Final Assignment

Owen Craven, Thomas Culwell, Bowen Mince

Literary Connection (Demographics)

Ben Hayes—Predicting Criminal Recidivism with R. (n.d.). Retrieved May 6, 2020, from http://benhay.es/posts/predicting-criminal-recidivism-r/

The prevalence of demographic factors in the COMPAS model presents an interesting ethical dilemma: How many factors are important, and to what degree might they propagate algorithm biases?

While the data can be sorted to find associations between race and recidivism, our model opts to not include a race factor amongst our primary variables.

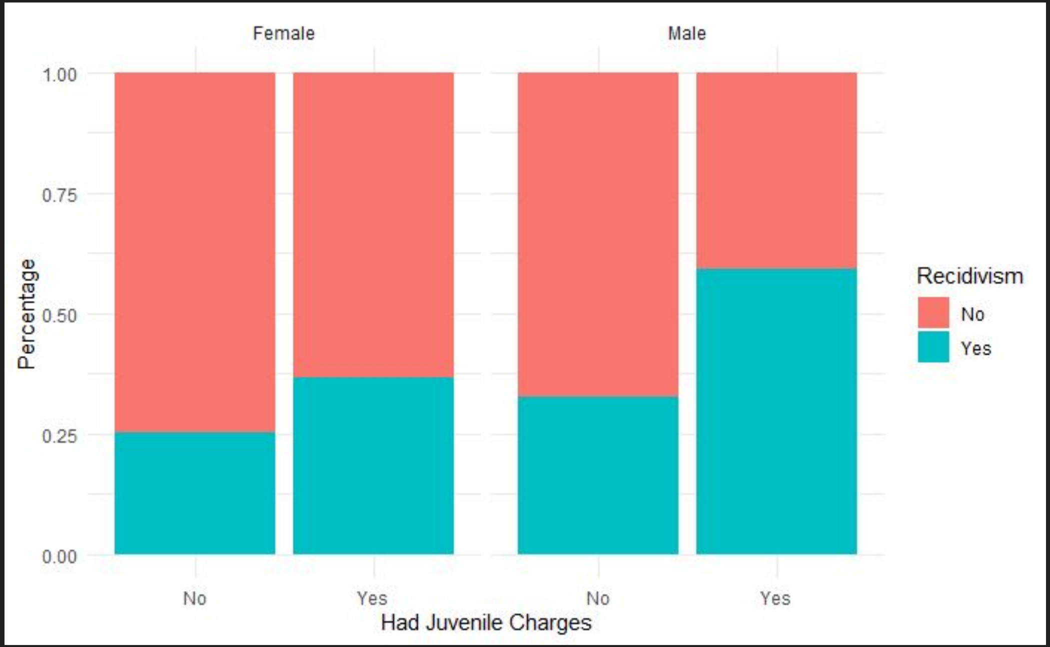
Literary Connection (Criminal History)

Criminal History and Recidivism of Federal Offenders. (2017, March 8). United States Sentencing Commission. https://www.ussc.gov/research/research-reports/criminal-history-and-recidivism-federal-offenders

The USSC attributes criminal history with strong indication of recidivistic tendencies, predicting rates of 30% to 85% in proportion to previous arrest counts.

Individuals without previous arrest counts saw 10% lower recidivism rates without prior contact with the criminal justice system.

