DS at Scale Capstone: Blight in Detroit

Problem description

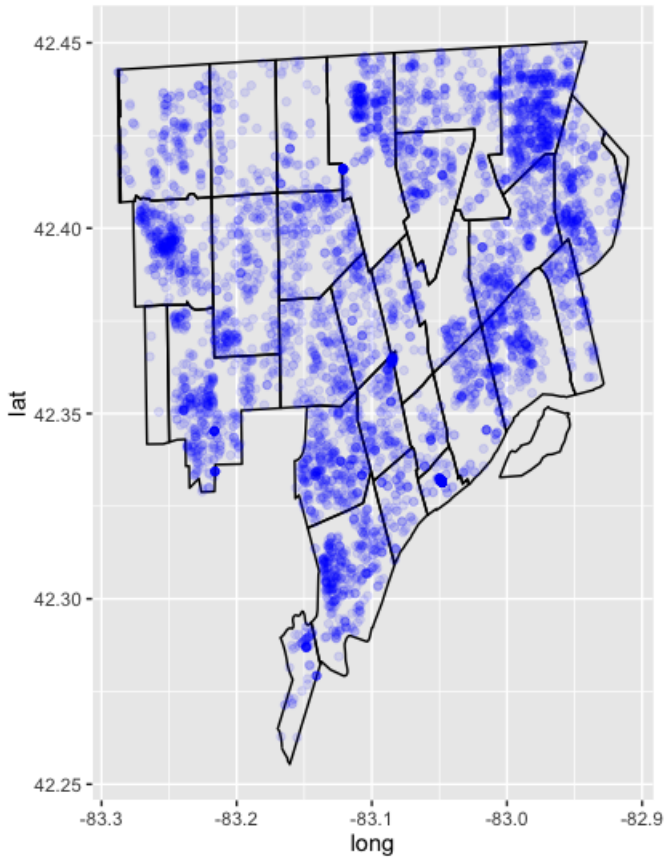
The purpose of this project was to build a binary classification model of blighted buildings in Detroit. The goal was to predict if a building was blighted or not using publicly available data.

The datasets used in the project included parcel information, permit violations, municipal service requests, crime, and demolitions in the Detroit area. The response variable was generated by considering a building as "blighted" if at least one "Dismantle" demolition permit was associated with that building. The remaining datasets were used to create features to help predict if a building was blighted or not.

Data

The data for this project came from five different sources: detroit-demolition-permits.tsv, detroit-311.csv, detroit-crime.csv, detroit-blight-violations.csv, and Parcel_Points_Ownership.csv. These files contain information about demolition permits, crime reports, blight violations, 311 calls, and parcel information respectively. Information about how they can be obtained is listed in the appendix.

In all of these data files incidents were recorded along with latitude and longitude (lat/lon) where the incident occurred. As part of the data cleanup process, incidences with lat/lon well outside of the Detroit area were discarded. Namely, records where the latitude was less than 42.25 or greater than 42.5 were discarded; and records where the longitude was less than -83.3 or greater than -82.9 were discarded. Similarly, records where lat/lon was missing were discarded. A plot showing the locations of the blighted buildings after this data cleanup is show below.



Methods

Deriving Buildings

The Parcel Points Ownership dataset was used as the ground truth for what defined a building. This dataset provides the locations as lat/lon pairs and addresses of properties in Detroit.

In all the datasets, the lat/lon pairs were converted to Universal Transverse Mercator (UTM), which changes the lat/lon pairs to points on a two-dimensional Cartesian grid. Because Detroit is a small geographic area, little relative geographic fidelity in terms of the distance between two points is lost through this conversion.

Following the generation of UTM coordinates, records in the demolition permits, crime reports, blight violations, and 311 calls datasets were associated with the same building identifiers as those in the parcel points ownership dataset if the record's geographic coordinates were within 30 meters of the building's location. If a record fell within the extents of multiple buildings, that is, if the record was within 30 meters of the center of multiple buildings in the parcel points dataset, that record was assigned to the closer building.

This process resulted in 381,047 buildings identified in the full dataset, 4,397 of which were identified as "blighted" according to demolition permit records.

Features

From the parcels dataset the building's last sale price, tax status, location, parcel size, year built, whether the land is known vacant, the equalized value, assessment value, and taxable value were considered. Not all of these values were available for all parcels, however, and so these features were only included when available.

Counts of blight violations were generated from the blight dataset. These counts were grouped by building identifier, the issuing agency, and violation code.

Counts of crimes associated with each building were generated, and these counts were grouped by each crime's category – larceny, drugs, fraud, etc.

Finally, 311 ticket counts grouped by issue type and rating were generated for each building.

All together, this resulted in 761 features that were included in the model.

