# Prediciting Housing Code Violations

## EDSP - Final Presentation

Maxwell Austensen

2017-05-08

# Recap

# Topic Motivation

Housing Code Violations cause serious harm to tenants, and are proxy for other harmful conditions

Currently the City and non-profit organizations are complaint-driven

Desire for resources to facilitate more proactive action

# Project Goal

Use available data sources to identify buildings likely to have serious housing code violations

```
getwd()
```

```
## [1] "/Users/Maxwell/repos/edsp17proj-austensen"
```

# Data

# Data Sources

Currently using publicly available data sources:

- History of violations, complaints, and litigation (HPD)
- Physical characteristics of buildings (DOF & DCP)

# Data Processing

- Download raw data and documentation files
- Select and clean variables
- Restrict to privately-owned rental units under HPD jurisdiction
- Adjust apartment-level violations by number of units
- Add census tract-level violation aggregates
- Reshape to wide building-level data set

# Descriptives

# Housing Maintanice Code Violations

**Focusing on only class C "Immediately Hazardous" (*serious*) violations**

- Peeling lead paint in dwellings where a child under 7 resides
- Inadequate supply of heat and hot water
- Broken or defective plumbing fixtures
- Defective plaster
- Defective faucets
- Rodents

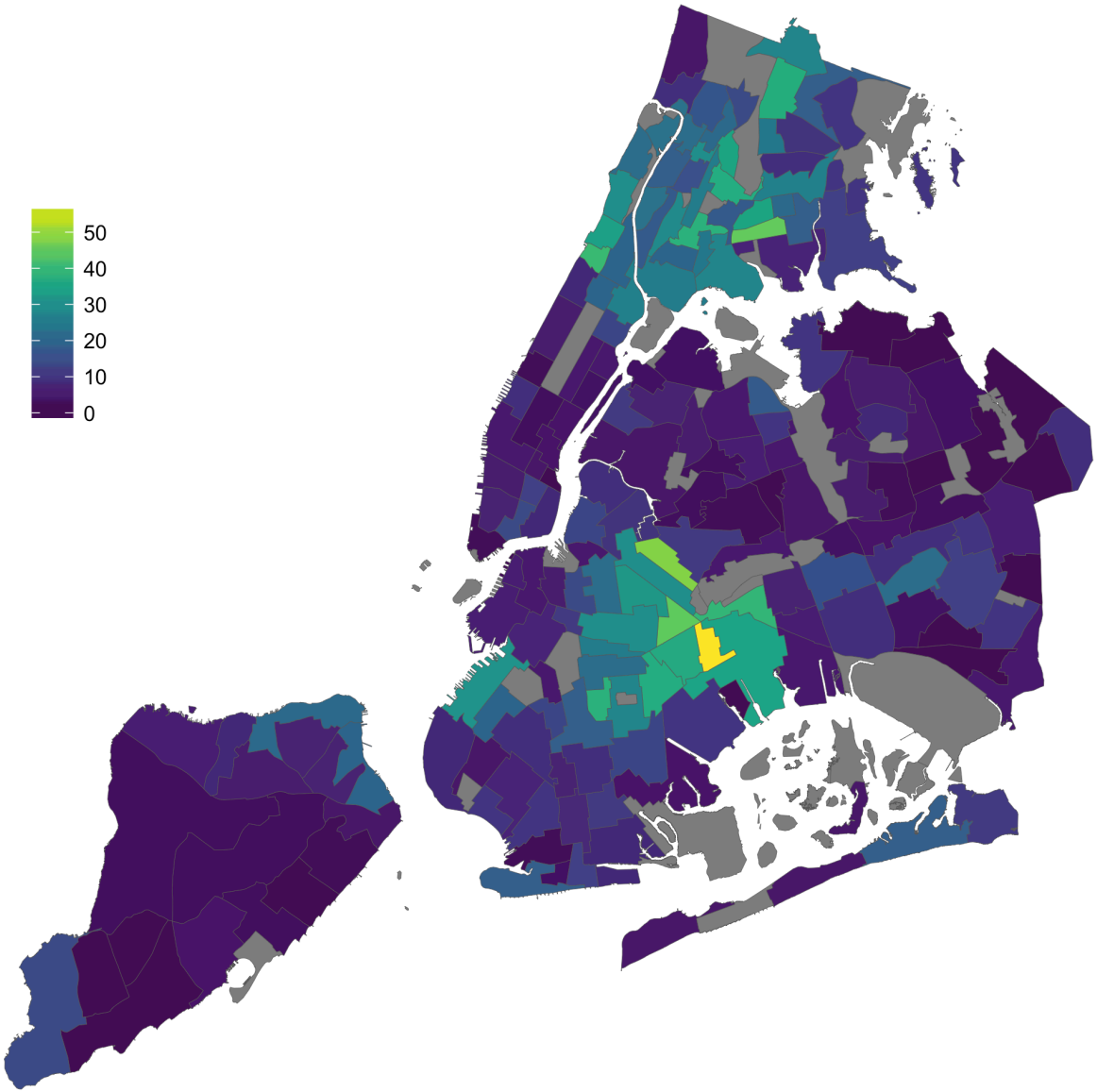
**Only 9.9% of buildings in sample had any serious violations in 2016.**

