

# Blight Prediction

*Alicia Brown*

*May 8, 2016*

## Contents

Overview . . . . .	1
The data . . . . .	1
Files . . . . .	1
Data preparation & Tools . . . . .	2
Buildings . . . . .	2
Features . . . . .	2
Models . . . . .	2
Training & Test datasets . . . . .	3
Tree Model . . . . .	3
Tree Model Analysis . . . . .	4

## Overview

The purpose of this project was to determine if any data features of a city can be used to predict blight for a given location. All code, ETL workflow and analysis is available on <https://github.com/aliciatb/blight>.

## The data

The data provided for this project include 311 Service Calls, Crime Incidents, Blight Violations and Permits for Demolition. The first 3 datasets provide the foundation for a Buildings dataset that consists of unique locations as well as the source of derived features used in the model creation like *Number of Crimes*, *Number of Blight Violations*, and *Number of Service Calls*. These datasets were downloaded from the course site, but are also available via capstone project repo on github. All of the data comes from Socrata powered Detroit Open Data Portal, <https://data.detroitmi.gov/>.

## Files

- Blight Violations
- 311 Service Calls
- Crime Incidents
- Demolition Permits

**Note** - I downloaded this Detroit Demolition dataset to use rather than one of building permits provided by instructor since it was cleaner data and contained only the essential fields needed to label known blight locations.

## Data preparation & Tools

The greatest challenge in the provided data from Socrata is within the Location column because it concatenates all of the fields used in the geocoding process, and when address fields are included, line breaks are entered into the field and cause havoc until they are removed from the data file. Before analysis of the data could be undertaken, all files were initially formatted using Excel PowerQuery for removal of aforementioned line breaks and standardization of the street number and addresses. Then the data was loaded into FME, a powerful ETL tool, to validate and standardize the geographic coordinates and create well formatted incident and unique building files. Exploratory analysis and model creation was performed in python notebooks and RStudio.

## Buildings

In order to derive a building, within the incident files, the latitude and longitude coordinates were rounded to 3 decimal points and then each file was individually joined with the demolition data that also had its latitude and longitude coordinates rounded to the same number of decimal points. Where there was a match in each file, then that building record was also labeled as blight. Any incident record that lacked a coordinate was excluded from the final dataset. One interesting discovery is that street addresses are not consistently captured across datasets and could not be relied on for aggregation due to variances in the same address.

## Features

Features drive the creation of prediction models because it is in their diversity that differences can be discovered that explain why one given building may be more prone to becoming blighted than another. In one of the readings for the course, Spatial Characteristics of Housing Abandonment, Dr. Morckel surmises that housing abandonment is a result of 3 key conditions - market conditions, gentrification and physical neglect. For this project, we are mostly focusing on the data evidence of neglect.

The first features added to the building dataset include a count of total 311 calls, crime incidents and blight violation citations for a given building. No filtering was performed on any of the incident datasets, because I didn't want to presume that calls about infrastructure or non-violent crimes are not related to a geographic inclination towards neglect.

A second set of features were added from a Property Values dataset found on the Detroit Data Portal that included appraised and taxed values, sales price, tax status, and whether it had been improved at any point. It also included well formatted street address and latitude and longitude coordinates which helped to further reduce the overall building dataset since any buildings that lack features are not useful in generating prediction models.

Features that were not included in this project that would be interesting to add so that the model accounts for possible gentrification and economic conditions could include - \* Building Permits - alterations and other types that may signal gentrification in progress in neighborhood \* Zillow Zestimate for an address \* American Community Survey annual estimates on income, mortgage and rental at Census tract block level

It would also be interesting to consider time as a factor and perhaps calculate incident counts at 90 day, 180 day and annual intervals before demolition.

## Models

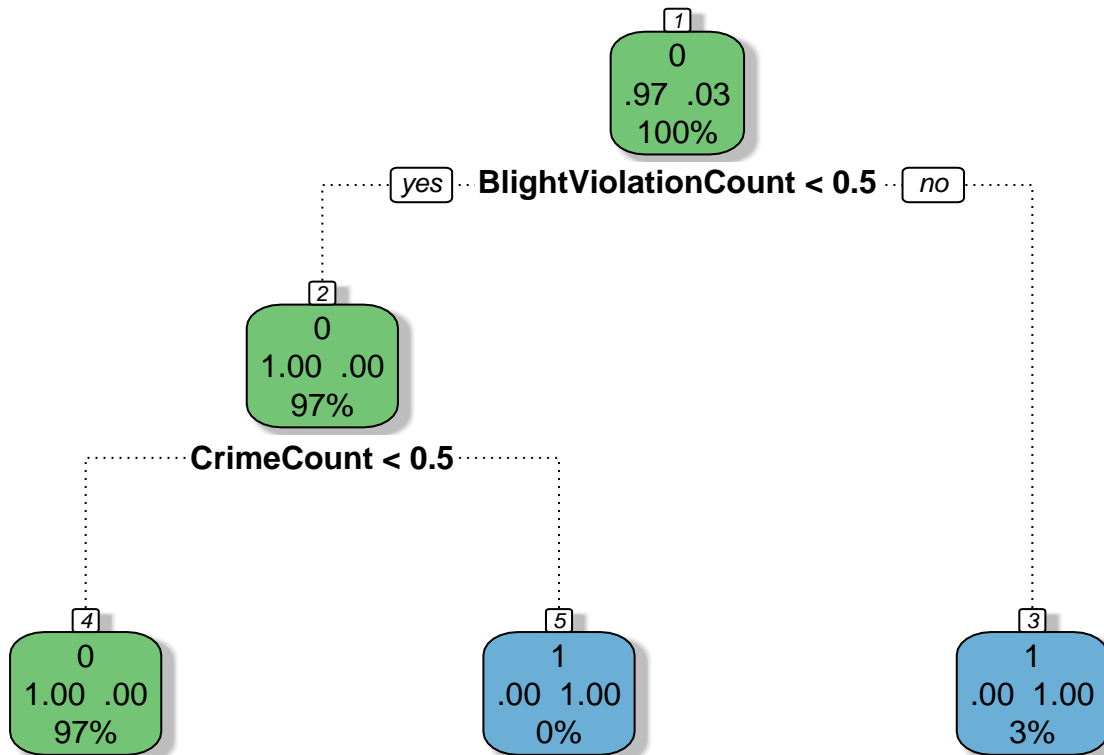
For the prediction of whether a location is blighted or not, I used logistic regression and classification tree models to answer that questions. Logistic regression models the probability that an outcome belongs to a category. Classification trees identify feature importance and provide a visual representation of the feature path taken to form the best performing model. The models will be judged on their Accuracy and Kappa values. Kappa values evaluate whether the accuracy is due to random chance.

