Big Data Analytics Individual Assignment 1

To attempt to predict recidivism is a difficult but very important problem. The difficulty comes from the multitude of variables that could be analyzed to see what trends we could conclude with from the data, with the importance being that if done correctly, we could help predict what criminals were to be more likely to reoffend. With that knowledge we could not only protect overall society, but we could help to understand the why, which could help us to further reduce repeat offenders.

When looking at the dataset provided, I wanted to run a logistical regression to understand the variety of variables and their relationship to the dependent variable, "arrest". First I created a variable based on the weeks data for each inmate that I called "sumweeksworking" which summed all of the "1" values for each inmate across the sample. What I wanted to know from this variable was if there was a positive correlation between the number of weeks an inmate worked that year and if they were arrested. From there I ran a base model with all the variables included and found that many of the variables had very high p-values that were not statistically significant. After removing several variables, the model below shows the variables that I found to be most significant when running the regression, and while there was only one variable (sumweeksworking) that reached statistical significance, I felt it was important to analyze the other variables as well to draw conclusions from them (Exhibit 1).

Exhibit 1: Model and Logistic Regression

```
Call:
glm(formula = arrest ~ agehigh + edhigh + priorlow + sumweeksworking,
    family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4438  -0.7553  -0.4087   0.9326   2.5173

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.60743    0.37021   1.641   0.1008
agehigh         -1.68101    1.07324  -1.566   0.1173
edhigh          -1.16118    0.66503  -1.746   0.0808 .
priorlow        -0.63204    0.38828  -1.628   0.1036
sumweeksworking -0.05705    0.01098  -5.194 2.06e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the data provided, the most significant indicators of recidivism when looking at the training data include the sum of weeks worked in the year of data, if the criminal was older, if the criminal had a higher education level, and if the criminal had fewer prior arrests.

When looking at the confusion matrices of the training data and validation data, we see a slightly higher accuracy on both versus the No Information Rate, meaning that the models on both data sets are better at predicting recidivism (Exhibits 2 and 3). We also can see that the accuracy on the training data is slightly higher, inferring that the model may not be as strong at predicting recidivism on larger data sets.

As well, both models have high Specificity and low Sensitivity, meaning that while the model is fairly successful at predicting who will not return to jail, the model is not accurate at predicting who will return.

What we can recommend from this is that looking at how long someone is working at job after being arrested can be a way to predict who will return to jail, so maybe by providing convicts with a program that helps them return to the work force after being released from jail will reduce recidivism. Another recommendation is that it would be smart to further analyze the correlations between some of these factors, including the correlation between age and education and recidivism. While a correlation with recidivism and low prior arrest totals makes logical sense, there is not significant evidence to support that age and education play a strong factor in recidivism, and their higher significance may be due to the sample size. I would also recommend re-evaluating other variables such as financial aid and its correlation to recidivism – just because it was not statistically significant in this data set does not mean that it would not be significant if this test was expanded onto more convicts.

Exhibit 2: CM on Training Data

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 188  57
         1   5   9

               Accuracy : 0.7606
                 95% CI : (0.7039, 0.8113)
    No Information Rate : 0.7452
    P-Value [Acc > NIR] : 0.312

                  Kappa : 0.1491

 Mcnemar's Test P-Value : 9.356e-11

            Sensitivity : 0.13636
            Specificity : 0.97409
         Pos Pred Value : 0.64286
         Neg Pred Value : 0.76735
             Prevalence : 0.25483
         Detection Rate : 0.03475
   Detection Prevalence : 0.05405
      Balanced Accuracy : 0.55523

       'Positive' Class : 1
```

Exhibit 3: CM on Validation Data

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 115  37
         1  10  11

               Accuracy : 0.7283
                 95% CI : (0.6556, 0.7931)
    No Information Rate : 0.7225
    P-Value [Acc > NIR] : 0.4711813

                  Kappa : 0.1804

 Mcnemar's Test P-Value : 0.0001491

            Sensitivity : 0.22917
            Specificity : 0.92000
         Pos Pred Value : 0.52381
         Neg Pred Value : 0.75658
             Prevalence : 0.27746
         Detection Rate : 0.06358
   Detection Prevalence : 0.12139
      Balanced Accuracy : 0.57458

       'Positive' Class : 1
```