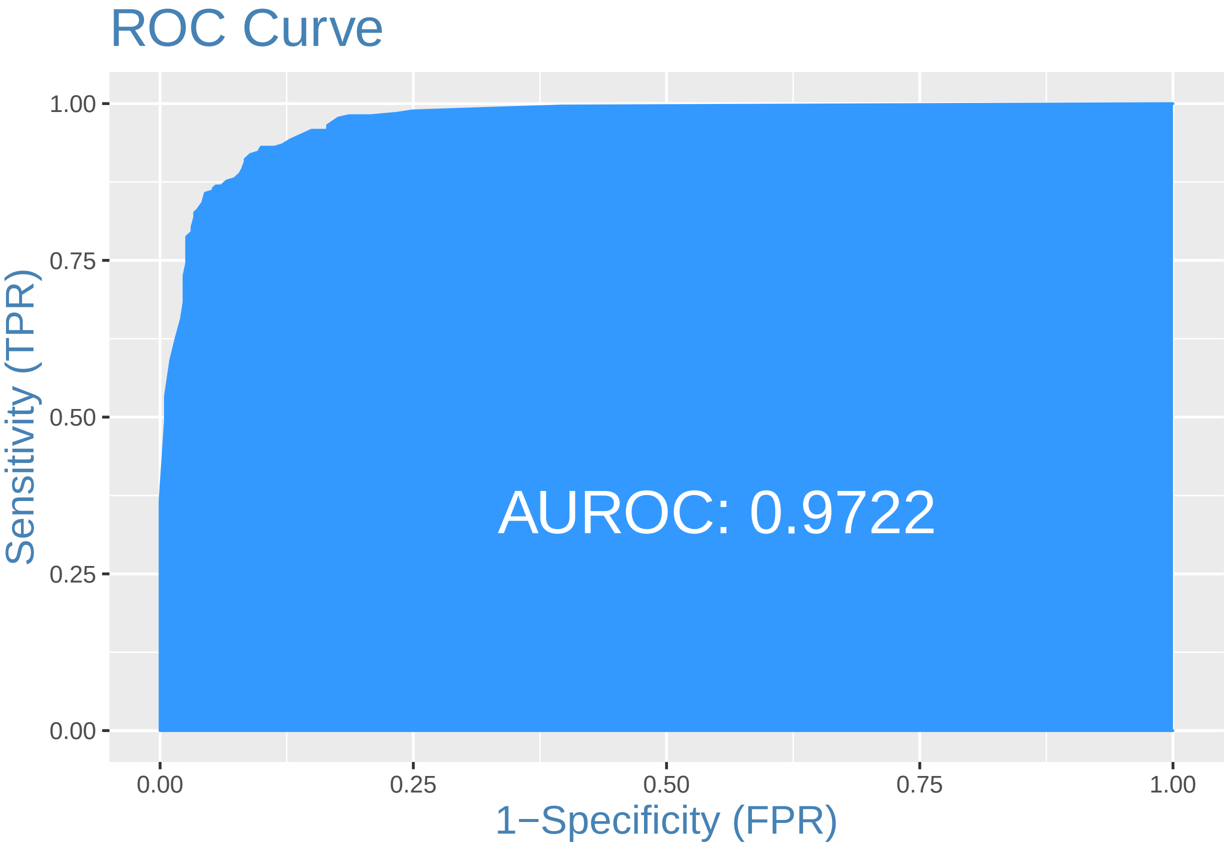
```
## [1] 34
```

The number of predictors in the forward stepwise selection model without CV is 34, which is less than the model with CV.

```
best_probs_noreg_test <- predict(step_best_final, test, type="response")
best_pred_noreg_test <- ifelse(best_probs_noreg_test>0.5, 1, 0)   # 1 means vote for Romney
sum(best_pred_noreg_test == test$vote12)/nrow(test)
```

```
## [1] 0.9109375
```

The testing accuracy is 91.09%, which is slightly higher than the forward stepwise selection model without CV, lasso model, and ridge model. Currently, this is the model with the highest accuracy score.



The AUROC for the forward stepwise selection model with CV is 97.22%, which is the highest score compared to forward stepwise selection model without CV, lasso and ridge.

