PREDICTING RECIDIVISM IN JOHNSON COUNTY

Ran Bi MSCAPP'18
Yanning Cui MAPSS'17
Ruochen Qiu MPP'18

INTRODUCTION

o   Recidivism is an important measure of the criminal justice system's effectiveness in reducing crime.

o   More than **40%** offenders return to state prison **within three years** after their release[1].

o   If states could reduce their recidivism rates by just **10 percent**, they could save more than **$635 million** combined in one year alone in averted prison costs.

1. Pew Center on the States. "State of recidivism: The revolving door of America's prisons." (2011). http://www.pewtrusts.org/en/research-and-analysis/reports/0001/01/01/state-of-recidivism

RELATED WORK

o Statistics from the Bureau of Justice tracked a sample of 404,638 former prison inmates from 30 states for five years following their release in 2005 and discovered that:

- Within three years of release, about 67.8 percent of released prisoners were rearrested
- Within five years of release, about 76.6 percent of released prisoners were rearrested
- Property offenders were the most likely to be rearrested, with 82.1 percent of released property offenders arrested for a new crime compared with 76.9 percent of drug offenders, 73.6 percent of public order offenders and 71.3 percent of violent offenders

o Three successful example of attacking recidivism: Oregon, Michigan and Missouri

o DSSG (Data Science for Social Good) at the University of Chicago used machine learning to prioritize outreach to individuals most at risk of being booked into jail within the next year. They argued that mental health problems might contribute in some individuals to an increased risk of recommitting crimes.

https://www.nij.gov/topics/corrections/recidivism/Pages/welcome.aspx
Salomon, Erika, et al. "Reducing Incarceration through Prioritized Interventions." (2017).
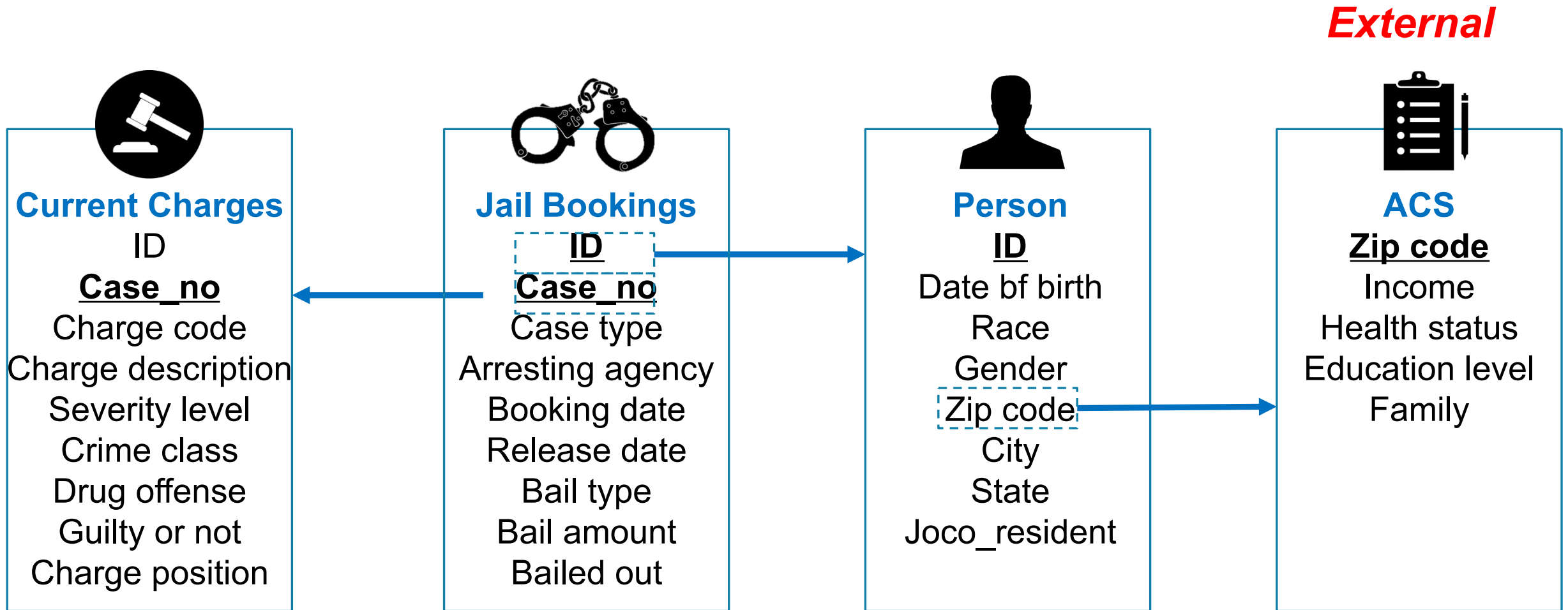
OUR GOAL

o Identify **top 200 individuals** who are most likely to return to jail during a 12-month period following his/her release

o Build a machine learning model to predict these high-risk individuals and find out what **factors** might contribute to the recidivism

