Predicting Detroit blight

Abstracts

Identifying buildings at risk of being abandoned is becoming a point of interest for US cities such as Detroit, in order to be able to take action to prevent these events to happen. The economical impact of this politic can be important: for example Detroit is spending $1bn in order to finance residential blight removal.

In this study, I created a model to predict if a building in Detroit will be blighted using data provided by Data Driven Detroit such as criminal incidents, permit violations, call to 311 and blight information. I also used additional demographic information such as median population income, median house value and racial information. Using machine learning, I was able to build a model that predict building blighting with **an accuracy of 0.89** and an **area under ROC of 0.89.**

Interestingly, this study shows the importance of spatial environment in building abandonment as neighborhood parameters were found to be more important in the predictive model than information related to the building itself.

Data are available at : https://github.com/David-Desmaisons/datascience_detroit_bligth

Methods

1. Tools

For this study, I used the python stack: scikit-learn for machine learning, pandas for file processing, fuzzywuzzy for string processing, geopy and geohash for processing geographic data, beautifulsoup for web scraping, bokeh and matplotlib for data visualization, Jupyter notebook to explore results. SQLite database was used to process and transform data. When needed, I used google maps and google fusion table to visualize data. Finally, I used google developers api for geolocation.

2. Building definition

One challenge of this study was to be able to define a building in order to relate different event to predefined buildings.

2.1. Removing incoherent data

Carefully checking the data of the files *detroit-blight-violations*, *detroit-311*, *detroit-crime*, *detroit-demolition-permits*, I was able to identified many inconsistent data including: missing or incoherent longitude and latitude, coordinates out of Detroit City, same location referenced up to thousand times with different addresses. These data were filtered from original data set in order to not compromise the final result.

2.2. Considering location

The second step was to consider the urbans topology of Detroit. After examining maps of Detroit, I considered that events occurring in a 10 m range may be assigned to the same local. For each location, the corresponding geohash was computed, which was used to filtered closeby locations. Real distance was then computed and location within 10 m distance was assigned the same location Id.

2.3. Considering street name and number

Finally in order to identify building and to better verify data integrity, I labelled the building by street names and street numbers. Street names were extracted using Levenshtein distance to a list of Detroit streets. Data having same localization (location id) and different street name were analyzed and clean-up (as it appears that). Street number were also corrected when needed comparing street numbers (as it appears that many building labelled with a 0 as street number).

That done, building could be considered as an unique combination of street name and street number: using this methodology, I found 3992 demolished buildings from the *detroit-demolition-permits* file.

3. Data set buildings

The buildings in the training set were selected from buildings identified in the *detroit-blight-violations* files, with a proportion of 50% non blight and 50% blight. As the total number of blight buildings referenced in this file was 2053, the total number of training set rows is 4106. The rationale behind this decision is 1) that many bligh building have no blight violation or any other events associated and this leads to a bias in training the machine (i.e building with no label will be considered as blight which obviously is not correct), 2) it makes sense to use this model for further blight prediction using data coming from *detroit-blight-violations* as an input.

4. Comparing models

Before model construction parameters were normalized using l2 norm. Models were trained using a 5 folds cross validation methodology to avoid overfitting. The best model was defined by the highest value of area under the ROC curve (AUC). For each dataset, I benchmarked the following models: decision tree, random forest, ada boost classifier, gradient boosting classifier, Logistic regression, and SGD classifier**.** All computation was done using ski-learn Python machine learning library.

Results

Features selection

For each building a list of a relevant features was created and model accuracy was computed using best ranking machine learning model . This process was repeated iteratively as feature selection was improved. For all the datasets, gradient boost model was found to have the best AUC scoring.

1. Naive model

A first and naive approach was to consider as features the number of events that occurred in a given building counting for each building the number of blight violation, the number of call to 311 and the number of crime. This approach results as expected in a poor model with an accuracy 0.51 and an AUC of 0.51.

