Blight Prediction

Alicia Brown

May 8, 2016

Contents

Overview . . . 1
The data . . . 1
Files . . . 2
Data preparation & Tools . . . 2
Buildings . . . 2
Features . . . 2
Models . . . 3
Training & Test datasets . . . 3
Logistic Regression Model . . . 3
Logistic Model Analysis . . . 4
Tree Model Analysis . . . 5
Bagging Model . . . 5
Bagging Model Analysis . . . 6
Boosting Model Analysis . . . 8
Other Models . . . 8
Model Comparison . . . 8
Conclusion . . . 10
Resources & Readings . . . 10

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 accross datasets and could not be relied on for aggregation due to variances in the same address.

Features

Features drive the creation of predicton models because it is in their diversity that differences can be discovered that explain why one given building may be more prone to becoming blightful 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

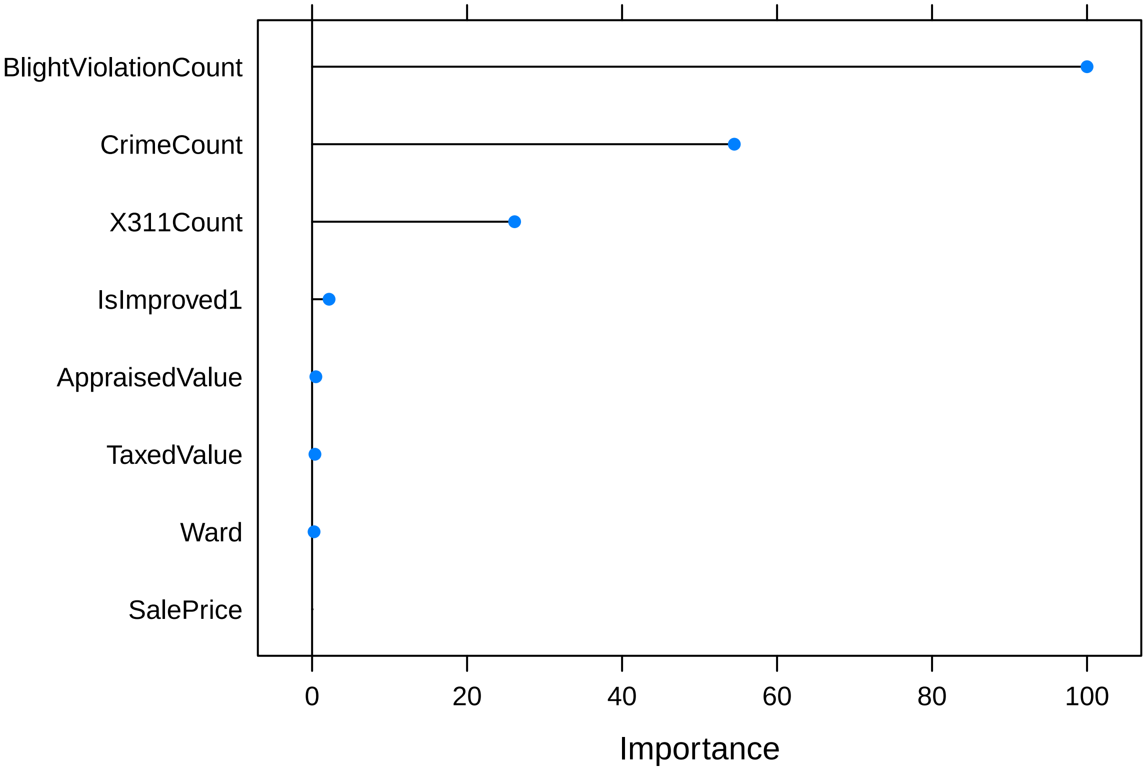
For the prediction of whether a location is blighted or not, I used logistic regression and classification tree models to answer that questions. Logistic regression models the probability that an outcome belongs to a category. Classification trees identify feature importance and provide a visual representation 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.

