Buildings blight in Detroit

Roman Kislenok

May 1st, 2016

Abstract

In this study we try to fit a simple model that assign to a buildings a probability of being dismantled in a near future by using blight violations, crimes, and demolition permits events for this building and buildings nearby.

The repeatability

All code for this project including this particular report are available at Git Hub [1].

On definition of the buildings

Our data-sets generously provided by Detroit Open Data [2] including blight violations, crimes, and demolition permits have addresses and coordinates. But it's easy to check that coordinates can point to a road nearby particular address associated with event or sometimes to some distant places in the Detroit (or even outside). We could come up with some interesting approaches (like geohashing and clustering) to unite events into buildings, but really we need not to. We'll use points on the map as a proxy for buildings. We want to think about "the building area" and include all events that lie inside specific radius R from point on the Detroit map in not so distance past (t).

Both radius and time-span are to be estimated.

Spliting data

We think that demolition permits points to the real buildings as city should pay for demolition and, we think, double check this data. We came out with 5687 Dismantled buildings.

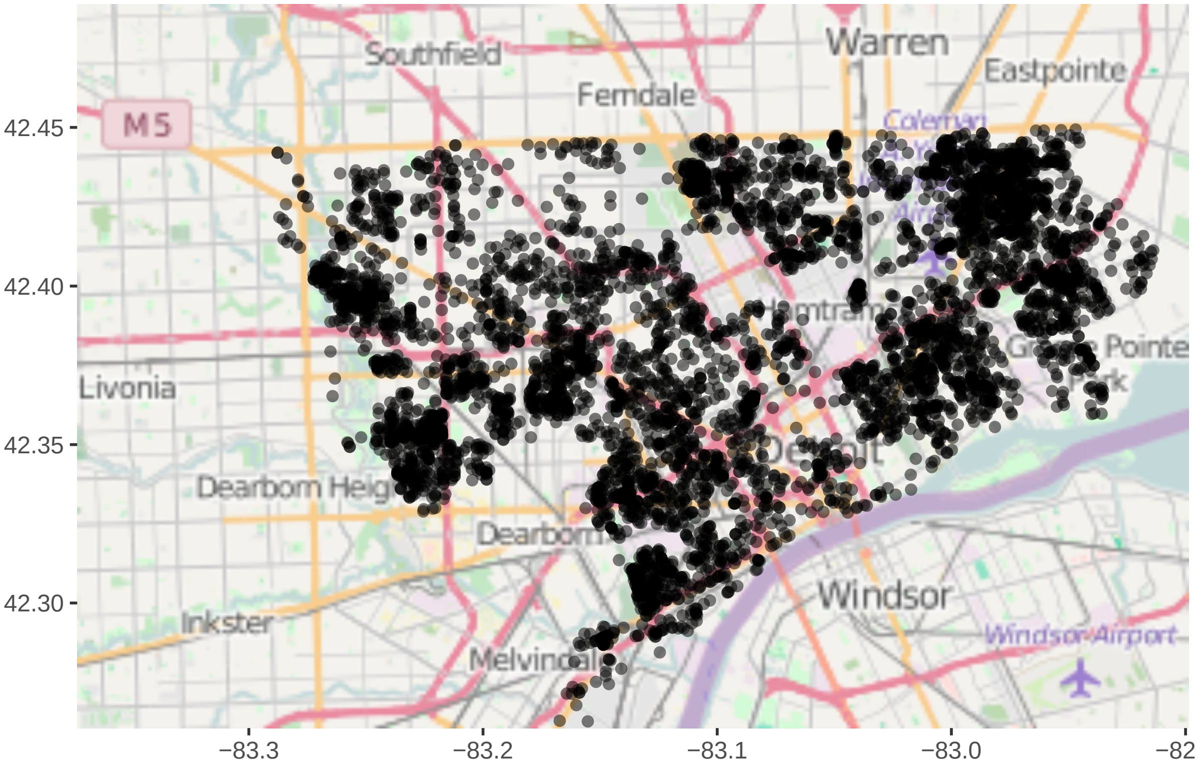
All other points are considered to be potential not demolished buildings. The proportion of demolished buildings to normal estimated to be 1 in 20, this estimation has to be questioned in future research!

