

# Modeling Building Fire Risk Across the City

## ANALYZE BOSTON Open Data Challenge

SEAN P. MCKENNA

(4 May 2017)

---

## 1 Background and Objective

The overarching objective of the ANALYZE BOSTON Open Data Challenge is to “demonstrate the potential of open data to deliver new insights and help to make life better for everyone who lives and works in Boston.” The Challenge has been organized into five tracts, and this particular submission is aimed at Track 4, *Identifying Fire Risks*. Specifically, the hope for this track is to “develop ways for the Boston Fire Department, Inspectional Services Department, and other City agencies to identify locations at high risk for fires and other dangerous events.” To that end, the desire is to develop ways to “enable the City to better direct its preventive outreach efforts (including safety inspections, smoke and carbon monoxide detector installation, and fire safety education) to address these hazards before they turn into tragedies.”

The current submission is intended to be a proof-of-concept that looks to develop a machine-learned model based on recent historical data from Inspectional Services Department (ISD) violations, Assessing Department (AD) building/property data, and incident reports from the Boston Fire Department (BFD) to provide measures of fire risk for addresses across the City. One of the key takeaways from this project is that, while the objective seems relatively straightforward given the available data, the reality is much more complex. There are a lot of useful, readily available datasets, from a number of sources, and the key to this project is linking these datasets in the most logical, valid, and effective way.

## 2 Data Use

The idea behind this submission was to link BFD incident data with ISD violation data and AD property data to create a model that would provide a measure of fire risk for properties across the City. The details of the data sources used are as follows.

### 2.1 Boston Fire Department Data

BFD incident data from 2012 to 2016 and the first three months of 2017 was used. Within this data, a subset of incident types were considered as fire-related and used to build the outcome variable (see Table 1).

### 2.2 Boston Inspectional Services Department Data

ISD code violation data starting from 2000 was used in the hope that violations would provide a sense of how likely a property might be to experience a fire incident. From the ISD violation

Code	Incident Type
111	Building fire. Excludes confined fires (113–118).
112	Fire in structure, other than in a building. Included are fires on or in piers, quays, or pilings: tunnels or underground connecting structures; bridges, trestles, or overhead elevated structures; transformers, power or utility vaults or equipment; fences; and tents.
113	Cooking fire involving the contents of a cooking vessel without fire extension beyond the vessel.
114	Chimney or flue fire originating in and confined to a chimney or flue. Excludes fires that extend beyond the chimney (111 or 112).
115	Incinerator overload or malfunction, but flames cause no damage outside the incinerator.
116	Fuel burner/boiler, delayed ignition or malfunction, where flames cause no damage outside the fire box.
117	Commercial compactor fire, confined to contents of compactor. Excluded are home trash compactors.
118	Trash or rubbish fire in a structure, with no flame damage to structure or its contents.
100	Fire, other.

Table 1: Boston Fire Department Incident Codes Used in the Model [1]

Label	Explanation	Count
<b>rules</b>	Anything related to permitting, process, etc.	
<b>maint</b>	Violations related to lack of maintenance	9205
<b>trash</b>	Violations related to trash (storage, placement, etc.)	32292
<b>safety</b>	Anything related to safety, fire or otherwise	2869
<b>vandal</b>	Anything related to vandalism not addressed	514
<b>neg</b>	Anything that could be considered neglect	13092

Table 2: Labels used to Categorize ISD Violations

data, we used the **Code** (nature of the violation) and the **Value** (fee for the violation). Of all the codes, there were 583 that were unique. Since many of these violation codes are related or very similar, all codes were mapped (manually) to six broad categories as shown in Table 2.

### 2.3 Boston Assessing Department Data

Property assessment data from 2014 to 2017 was incorporated to augment the information being used to predict fire risk. The variables that were available in all years<sup>1</sup> of the assessment data believed to be relevant were: **PTYPE**, **LU**, **OWN\_OCC**, **AV\_BLDG**, **YR\_BUILT**, **STRUCTURE\_CLASS**, **R\_EXT\_FIN**, **S\_EXT\_FIN**, **GROSS\_AREA**. The **PTYPE** occupancy codes were reduced to a set of categorical factors as shown in Table 3. Similarly, the **LU** land use codes were reduced to a set of categorical factors as shown in Table 4.

<sup>1</sup>The format of the assessment data changed between 2015 and 2016.

PTYPE Code	Property Type Label
< 100	MultiUse
100 – 110	Residential
111 – 140	Apartment
300 – 399	Commercial
400 – 465	Industrial
900 – 929	ExemptOwn
937 – 999	Exempt

Table 3: Labels used to Categorize Assessment Data PTYPE Codes

LU Codes	Property Type Label
R1, R2, R3, R4, A	res
RL, CP, AH, CL	other
RC	mix
CM, CC	condoBldg
CD	condo
C, I	nonres
E, EA	exempt

Table 4: Labels created to Categorize Property Type Data LU Codes

### 3 Modeling Approach

Without going into detail, the steps of the modeling approach can be summarized as follows.