Logistic Regression Model

```
##
## Call:
## NULL
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.189e-05  -7.760e-06  -7.210e-06  -7.090e-06   5.133e-04
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.063e+00  4.398e+02  -0.005    0.996
## CrimeCount            3.779e+01  1.007e+03   0.038    0.970
## X311Count             1.079e+01  5.990e+02   0.018    0.986
## BlightViolationCount  1.270e+02  1.845e+03   0.069    0.945
## Ward                 -5.943e-02  2.987e+02   0.000    1.000
## SalePrice            -4.230e-02  1.792e+03   0.000    1.000
## IsImproved1          -3.525e-01  2.301e+02  -0.002    0.999
## AppraisedValue       -1.195e+01  3.313e+04   0.000    1.000
## TaxedValue            9.641e+00  3.488e+04   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3.4638e+04  on 120738  degrees of freedom
## Residual deviance: 1.1487e-05  on 120730  degrees of freedom
## AIC: 18
##
## Number of Fisher Scoring iterations: 25
```



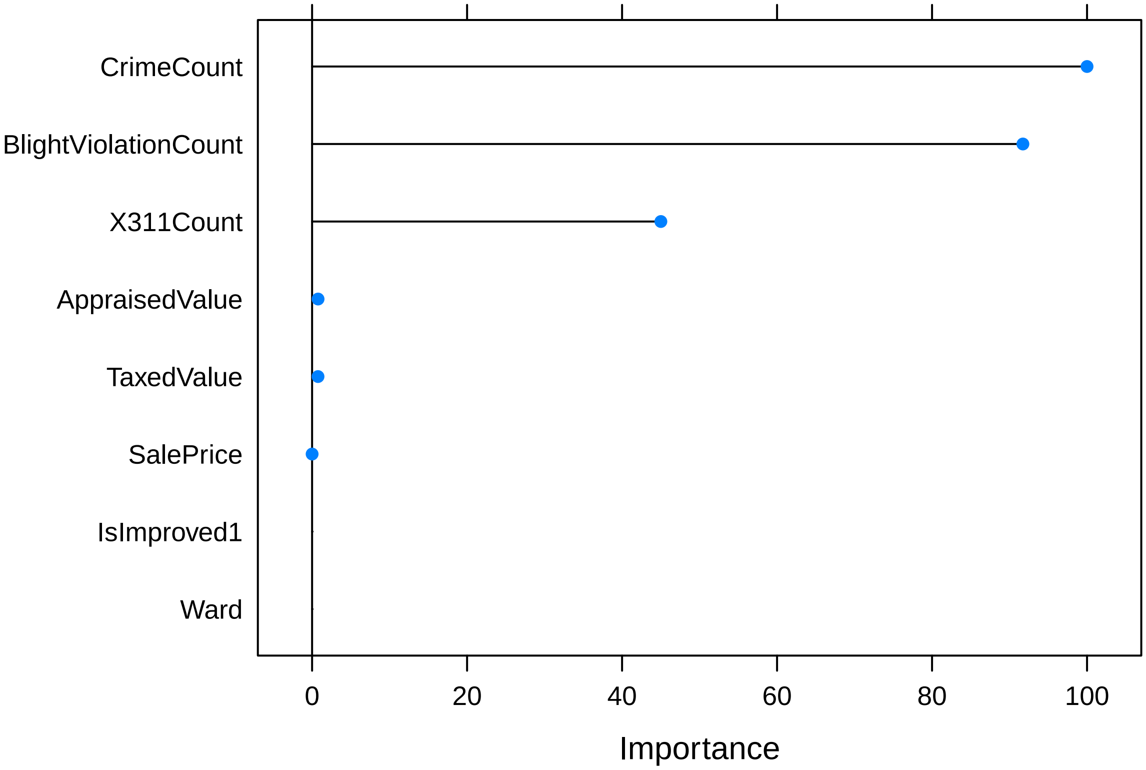
Logistic Model Analysis

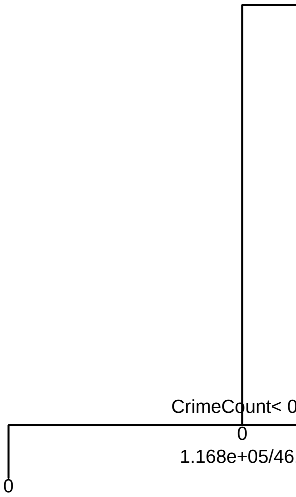
This model indicates that the top 3 fields that have the most predictive value are **Blight Violation Count**, **Crime Count** and **311 Call Count**. The overall accuracy of the model is 100% with a Kappa value of 100%. ### Tree Model The first model is a simple tree that reveals the decisions that most frequently indicated what factors led to blight classfication.

```
## CART
##
## 120739 samples
##      8 predictor
##      2 classes: '0', '1'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 120739, 120739, 120739, 120739, 120739, 120739, ...
## Resampling results across tuning parameters:
##
##   cp           Accuracy   Kappa
##   0.006619145  0.9998777  0.9980571
##   0.110743381  0.9982622  0.9710488
##   0.882637475  0.9787886  0.3730755
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was cp = 0.006619145.
```

```
## n= 120739
##
```

```
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 120739 3928 0 (0.9674670156 0.0325329844)
##   2) BlightViolationCount< 0.5 117272  461 0 (0.9960689679 0.0039310321)
##     4) CrimeCount< 0.5 116837   26 0 (0.9997774678 0.0002225322) *
##     5) CrimeCount>=0.5 435    0 1 (0.0000000000 1.0000000000) *
##   3) BlightViolationCount>=0.5 3467    0 1 (0.0000000000 1.0000000000) *
```