We prepared our data-sets (details provided in Table 1):

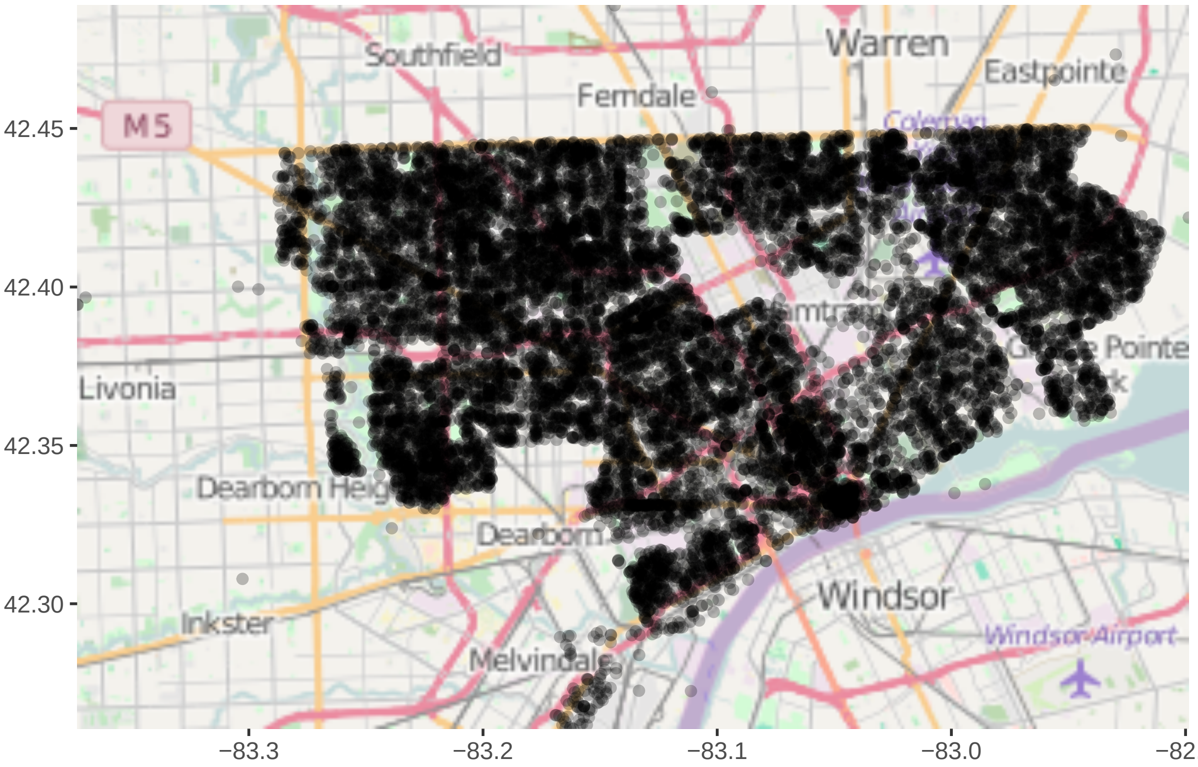
- 90% of the buildings from demolition permits goes to train data-set, add normal buildings to have 50%-50% proportion;
- 5% of the buildings from demolition permits goes to validation data-set, add normal buildings to represent "real" situation of 1 to 20;
- 5% of the buildings from demolition permits goes to test data-set, add normal buildings to represent "real" situation of 1 to 20;

Table 1: Train, test, and validation datasets

Label	Train	Validation	Test
Dismantle	5117 (50%)	285 (4.76%)	285 (4.76%)
Normal	5117 (50%)	5700 (95.24%)	5700 (95.24%)



Detroit demolishion permits − our buildings labeled ‘Dismantle‘



Our buildings labeled ‘Normal‘

Validation data-set we’ll use to select features. Both train and validation data-sets we’ll use to train our final model.