Among these properties:

- The average adjusted number of serious violations was 1.7.
- Only 47.2% also had a serious violation in the previous year.

Adjusted Number of Serious Housing Code Violations
per 1,000 Privately Owned Rental Units

Neighborhood Tabulation Areas, 2016




*Sources: NYC HPD, MapPLUTO, NYC DOF Final Tax Roll File*

# Models

# Modeling strategy

**Outcome:** Binary indicator of whether a building had any serious violations

**Training Data:** 2013-14 data to predict violations in 2015

**Test Data:** 2014-15 data to predict violations in 2016

**Classes are highly unbalanced:**

- Each year ~90% of buildings do not have any serious violations
- Improvements over no-information accuracy are constrained
- Model evaluation will emphasize precision and recall

**Past Violation**

Predict violation if building had violation in previous year

**Logistic Regression**

Selected model using step-wise algorithm with AIC, removing number of buildings and tract-level serious violations from 2 years prior

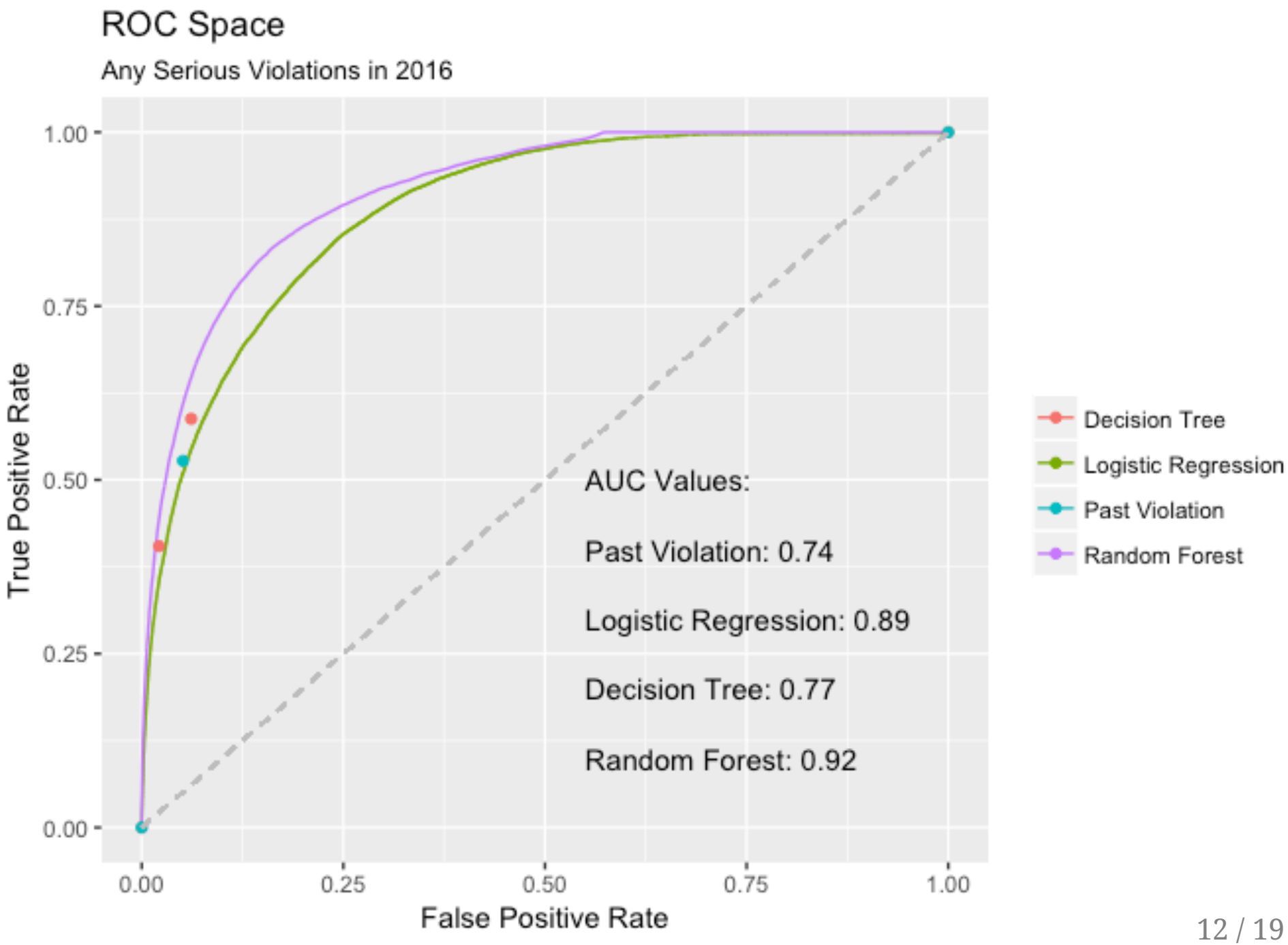
**Decision Tree**

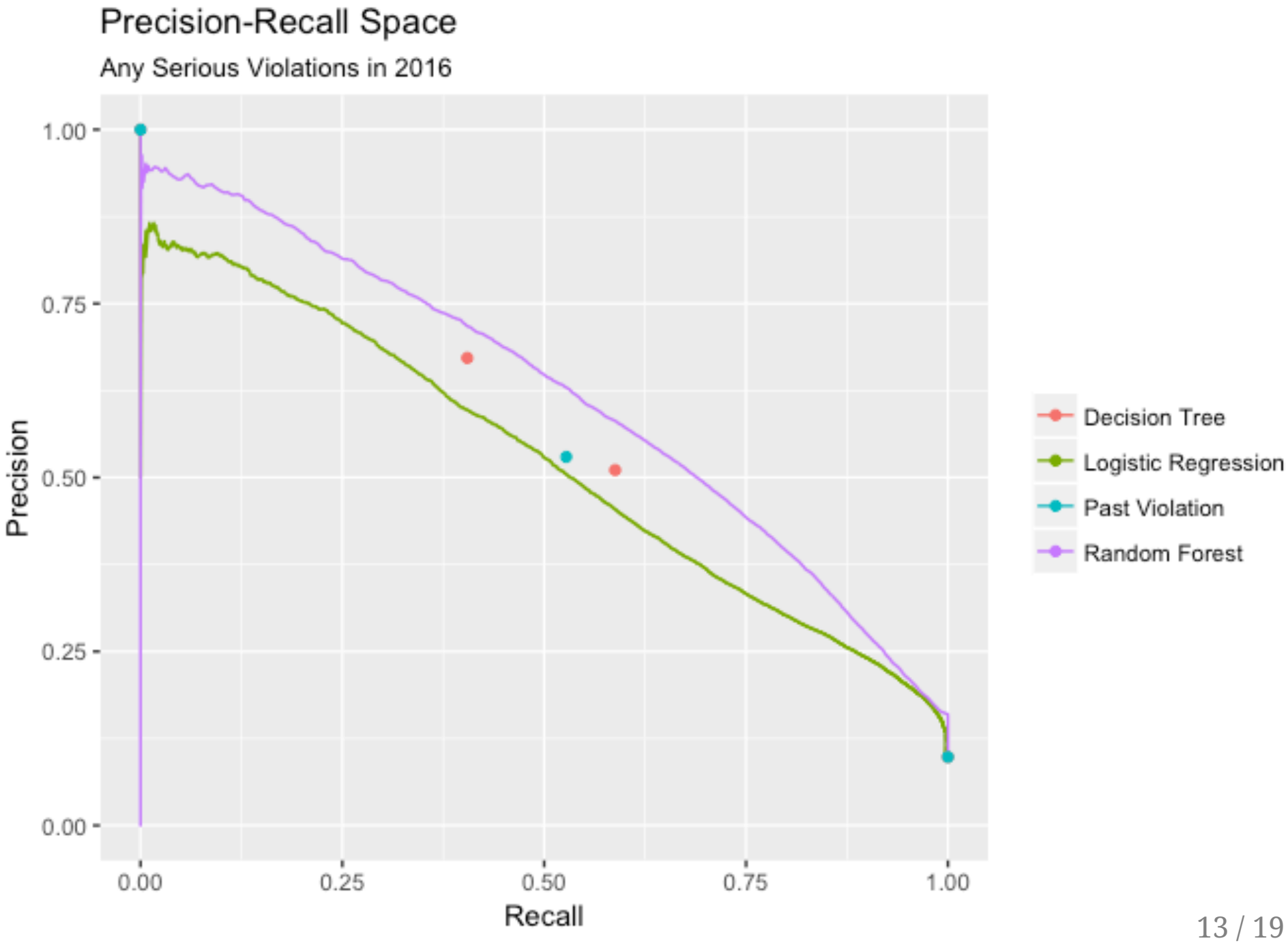
Not significantly higher accuracy compared to the logistic model