Model

As mentioned earlier, only around 1% of all the buildings identified were blighted. To help alleviate problems training a model on such skewed classes, a smaller, balanced dataset where 50% of the buildings were identified as blighted and 50% of the buildings were identified as not blighted was created. This smaller dataset consisted of the 4,397 blighted buildings, along with 4,397 randomly selected non-blighted buildings from the larger dataset.

The 8,794 buildings in the smaller dataset were split up into two groups: 80% of the data in this smaller dataset was used to train the model, with 20% of the data, or 1,758 buildings, taken out as test data to assess the model's performance after training.

A generalized boosted regression model was created to predict if buildings were blighted. The model was tuned over a grid with feature interaction depths of 1, 3, and 5; shrinkage values of .05, .1, and .2; and the total number of trees fit ranging from 100 to 1500 in increments of 100. Model performance on the training data was assessed using 10-fold repeated cross validation with five repeats.

In this process, area (AUC) under the receiver operating curve (ROC) was used as the metric for model assessment. Ideally, the model would be assessed by taking into account a suitable loss function when finding false negatives/false positives. However, since these losses are not immediately clear, AUC provides a reasonable substitute. A key feature of AUC is that it is independent of the fraction of the population that is blighted versus non-blighted. This feature of AUC makes it more suitable than accuracy as a metric for model assessment in this case because of the extreme class imbalance of blighted versus non-blighted buildings.

Results

On the test set the model achieved accuracy of 0.78, sensitivity of 0.76, and specificity of 0.81. The confusion table for the test set is shown below.

Test data:

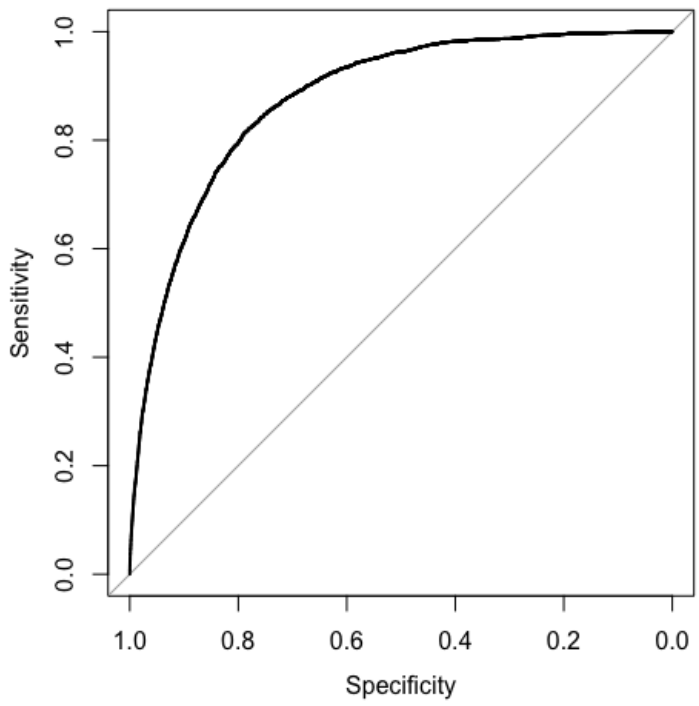
	Blight free	Blighted
Blight free	667	168
Blighted	212	711

On the full data set the model achieved accuracy of 0.76 with a 95% confidence interval of (0.7619, 0.7646), sensitivity of 0.76, and specificity of 0.84. The confusion table for the complete dataset is shown below.

Full dataset:

	Blight free	Blighted
Blight free	287,146	721
Blighted	89,508	3,676

The AUC for the final model on the full data set was 0.88, indicating good predictive power. The ROC for the complete dataset, which shows the trade-off in sensitivity and specificity as the discrimination threshold is varied, is shown below.



Discussion

Considering the data's limitations and the problem domain's inherent uncertainty, this relatively simple model did a good job identifying blighted buildings. The default discrimination threshold of 0.5 results in many false positives, however, and the depending on the costs of false positives for a particular city, may need to be adjusted using the ROC.

Currently the model does not take any spatial, temporal, or economic effects into consideration. Previous literature suggests that these effects may be important predictors for blight, and that instances of blight may not be independent – that is, instances of blight may causally drive new blight resulting in blight "clusters." Future work may strengthen the model by adding temporal and economic predictors, and by incorporating a spatial dependence structure for instances of blight.

Appendix

Data

Data for 311 calls, blight violations, crime, and demolitions is available at the Data Science at Scale github page: https://github.com/uwescience/datasci_course_materials/tree/master/capstone/blight

Parcel data is available from the Detroit Open Data Portal here: https://data.detroitmi.gov/Property-Parcels/Parcel-Points-Ownership/eijm-6nr4/data#Export

Reproducing results

All of the code used to generate this report is available at: https://github.com/zcollab/ds_capstone. To reproduce the model first download the data as described above, and then run transformer.R in the data directory. This file generates the building identifiers. Then run rforest.R in the models/rforest directory to train the model. The model will be saved as "gbm_fit_auc_grid.rds".