Feature building and selection

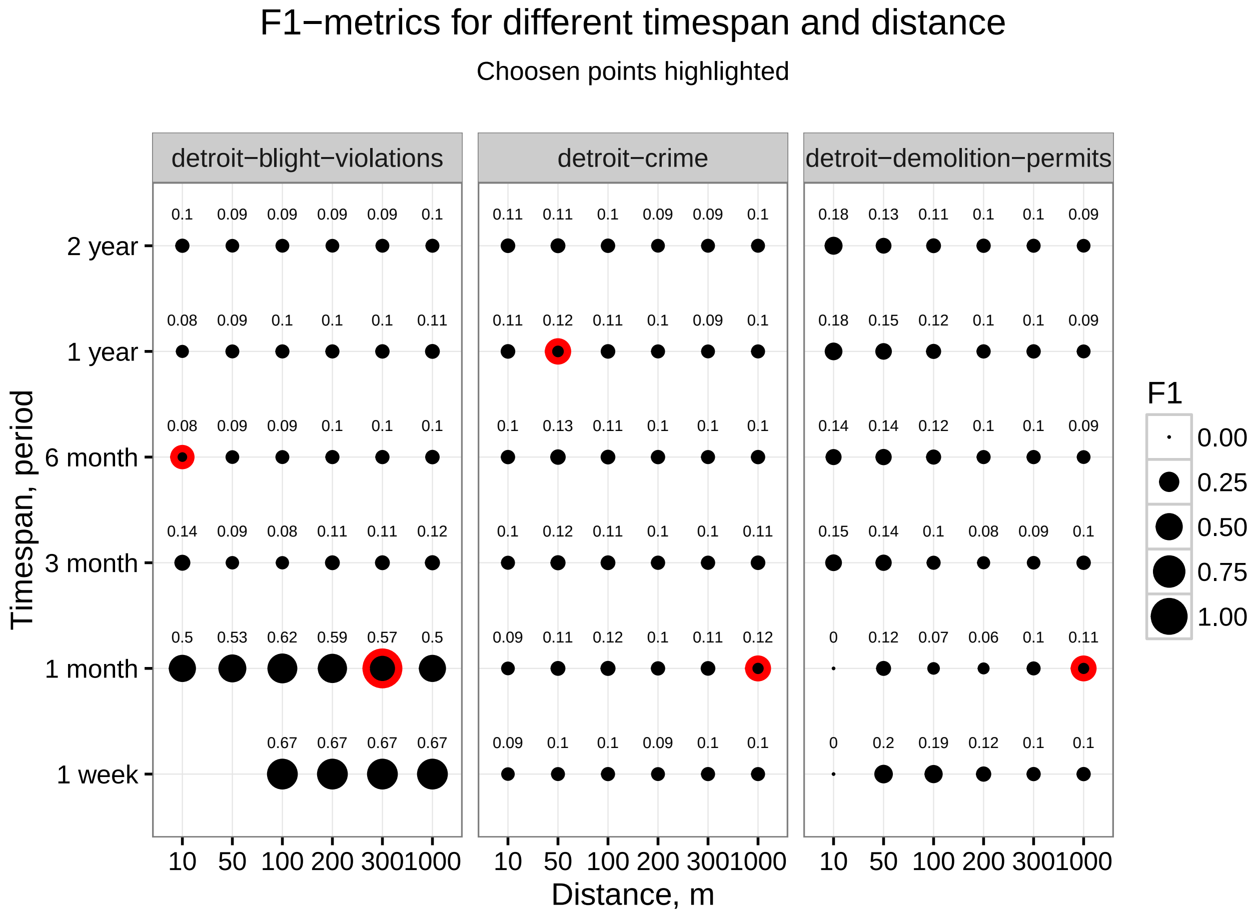
Each "Dismantle" labeled buildings has the associated *state* date (of particular demolition permit) and for all "Normal" labeled buildings we set *state* date to the maximum date of available demolition permits. It's done so, because we know that buildings are still "Normal" up to the last entry in demolition permits data-set. Each feature would be the amount of different events in a particular distance from the building in not so distant past from *state* date. To select distance and timespan we train a simple logistic regression with one feature using train data-set and evaluate results using validation data-set.