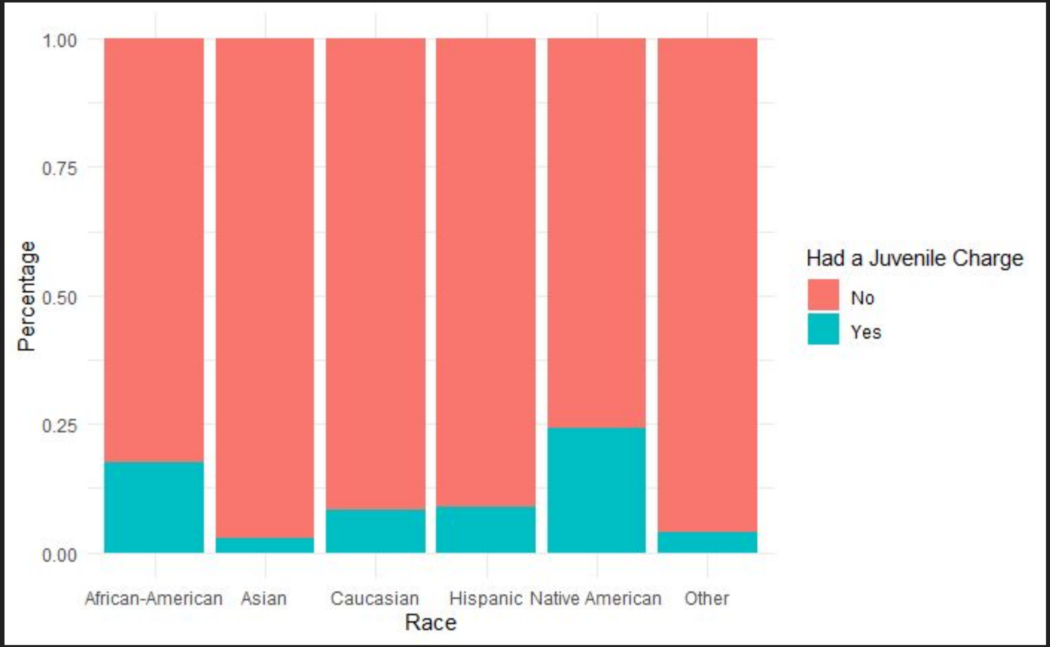
Juvenile Charges




```
CrimeDataTrain <- mutate(CrimeDataTrain, Juvenile_Charge = ifelse(juv_fel_count + juv_misd_count + juv_other_count > 0, "Yes", "No"))
```

12% of people in the data set had some sort of Juvenile charge.

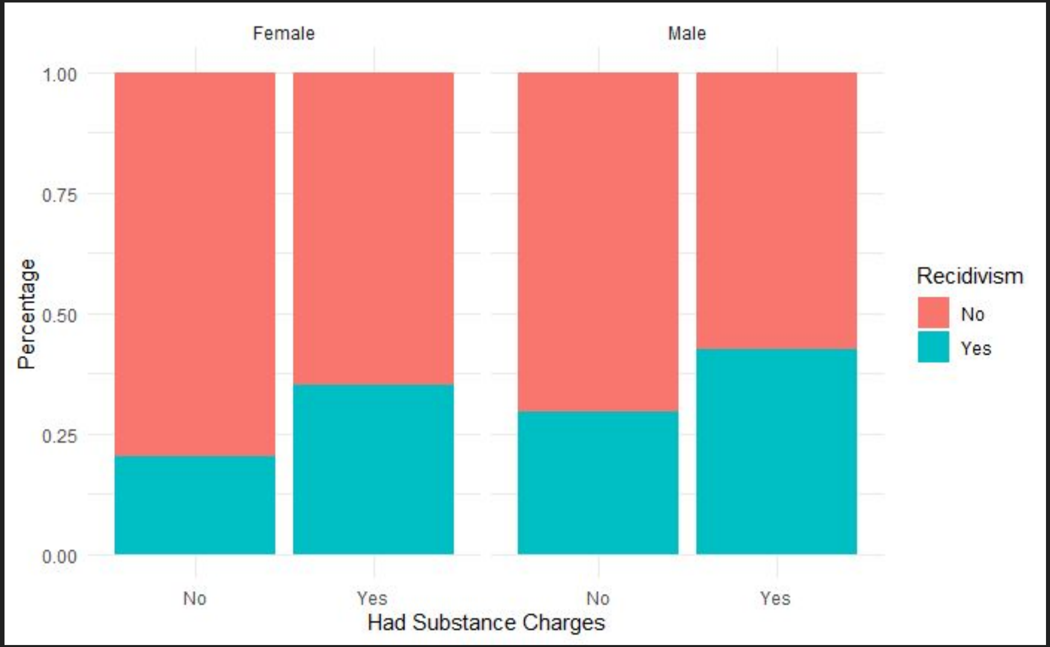
Race and Juvenile Charges




```
ggplot(data = CrimeDataTrain) +
geom_bar(aes(x = race, fill =
as.factor(Juvenile_Charge)), position = "fill") +
labs(x = "Race", y = "Percentage", fill = "Had a
Juvenile Charge") + scale_fill_discrete(labels
= c("No", "Yes")) + theme_minimal()
```