2. Adding environmental information

Heat map of blight incident (see Fig 1.) revealed that blight has a tendency to cluster as described in other studies (Morckel 2014). In order to improve the model, I decided to included to the features environmental information coming from building neighborhood that might also cluster. To do so I used geohashes, that allowed to slice building environment into regions of differents sizes (see Fig. 2). For this, I considered events occurring in the same geohash ranges using the precisions of 6, 7 or 8 digits (corresponding approximately and respectively to size of : 1.22km×0.61km, 153m×153m, 38.2m×19.1m, see Fig 2)

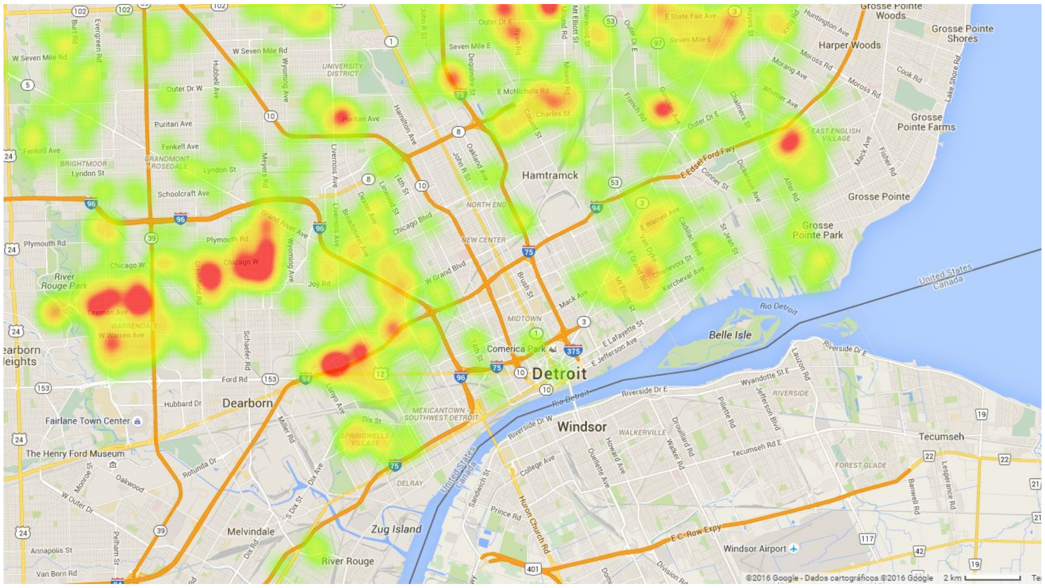


Fig 1. Buildings blighted heatmap showing clustering (google fusion table)

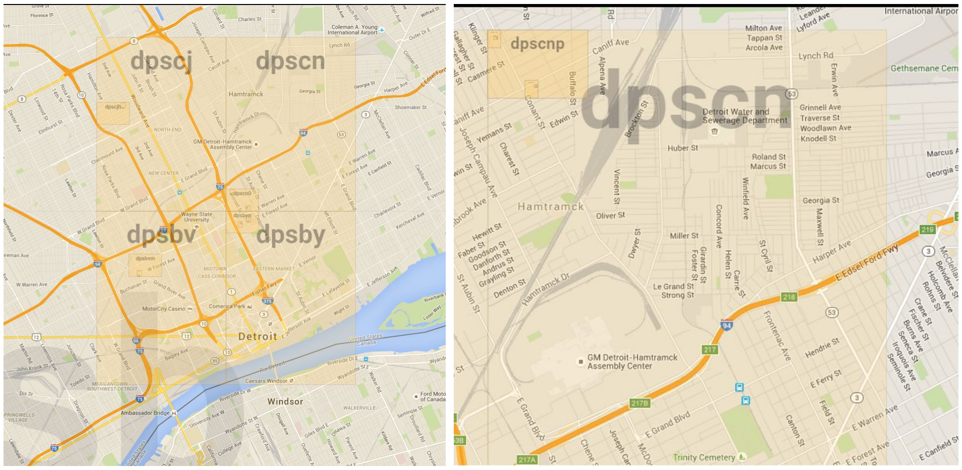

dpscj
dpscn
dpsbv
dpsby
Hamtramck
Detroit
Windsor
dpscnp
dpscn
Detroit Water and Sewerage Department
GM Detroit-Hamtramck Assembly Center
Hamtramck