We use events from:

- "detroit-blight-violations" data-set provided on Coursera web site, similar data-set can be obtained from Detroit Open Data;
- "detroit-crime" data-set obtained from Detroit Open Data, originally provided file has events only since 2015;
- "detroit-demolishion-permits" data-set provided on Coursera web site, similar data-set can be obtained from Detroit Open Data.

We intentionally skip "detroit-311" data-set as it only contains data since the middle of 2014.

The results are presented on a pictures below. Combinations of distance and timespan chosen to "cover" some greater area of parameters with similar values.



F1−metrics for different timespan and distance

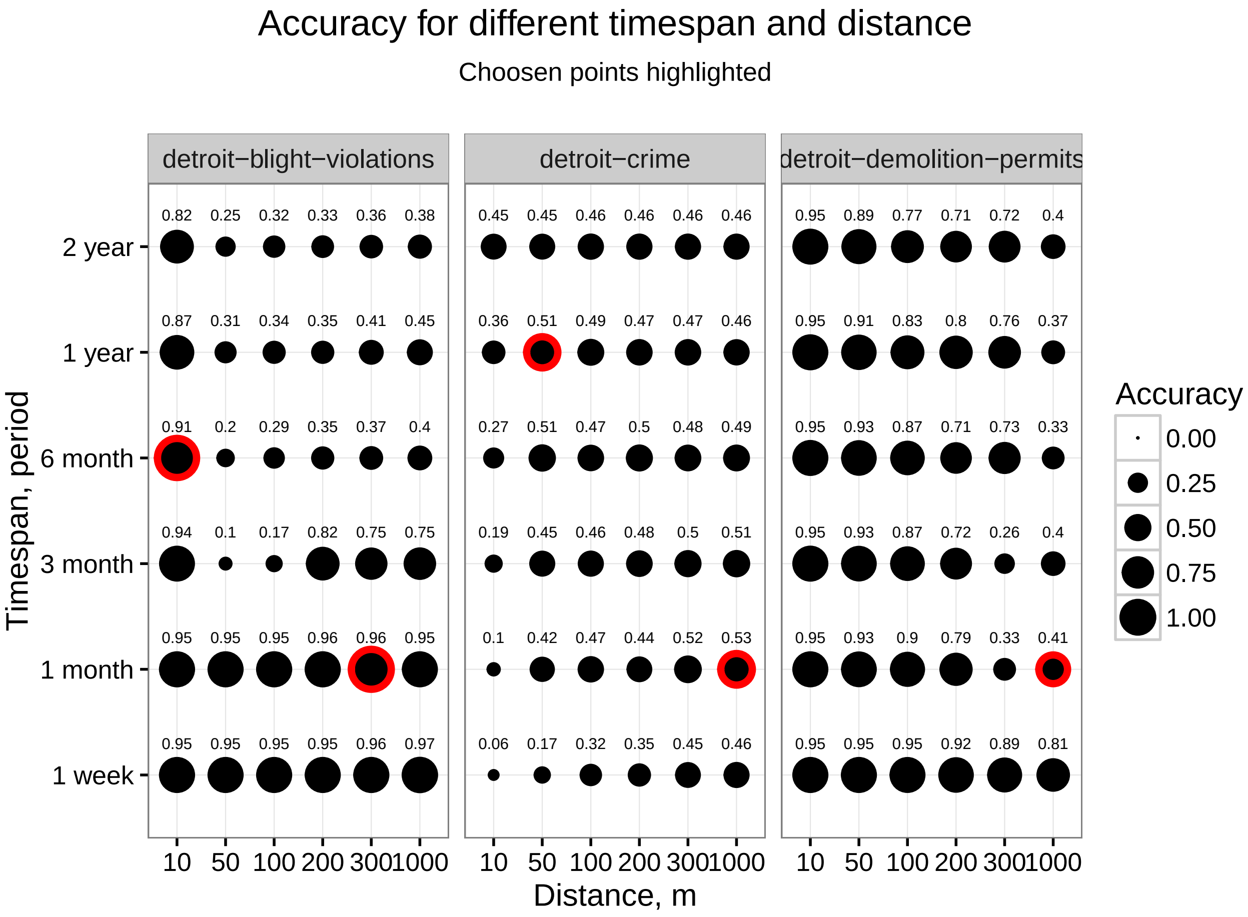
Choosen points highlighted

To choose parameters we use accuracy, F1 metrics and common logic (not documented here). Here we should note that accuracy is expected to be high - almost all buildings are Normal in validation data-set, so accuracy of 0.96 should not confuse you. F1 is more balanced metrics and much more appropriate here.