Tree Model Analysis

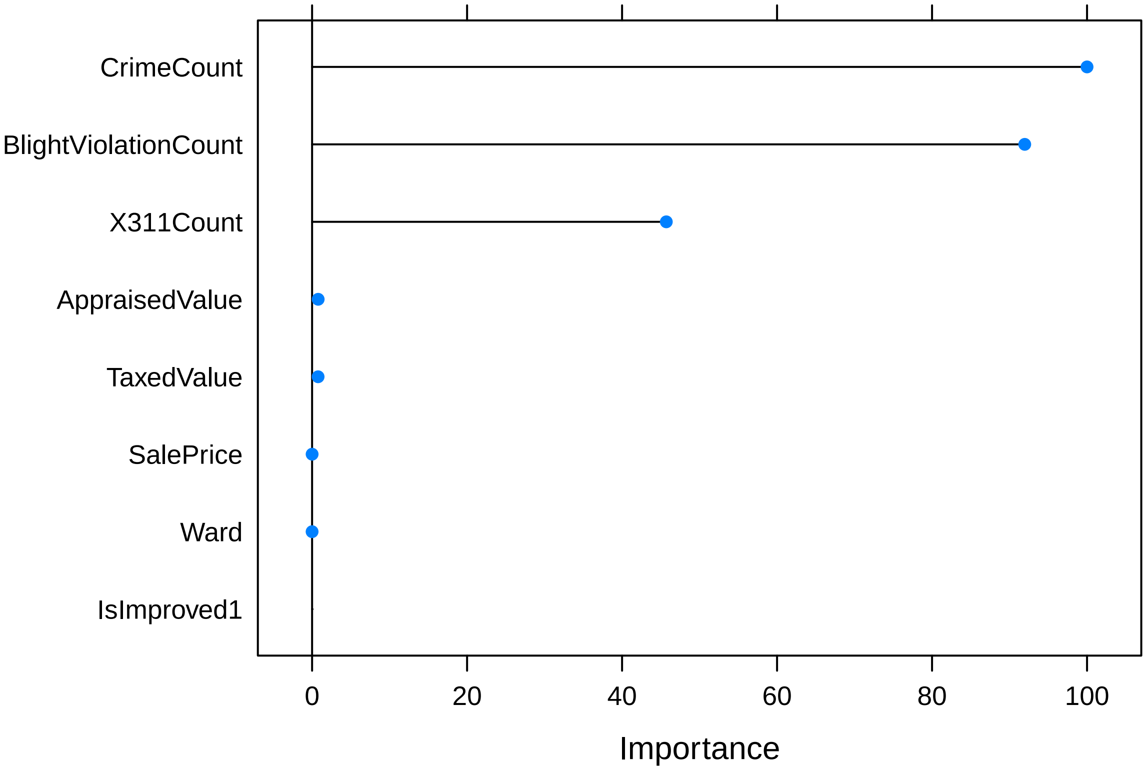
This model indicates that the top 3 fields that have the most predictive value are **Crime Count**, **Blight Violation Count** and **311 Call Count**. The overall accuracy of the model is 99.98% with a Kappa value of 99.72%.

Bagging Model

The next model involves Bootstrap Aggregating where the data is randomly resampled multiple times and the average is returned.

```
## Bagged CART
##
## 120739 samples
##      8 predictor
##      2 classes: '0', '1'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 120739, 120739, 120739, 120739, 120739, 120739, ...
## Resampling results:
##
##   Accuracy  Kappa
##   1         1
##
##
```