Fig 2. Illustration of geohashes, city of Detroit (from [http://geohash.gofreerange.com/)](http://geohash.gofreerange.com/)

This second model includes the 11 following features:

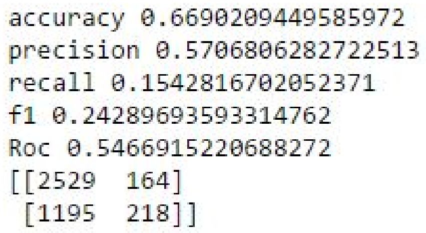
-number of blight violation, number of 311 call that occurred in the building

-number of crime, number of blight violation, number of 311 call, that occurred in the immediate surrounding of the house (geohash 8 digits)

-number of crime, number of blight violation, number of 311 call, that occurred in an intermediate surrounding of the house (geohash 7 digits)

-number of crime, number of blight violation, number of 311 call, that occurred in a larger surrounding of the house (geohash 6 digits)

This model result in an improved accuracy of 0.66, and an AUC of 0.55 as shown on Fig 3.Illustrating the corresponding confusion matrix.



```
accuracy 0.6690209449585972
precision 0.5706806282722513
recall 0.1542816702052371
f1 0.24289693593314762
Roc 0.5466915220688272
[[2529  164]
 [1195  218]]
```

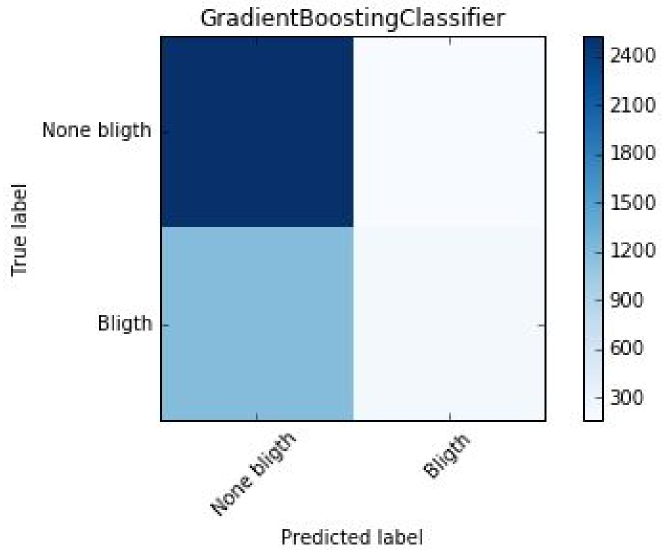


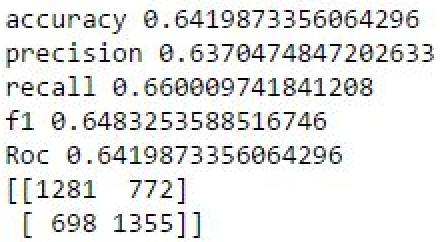
Fig 3. Model 2 confusion matrix

3. Adding demographical data

As parameters such as demographic data (for example population age, unemployment), building market conditions are reported to influence building abonment (Morckel 2014), I decided to include such data to the model. To do so, first I used google geolocation API to retrieve zip code information from the buildings in the data set. I then extracted demographic and house market data by zip code from www.city-data.com/ using web scrapping.

With this additional methodology, were added to the model: average number of persons per household, median house value, income per household, percentage of family per household, percentage of the population below poverty, percentage of renters, unemployment percentage, percentage of white population, median age.

This data improved the model to an accuracy of 0.64 and an ROC AUC of 0.64 which is a good improvement from model 2 as shown in Fig 4



```
accuracy 0.6419873356064296
precision 0.6370474847202633
recall 0.660009741841208
f1 0.6483253588516746
Roc 0.6419873356064296
[[1281  772]
 [ 698 1355]]
```

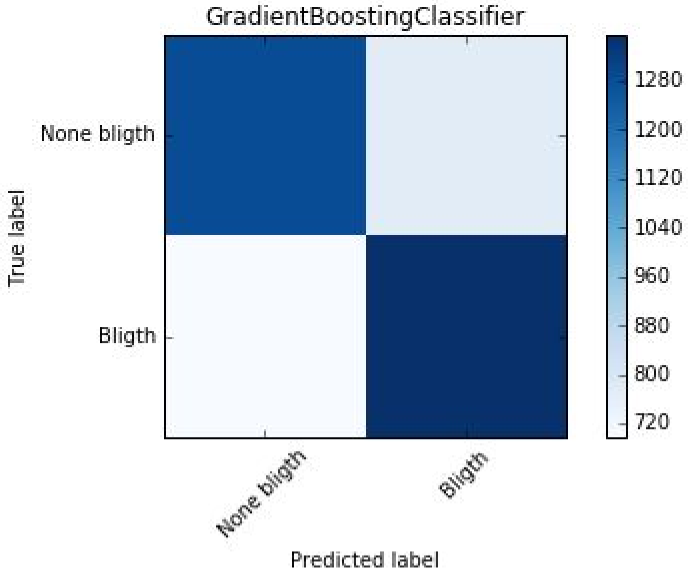


Fig 4. Model 3 confusion matrix

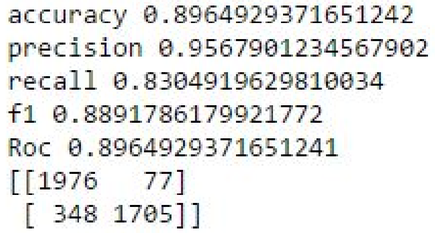
4. Further improvements

Finally, additional features were added to model 3. To do so, I used the following approaches: 1) taking into account the time dimension of events considering events that occurred in a given time window before blighting, 2) taking into account description of the different events (for example blight fine can be paid or unpaid, blight violations have different categories), 3) considering larger region as environment using 5 digits precision geohash corresponding to region of approximately 4.9km x 4.9km.

Features in final model includes:

- Number of blighted buildings in the same region (geohash 5) a year before bligthing or before 2015-08-28 (date of the last blighting)
- Number of blighted buildings in the same region (geohash 6) a year before bligthing or before 2015-08-28 (date of the last blighting)
- Number of crime, blight violation and 311 calls that occurred in the same region (geohash 5)
- Number of crime, blight violation and 311 calls that occurred in the same region (geohash 6)
- Number of crime, blight violation and 311 calls that occurred in the same region (geohash 7)
- Number of crime, blight violation and 311 calls that occurred in the same region (geohash 8)
- Number of unpaid or paid blight, violation 311 calls that occurred in the building
- Average number of persons per household
- Median house value
- Average income per household
- Percentage of family per household
- Percentage of the population below poverty
- Percentage of renters
- Unemployment percentage
- Percentage of white population
- Median age of the population

This model greatly improved as its accuracy and its AUC ROC are both > 0.89 (see Fig. 6)



```
accuracy 0.8964929371651242
precision 0.9567901234567902
recall 0.8304919629810034
f1 0.8891786179921772
Roc 0.8964929371651241
[[1976   77]
 [ 348 1705]]
```