o Help local government identify these individuals as early as possible and execute **customized interventions** based on their personal characteristics and circumstances
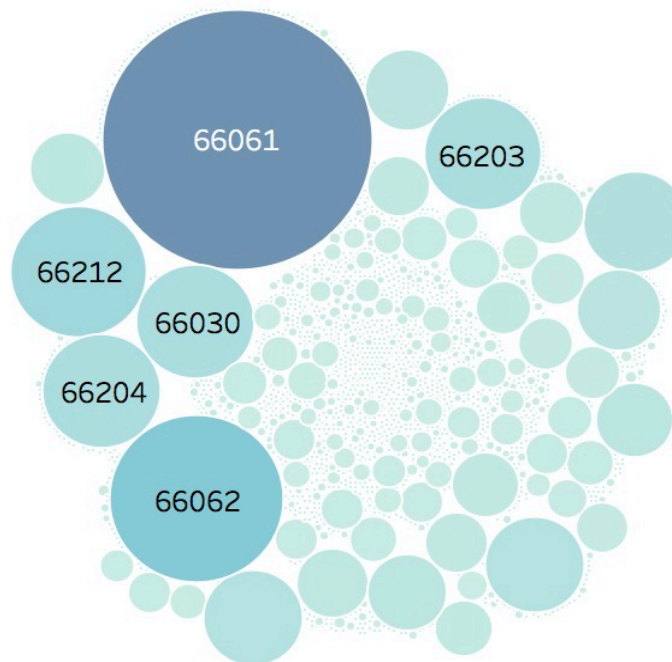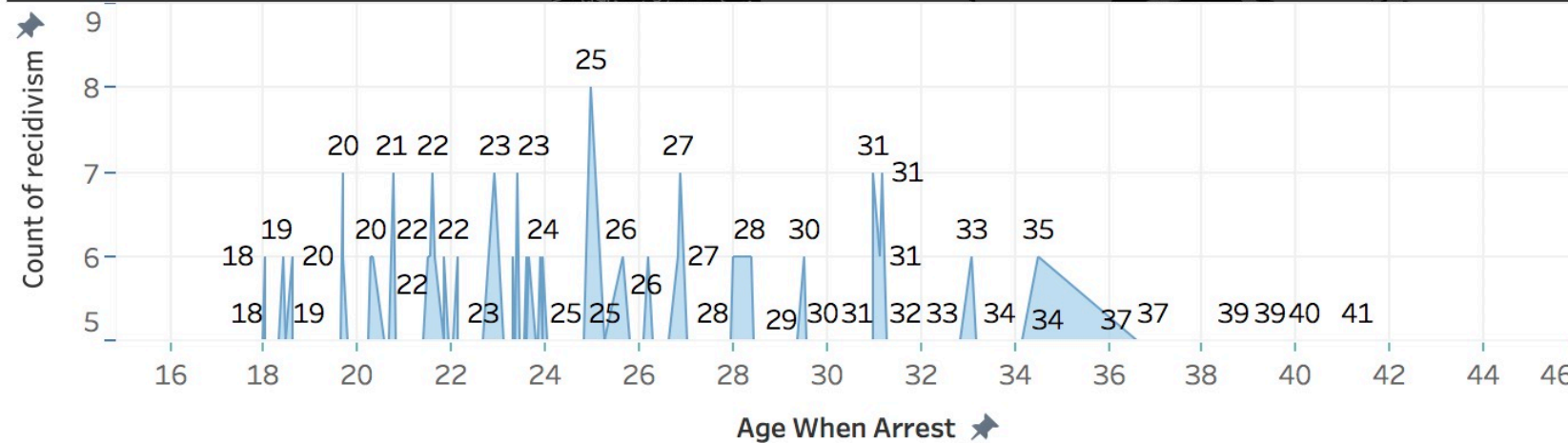