Big Data Analytics: Individual Assignment 1

Name: Akshay Kharbanda
Student No.: 251267346

Approach

- Upon inspecting the arrest_data.text file, several convicts had quite a few 'NA' entries for the employment status data for the 52 weeks. To handle this missing values, the NA was replaced by the most common value of the employment status in that week using the 'mode' function.
- Before performing logistic regression, I checked the possible multicollinearity issues amongst the various variables in our dataset. For this, the 'vif' function was used to find out variables with VIF more than 10, indicating high multicollinearity. Based on this result, most of the 'emp' variables showed high multicollinearity, which meant that these variables for employment status of the convicts at each week following their release were highly correlated and likely represented redundant information.
- To fix the same, I performed PCA on these 52 'emp' variables to capture the most important patterns in the employment status data along with reducing the dimensionality. The first principal component was used as a variable named 'emp_comp' to represent the 52 'emp' variables.
- Logistic regression was performed to predict 'arrest' using the 'emp_comp' variable along with other independent variables, namely, 'race', 'wexp', 'mar', 'paro', 'prio', 'educ' and 'fin'. The result is shown in Exhibit-1.
- Confusion matrix for the predicted values on validation data (Exhibit-2) was derived to evaluate the performance of the model by checking the sensitivity and specificity values.
- Further, 'roc' was used to derive the classification's optimal sensitivity and specificity values (0.3 for this case).

Results & Insights

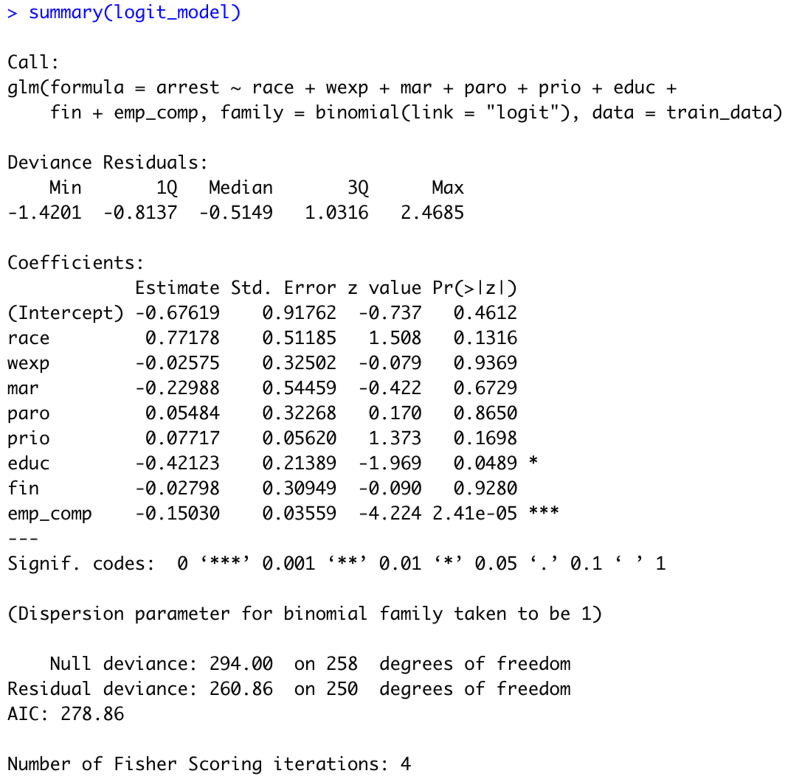
- Based on the result of the logistic regression (Exhibit-1), the variables that are found to be statistically significant (at the 0.05 level) for reducing recidivism were **education** (educ) and **employment status** after release (emp_comp variable, i.e., first component of the PCA for emp data) based on p-values.
- "fin" was found to be not statistically significant, suggesting that providing financial aid is not strongly associated with reducing the likelihood for arrest.
- Null deviance of 294 versus residual deviance of 260.86 suggests that this model explains only some of the variation in the data, but a lot of it is still not explained.
- Based on the result of the confusion matrix on the validation data (Exhibit-2), we see that the model correctly predicted the output on the validation data 68.21% of the times.
- The model sensitivity is 0.72,which means that the model performs well in identifying convicts who are at high risk of recidivism/rearrest (72% of the times), while the specificity is 0.58 (low), meaning that this model is not great at identifying convicts who are not re-arrested (58% times)

