Big Data Analytics Individual Assignment 3

Overall I see financial aid as an effective tool in reducing the amount of recidivism. As per my analyses, it is shown that based on the data provided it reduces the rate of recidivism by 20-30%, which is a strong number considering the context of the situation. Thus, I would expand this to increase financial aid to more convicts as they are released from prison, as it will considerably reduce the amount of recidivism, help them start new lives, and help protect the community as a whole.

When analyzing the variables from the dataset, it was found by looking at different models that the variables with the highest impact on recidivism include agehigh, edhigh, priorlow, sumweeksworking and financial. The two most important variables to look at from my point of view are sumweeksworking and financial, as they focus on what happens to the convicts after leaving prison. The significance of these variables is important as they show that if we as a community can support convicts with work opportunities and financial aid as they leave prison, then they will be less likely to reoffend.

For the logistic regression, I wanted to analyze the data and specifically, the arrest data, against the many independent variables, to see which variables had the strongest relationship. I created a variable called "sumweeksworking" that summarized the data from the weeks data from the original data file. This told me the total amount of weeks they spent working of the 52-week period. With this variable specifically, the goal was to see whether there was a correlation between the length the inmate was working for during that period and if they were likely to reoffend. What I found when I ran the model with all the independent variables was that there were many variables that had very high p-values and were not statistically significant. Because of the number of variables and the high p-values, I began removing them to not overfit the model and to reduce multicollinearity. After removing variables with high p-values, I found the most statistically significant independent variables when running the regression included agehigh, edhigh, priorlow and sumweeksworking, and while sumweeksworking was the only value that reached true statistical significance, the other variables included proved to help with my analysis of the data and my recommendations.

```
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
```

When doing the exponential survival analysis, I wanted to see the importance of whether financial aid was effective in reducing recidivism. A survival analysis in this case is used to predict when the population will reoffend. Because I strictly wanted to see the effects of financial aid on arrest data, I used it as my only independent variable and found that when running a survival test on the financial variable when they do receive financial aid (e^-.363) = .696. This means that if a person recieves financial aid, then they have a 30% less likely chance of re-offending.

```
Call:
survreg(formula = Surv(week, arrest) ~ fin, data = arrestdata,
    dist = "exponential")
             Value Std. Error     z      p
(Intercept) 4.989      0.123 40.53 <2e-16
fin         0.363      0.190  1.91  0.056

Scale fixed at 1

Exponential distribution
Loglik(model)= -700.1   Loglik(intercept only)= -702
        Chisq= 3.71 on 1 degrees of freedom, p= 0.054
Number of Newton-Raphson Iterations: 5
n= 432
```

I then ran a survival analysis again using a Weibull distribution to see if the dataset was distributed closer to a Weibull distribution versus an exponential distribution. I also wanted to evaluate financial aid in the same way as the exponential survival analysis, so I inlcuded it as the only indepenent variable. I found that the financial variable (0.2711) as well as the survival test for when they do have financial aid (e^-.2711 = .763) were slightly different, and that the p-value was slightly lower, which suggests that the Weibull distribution is more accurate than the exponential distribuion to this dataset.

```
Call:
survreg(formula = Surv(week, arrest) ~ fin, data = arrestdata,
    dist = "weibull")
              Value Std. Error     z       p
(Intercept)  4.6890     0.1155 40.60 < 2e-16
fin          0.2711     0.1405  1.93 0.05365
Log(scale)  -0.3135     0.0902 -3.48 0.00051

Scale= 0.731

Weibull distribution
Loglik(model)= -694.7   Loglik(intercept only)= -696.6
        Chisq= 3.88 on 1 degrees of freedom, p= 0.049
Number of Newton-Raphson Iterations: 6
n= 432
```

I then further analyzed the data using a cox proportional-hazards model. This model allowed me to analyze how specific independent variables influenced the rate of recividism. Cox tells us how much a specific variable increases or decreases the hazard rate. For this model I analyzed the variables from the logistic regression, as well as financial aid. I found that the most significant variables in decreasing the hazard rate (recividism), was edhigh and age low, but that overall each of these variables decreased the rate of recidivism.

```
Call:
coxph(formula = Surv(week, arrest) ~ fin + agehigh + edhigh +
    priorlow + sumweeksworking, data = arrestdata)

                    coef exp(coef)  se(coef)      z       p
fin             -0.301819  0.739472  0.190283 -1.586   0.113
agehigh         -0.653380  0.520284  0.509980 -1.281   0.200
edhigh          -0.782650  0.457193  0.421003 -1.859   0.063
priorlow        -0.245386  0.782402  0.213214 -1.151   0.250
sumweeksworking -0.056131  0.945415  0.007685 -7.304 2.8e-13

Likelihood ratio test=94.06  on 5 df, p=< 2.2e-16
n= 432, number of events= 114
```