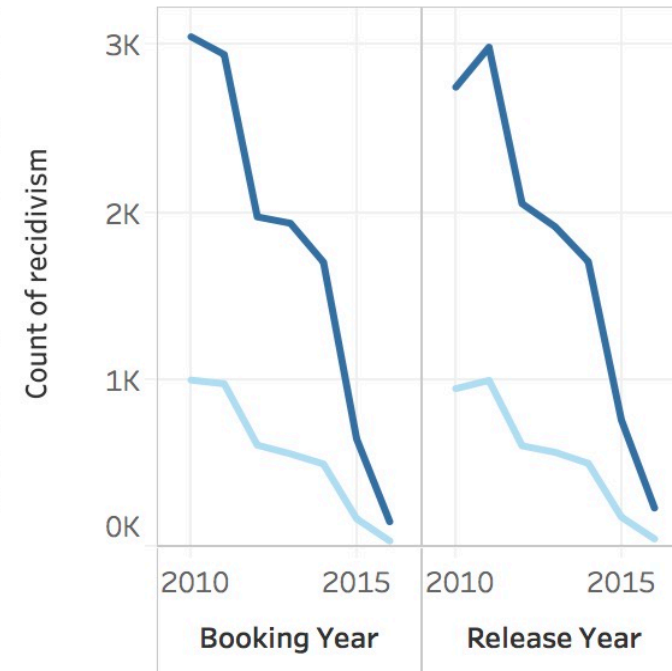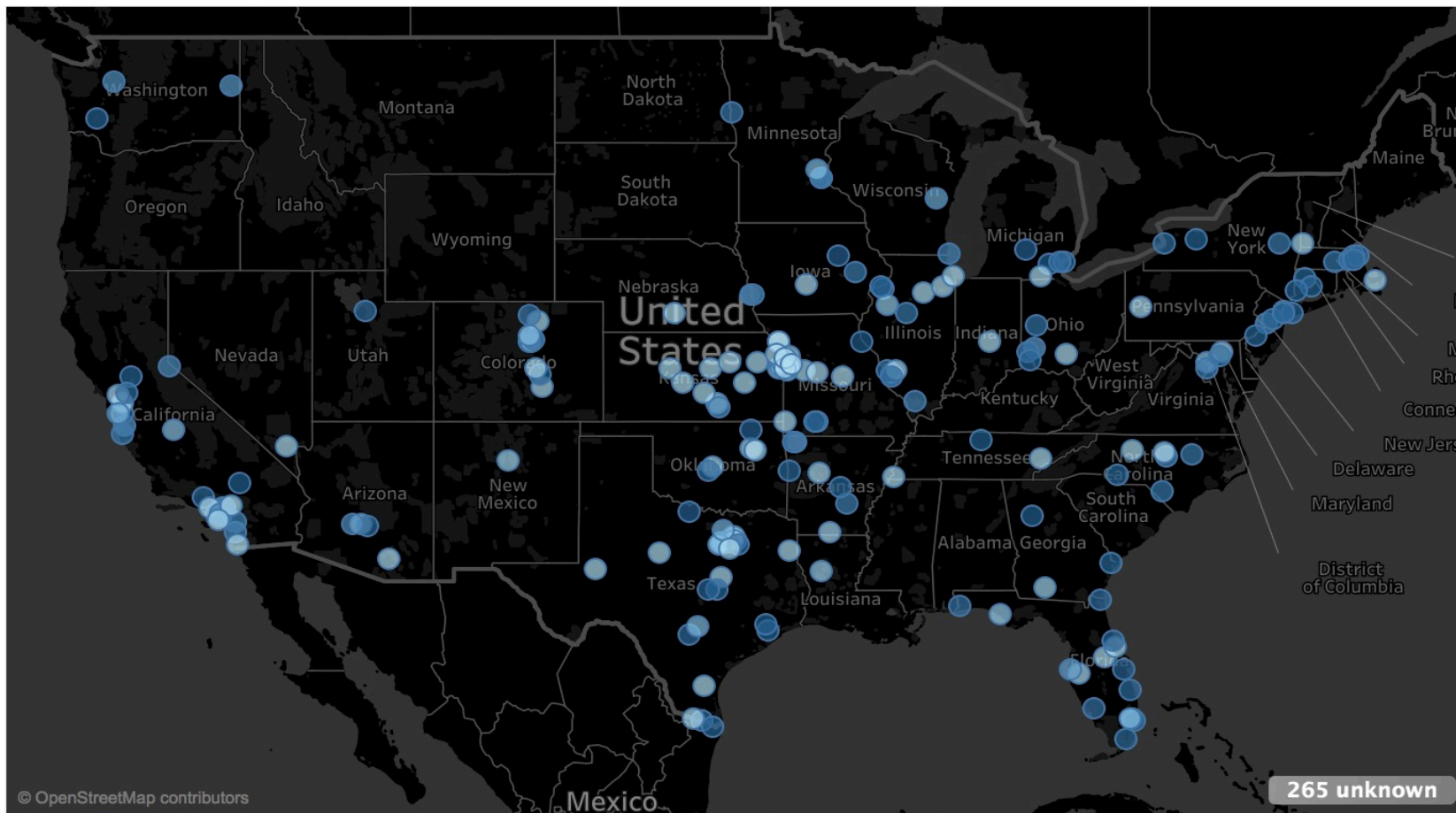
OVERVIEW
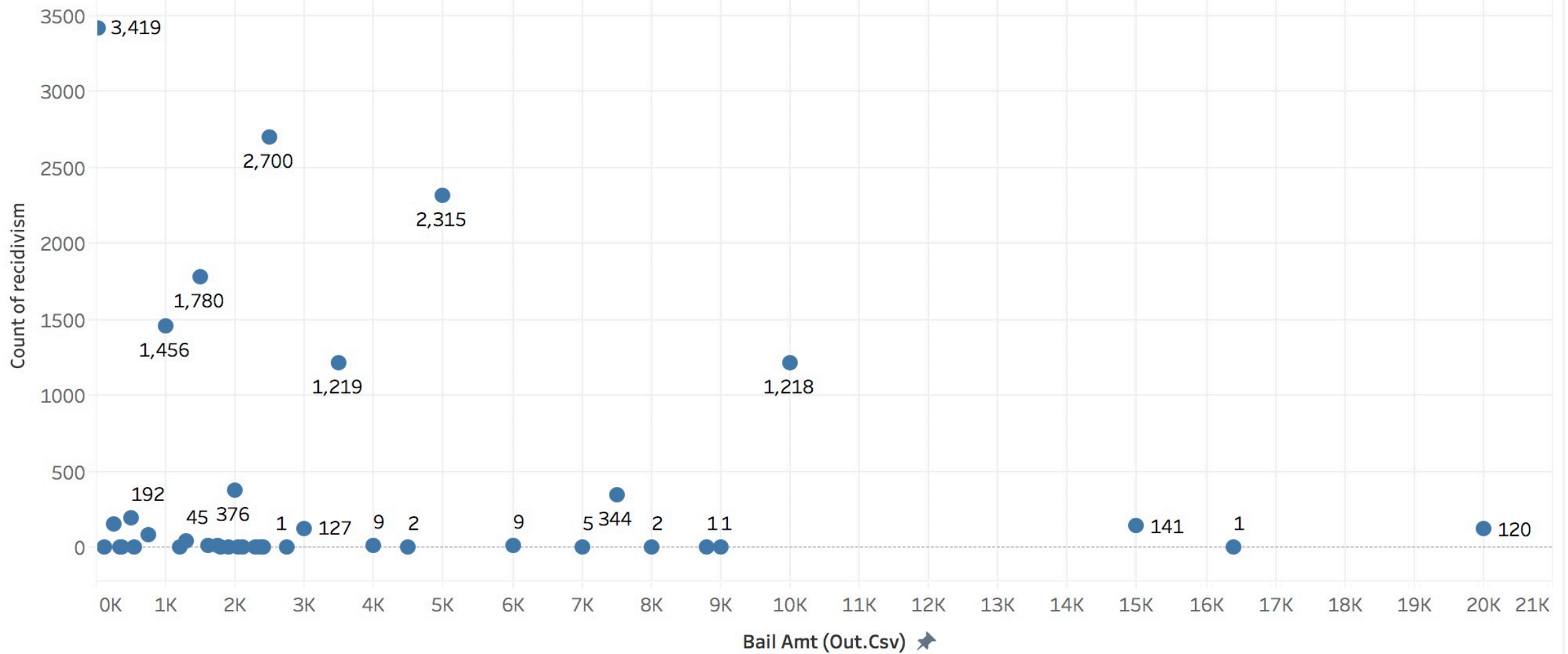
o   Our study estimates the recidivism patterns of based on data from Johnson county Criminal Justice Department