Recommendations

- Looking at the statistical significance of the various variables for explaining the effect on arrests, it can be ascertained that more emphasis should be given to improving education levels and employment opportunities for the released convicts, possibly through vocational training and learning programs for their personal and professional development.
- Based on the results, I recommend re-evaluating this financial-aid program to better target those at risk of recidivism better.
- It would be important to collect more information about the convicts in order to determine the best predictor variables leading to rearrests and improve the accuracy of the model. For example, information about their household income, social background, substance abuse, mental health, etc., could be a better predictor for rearrests, but that data has not been collected.

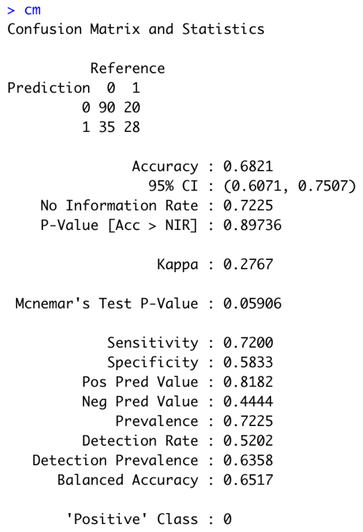
Exhibits

Exhibit-1



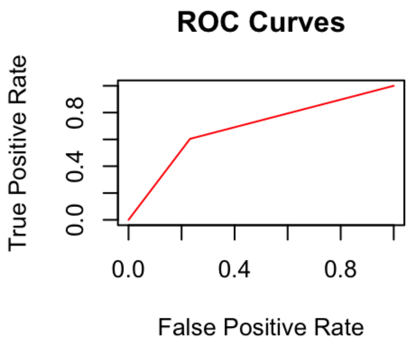
```
> summary(logit_model)

Call:
glm(formula = arrest ~ race + wexp + mar + paro + prio + educ + 
    fin + emp_comp, family = binomial(link = "logit"), data = train_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.4201  -0.8137  -0.5149   1.0316   2.4685  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.67619    0.91762  -0.737   0.4612    
race         0.77178    0.51185   1.508   0.1316    
wexp        -0.02575    0.32502  -0.079   0.9369    
mar         -0.22988    0.54459  -0.422   0.6729    
paro         0.05484    0.32268   0.170   0.8650    
prio         0.07717    0.05620   1.373   0.1698    
educ        -0.42123    0.21389  -1.969   0.0489 *  
fin         -0.02798    0.30949  -0.090   0.9280    
emp_comp    -0.15030    0.03559  -4.224 2.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 294.00  on 258  degrees of freedom
Residual deviance: 260.86  on 250  degrees of freedom
AIC: 278.86

Number of Fisher Scoring iterations: 4
```

Exhibit-2



```
> cm
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 90 20
         1 35 28

               Accuracy : 0.6821
                 95% CI : (0.6071, 0.7507)
    No Information Rate : 0.7225
    P-Value [Acc > NIR] : 0.89736

                  Kappa : 0.2767

 Mcnemar's Test P-Value : 0.05906

            Sensitivity : 0.7200
            Specificity : 0.5833
         Pos Pred Value : 0.8182
         Neg Pred Value : 0.4444
             Prevalence : 0.7225
         Detection Rate : 0.5202
   Detection Prevalence : 0.6358
      Balanced Accuracy : 0.6517

       'Positive' Class : 0
```

Exhibit-3



R-code used for this assignment:
https://github.com/Akshayy99/Big_Data_Analytics_IA1/blob/main/assignment-1.R