Course Project Phase 2

PAULA MCCREE BAILEY

BAN 502

The Data

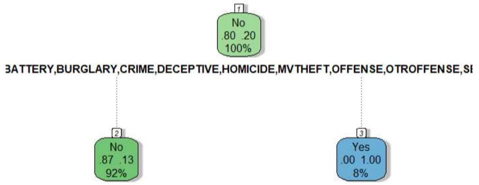
The dataset comes from the City of Chicago's Data Portal and contains information about crime in Chicago in 2018.

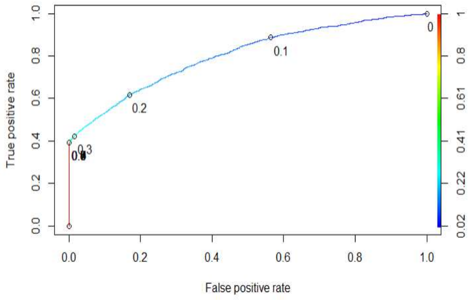
We reduced the size of the dataset by sampling 15,000 observations from the original dataset, which contained 267,000 observations.

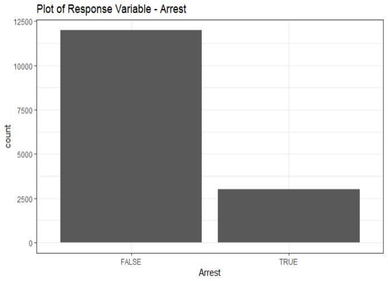
We built models to predict the likelihood that a person would be Arrested.

We used various models including Classification tree, Logistic regression, Stacking, Random forest, K-fold regression, and Regression with stepwise (Backwards and Forward).

Three will be discussed in this presentation : Classification, Logistic regression, and Stacking.


1
No
.80 .20
100%
BATTERY,BURGLARY,CRIME,DECEPTIVE,HOMICIDE,MVTHEFT,OFFENSE,OTROFFENSE,SI
2
No
.87 .13
92%
3
Yes
.00 1.00
8%





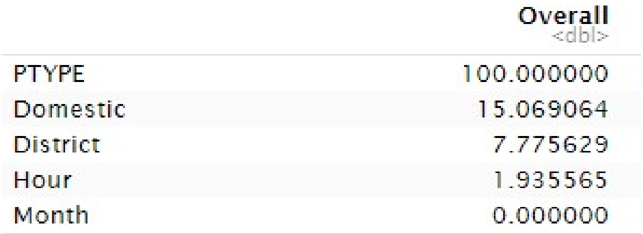
The Variables

Initially, we planned to use the following variables as predictors for Arrest (response): Primary Type, Domestic, District, FBI Code, Hours, Ward, and Date.

As mentioned, FBI Code is a redundant variable to Primary type, but has more detailed levels for crime. This is the same situation with Ward, which is redundant to District. This additional information is not necessary to determine the ability to predict Arrest.

As for the date, specifically the month, it was thought that certain months offered a greater predictor. For example, it is thought is crime rises during the summer months. This is not the case. See diagram to right.

Month does not provide any predictability for Arrest.



```
                         Overall
                           <dbl>
PTYPE                 100.000000
Domestic               15.069064
District                7.775629
Hour                    1.935565
Month                   0.000000
```

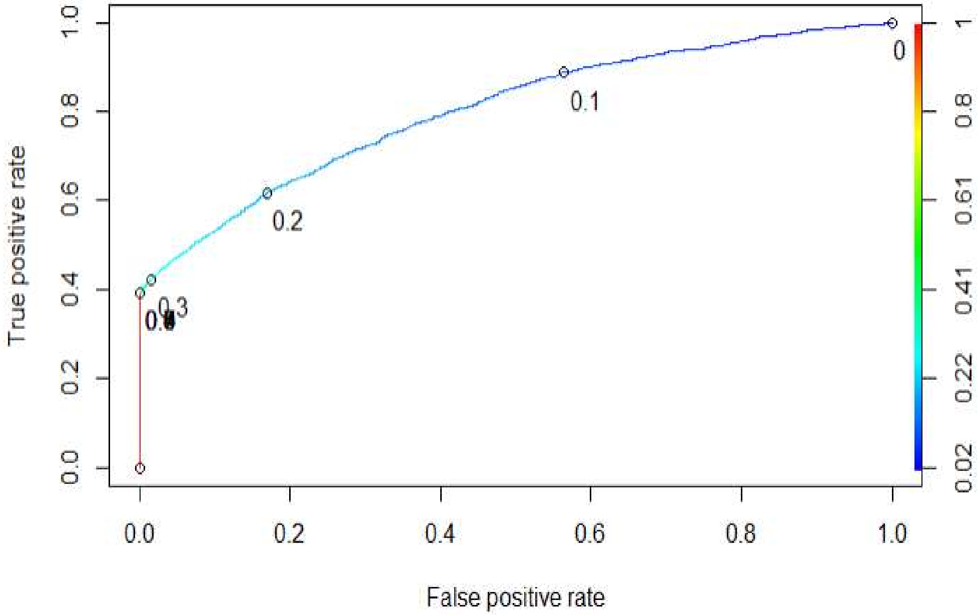
Our final predicator variables ended up consisting of Primary Type, Domestic, Hours, and District.

The results above indicate how well the predictor variables especially PTYPE (Primary Type) determined if a person was likely to be arrested.

The greater the number the greater influence this variable had on the Arrest, the response variable.

Although, Hour has a low predictability, combined with the other predictors it increased the response for being arrested.

Model: Logistic Regression



This Model uses the entire data sample set. There is no splitting of the dataset. This model is good. The area under the curve (AUC) measures the strength of the model, which was 80.26%.

When we used this threshold to tested the accuracy of the model , the accuracy rate declines 73.32%.

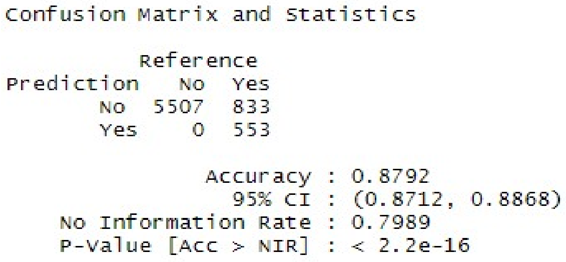
However, by increasing the threshold to 0.5, we can increase the accuracy to 87.78%. The threshold is just a cutoff point for our dataset.

Overall this model is good. It is important to remember that the dataset is skewed in favor of not being arrested.

Recall from the slides in first presentation – almost 12,000 out of 15,000 were not arrested.

Model: Stacking

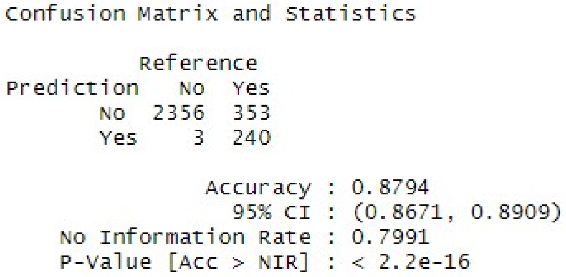
Training



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  5507  833
       Yes    0  553
                                          
               Accuracy : 0.8792          
                 95% CI : (0.8712, 0.8868)
    No Information Rate : 0.7989          
    P-Value [Acc > NIR] : < 2.2e-16       
```

Testing



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  2356  353
       Yes    3  240
                                          
               Accuracy : 0.8794          
                 95% CI : (0.8671, 0.8909)
    No Information Rate : 0.7991          
    P-Value [Acc > NIR] : < 2.2e-16       
```

This Model splits the dataset into a training (.70) and testing (.30) groups. In Stacking, we used General linear regression, Random tree, Classification tree, and Net neutral models to produce the best model.

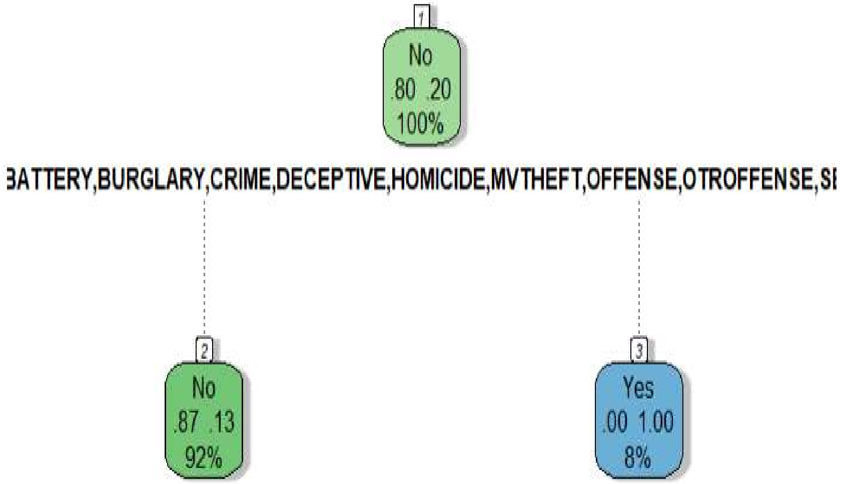
This model determined that the linear regression model was the best of the four models.

The training model performs at 87.92% regarding predicting the response. The testing data also performed a similar rate of 87.94%.

The naïve model – assumed no one is arrested only resulting in ~80.0%, which is much less, but much better than the training and testing models.

Recall from the slide in first presentation – almost 12,000 out of 15,000 were not arrested.

Model: Classification Tree


1
No
.80 .20
100%
3ATTERY,BURGLARY,CRIME,DECEPTIVE,HOMICIDE,MVTHEFT,OFFENSE,OTROFFENSE,SI
2
No
.87 .13
92%
3
Yes
.00 1.00
8%

This Model also splits the dataset into a training (.70) and testing (.30) groups. It uses all predictor variables to determine which one is best. In this situation, Primary Type influences the response variable Arrest.

The type of crime commended determines if the person is likely to be arrested or not.

The training model performs at 87.7% regarding predicting the response. The testing data performed at a rate of 87.97%.

The naïve model – assumed no one is arrested only resulting in ~80.0%. Again, it is lower than the models above, but not better.

Summary

Model Used	Performance Training & Testing Data	Best Model
Classification Tree		
Training	87.70%	N/A
Testing	87.97%	
Logistic Regression		
	80.27%	N/A
Adjusting Threshold	87.78%	
Stacking		
Training	87.92%	Logistic Regression
Testing	87.94%	

All three methods provided similar results and preformed well with the Chicago dataset. It is important to recognized the sample was skewed towards not being arrested. If we selected one model, Stacking would be the best model for our dataset.