o   Use a 5-year follow-up period

o   28,579 jail bookings; 22,110 unique individuals; time span 2010-2016

o   We also utilize external data - American Community Survey 5-Year Data[1] (2009-2015)

o   We trained our model with data from jail system (booking, charge, and demographic information of individuals) as well as information of income, marital status, education level, and health insurance coverage from ACS.

1. https://www.census.gov/data/developers/data-sets/acs-5year.html

DATA SOURCES

External

Current Charges
ID
Case_no
Charge code
Charge description
Severity level
Crime class
Drug offense
Guilty or not
Charge position

Jail Bookings
ID
Case_no
Case type
Arresting agency
Booking date
Release date
Bail type
Bail amount
Bailed out

Person
ID
Date bf birth
Race
Gender
Zip code
City
State
Joco_resident

ACS
Zip code
Income
Health status
Education level
Family

Count of recidivism vs. Bail Amt (Out.Csv)

Data labels on scatter plot: 3,419 · 2,700 · 2,315 · 1,780 · 1,456 · 1,219 · 1,218 · 192 · 45 · 376 · 1 · 127 · 9 · 2 · 9 · 5 · 344 · 2 · 11 · 141 · 1 · 120

Bail Type

Bail Type	Count
CA	750
GPS	280
MISS	3,323
ORCD	615
PR	3,083
SUR	8,211

Case Type

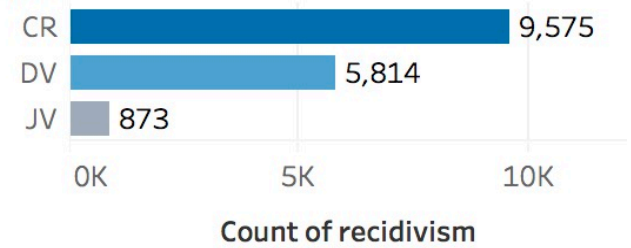| Case Type	Count of recidivism
CR	9,575
DV	5,814
JV	873

LABEL & FEATURE ENGINEERING

Label: Recidivism
- Recidivism = 1 if the individual is re-booked to jail within one year after releasing

Our Challenge
- Available data are mostly categorical

Feature Engineering Strategies
- Categorical to Binary
- **Categorical Histogram**
- Features for missing values
- Discretization
- Date/Time Features
- **Aggregates over different periods**

- Guilty Finding Count
- Trial Occurred Count
- Charge Description Count
- Missing Value

- Jail Booking Count Last Year / Month
- Average Bail Amount Last Year / Month
- In Jail Length & Length Buckets

- Arresting Age & Age Buckets
- Race Dummy
- Residency Dummy

- Income Buckets
- Age * Health Status Interaction

MODELLING

Classifiers	Hyperparameters	Rational
Decision Tree	criterion, max_depth, min_samples_split, min_samples_leaf	Simulate expert heuristics as evaluation baseline (Simple 1 or 2 deep Decision Tree)
K Nearest Neighbors	n_neighbors, weights, algorithm	Robust to noisy data, generally fit for classification questions
Logistic Regression	penalty, c, class weight	Mostly used by existing recidivism risk assessment tools[1]
Random Forest	n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf	Considers the nonlinear effects of a large number of variables with complex interactions
Gradient Boosting	n_estimators, learning_rate, subsample	Ensemble of weak prediction model; Reduce overfitting issues