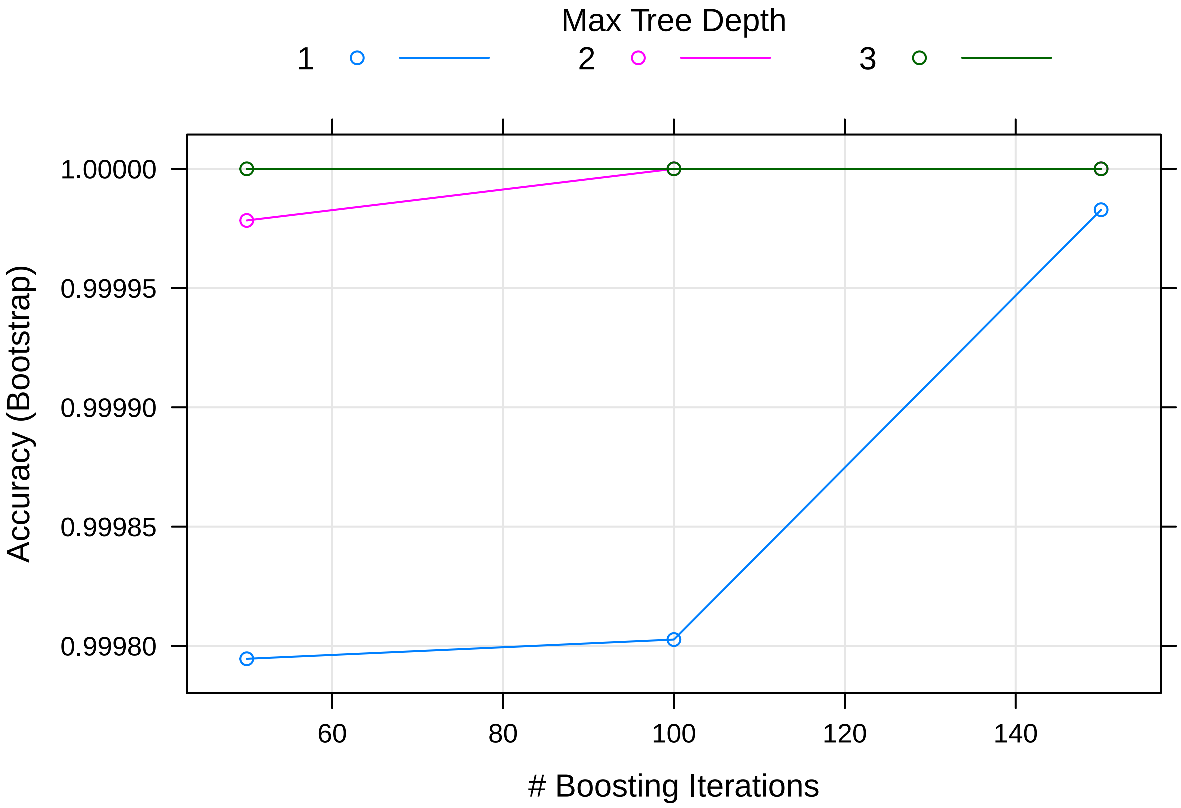
```
##
## Bagging classification trees with 25 bootstrap replications
```