**Random Forest**

Significantly higher accuracy than all other models, and allows for specifying a threshold to balance the trade off between precision and recall

| Statistic | Past Violation | Logistic Regression | Decision Tree | Random Forest |
|---|---|---|---|---|
| Accuracy | 0.907 | 0.904 | 0.905 | 0.923 |
| Precision | 0.531 | 0.516 | 0.519 | 0.644 |
| Recall | 0.529 | 0.528 | 0.592 | 0.505 |


ROC Space
Any Serious Violations in 2016
True Positive Rate
False Positive Rate
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
AUC Values:
Past Violation: 0.74
Logistic Regression: 0.89
Decision Tree: 0.77
Random Forest: 0.92
Decision Tree
Logistic Regression
Past Violation
Random Forest


Precision-Recall Space
Any Serious Violations in 2016
1.00
0.75
0.50
0.25
0.00
Precision
0.00
0.25
0.50
0.75
1.00
Recall
Decision Tree
Logistic Regression
Past Violation
Random Forest

# Variable Importance

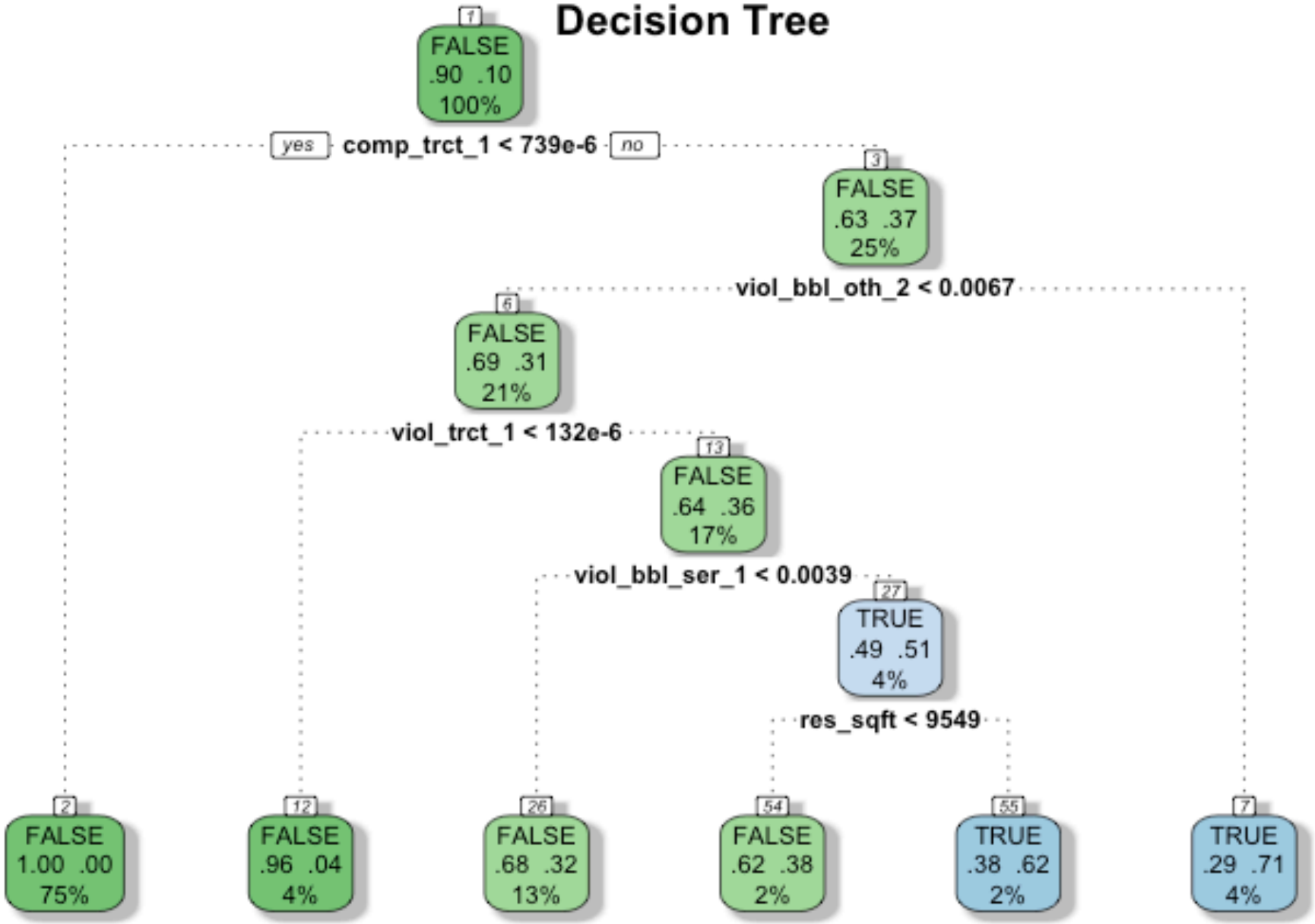
The following were associated with increased likelihood of violations:

**HPD data sources:**

- Complaints in previous year (both building- & tract-level)
- Violations in previous years (both serious & lesser categories, and building- & tract-level)
- Litigation against owner in previous year

**Building Characteristics:**

- Lower assessed value
- Older/Less recently renovated
- Larger buildings (# floors, # units, lot area)
- Smaller units
- Mixed-use buildings
- Full below-grade basement


Decision Tree
1
FALSE
.90 .10
100%
yes
comp_trct_1 < 739e-6
no
3
FALSE
.63 .37
25%
viol_bbl_oth_2 < 0.0067
6
FALSE
.69 .31
21%
viol_trct_1 < 132e-6
13
FALSE
.64 .36
17%
viol_bbl_ser_1 < 0.0039
27
TRUE
.49 .51
4%
res_sqft < 9549
2
FALSE
1.00 .00
75%
12
FALSE
.96 .04
4%
26
FALSE
.68 .32
13%
54
FALSE
.62 .38
2%
55
TRUE
.38 .62
2%
7
TRUE
.29 .71
4%

# App Prototype

# Next Steps

## Incorporate More Data Sources

- Housing Data Collective
- Neighborhood-level survey data

## Develop Prediction Models Further

- Tuning model parameters
- Try different options for training/test splits
- Try techniques to deal with class imbalance
- Try regression with adjusted violations count

## Continue App Development

- Test options optimizing performance
- Polish design elements
- Add tab with methods and model info

# Thanks!