https://www.nij.gov/journals/271/pages/predicting-recidivism.aspx
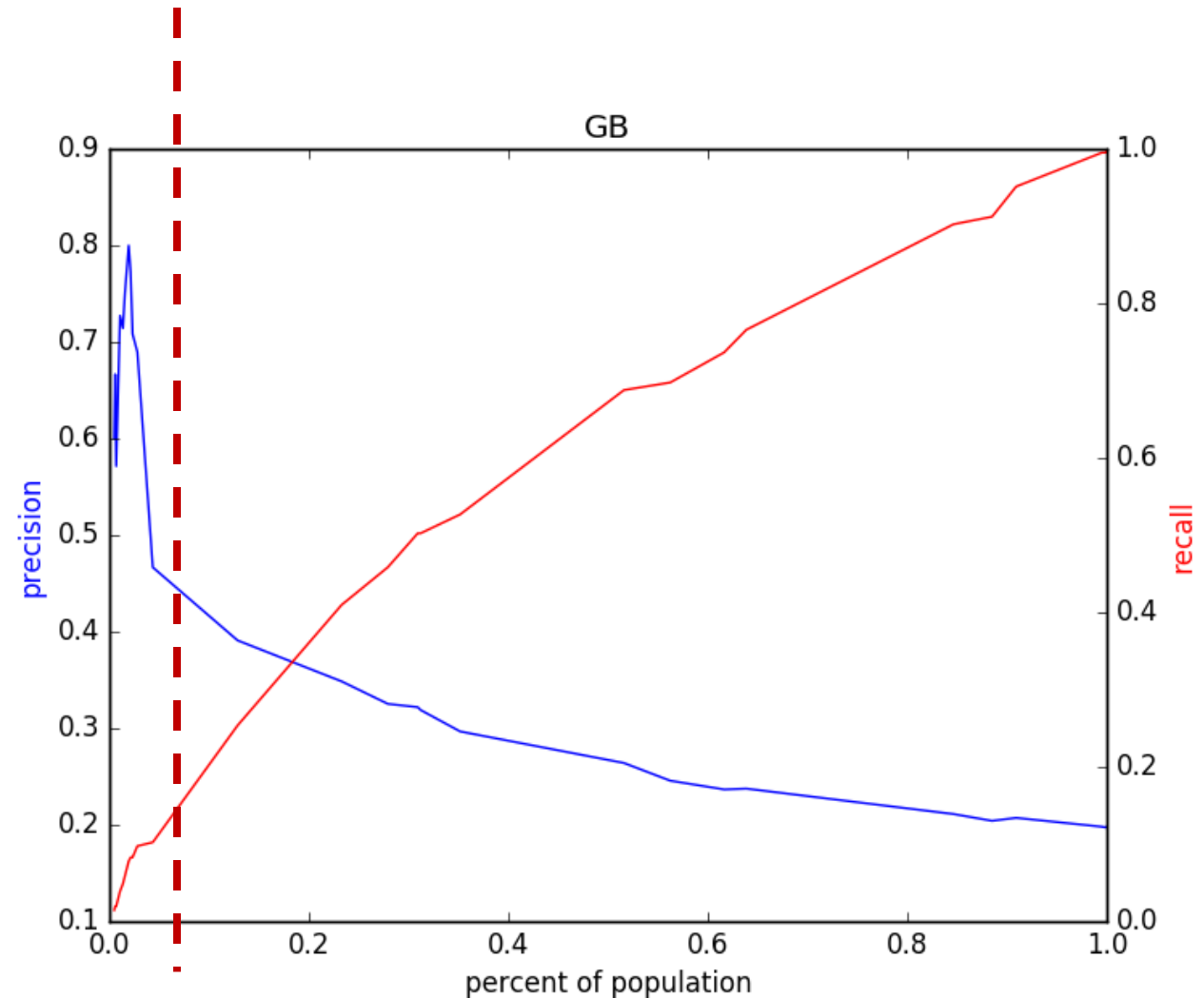
MODEL SELECTION
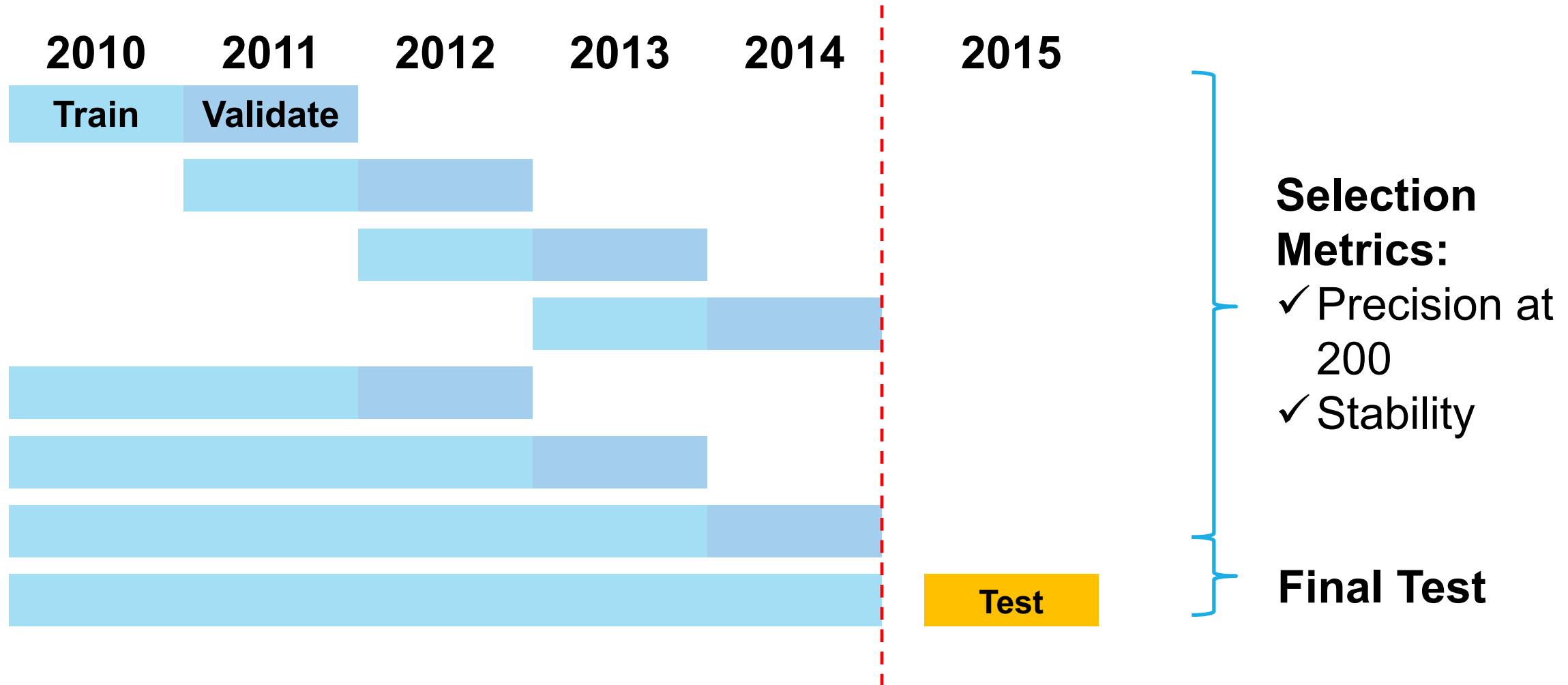
EVALUATION - METRICS

Best Model:

- GradientBoostingClassifier

  {'learning_rate': 0.001, 'max_depth': 5,
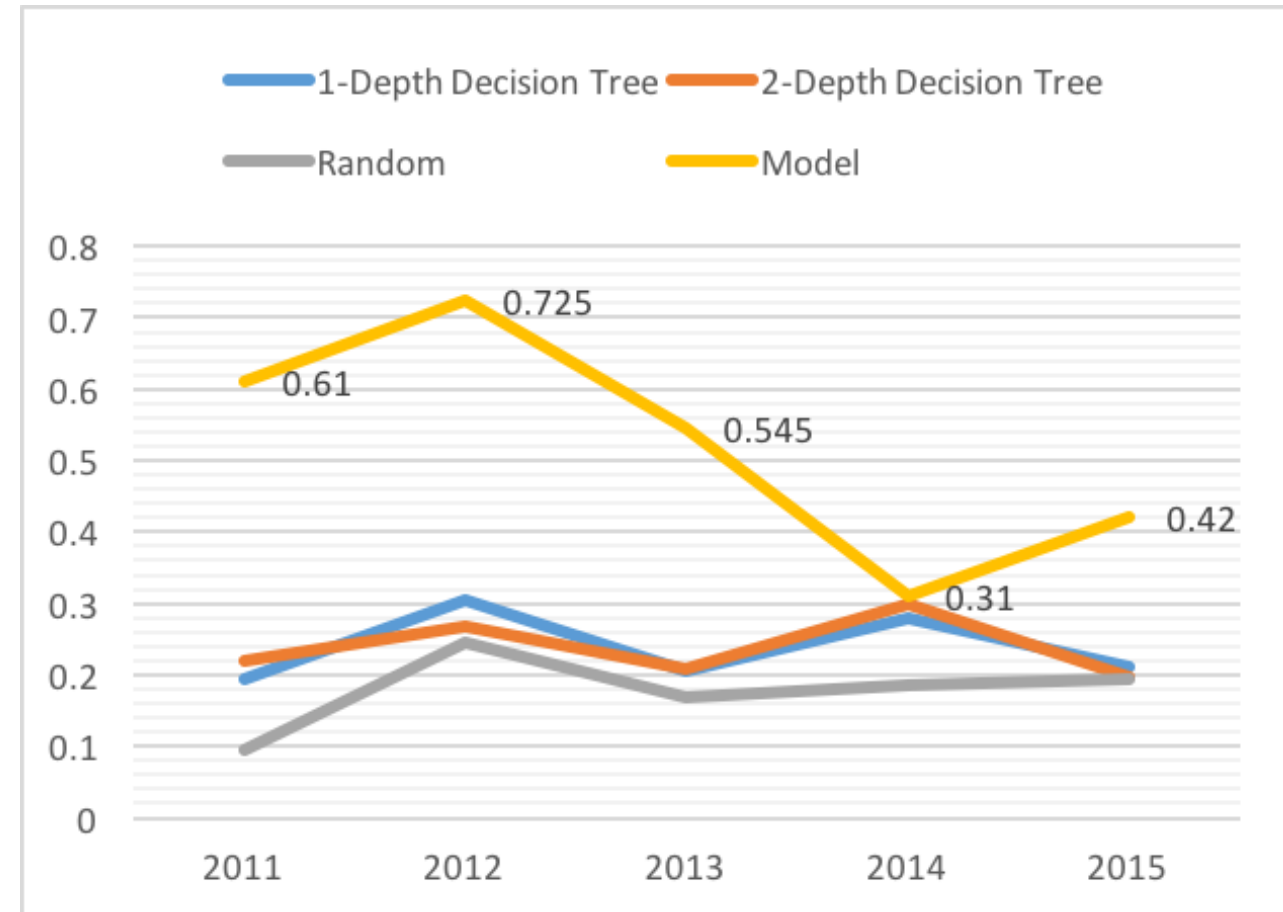
  'n_estimators': 10, 'subsample': 1.0}

Metrics Score:

- **Precision at 200: 0.42**

- AUC-ROC: 0.6444

EVALUATION - BASELINE

Baseline Cases:

- **Simple Heuristics**
  - Simulated by a decision stump
  - Classify individuals who had one booking in the last year as at risk

- **Expert Heuristics**
  - Simulated by a simple 2 depth decision tree
  - Prior booking + More than 1 charge in the case

- **Random Selection**

Legend: 1-Depth Decision Tree, 2-Depth Decision Tree, Random, Model

Data points (Model): 2011: 0.61, 2012: 0.725, 2013: 0.545, 2014: 0.31, 2015: 0.42