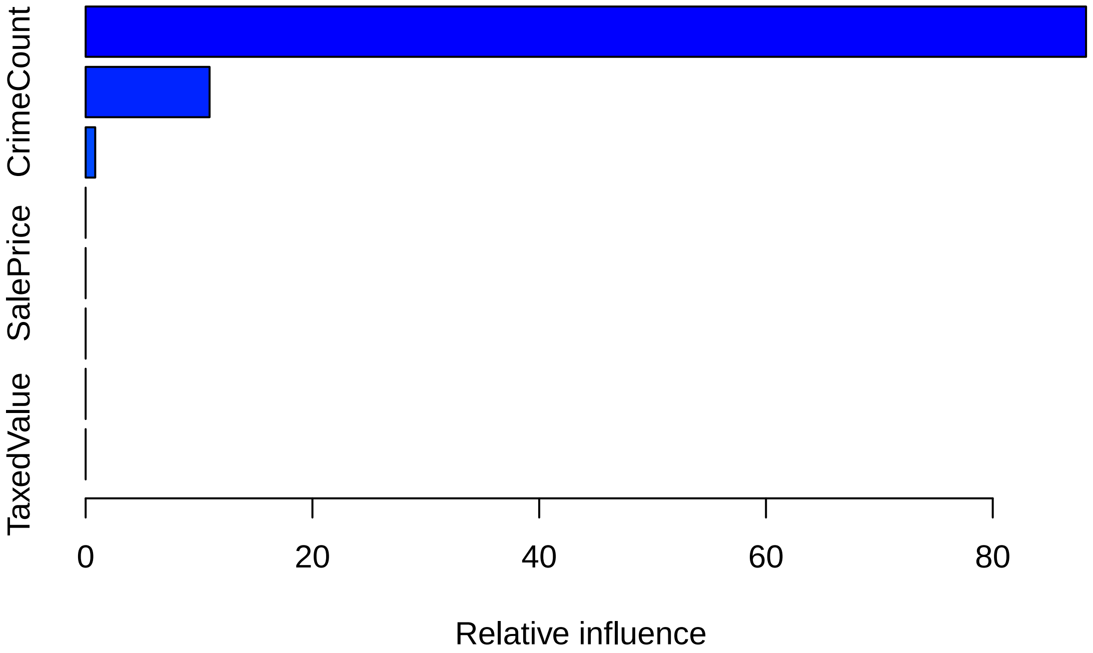


Bagging Model Analysis

This model indicates that the top 3 fields that have the most predictive value are **Blight Violation Count**, **Crime Count** and **311 Call Count**. The overall accuracy of the model is 100% with a Kappa value of 100%. ### Boosting Model This model will combine weak classifiers so they can contribute to creating a more powerful model.

```
## Stochastic Gradient Boosting
##
## 120739 samples
##      8 predictor
##      2 classes: '0', '1'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 120739, 120739, 120739, 120739, 120739, 120739, ...
## Resampling results across tuning parameters:
##
##   interaction.depth  n.trees  Accuracy   Kappa
##   1                   50      0.9997946  0.9967242
##   1                  100      0.9998027  0.9968496
##   1                  150      0.9999828  0.9997302
##   2                   50      0.9999784  0.9996561
##   2                  100      1.0000000  1.0000000
##   2                  150      1.0000000  1.0000000
##   3                   50      1.0000000  1.0000000
##   3                  100      1.0000000  1.0000000
##   3                  150      1.0000000  1.0000000
##
## Tuning parameter 'shrinkage' was held constant at a value of 0.1
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10
## Accuracy was used to select the optimal model using  the largest value.
## The final values used for the model were n.trees = 50, interaction.depth
##  = 3, shrinkage = 0.1 and n.minobsinnode = 10.
```





```
##                                       var      rel.inf
## BlightViolationCount BlightViolationCount 8.822745e+01
## CrimeCount                     CrimeCount 1.092832e+01
## X311Count                       X311Count 8.442137e-01
## AppraisedValue             AppraisedValue 7.277541e-06
## SalePrice                       SalePrice 1.538905e-06
## Ward                                 Ward 0.000000e+00
## IsImproved1                   IsImproved1 0.000000e+00
## TaxedValue                     TaxedValue 0.000000e+00
```

Boosting Model Analysis

This model indicates that the top 3 fields that have the most predictive value are **Blight Violation Count**, **Crime Count** and **311 Call Count**. The overall accuracy of the model is 100% with a Kappa value of 100%.

Other Models

A Random Forest model is the most powerful of the tree classification models that involves bagging where it also resamples the feature combinations. However, the complexity of its algorithm causes long processing and in this case caused memory exceptions on 2 different and otherwise powerful macbookpro and imac computers, so was unable to run this model successfully.