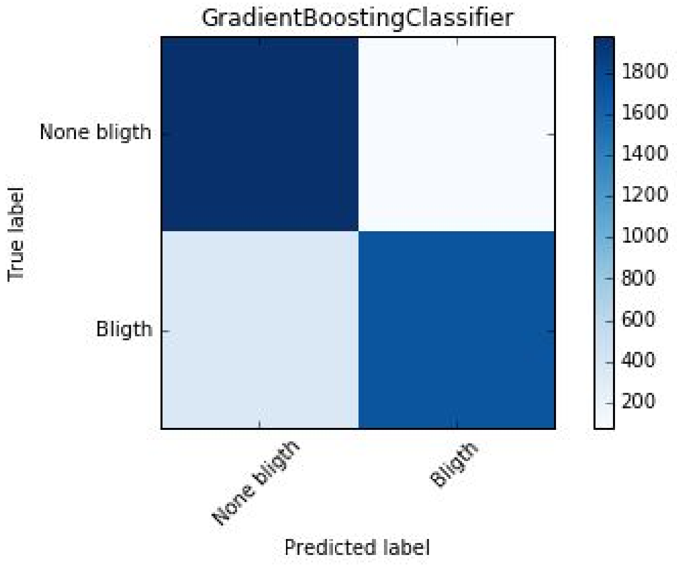


Fig 6. Final model confusion matrix

Fig. 6 illustrates the relative importance of the feature of the final model. The massive importance of the number of blighted buildings in the same region -geohash 5- (Blight_5) is shown (as its weight is almost 53%). Note that the 3 following features by importance are also related to spatial environment: number of violation(7.4%), number of crime (4.7%) and number of call to 311 (3.5%) for a region of geohash 5. The two next features are related to the building itself: number of call to 311 (3.5%) and number of unpaid blight violation (3.3%). Then comes other spatial information: number of violation in the immediate neighborhood- geohash 8 (3.3%) , number of crime in a Geohash 6 range (3%) and number of blighted buildings in the same region -geohash 6 (2%). All other features have an importance inferior to 2%.

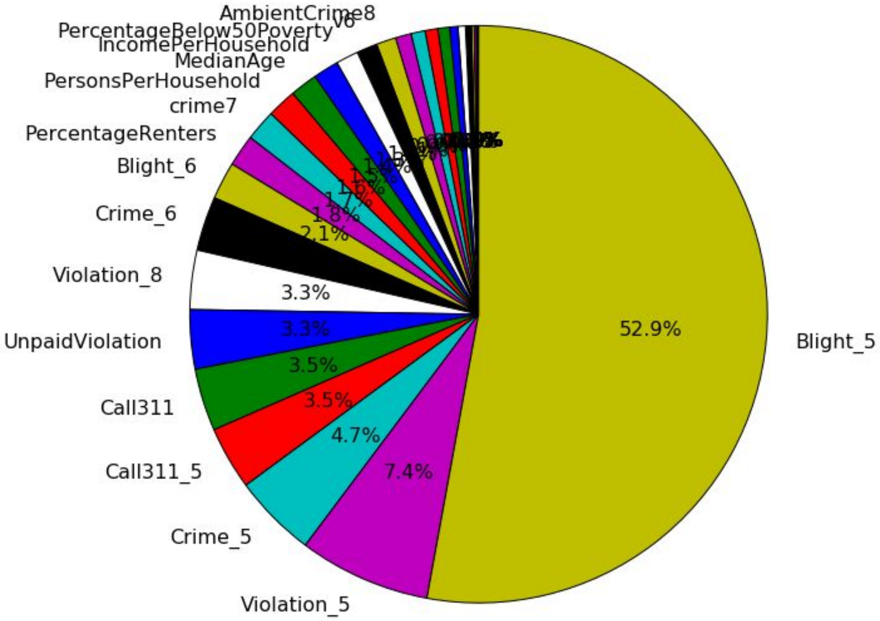


Fig 6. Final model features importance

In conclusion, I found the results of this study very promising in providing a tool that could be used to anticipate building abandonment in Detroit.

Reference:

Spatial distribution of abandonment Victoria C Morckel Applied Geography 48, 8-16