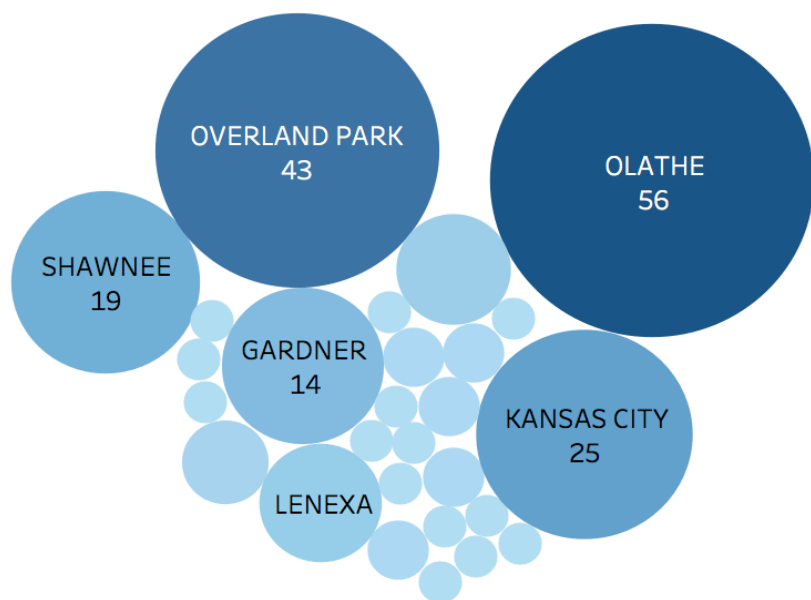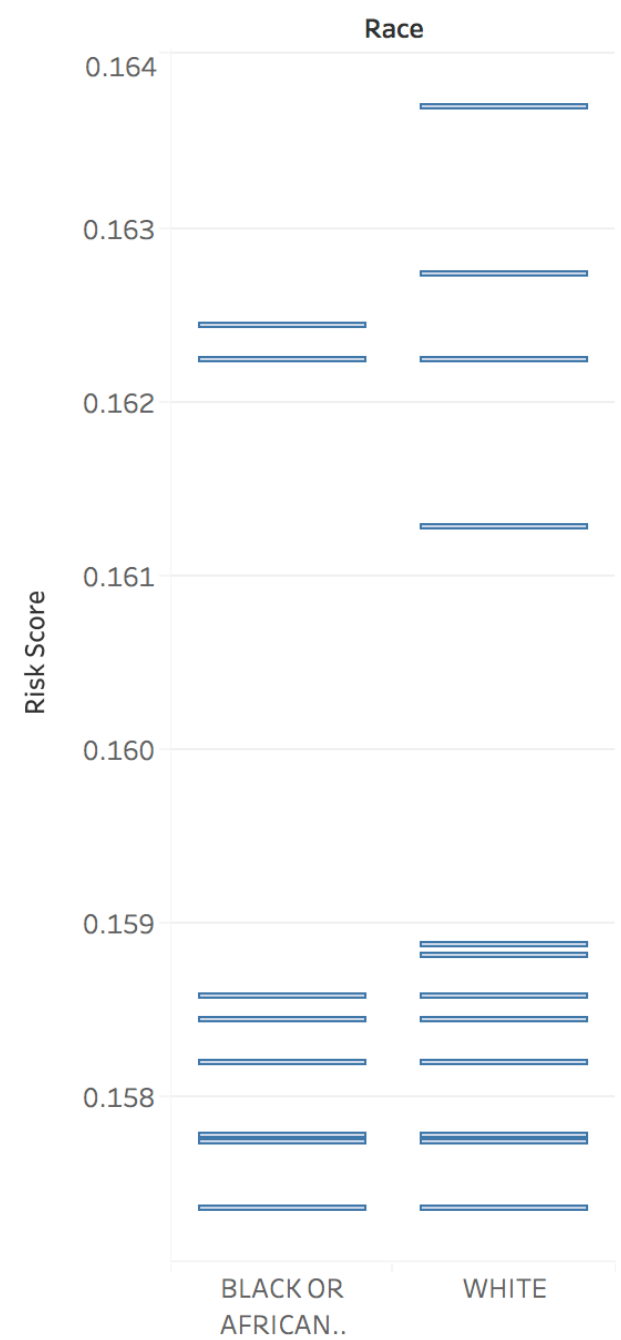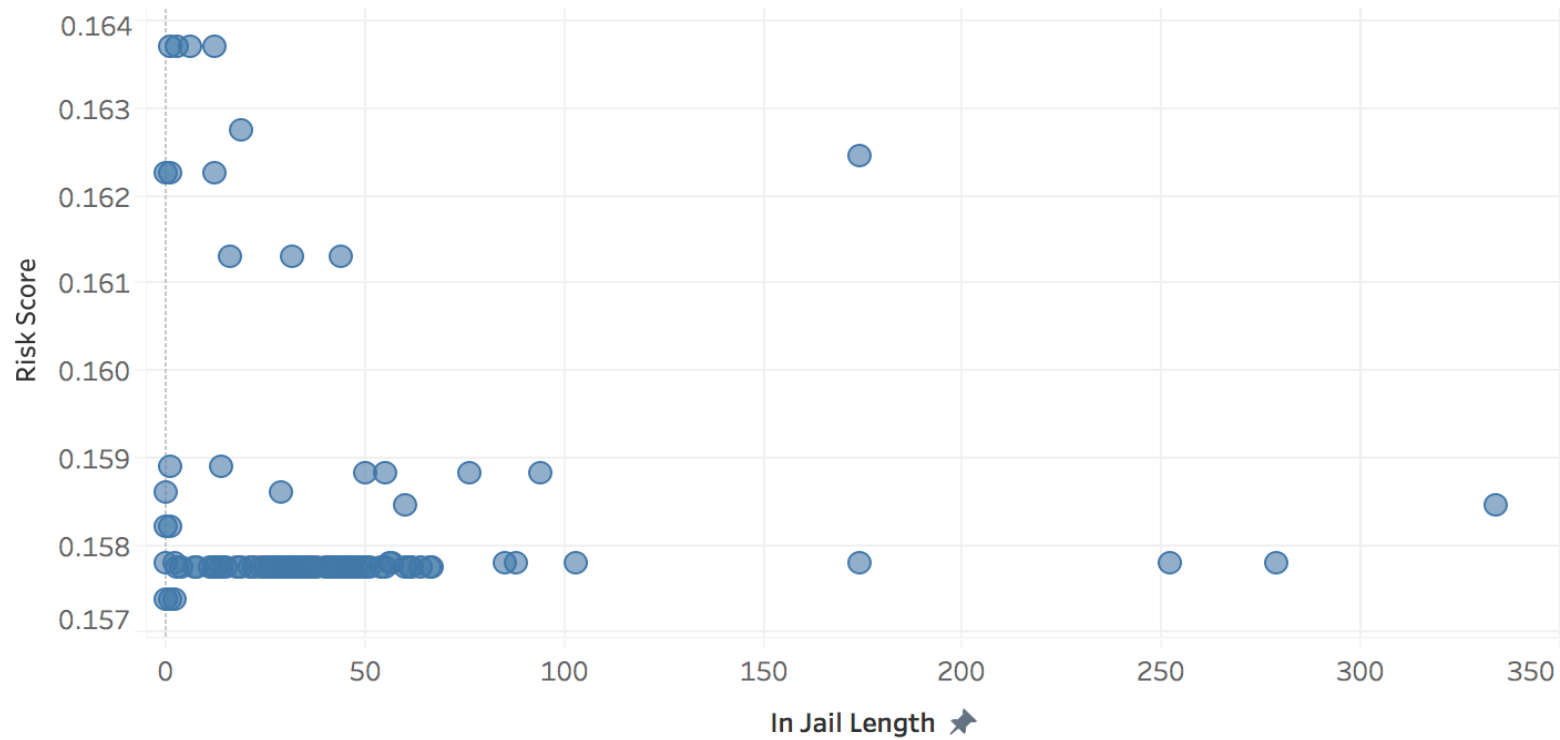**Precision for the top 200 individuals at risk**