Model Comparison

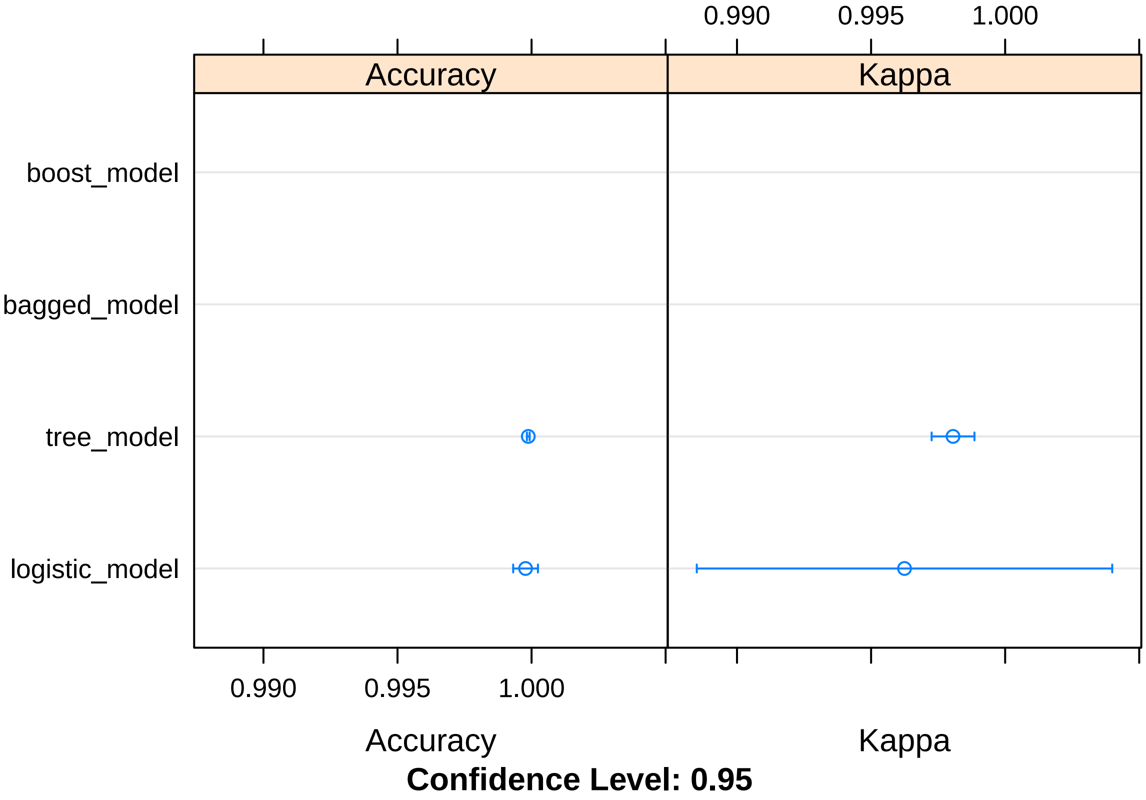
It is easy to compare the 4 models in R and generating a visualization on how they compare using Caret package [resample methods] (http://www.inside-r.org/packages/cran/caret/docs/as.matrix.resamples).

```
# compare
results = resamples(list(tree_model = tree_model,
                         bagged_model = bagged_model,
                         boost_model = boost_model,
                         logistic_model = logistic_model))

# compare accuracy and kappa
summary(results)
```

```
##
## Call:
## summary.resamples(object = results)
##
## Models: tree_model, bagged_model, boost_model, logistic_model
## Number of resamples: 25
##
## Accuracy
##                  Min. 1st Qu. Median   Mean 3rd Qu. Max. NA's
## tree_model     0.9997  0.9998 0.9998 0.9999       1    1    0
## bagged_model   1.0000  1.0000 1.0000 1.0000       1    1    0
## boost_model    1.0000  1.0000 1.0000 1.0000       1    1    0
## logistic_model 0.9944  1.0000 1.0000 0.9998       1    1    0
##
## Kappa
##                  Min. 1st Qu. Median   Mean 3rd Qu. Max. NA's
## tree_model     0.9953  0.9961 0.9972 0.9981       1    1    0
## bagged_model   1.0000  1.0000 1.0000 1.0000       1    1    0
## boost_model    1.0000  1.0000 1.0000 1.0000       1    1    0
## logistic_model 0.9062  1.0000 1.0000 0.9962       1    1    0
```

```
# plot results
dotplot(results)
```



Conclusion

Three of the the four models returned accuracy and kappa values of 100%, while the 4th scored 99.9% accuracy and 97% kappa values and all indicated that blight violation and crime counts are key factors in whether a building is blight, with 311 call counts also being a significant indicator. It isn't surprising that these 3 attributes are excellent indicators of blight, but their usefulness in building a model to predict the probability of blight can be questionable if the goal of a city is to become aware of potential blight early enough to be able to counter whatever conditions are truly the underlying cause of blight. For this reason, I think it is very important to build models using economic and gentrifying features in addition to physical neglect and crime statistics because they may be able to reveal problem areas early. I would also like to account for time in models and see if there is a significant difference in incident counts at different time intervals.

In addition to adding more economic features in model building, I am excited about the possibility of a new feature that FME has added to its latest release is the RCaller transformer that I hope will allow it to leverage the Caret library which is what I used entirely for my models. This would make it possibly to create a workflow that could analyze the data for any city, provided the datasets conform to a general schema for 311 calls, crime incidents and blight violation counts, on a regular schedule and predict what properties are likely to be blight. If the data can indeed be standardized across a group of cities, and more known blight can be used to train the models, this would be an amazing opportunity to empower cities to proactively address potential issues while conditions are manageable rather than only be able react to them with bulldozers.

Resources & Readings

- Detroit Blight Removal Task Force Plan
- Course Readings