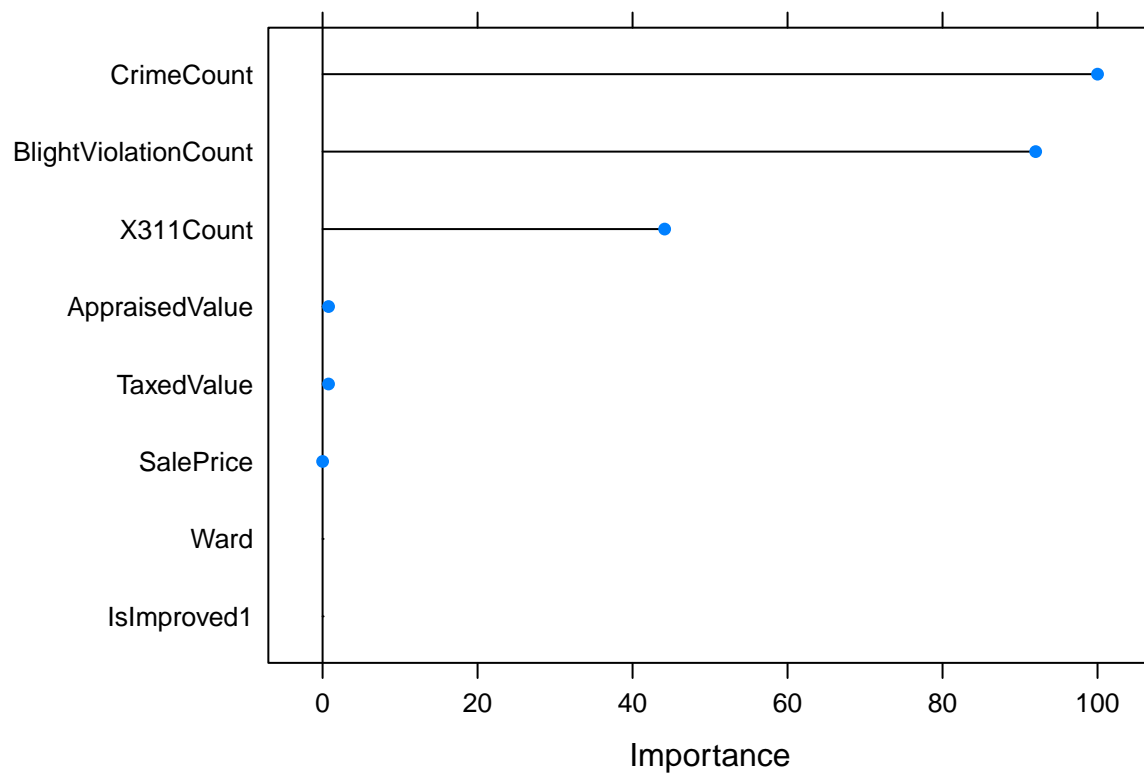
## Training & Test datasets

In order to estimate the accuracy of a model, one must divide the data into training and test datasets. I have allocated 75% of the data for training the models. I have also set both blight and IsImproved fields to factors and dropped fields from the dataset like Address, Latitude, Longitude, Parcel Number, Tax Status and Year that Residence was built so that the models evaluated only the fields - IsImproved, Appraised Value, Taxed Value, Ward Number, and number of 311 calls, crimes, and blight violations as indicators of blight. Originally, Tax Status was included, but its importance to all models was not significant.

## Tree Model

The first model is a simple tree that reveals the decisions that most frequently indicated what factors led to blight classification.





```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##           0 38937  15
##           1    0 1294
##
##           Accuracy : 0.9996
##           95% CI : (0.9994, 0.9998)
##           No Information Rate : 0.9675
##           P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.994
##           McNemar's Test P-Value : 0.0003006
##
##           Sensitivity : 1.0000
##           Specificity : 0.9885
##           Pos Pred Value : 0.9996
##           Neg Pred Value : 1.0000
##           Prevalence : 0.9675
##           Detection Rate : 0.9675
##           Detection Prevalence : 0.9678
##           Balanced Accuracy : 0.9943
##
##           'Positive' Class : 0
##
```

## Tree Model Analysis

This model indicates that the top 3 fields that have the most predictive value are **Crime Count**, **Blight Violation Count** and **311 Call Count**. The overall accuracy of the model is 99.98% with a Kappa value of 99.72%.