By race, African-Americans and Native Americans had more juvenile charges. (High proportion for Native Americans possibly a result of small sample size)

Substance Charges and Sex




```
CrimeDataTrain <- mutate(CrimeDataTrain, Substance = ifelse(Drugs > 0 | Alcohol > 0 | Type_Tobacco > 0, "Yes", "No"))
```

50% percent of people had charges that were substance related.

Building a Model

```
CrimeDataTrain <- filter(CrimeDataTrain, !is.na(in_date))

CrimeDataTrain <- filter(CrimeDataTrain, !is.na(violent_charge))

CrimeDataTrain <- filter(CrimeDataTrain, !is.na(Substance))

set.seed(123)

traincrime <- sample(1:nrow(CrimeDataTrain), 6000)

data.train <- (CrimeDataTrain[traincrime,])

data.test <- (CrimeDataTrain[-traincrime,])
```

We decided to look at a random 6000 individuals out of 7819 we had after filtering

Model Comparisons

Logistic Regression

```
TrainLogit <- glm(is_recid ~ Juvenile_Charge + priors + charges + arrests + age + convicted + Substance + sex, data = data.train, family = "binomial")
```

Training Data

	0 Pred	1 Pred
0 Obs	2601	1302
1 Obs	754	1343

Sensitivity: 63.4%

Specificity: 67.8%

Testing Data

	0 Pred	1 Pred
0 Obs	832	392
1 Obs	235	360

Sensitivity: 60.5%

Specificity: 68%

Random Forest

```
forest1 <- randomForest(is_recid ~ Juvenile_Charge + priors + convicted + arrests + charges + age + Substance + sex, data = data.train, importance=TRUE, ntree=100, mtry = 4, do.trace=TRUE)
```

Training Data (32.5% cutoff)

	0 Pred	1 Pred
0 Obs	3725	178
1 Obs	257	1840

Sensitivity: 87.7%

Specificity: 95.4%

Testing Data (31.5% cutoff)

	0 Pred	1 Pred
0 Obs	799	425
1 Obs	211	384

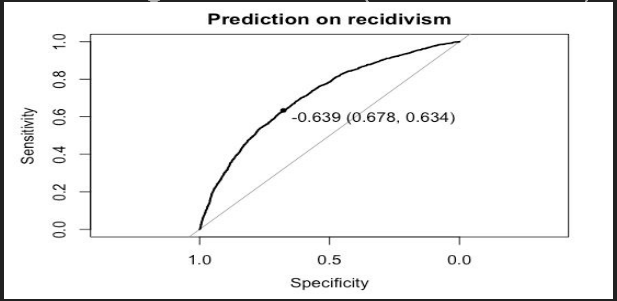
Sensitivity: 64.5%

Specificity: 65.3 %

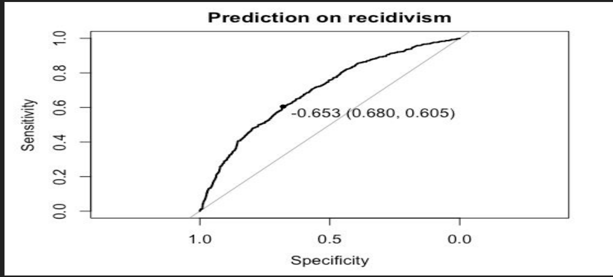
Model Comparisons

Logistic Regression

Training Data ROC (AUC = 0.711)

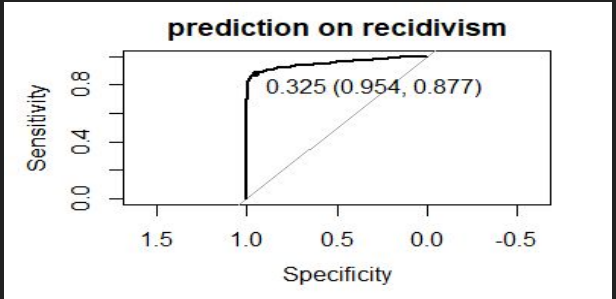


Testing Data ROC (AUC = 0.694)

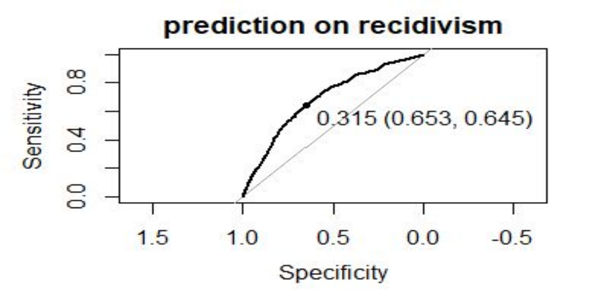


Random Forest

Training Data ROC (AUC = 0.95)



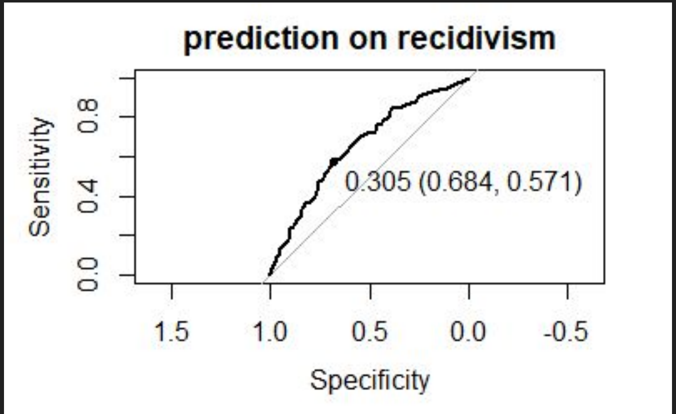
Testing Data ROC (AUC = 0.69)



Comparison with Whites and Non-Whites with Forest Model

Whites (30.5% cutoff)

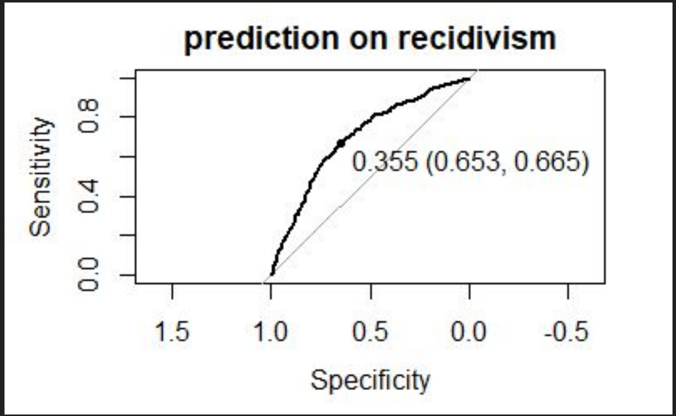
	0 pred	1 pred
0 obs	324	150
1 obs	81	108



AUC = .658

Non-Whites (35.5% cutoff)

	0 pred	1 pred
0 obs	490	260
1 obs	136	270



AUC = .696

Have a Wonderful Summer :)