In summary, the best model for the election dataset is the forward stepwise selection model with CV.

```
##                                                     lasso        ridge
## (Intercept)                                   -0.93568242 -1.302116248
## ...1                                                   NA  0.044323892
## abortionMore conds                                     NA  0.081105591
## abortionNever                                          NA  0.102736444
## abortionSome conds                             0.11053545  0.196979569
## congappAppWeak                                         NA -0.004183322
## congappDisappStr                                       NA  0.007585518
## congappDisappWk                                        NA  0.067184401
## dem                                           -1.01960992 -0.619887888
## deathpenAppStrng                               0.09509861  0.148790108
## deathpenOppose                                         NA -0.051638943
## deathpenOppStrng                              -0.03695456 -0.125023833
## age30-39                                               NA -0.013458834
## age40-49                                               NA  0.002255430
## age50-59                                               NA -0.002958100
## age60-69                                               NA  0.021417689
## age70-older                                            NA  0.025630376
## educHS or less                                         NA -0.094547938
## educSome coll                                          NA -0.020406639
## blackYes                                      -0.48503386 -0.354716325
## hispanicYes                                   -0.10954981 -0.143870530
## incomeQuint2                                           NA  0.002688703
## incomeQuint3                                           NA -0.002364358
## incomeQuint4                                           NA  0.022965907
## incomeQuint5                                   0.05052678  0.129165460
## veteran2. No                                           NA  0.005468809
## econ_pastSame                                  0.19020470  0.129067248
## econ_pastWorse                                 0.61286037  0.426924371
## unemp_past2. About the same                            NA  0.131314921
## unemp_past3. Worse                             0.17763583  0.305082576
## envir_gwarm2. Probably hasn't been happening   0.07639220  0.139803545
## religion2. Other Christian                     0.07337986  0.093637068
## religion3. Other religion                              NA  0.022315875
## religion4. Not religious                      -0.01977507 -0.082154894
## gay_marryYes                                  -0.37181244 -0.244416124
## gay_adopt1                                             NA -0.116453144
## genderMale                                             NA -0.001058323
## gun_control2. Easier                                   NA  0.093551102
## gun_control3. Keep these rules about the same  0.15531017  0.199114120
## aca_app2. Favor moderately                    -0.08258346 -0.198760112
## aca_app3. Favor a little                               NA -0.021917049
## aca_app4. Neither favor nor oppose             0.15493691  0.106495588
## aca_app5. Oppose a little                      0.05397792  0.086462620
## aca_app6. Oppose moderately                    0.35911343  0.256784721
## aca_app7. Oppose a great deal                  0.88654815  0.521426200
## immig_citizen2. Oppose                                 NA  0.113833878
## immig_citizen3. Neither favor or oppose                NA  0.017846289
## immig_jobs2. Very                                      NA  0.055193811
## immig_jobs3. Somewhat                                  NA -0.042075691
## immig_jobs4. Not at all                                NA -0.049706092
## marriedYes                                             NA  0.080194188
## owngun2. No                                   -0.03121594 -0.102019760
## govwaste2. Waste some                                  NA -0.104210266
## govwaste3. Don't waste very much                       NA -0.065541279
## usworld_stay2. Disagree                        0.00045615  0.083795086
## trust_social2. Most of the time                        NA  0.058201972
## trust_social3. About half the time                     NA  0.021856973
## trust_social4. Some of the time                        NA -0.045651621
## trust_social5. Never                                   NA -0.092028384
## govcorrpt2. Most                                       NA -0.003617818
## govcorrpt3. About half                                 NA  0.037315015
## govcorrpt4. A few                                      NA -0.025144530
## govcorrpt5. None                                       NA -0.060937620
##                                                     best     best_cv
## (Intercept)                                   -4.1437740 -4.36325273
## ...1                                                  NA          NA
## abortionMore conds                             0.5098186  0.61736638
## abortionNever                                  0.9457779  1.10884991
## abortionSome conds                             0.9735224  0.96145134
## congappAppWeak                                        NA          NA
## congappDisappStr                                      