Humza Malik 251264428

Introduction

In the dataset, we are provided with employment status of convicts for the next 52 weeks along with some other variables that have a relationship with them getting arrested again once they are released from prison. The goal of the assignment is to evaluate the recidivism program put in place by the prison and make some recommendations to increase the effectiveness of this program.

Logistic Regression

To understand the relationship between our dependant variable (i.e., arrest) and other variables in the dataset, we first create a correlation plot. Looking at figure 1, we can see that variables like age, education and employment have the strongest relationship with arrest.

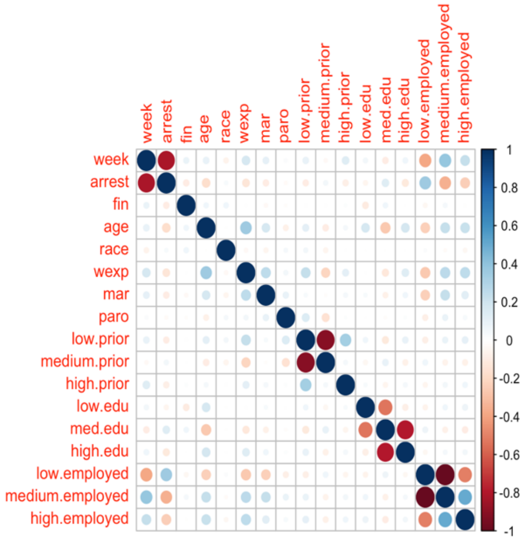
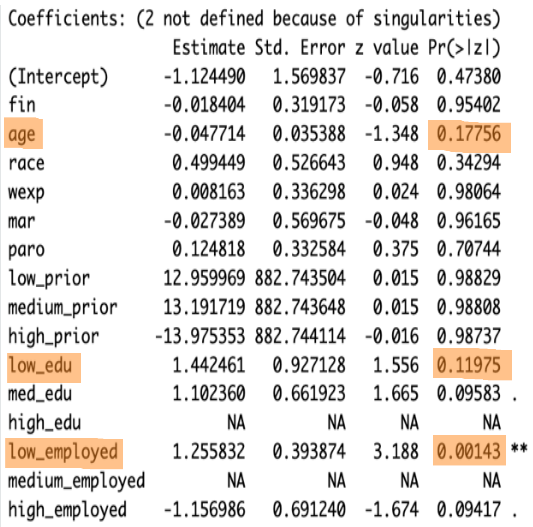


Figure 1: Correlation



```
Coefficients: (2 not defined because of singularities)
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.124490   1.569837  -0.716  0.47380
fin             -0.018404   0.319173  -0.058  0.95402
age             -0.047714   0.035388  -1.348  0.17756
race             0.499449   0.526643   0.948  0.34294
wexp             0.008163   0.336298   0.024  0.98064
mar             -0.027389   0.569675  -0.048  0.96165
paro             0.124818   0.332584   0.375  0.70744
low_prior       12.959969 882.743504   0.015  0.98829
medium_prior    13.191719 882.743648   0.015  0.98808
high_prior     -13.975353 882.744114  -0.016  0.98737
low_edu          1.442461   0.927128   1.556  0.11975
med_edu          1.102360   0.661923   1.665  0.09583 .
high_edu               NA         NA      NA       NA
low_employed     1.255832   0.393874   3.188  0.00143 **
medium_employed        NA         NA      NA       NA
high_employed   -1.156986   0.691240  -1.674  0.09417 .
```

Figure 2: Logistic Regression

Running a logistic regression would help establish a clearer relationship between arrest and these variables, as shown in figure 2 and highlighted in orange. While employment is statistically significant and should be looked at, I have chosen to select age and education too, based on their low P values, to make my recommendations more thorough and holistic.

Recommendations

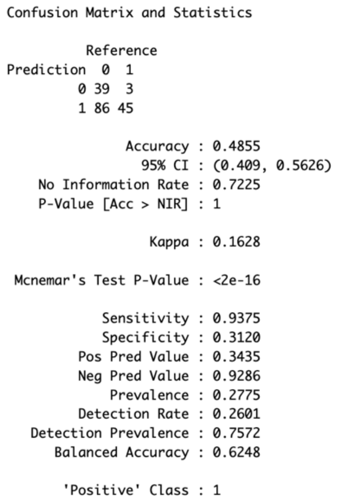
1. Employment: The prison should actively work to help convicts get integrated back into society by providing them with employment opportunities once they are

released. To make convicts stay committed to these jobs, convicts should be encouraged to work in prison on different jobs. The prison can also help these convicts get in touch with organizations that have a track record of helping felons with career opportunities.

2. Education: Raising education level among convicts also makes them less likely to get arrested again. The prison can intervene here by providing them with vocational training, workshops and skill development. The prison can also develop partnerships with state and federal level ministries, and community colleges to ensure felons continue their education once they are released.
3. Age: Since most of the convicts are young, they have a higher probability of getting arrested again. To minimize this, the prison could offer counselling services and induct young convicts into state-run social programs that ease their transition into society once they are released.

Methodology & Analysis

To ensure the results of the logistic regression were credible, we created a confusion matrix whose results can be seen in figure 3.



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 39  3
         1 86 45

               Accuracy : 0.4855
                 95% CI : (0.409, 0.5626)
    No Information Rate : 0.7225
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1628

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9375
            Specificity : 0.3120
         Pos Pred Value : 0.3435
         Neg Pred Value : 0.9286
             Prevalence : 0.2775
         Detection Rate : 0.2601
   Detection Prevalence : 0.7572
      Balanced Accuracy : 0.6248

       'Positive' Class : 1
```

Figure 3: Confusion Matrix

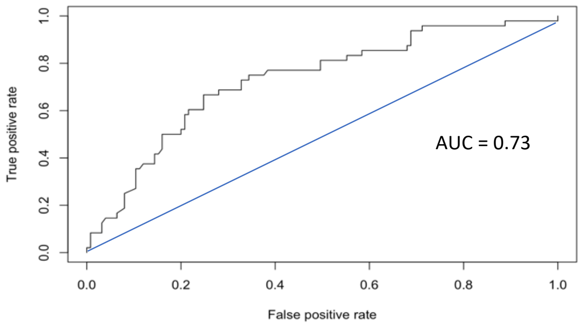


Figure 4: ROC & AUC

Looking at the sensitivity rate, we were able to predict correctly 93% of convicts who did not get arrested again. This adds credence to our logistic regression results.

Similarly, to help us decide which probability threshold level is optimal to classify convicts being arrested again or not, we used ROC and AUC whose results can be seen in figure 4. This allows us to compare the results of different threshold levels on the sensitivity rate, without having to create multiple confusion matrices.

With an AUC of 0.73, which is close to 1, we can see that the model has a desirable level of separability, making our results from the logistic regression more robust.