RESULT

1. The individuals who had **multiple charges per case** are more likely to return.

2. The individuals who **aged between 19 – 24** have relatively high risk to return.

3. The individuals with a case type of **criminal** are more likely to return, compare with **domestic violence** and **juvenile**.
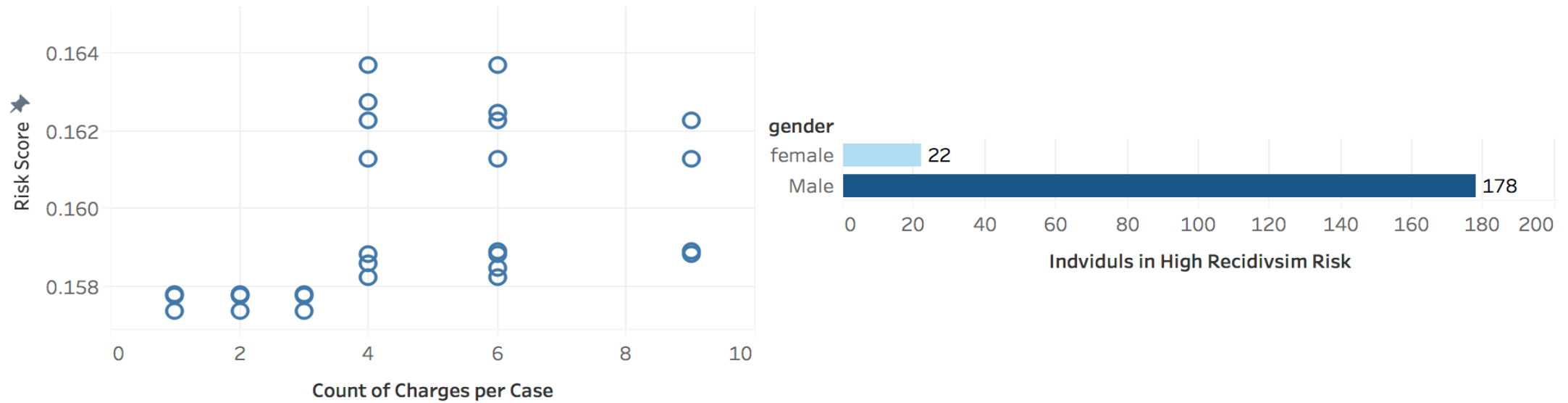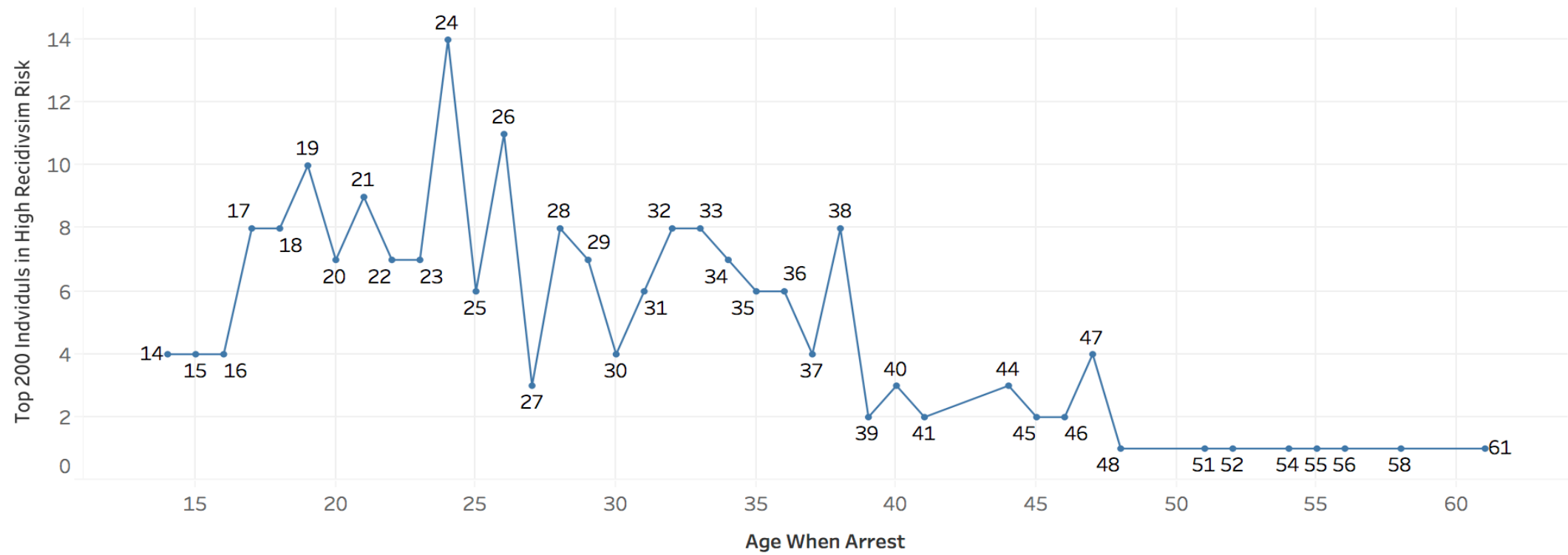
RESULT

4. Most individuals spent less than one day in jail and those who stay shorter have higher risk to return.

5. Most of the recidivists are from Olathe, Overland Park, Kansas City and Shawnee (supervision, risk assessment, job skills, correctional program)

6. The individuals who are bailed out through **surety bond** are more likely to return.

Risk Score (scatter plot)

Y-axis: Risk Score (0.157 to 0.164)
X-axis: In Jail Length 📌 (0 to 350)

Race

Y-axis: Risk Score (0.159 to 0.164)
X-axis: BLACK OR AFRICAN.. | WHITE

City bubbles

OVERLAND PARK 43
OLATHE 56
SHAWNEE 19
GARDNER 14
KANSAS CITY 25
LENEXA

Bail Type

Bail Type	
Null	45
CA	6
GPS	3
ORCD	7
PR	35
SUR	104

POLICY RECOMMENDATION

- **Juvenile Program**
  - Bullying prevention, afterschool recreation, etc.

- **Community Support Project**
  - Conflict resolution and violence prevention
  - job oriented skill building

- Put limitations on individuals who tend to be bailed out through SUR type bond

- Pay more attention to those who have multiple charges in one case or spent less than one day in jail (supervision, mental health intervention)

SAMPLE EXPERIMENT DESIGN

Randomly divide the top 200 individuals with high risk of recidivism into two groups:

Treatment Group

- Phase 1 : Provide inmates with targeted case management and services during incarceration, to get them prepared for fitting back in with normal life (ex. job skill training)
- Phase 2 : Provide inmates with up to 12 months of supportive services in the community after they are released

Control Group (no customized intervention)

- Collect one-year recidivism rate from both groups, compare the result to see if the intervention works

LIMITATIONS

o Limited data sources: less than 30,000 data entries (if we can have access to mental health data or other personal information of the inmates, the model will preform better)

o The historical data collection in criminal justice may inevitable biased in the process.

o ACS data: ACS data (income, marital status, education level, health insurance coverage) only goes down to zip code level, which may not precisely reflect each individual's circumstances