Springboard Capstone 1: Bond Factors for CT Pretrial Detainees

Misty M. Giles

Overview

Problem:
Connecticut placed undue financial burden on citizens arrested for non-violent misdemeanors and intended to reform cash bonds with 2016's "An act concerning pretrial justice reform."

While researching the effect of the legislation, I found evidence of gender bias and created a model to predict gender of a detainee.

Data - Source

CT Department of Corrections:

- Anonymized demographic data on "individuals being held in Department of Corrections facilities while awaiting trial" claiming to be updated "nightly"

- About 2.5 million rows for 30 thousand detainees between 1 July 2016 and 30 June 2018 (combined misdemeanor and felony)

- https://data.ct.gov/Public-Safety/Accused-Pre-Trial-Inmates-in-Correctional-Faciliti/b674-jy6w

Progress - Data Wrangling (condensing)

Condensing:

- One row per detainee per arrest

- Preserved most recent information per arrest (info as of last date available)

- New structure: approx 40 thousand arrest rows for 30 thousand individuals (average 1.4 stays each)

Filtering:

- Removed rows where misdemeanor or felony couldn't be determined

- New structure: approx 29 thousand rows for 23 thousand individuals (average 1.3 stays each)

Progress - Data Wrangling (missing values)

Days detained:

- Calculated, estimate

- Subtract last date inmate recorded in system from latest admission date

- Observed at least one identifier with two, two-week gaps between contiguous rows with no change in admission date

Offense classification:

- Derived

- Last two characters of "offense" value, when available

- One quarter without code

- 60% without code stated "probation violation"

Progress - Data Wrangling (outliers)

Bond amounts:

- 93 arrests with bond amounts less than $100

- CT: "for particularly low (less than $100), this bond amount may be considered a place holder [sic] value"

- Law put in place to reduce cash bonds for lower-income individuals, so none cut from sample

Latest admission dates:

- 619 admission dates predate 2014

- CT: "if a date is more than one year old, an inmate should not be considered to have been held for the entire duration"

- Could be individual's original entry into CT DOC, so capped at 3 years (2 years in sample plus 1 year allowance)

Progress - Exploration

Misdemeanor demographics:

- 7709 arrests for 6562 individuals
- Age range: 16 to 86
- Median age: 34
- 83% male
- 43% white

Charging information:

- Bond range: $1 to $1,025,000
- Median bond: $7,550
- Median days detained: 20
- Most common offense: "Failure to appear, second degree AM" (23%)

Modeling

Gender Classification - Logistic Regression

Original model:

- Accuracy: 83%

- Recall: 100%

- Precision: 83%

- Area under ROC curve: 58%

- Model assumes all detainees are male

After rebalancing/tuning:

- Accuracy: 83%

- Recall: 56%

- Precision: 89%

- AUROC: 65%

- No longer assuming male, increased area under curve shows improvement of prediction power

Gender Classification - Random Forest

Model:

- Accuracy: 80%

- F1 score: 86%

- Recall: 89%

- Precision: 84%

- AUROC: 61%

- All scores closer to desired results

Gender Classification - Model Choice

Tuned Logistic Regression:

- Accuracy: 83%

- Recall: 56%

- Precision: 89%

- AUROC: 65%

Tuned Random Forest:

- Accuracy: 80%

- Recall: 89%

- Precision: 84%

- AUROC: 61%

With slight decreases elsewhere, the random forest model returns the correct result nearly 60% more often than logistic regression.

Further opportunities:

1. Use NLP techniques on offense column to further determine impact on non-violent offenders

2. Try binning and/or other curves to create predictive model for bond

3. Other?

____