Our chosen predictors listed in the Table 2.

Table 2: Choosen predictors

Dataset	Distance, m	Timespan
detroit-blight-violations	300	1 month
detroit-blight-violations	10	6 month
detroit-crime	1000	1 month
detroit-crime	50	1 year
detroit-demolition-permits	1000	1 month



Model

We fit result model to train data-set using binomial model due to its great interpretability. First fit (in Table 3) using all selected features has meaningless coefficient for `detroit-blight-violations_10_6 month`, to be excluded from the final model (Table 4).

Our model predicts the odds of a building being Dismantled increase with amount of blight violations in 300m in last month. But it's quite strange that crimes amount in 1km last month, crimes amount in 50m in last year, and amount of demolition permits in 1km last month all has negative impact on odds for building being Dismantled.

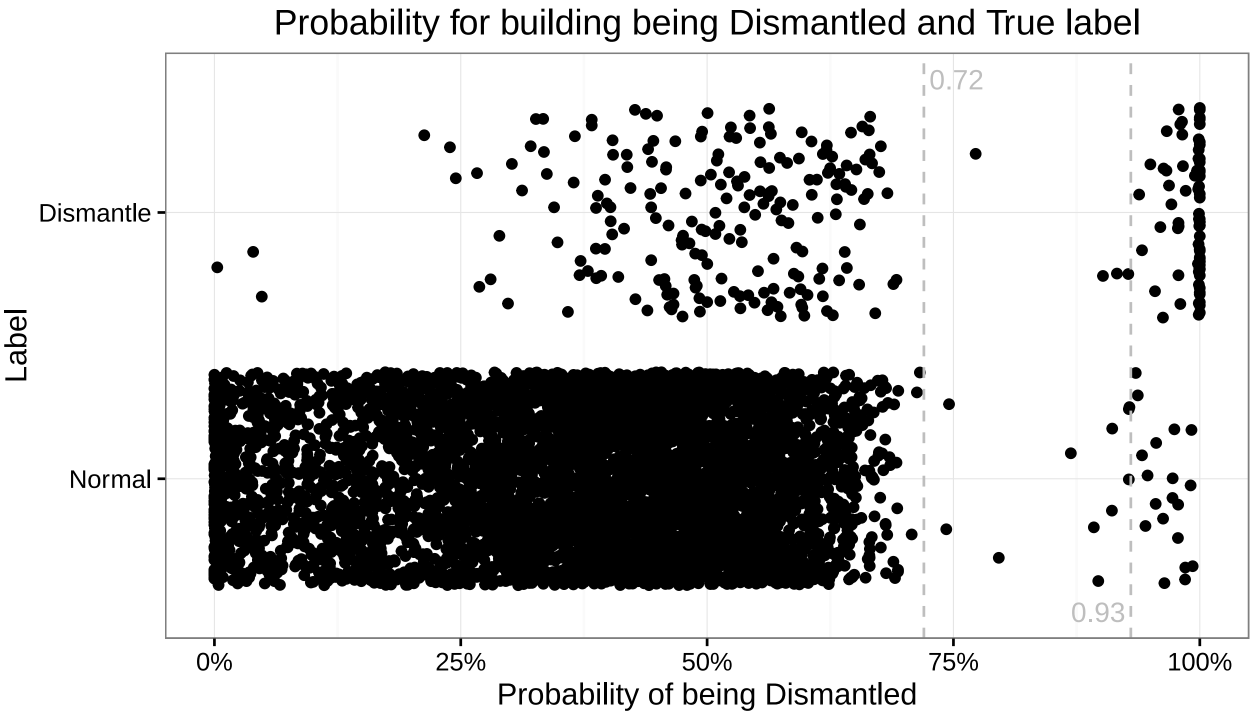
Table 3: The model with all selected features