1. Load all the BFD datasets, combine them, and subset records to only include fire-related incidents (as defined for this analysis) (F).
2. Load ISD violation data (V).
3. For ISD records with ranges of street numbers, where the range is 10 or less, insert records to fill in these ranges; cases where the range is greater than 10, use just the first number.
4. Load all the AD datasets and merge them such that the most recent assessment data is kept (A).
5. Various data cleaning/manipulation operations.
6. For AD records with ranges of street numbers, follow same logic as that outlined in step 3.
7. Merge code violation data with building assessment data (using street address) ( $VA = V \cup A$ ).
8. Combine VA with the fire incident data for records where the fire incident post-dates the assessment/violation data ( $VAF = VA \cup F$ , VA earlier than F.)
9. Map the detailed violation codes in VAF to the six high-level violation categories (Table 2).

Feature	Explanation	Type
LU	High-level property type	Categorical (7 levels)
Value	Median ISD fee	Normalized float
OWN_OCC	Owner occupied	Categorical (2 levels)
violationCount	Total number of ISD violations	Normalized float
AV_BLDG	(Assessed building value)/(property area measure)	Normalized float
YR_BUILT	Year property was built	Normalized float
trash	Total “trash” violations	Integer
safety	Total “safety” violations	Integer
neg	Total “neglect” violations	Integer
maint	Total “maintenance” violations	Integer
vandal	Total “vandalism” violations	Integer

Table 5: Description of Features for a Given Address used in Model

10. Group all records in `VAF` by address (`gVAF = grouped(VAF)`).
11. Map the `PTYPES` in `gVAF` to the categories in Table 3.
12. Map the `LU` in `gVAF` to the categories in Table 4.
13. Handle some data outliers.
14. Make factors out of categorical variables and normalize numerical variables.
15. Split `Gvaf` into training and test sets (70:30 split).
16. Train a Random Forest model with 10-fold cross-validation, splitting parameter of 4, and 100 trees<sup>2</sup>.
17. Run the model from step 16 on the test set data.
18. Compute performance metrics.

The final set of features used in the Random Forest model are summarized in Table 5. It is important to note that a number of different models were explored (not deeply, of course), including a neural network, support vector machine, gradient boosting machine, and a simple logistic classifier.

## 4 Findings

Overall, the results were somewhat disappointing. Many variations were explored, in very ad hoc ways, and nothing seemed to result in a robust model, *i.e.*, a model that was able to identify properties that would experience a fire incident while minimizing false alarms. That said, the model does show some promise. Before discussing those findings, Table 6 shows the output of the Random Forest variable importance estimation.

Considering how the model predictions would be used operationally, consider these results:

---

<sup>2</sup>Number of trees did not seem to be a factor.

Variable	Importance
AV_BLDG	100.0
OWN_OCC	45.7
Value	44.2
YR_BUILT	34.1
LU	27.8
violationCount	27.1
neg	24.3
trash	20.8
maint	20.4
safety	10.3
rules	8.2
vandal	0.0

Table 6: Random Forest Mode Variable Importance Estimates

- Of the 10365 test samples, 957 had fires (9.2%).
- Of those 957, the model identified 232 as fire-prone (117 are actually false positives).
- If the BFD and/or ISD visited those 232 properties, 115 would hopefully have their situation corrected and a fire incident averted.
- Without the model predictions, if the City randomly visited that same number of properties (232), the number of properties that would have a fire incident averted would be 9.2% of 232 or about 21.
- In this sense, the model provides a 5-fold improvement over random inspections.

#### 4.1 Detailed Performance Output

```
fitted.results fire notfire
fire      115      117
notfire   842     9291

Accuracy : 0.9075
95% CI : (0.9017, 0.913)

Kappa : 0.1633
McNemar's Test P-Value : <2e-16

Sensitivity : 0.12017
Specificity : 0.98756
Pos Pred Value : 0.49569
Neg Pred Value : 0.91691
Prevalence : 0.09233
Detection Rate : 0.01110
Detection Prevalence : 0.02238
Balanced Accuracy : 0.55387
```

## 5 Areas for Further Analysis and Improvement

- Consider other risks such as fire in mobile property used as a fixed structure, outside rubbish or special outside fire, overpressure rupture/explosion/overheat (no fire), and hazardous condition (no fire).
- Improved handling of street numbers – ranges, etc.
- Deeper exploration of which property assessment data should be considered (and in what way) and how to bridge the data content change from 2015 to 2016.
- More thorough and standardized way to map the hundreds of IDS violation codes to a more meaningful and compact list.
- Overall code efficiency, structure, handling anomalies, and so on.
- Additional modeling analysis to assess other algorithms (*e.g.*, SVM, ANN, GBM) and subsequent tuning.
- Interactive map-based visualization tool.
- Output of relevant information to user(s) based on use cases.

## 6 Summary

This proof-of-concept has demonstrated the promise of a data-driven, model-based approach to identifying high-risk locations in the City.

## References

- [1] NFIRS. BOSTON FIRE DEPARTMENT INCIDENT CODES (NFIRS – National Reporting Codes), <https://www.facebook.com/notes/boston-fire-fighters-local-718-iaff/boston-fire-department-incident-codes-nfirs-national-reporting-codes/10150267180342961/>, 2009. [Online; accessed 25-April-2017].