Data is from compas-scores-two-years.csv

7214 individuals, 53 columns containing both numerical and categorical information. Columns are explained below along with some basic statistics about their contents.

The COMPAS score is either low (1-3), medium (4-6), or high (7-10). According to the ProPublica article, both medium and high are predictive of recidivism. When performing our analysis we could predict arrest, violent recidivism, misdemeanor, felony, high risk of violence, and high risk of arrest.

Throughout the data set, the following abbreviations are used.

c stands for COMPAS and refers to the initial offense when the COMPAS score was calculated.

r stands for recidivism and refers to the next offense after the COMPAS score was calculated.

v stands for violent in order to differentiate violent crimes from non-violent ones (vr is therefore violent recidivism) because there are separate predictions for violent crimes.

Categorical columns:

'name': Name

'first': First name

'last': Last name

'compas\_screening\_date',

'sex': Male or Female

'dob': Date of birth

'age\_cat': Greater than 45, 25-45, Less than 25

'race': Other, Black, White, Hispanic, Native American, Asian

'c\_jail\_in': Date that they checked into jail the first time (when the COMPAS score was calculated)

'c\_jail\_out': Date checking out of jail

'c\_case\_number': 13 digit alphanumeric string referring to the case where the COMPAS score was calculated

'c\_offense\_date': Date of COMPAS scoring offense

'c\_arrest\_date': Date of COMPAS scoring arrest

'c\_charge\_degree': Felony or Misdemeanor for COMPAS offense

'c\_charge\_desc': Description of the COMPAS offense charges, e.g. Grand Theft in the 3rd Degree

'r\_case\_number': 13 digit alphanumeric string referring to the recidivist case

'r\_charge\_degree': Felony or Misdemeanor for recidivist offense

'r\_offense\_date': Date of recidivist offense

'r\_charge\_desc': Description of the recidivist offense

'r\_jail\_in': Date for checking into jail for recidivist offense

'r\_jail\_out': Date for checking out of jail for recidivist offense

'vr\_case\_number': Violent recidivist case

'vr\_charge\_degree',

'vr\_offense\_date',

'vr\_charge\_desc',

'type\_of\_assessment': Risk of recidivism

'score\_text': Low, medium, or high

'screening\_date',

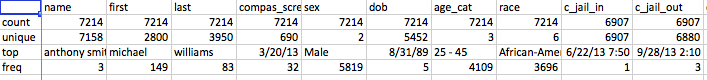
'v\_type\_of\_assessment': Risk of violence

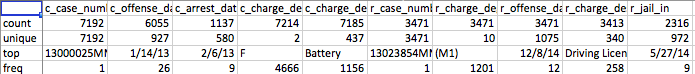
'v\_score\_text',

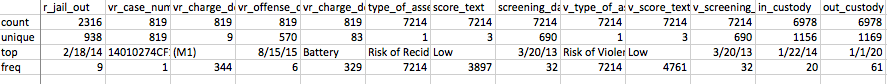
'v\_screening\_date',

'in\_custody',

'out\_custody':







Numerical columns:

'id': Id of the person

'age': Age

'juv\_fel\_count': Number of juvenile felonies

'decile\_score': COMPAS score

'juv\_misd\_count': Number of juvenile misdemeanors

'juv\_other\_count': Number of other juvenile crimes

'priors\_count': Number of prior arrests

'days\_b\_screening\_arrest',

'c\_days\_from\_compas',

'is\_recid': Whether or not the person was arrested again within 2 years

'r\_days\_from\_arrest': How long until the recidivist arrest

'violent\_recid': Not used

'is\_violent\_recid': Whether or not the recidivism was violent

'decile\_score.1': Same as decile\_score

'v\_decile\_score': COMPAS score for judging violent offenses

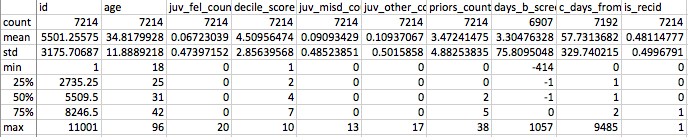
'priors\_count.1',

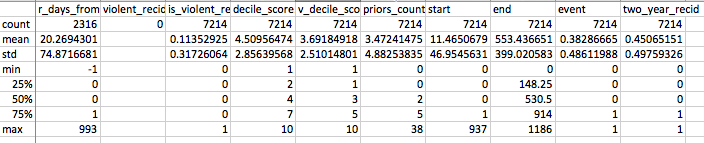
'start',

'end',

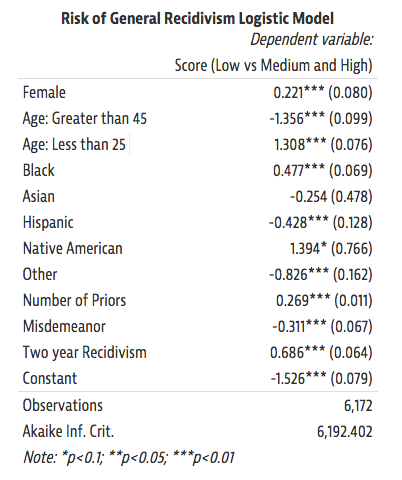
'event',

'two\_year\_recid': Slightly different but similar to is\_recid??? Not sure exactly what





ProPublica used these general factors in their logistic regression model.



Proposed binarized categories:

Is male

Age: <=18, 18-22, <=20, <=22, <=25, 24-30, 24-40, >=30, >=40

Black

Asian

Native\_american

Other  
# juvenile felonies - 0, 1-3, >=3, >=5  
# total juvenile crimes - same thing  
# total past crimes - 0 , 1-3, >=3, >=5, >=10  
# total violent crimes...  
age first jail time - <=18, <=24, 18-24, >=25, >=40