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9233449	0.1189328	7.763587	0.0000000
`detroit-blight-violations_300_1 month`	3.4433822	0.3467193	9.931325	0.0000000
`detroit-blight-violations_10_6 month`	-0.1145566	0.0629033	-1.821153	0.0686127
`detroit-crime_1000_1 month`	-0.0036892	0.0010788	-3.419890	0.0006289
`detroit-crime_50_1 year`	-0.1487948	0.0125172	-11.887193	0.0000000
`detroit-demolition-permits_1000_1 month`	-0.0363033	0.0061887	-5.866035	0.0000000

Table 4: The model including features with meaningful coeficients

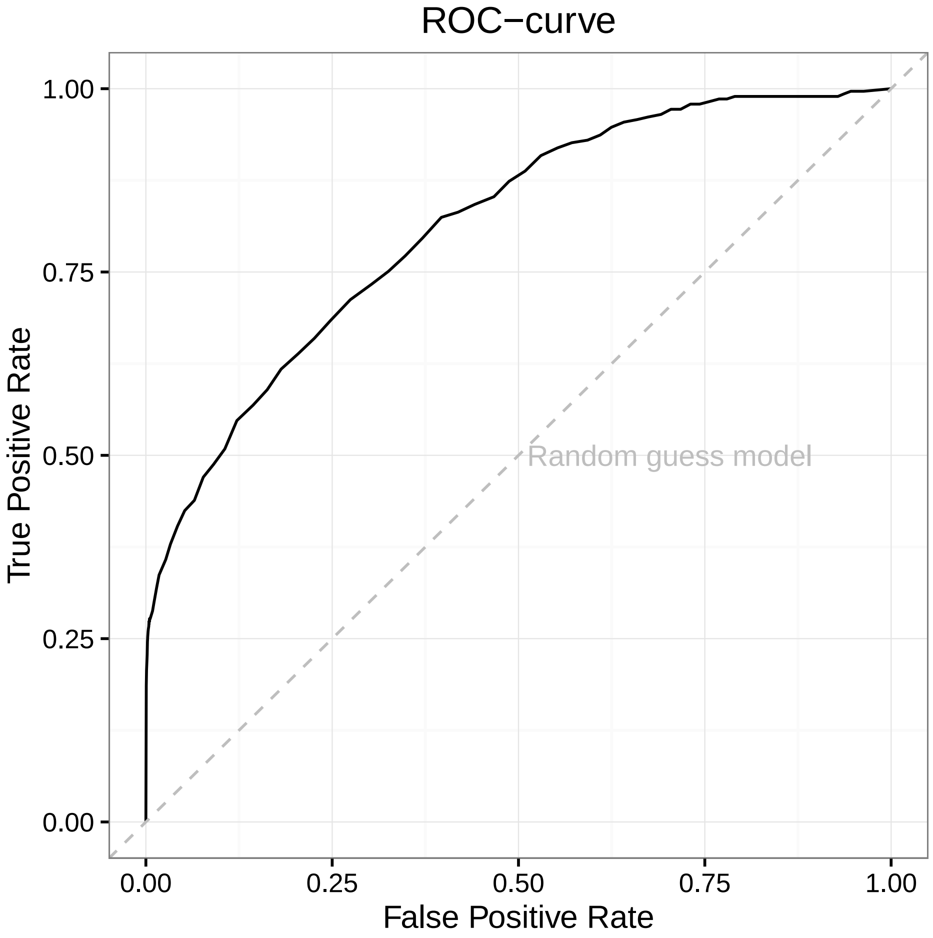
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9095647	0.1172567	7.757035	0.0000000
`detroit-blight-violations_300_1 month`	3.4122923	0.3397886	10.042398	0.0000000
`detroit-crime_1000_1 month`	-0.0036751	0.0010661	-3.447147	0.0005688
`detroit-crime_50_1 year`	-0.1482185	0.0123557	-11.995919	0.0000000
`detroit-demolition-permits_1000_1 month`	-0.0363330	0.0061217	-5.935119	0.0000000

