Modeling Building Fire Risk Across the City

ANALYZE BOSTON Open Data Challenge

Sean P. McKenna

(9 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. This totalled 225k records. Within this data, a subset of incident types were considered as fire-related (42k records) and used to build the outcome variable (see Table 1).

2.2 Boston Inspectional Services Department Data

ISD code violation data starting from 2000 (370k records) was used in the hope that violations would provide a sense of how likely a property might be to experience a fire incident. From

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]

the ISD violation 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 (183k records) was incorporated to augment the information being used to predict fire risk. The variables that were available in all years[1] 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

[1]The format of the assessment data changed between 2015 and 2016.

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

Table 2: Labels used to Categorize ISD Violations

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

were reduced to a set of categorical factors as shown in Table 4.

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, impute the street numbers contained in the range, counting by 2.
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, impute the street numbers contained in the range, counting by 2.
7. Merge code violation data with building assessment data (using street address) (VA = V $\bowtie\!\!\!\sqsubset$ 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.)

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

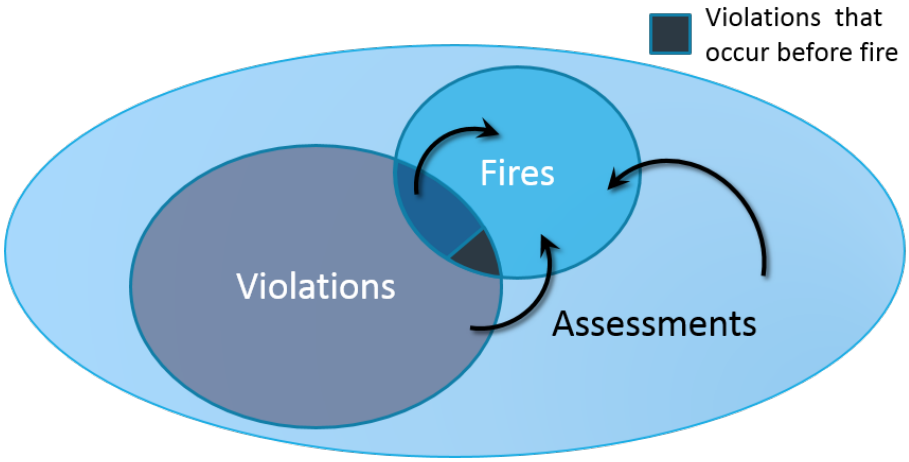

Figure 1: Visual summary of open data used. Assessment data served as the complete data space (in terms of unique property addresses.) Within **Assessments**, some properties had **Violations** and some had **Fires**. The intersection of **Violations** and **Fires** had both. Of those, cases where the fire preceded the violation were removed. Removing cases where the violation occurred after the fire incident is consistent within an operational context; whether those cases can be used in model building was not deeply explored.

9. Map the detailed violation codes in `VAF` to the six high-level violation categories (Table 2).
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 3[2], and 200 trees[3].
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.

[2]Number of variables randomly sampled as candidates at each split.
[3]Number of trees did not seem to be a factor.

Feature	Explanation	Type
LU	High-level property type	Categorical (7 levels)
Value	Median ISD fee	Normalized float
OWN_OCC	Owner occupied	Categorical (2 levels)
City	Neighborhood	Categorical (17 levels)
violationCount	Total number of ISD violations	Normalized float
AV_BLDG	(Assessed building value)/(property area measure)	Normalized float
area	Building area (combination of GROSS and LIVING	Normalized float
Kitchen	Number of kitchens at property	Normalized float
YR_BUILT	Year property was built	Normalized float
rules	Total "rules" 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

4 Findings

Overall, the results were fair. Many variations were explored, in very ad hoc ways. That said, the model does show some promise. Before discussing those findings, Table 6 shows the output of the Random Forest variable importance estimation.

Variable	Importance
area	24.95
violationCount	22.47
YR_BUILT	16.49
AV_BLDG	15.47
Kitchen	14.85
City	13.89
OWN_OCC	11.63
rules	4.37
Value	3.87
safety	3.21
trash	3.18
neg	2.82
maint	2.22
LU	2.20
vandal	1.03

Table 6: Random Forest Mode Variable Importance Estimates

In terms of performance, looking at accuracy alone is misleading. Instead, we consider the confusion matrix, the F1-score, and the area under the ROC curve (AUC). Here are the detailed results.

```
fitted.results  fire notfire
       fire      281     162
       notfire  2365   25729

               Accuracy : 0.9114
           Sensitivity : 0.106198
           Specificity : 0.993743
         Pos Pred Value : 0.634312
         Neg Pred Value : 0.915818
             Prevalence : 0.092722
         Detection Rate : 0.009847
   Detection Prevalence : 0.015524
      Balanced Accuracy : 0.549971
```

These results show that of the 2646 fire incidents, the model identifies 281, with 162 false positives. The F1-score is the last line, 0.55. The AUC was computed to be 0.74.

Considering how the model predictions might be used operationally, one approach is to look at adjusting the threshold for classification such that the false alarm rate is roughly 50%. In that case, the results become:

```
           fire notfire
  fire      552     603
  notfire  2094   25288

               Accuracy : 0.9055
           Sensitivity : 0.20862
           Specificity : 0.97671
         Pos Pred Value : 0.47792
         Neg Pred Value : 0.92353
             Prevalence : 0.09272
         Detection Rate : 0.01934
   Detection Prevalence : 0.04047
      Balanced Accuracy : 0.59266
```

Given these results, consider the following scenario:

→ Of the 28,537 test samples, 2646 had fires (9.3%).
→ Of those 2646, the model identified (546+536)=1082 as fire-prone (536 of which are the false positives).
→ If the BFD and/or ISD visited those 1082 properties, 546 would hopefully have their situation corrected, and a fire incident averted.
→ Without the model predictions, and if the City randomly visited that same number of properties, the number of properties that would have a fire incident averted would be 9.3% of 1082, or about 100.
→ In this sense, the model provides a 5-fold improvement over random inspections.

5 Areas for Further Analysis and Improvement

- Consider other data sets to capture additional features
- Better understanding of fire incidents in Boston (*e.g.*, talk with BFD)
- 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

As a proof-of-concept, done in limited time, this analysis is not

- Rigorous
- Complete
- Informed by domain experts
- Optimal

However, this analysis is

- A good exploration of Boston's Open Data
- A demonstration of the potential of such data
- A first step toward developing ways in Boston can use this data to better direct preventive efforts and increase safety across 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].