NA          NA
## congappDisappWk                                       NA          NA
## dem                                           -2.4026797 -2.31034129
## deathpenAppStrng                               0.6388472  0.35136874
## deathpenOppose                                 0.2286725  0.05962663
## deathpenOppStrng                              -0.4151118 -0.59735434
## age30-39                                              NA          NA
## age40-49                                              NA          NA
## age50-59                                              NA          NA
## age60-69                                              NA          NA
## age70-older                                           NA          NA
## educHS or less                                -0.5973549 -0.48599336
## educSome coll                                 -0.1705721 -0.02348075
## blackYes                                      -2.8688248 -2.67465971
## hispanicYes                                   -0.8938859 -0.64259333
## incomeQuint2                                   0.3533962  0.59215159
## incomeQuint3                                   0.3009105  0.55007328
## incomeQuint4                                   0.4503326  0.75525008
## incomeQuint5                                   0.9361708  1.11660923
## veteran2. No                                          NA          NA
## econ_pastSame                                  0.9621199  0.90685611
## econ_pastWorse                                 2.0727225  1.94027210
## unemp_past2. About the same                    0.5548666  0.56085618
## unemp_past3. Worse                             0.8297521  0.98953028
## envir_gwarm2. Probably hasn't been happening   0.4027137          NA
## religion2. Other Christian                     0.4870508  0.36332303
## religion3. Other religion                      0.4474086  0.37632153
## religion4. Not religious                      -0.1812039 -0.39382061
## gay_marryYes                                  -0.8481541 -0.86527619
## gay_adopt1                                            NA          NA
## genderMale                                            NA          NA
## gun_control2. Easier                           0.5062330  0.37436058
## gun_control3. Keep these rules about the same  0.5703571  0.46466090
## aca_app2. Favor moderately                     0.4903494  0.67384810
## aca_app3. Favor a little                       1.2384568  1.36229282
## aca_app4. Neither favor nor oppose             1.9129731  1.93809500
## aca_app5. Oppose a little                      2.1394910  2.58084730
## aca_app6. Oppose moderately                    2.5488809  2.71386602
## aca_app7. Oppose a great deal                  3.6239675  3.78973674
## immig_citizen2. Oppose                                NA          NA
## immig_citizen3. Neither favor or oppose               NA          NA
## immig_jobs2. Very                                     NA  0.12423849
## immig_jobs3. Somewhat                                 NA -0.32257201
## immig_jobs4. Not at all                               NA -0.08289446
## marriedYes                                            NA          NA
## owngun2. No                                           NA          NA
## govwaste2. Waste some                                 NA          NA
## govwaste3. Don't waste very much                      NA          NA
## usworld_stay2. Disagree                        0.5512666  0.49931134
## trust_social2. Most of the time                       NA  0.49605271
## trust_social3. About half the time                    NA  0.45131568
## trust_social4. Some of the time                         NA  0.26880571
## trust_social5. Never                                     NA -0.82550816
## govcorrpt2. Most                                         NA          NA
## govcorrpt3. About half                                   NA          NA
## govcorrpt4. A few                                        NA          NA
## govcorrpt5. None                                         NA          NA
```

In the end, we combine all the coefficients chosen by four models together into a table. The magnitude of the coefficients for best(forward stepwise selection) and best_cv(forward stepwise selection with cross validation) are pretty large since they don't perform regularization. Ridge contains all the available predictors since it doesn't perform variable selection. Lasso regression performs variable selection and regularization (variance reduction) at the same time. Lasso has the least number of predictors. Both lasso and best don't pay attention to predictor such as age. In addition to that, lasso also ignores predictors such as immig_citizen, immig_jobs, trust_social, and govcorrpt.

In summary, ridge regression only performs regularization; forward stepwise selection only performs variable selection; lasso regression performs both the regularization and variable selection.

The testing accuracy for all four models are slightly lower compared to the in-class models. The reason might be due to the randomness when split the testing and training dataset. The magnitude of coefficients and predictors choice are consistent. Predictors included in the lasso, ridge, and forward stepwise selection and values of coefficients are similar. Moreover, lasso has the least number of predictors in the model.

Code Appendix:

```r
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(caret)
library(glmnet)
library(MASS)
library(InformationValue)
election <- read_csv("/Users/justinli/Desktop/Election12_Large.csv")
election <- na.omit(election)
election$abortion <- as.factor(election$abortion)
election$congapp <- as.factor(election$congapp)
election$deathpen <- as.factor(election$deathpen)
election$age <- as.factor(election$age)
election$educ <- as.factor(election$educ)
election$income <- as.factor(election$income)
election$econ_past <- as.factor(election$econ_past)
election$unemp_past <- as.factor(election$unemp_past)
election$religion <- as.factor(election$religion)
election$gun_control <- as.factor(election$gun_control)
election$aca_app <- as.factor(election$aca_app)
election$immig_citizen <- as.factor(election$immig_citizen)
election$immig_jobs <- as.factor(election$immig_jobs)
election$govwaste <- as.factor(election$govwaste)
election$trust_social <- as.factor(election$trust_social)
election$govcorrpt <- as.factor(election$govcorrpt)
election$black <- as.factor(election$black)
election$hispanic <- as.factor(election$hispanic)
election$veteran <- as.factor(election$veteran)
election$envir_gwarm <- as.factor(election$envir_gwarm)
election$gay_marry <- as.factor(election$gay_marry)
```

```
election$gay_adopt <- as.factor(election$gay_adopt)
election$gender <- as.factor(election$gender)
election$married <-as.factor(election$married)
election$owngun <- as.factor(election$owngun)
election$usworld_stay <- as.factor(election$usworld_stay)
election$vote12 <- ifelse(election$vote12 == "Obama", 0, 1)
election$vote12 <- as.factor(election$vote12)
set.seed(123)
split = sample(c(rep(0, 0.8*nrow(election)), rep(1, 0.2*nrow(election))))
train = election[split == 0, ]
test = election[split == 1, ]
dim(train)
dim(test)
folds <- createFolds(train$vote12, k=10)
lambda_seq <- exp(seq(-7,-1,0.1))
lambda_seq
trainCtrl <- trainControl(method = "cv", index = folds)
lasso_mod_cv <- train(vote12~.,
             data = train,
             method = 'glmnet',  # alpha=1 for lasso
             preProc = c("scale"), # scale data   # standardize the dataset
             trControl = trainCtrl,
             tuneGrid = expand.grid(alpha = 1, lambda = lambda_seq),
             metric =  "Accuracy"  # RMSE for regression
                                   # Accuracy for classification
)
ggplot(lasso_mod_cv)
lasso_mod_best <- lasso_mod_cv$finalModel$lambdaOpt
lasso_mod_best
coef(lasso_mod_cv$finalModel, lasso_mod_cv$finalModel$lambdaOpt)
lasso_probs <- predict(lasso_mod_cv, train, type="prob")
lasso_pred <- ifelse(lasso_probs[,2]>0.5, 1, 0) # 1 means vote for Romney
sum(lasso_pred == train$vote12)/nrow(train)
plotROC(train$vote12, lasso_probs[,2], Show.labels=F)
# (data, outcome)
lasso_probs_test <- predict(lasso_mod_cv, test, type="prob")
lasso_pred_test <- ifelse(lasso_probs_test[,2]>0.5, 1, 0)  # 1 means vote for Romney
sum(lasso_pred_test == test$vote12)/nrow(test)
plotROC(test$vote12, lasso_probs_test[,2], Show.labels=F)
# (data, outcome)
trainCtrl <- trainControl(method = "cv", index = folds)
ridge_mod_cv <- train(vote12~.,  # can add interactions/transformations in formula
                      data = train,
                      preProc = c("scale"), # scale data
                      method = 'glmnet',
                      trControl = trainCtrl,
                      tuneGrid = expand.grid(alpha = 0, lambda = lambda_seq),  # ridge
                      metric =  "Accuracy" # can change to RMSE for regression
)
ggplot(ridge_mod_cv)
ridge_mod_best <- ridge_mod_cv$finalModel$lambdaOpt
coef(ridge_mod_cv$finalModel, ridge_mod_cv$finalModel$lambdaOpt)
ridge_probs <- predict(ridge_mod_cv, train, type="prob")
```

```r
ridge_pred <- ifelse(ridge_probs[,2]>0.5, 1, 0)
sum(ridge_pred == train$vote12)/nrow(train)
plotROC(train$vote12, ridge_probs[,2], Show.labels=F)
ridge_probs_test <- predict(ridge_mod_cv, test, type="prob")
ridge_pred_test <- ifelse(ridge_probs_test[,2]>0.5, 1, 0)
sum(ridge_pred_test == test$vote12)/nrow(test)
plotROC(test$vote12, ridge_probs_test[,2], Show.labels=F)
# logistic regression with just the intercept
glm1 <- glm(vote12~1, data=train, family="binomial")
glm2 <- glm(vote12~., data=train, family="binomial")
step_best <- stepAIC(glm1,
                     direction="forward",
                     scope=list(upper=glm2, lower=glm1),
                     k=2)   #k = 2:AIC, k = log(n):BIC
coef(step_best)
best_probs <- predict(step_best, train, type="response")
best_pred <- ifelse(best_probs>0.5, 1, 0)   # 1 means vote for Romney
sum(best_pred == train$vote12)/nrow(train)
plotROC(train$vote12, best_probs, Show.labels=F)
best_probs_test <- predict(step_best, test, type="response")
best_pred_test <- ifelse(best_probs_test>0.5, 1, 0)
sum(best_pred_test == test$vote12)/nrow(test)
plotROC(test$vote12, best_probs_test, Show.labels=F)
coef_df <- data.frame(lasso = as.data.frame.matrix(coef(lasso_mod_cv$finalModel,
                                                   lasso_mod_cv$finalModel$lambdaOpt)),
                      ridge = as.data.frame.matrix(coef(ridge_mod_cv$finalModel,
                                                   ridge_mod_cv$finalModel$lambdaOpt))) %>%
  rename(lasso = s1, ridge = s1.1)  # rename columns

coef_df
coef_df[abs(coef_df) < 0.00001] <- NA
coef_df$best <- NA
coef_df[names(coef(step_best)), "best"] <- coef(step_best)
coef_df
max_steps = dim(train)[2]
best_cv <- rep(0, max_steps)
# All possible number of steps
for (j in 1:max_steps){
  # Iterate over folds
  for (i in 1:10){

    # Run selection up to j steps with k = 0 (no penalty)
    glm1 <- glm(vote12~1, data=train[-folds[[i]],], family="binomial")
    glm2 <- glm(vote12~., data=train[-folds[[i]],], family="binomial")
    step_best_cv <- stepAIC(glm1,
                            direction="forward", # forward stepwise selection
                            scope=list(upper=glm2,lower=glm1),
                            k=0, # no penalty for number of predictors
                            trace = 0, # no printed information
                            steps = j # the maximum number of steps to be considered)
    )

    # Get accuracy on withheld test set
    pred_probs <- predict(step_best_cv, train[folds[[i]],], type="response")
    pred_vals <- ifelse(pred_probs>0.5, 1, 0)

    best_cv[j] <- best_cv[j] + sum(pred_vals == train$vote12[folds[[i]]])
    # folds[[i]] accurately return indexes
    # count total accurate prediction for each step
  }
}
best_cv/nrow(train)
# Above values show the accuracy on training data for each step. For example, the first value 0.8496094

best_num_steps <- which.max(best_cv/nrow(train)) # CV on best number of steps
best_num_steps = 16
glm1 <- glm(vote12~1, data = election, family="binomial")
glm2 <- glm(vote12~., data = election, family="binomial")
step_best_final <- stepAIC(glm1,
                           direction="forward",
                           scope=list(upper=glm2,lower=glm1),
                           k=0,
                           trace = 0,
                           steps = best_num_steps)
coef(step_best_final)
length(coef(step_best_final))
length(coef(step_best))
best_probs_noreg_test <- predict(step_best_final, test, type="response")
best_pred_noreg_test <- ifelse(best_probs_noreg_test>0.5, 1, 0)   # 1 means vote for Romney
sum(best_pred_noreg_test == test$vote12)/nrow(test)
plotROC(test$vote12, best_probs_noreg_test, Show.labels=F)
#(data, outcome)
coef_df[names(coef(step_best_final)), "best_cv"] <- coef(step_best_final)
coef_df
```