Model threshold. Evaluating on a validation data-set using different thresholds for treating building as Dismantled we produce a nice convex ROC-curve. Using it with plot of model predicted probability to the true labels gives us optimal threshold for classification to optimize accuracy (conservative model).



The optimal threshold cut lie within interval of 0.72 - 0.93 and gives us accuracy of 0.961. It should be mentioned here that there are a lot of Dismantled buildings with lower probability, so given another optimization goal we can catch them by this model, of course with a lot of Normal buildings being misclassified. Also our result suggests that we should include more features in a future models.

As a result we want to stick with lower value for threshold - 0.72.



Results The final model is build using train and validation data-sets, labels assigned using threshold of 0.72. Coefficients are almost not changed from previous fit (compare Tables 4 and 5). Contingency table (Table 6) suggests that while having a great accuracy of 0.964 we have a great misclassification rate (also look at Table 7 for other metrics).

Table 5: The final model coeficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9671079	0.1107623	8.731380	0.00e+00
`detroit-blight-violations_300_1 month`	3.3360821	0.3092147	10.788887	0.00e+00
`detroit-crime_1000_1 month`	-0.0042314	0.0010053	-4.209120	2.58e-05
`detroit-crime_50_1 year`	-0.1461110	0.0116230	-12.570870	0.00e+00
`detroit-demolition-permits_1000_1 month`	-0.0380020	0.0058480	-6.498247	0.00e+00

Table 6: Contingency table for the final model

	Normal	Dismantle
Normal	5672	28
Dismantle	189	96

Table 7: Performance metrics for the final model

Metrics	Value
Accuracy	0.9637427
False negative rate (error rate)	0.0362573
True positive rate (recall)	0.3368421
Precision	0.7741935
F1	0.4694377

Other potential features to consider

Proposed process allow us to include other data sets from Detroit Open Data or any other sources with or without spatial information. Some data-sources to be considered includes:

- there are more information in data used:
  - payment status for blight violations - we should check whether not paying bills increase probability of building being blighted in future,
  - type of blight violation, different crime categories may have different implications on status;
- blight may be the result of the local economic circumstances, so it'll be great to use some data regard local business health and vacant jobs;
- it would be interested to investigate blight in regard to Schools and other education centers, Recreations, Parks and other public places;
- some social information about citizens.

References and links

- [1] https://github.com/kislenok-roman/datasci_course_materials/tree/master/capstone/blight
- [2] Detroit Open Data (https://data.detroitmi.gov)
- [3] D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf