working paper

16 37

Predictive Modeling of Surveyed Property Conditions and Vacancy

Hal Martin, Isaac Oduro,
Francisca Garca-Cobin, April Hirsh Urban, and
Stephan Whitaker

FEDERAL RESERVE BANK OF CLEVELAND

Working Paper 16-37 December 2016

Predictive Modeling of Surveyed Property Conditions and Vacancy

Hal Martin, Isaac Oduro, Francisca Garca-Cobin,
April Hirsh Urban, and Stephan Whitaker

Using the results of a comprehensive in-person survey of properties in Cleveland, Ohio, we fit predictive models of vacancy and property conditions. We draw predictor variables from administrative data that is available in most jurisdictions such as deed recordings, tax assessor's property characteristics, and foreclosure filings. Using logistic regression and machine learning methods, we are able to make reasonably accurate out-of-sample predictions. Our findings indicate that housing professionals could use administrative data and predictive models to identify distressed properties between surveys or among nonsurveyed properties in an area subject to a random sample survey.

Keywords: Vacancy, distressed properties, machine learning, predictive models, property surveys.

JEL codes: R31, C53, C25.

Suggested citation: Martin, Hal, Isaac Oduro, Francisca Garca-Cobin, April Hirsh Urban, and Stephan Whitaker, "Predictive Modeling of Surveyed Property Conditions and Vacancy," Federal Reserve Bank of Cleveland, Working Paper no. 16-37.

Hal Martin is at the Federal Reserve Bank of Cleveland (hal.martin@clev.frb.org). Isaac Oduro is at Case Western Reserve University (isaac.oduro@case.edu). Francisca Garca-Cobin Richter is at Case Western Reserve University (fxr58@case.edu). April Hirsh Urban is at Case Western Reserve University (april.urban@case.edu). Stephan Whitaker is at the Federal Reserve Bank of Cleveland (stephan.whitaker@clev.frb.org).

1 Introduction

Cities, housing authorities, and other local governments make extensive use of data on individual properties to support the provision of public services and target housing interventions. County governments record property transactions, tax receipts, and property characteristics for property tax valuations, Occasionally, agencies will supplement this data by commissioning a comprehensive, in-person survey of all properties. This analysis investigates whether machine learning techniques or standard econometric techniques could enable housing professionals to predict the findings of an in-person survey using administrative data already in hand. If highly accurate prediction is possible, service agencies could identify properties of concern between surveys. Additionally, agencies could opt for a sample survey costing a fraction of a comprehensive survey, use the sample to train a predictive model, and then use the predictive model to identify properties of concern among all non-surveyed properties.

The survey data used in this analysis covers 98,191 properties in the city of Cleveland, Ohio.[1] The Thriving Communities Institute conducted the survey from June through September of 2015 (Thriving Communities Institute, 2016). The Western Reserve Land Conservancy, the Cleveland Foundation and the City of Cleveland provided funding to cover the approximately $200,000 cost (Naymik, 2015). The 16 surveyors recorded 17 observations and an image for each property. We will focus on their assessment of whether or not the property was vacant, and the assignment of an overall condition letter grade. The negative externalities of vacant and blighted properties have been explored in a large and growing literature. Therefore, housing professionals would be interested in identifying or predicting these two statuses. In specifying predictive models, we draw 46 predictive variables from county property records, court filings, Census data, and US Postal Service records.

The remainder of this paper proceeds as follows. Section 2 reviews similar modeling projects already undertaken, and some related literature. Section 3 describes the data sources in more detail and our definitions of potential predictors. Section 4 describes the classification

[1] The count of surveyed residential structures is much higher, at approximately 125,000. However, many observations are dropped because they fail to merge with all the predictive variables from the administrative data.

techniques we consider. Section 5 presents our results, and section 6 concludes with a discussion of the selection of a model by practitioners and future directions.

2 Literature

In the recent literature, there have been some attempts to create models that predict vacant or distressed properties. Morckel (2014) compared models that estimate vacancy using a variety of proxy variables for vacancy and determined that the different proxy variables led to significantly different results. Hillier et al. (2003) distinguished between short-term and long-term vacancies, as well as abandonment, where abandonment is a designation of "imminently dangerous" by the City of Philadelphia's Department of Licenses and Inspections. They build a logistic regression to estimate abandonment, note the functional limits of their model, and suggest the possible application of alternative machine learning methods, such as neural networks, to the problem. Their study demonstrated the importance of physical, financial, and neighborhood level data in predicting housing distress. Appel et al. (2014) provided the basis for this study by using a random forest model to predict vacancy in Syracuse at the parcel level. Appel et al. demonstrated that machine learning methods, such as a random forest model, are highly accurate, although they did not compare their results to other methods, such as a logistic regression. In their partnership with Syracuse, Appel et al. demonstrated how modern data analytic approaches can inform city policy in fighting blight.

The numerous studies that have estimated the negative externalities of distressed properties make use of county recorder and tax assessor data like that used here. The most extensively studied negative externality is that of foreclosure (Immergluck and Smith, 2006; Schuetz et al., 2008; Harding et al., 2009; Leonard and Murdoch, 2009; Rogers and Winter, 2009; Rogers, 2010; Campbell et al., 2011; Groves and Rogers, 2011). Hartley (2014) investigated the interaction of vacancy and foreclosure by producing separate estimates in high- and low-vacancy neighborhoods. Mikelbank used a comprehensive survey conducted by the city of Columbus, Ohio, to estimate the externality of a property being "vacant and abandoned" (2008). The Columbus survey, like the Cleveland survey, included the surveyor's assessment of this extremely distressed status. Mikelbank estimated that for each additional vacant and abandoned property within 250 feet of a sale, the price was lower by 2 percent. The externality estimate declined to 1.5 percent at 250-500 feet and declined further at distances beyond that. Whitaker and Fitzpatrick 2011 use address-level US Postal Service vacancy data to estimate the negative externality of a vacant property, along with other indicators of distress. They find that an additional vacant property within 500 feet of a sale had a negative externality of 2.9 percent in low-poverty tracts and 0.8 percent in high-poverty tracts with the former being statistically significant. Whitaker and Fitzpatrick argue that tax delinquency is a viable proxy for housing neglect and blight because owners that are unable or unwilling to pay their property taxes are likely also unable or unwilling to maintain the property. They find that tax delinquent properties have negative externalities that are similar in magnitude to those of vacant homes.

3 Data

The City of Cleveland contains almost 180,000 parcels, of which about 159,000 are residential. These parcels were investigated by the Thriving Communities Institute in the summer of 2015 to determine their vacancy status and condition. The staff of 16 surveyors reported that over 12,000 of the residential parcels were vacant. The surveys recorded a wide variety of characteristics of each home before assigning an overall condition letter grade. Approximately 40 percent of the homes were deemed to be in "A" condition and another 40 percent in "B" condition. Sixteen percent were graded "C" and 3 percent "D." The lowest score of "F" was assigned to 1,531 properties corresponding to 1.4 percent of the observations. It is notable that the two outcomes we are trying to predict are uncommon in the case of vacancy and rare in the case of "F" grades.

While the survey's recorded details of the property's condition (such as peeling paint, missing gutters, cracked cement, etc.) are interesting we do not incorporate them in our study at this time. If we place ourselves in the position of a local government housing professional, we would not have current observations of these details except when we had recently completed an in-person survey. Our goal is to predict the status of properties that have not been recently surveyed. Therefore, the only two variables we use from the TCI survey are the vacancy indicator and property condition letter grade.

Table 1 lists the variables we will consider as predictors. The bulk of the predictive variables are drawn from data maintained by the County Fiscal Officer. This office performs the functions of recording deed transfers and maintaining property characteristic files for property tax assessment. We anticipate that many of the characteristics that make a property more valuable will also make the property less likely to be vacant or neglected. Larger, newer homes, with more amenities should be occupied and well maintained more often. If transfer data indicates that the home has changed owners multiple times in the recent past, it is likely in a flipping cycle, which has been associated with vacancy and poor maintenance (Coulton et al., 2010; Ford et al., 2013). Data indicating foreclosure filings are available from the Cuyahoga County Clerk of Courts. We have used the filing dates to create indicators if the property has been foreclosed upon in the last year or two years.

The City of Cleveland Department of Buildings and Housing provided data reflecting complaints they have received from the public and violations identified by housing inspectors. Violation and complaint data are used to derive a number of features, such as number of days since the last complaint, and the number of complaints and violations in the 12 months before the survey. The department also identifies which housing transactions are arms-length sales. This is important because the raw deed recording data includes thousands of "transactions" that are additions of family members to the title or transfers of properties into trusts.

The only data we utilize that is not publicly available is the address-level vacancy data from the US Postal Service. When postal carriers observe that a home has been vacant for 90 days, they record it as such in the USPS's main address database (these data do not include short-term or seasonal vacancies). This prevents mail addressed to the vacant home from accumulating at the property or needlessly being carried out and back each day. The address database, including vacancy status, is routinely audited and maintained at an accuracy level above 95 percent. The USPS makes its vacancy data commercially available to direct mailers. The companies can run their mailing lists through a software program that marks each record if the address is vacant. Mailings are not prepared for these addresses, so wasted printing and postage is avoided. We have subscribed to the vacancy data since April 2010. We run our list of Cuyahoga County addresses through the software, and create a panel of vacancy indicators.[2]

The two vacancy measures in the data differ considerably, but we do expect postal vacancy to be predictive of survey-reported vacancy. Whether local officials could use the postal vacancy measure as a substitute for survey-reported vacancy depends on the precise vacancy-related concern. For example, a neighborhood that has high turn-over might show few postal vacancies if units are re-occupied in less than 90 days. However, on the particular day of a survey, canvassers might identify many empty units that are not contributing to the vitality and safety of the neighborhood. Postal vacancy, summed over all months since April, 2010, should also be predictive of the lowest letter grade. The longer a property has been vacant, the more likely that it has been vandalized or deteriorated without an occupant to notice maintenance problems.

As shown by Hillier et al., neighborhood characteristics are valuable in predicting blight and help put the structure and owner data into context (2003). Include census block data from the US Census, demographic variables on race, age, income, education, and crime. The survey date during the summer of 2015 is not the same for each home, the data is also filtered such that no data is used from after the inspection. [3]

4 Methods

The data is split randomly into two sets, a test and training set, with 20% of the data going into the former and 80% to the latter. Models are built using the training data and verified using the test data. The first outcome variable is 0 for not vacant, or 1 for vacant as determined by the Thriving Communities Institute ground survey. The second outcome

[2] If a property has been vacant for a year, the USPS database stops populating the field *vacant* and begins recording a status in a variable labeled *no stat*. In our analysis, we code properties with *no stat*="Y" as vacant properties.

[3] The model is built using Python, and the code is available on Github.

variable is set equal to 1 if the surveyed condition was graded "F" and set to 0 otherwise. Each model returns a probability between 0 and 1 for the outcome, which is rounded to the nearest integer for the prediction. Multiple models were built in order to determine which model is the most accurate in application to this type of task. These models include a logistic regression, a random forest model, and a gradient boosted model.

4.1 Logistic regression

For this problem, binary logistic regression is used, as we are estimating an outcome, vacancy, that is binary, occupied or unoccupied. Logistic regressions assume variables have monotonic effects on the outcome, are quick to train, and produce models that are unlikely to overt. Of the three, logistic regression models are the easiest conceptually and yield results that are easy to interpret.

In the paradigm of machine learning, the problem of classification is comparable to that of predicting discrete outcomes using binary logistic models. When presented with a case, the question is: to which of two classes does this case belong? In the context of this study, the question is whether a residential structure is vacant or not.

4.2 Random forest model

Random forest is a classification technique developed in Breiman (2001) in which many decision trees are constructed around random subsets of the dataset. For each subset of observations randomly selected, the tree proceeds to evaluate which of a subset of available predictor variables best classifies the observations according to the outcome variable. A single tree considers multiple predictors in sequence. While a single tree contains a relatively weak predictive model, the combination of inputs from all trees in a large forest forms a model that makes better predictions about the outcome.

4.3 Gradient boosted model

Gradient boosted models, like random forest models, are also built from a number of decision trees. In the gradient boost technique (Friedman et al., 2000; Friedman, 2001), sequential additions are made to a naive or imperfect model in stages, using a decision tree at each stage to inform the addition to the model. Each trees parameters are optimized to minimize a loss function over the residuals of the previous stage's fit, with each stages tree adding to the previous stages accumulated changes to the model.

5 Results

5.1 Vacancy Models

Table 2 presents the parameters of a logistic model estimated with the surveyor's vacancy assessment as the dependent variable and all of the independent variables included. To ease comparison with the random forest and gradient boosted results, the coefficients have been sorted by the absolute value of the Z statistic. The postal vacancy variable has the highest Z score in the logistic model of the surveyor's assessment of vacancy. Among the other variables with the largest Z values are some that will later be identified as important for the random forest and gradient boosted techniques. These include the days since a complaint, the indicator of poor condition in the tax records, and tax delinquency.

In figure 1, we see the distribution of the predicted probability that each property in the test sample will be vacant. To identify specific properties that housing professionals should investigate, or to determine our model's type I and type II error, we must select a probability cut off. Probabilities above the cut off are rounded up to one and predicted to be vacant. The vertical line in figure 1 is placed at the cut off of .5 that we might default to. In table 3, we present the type I and type II error that arises from using a .5 cut off with our logistic model. When applied to the test data, the model is fairly successful at predicting the properties that surveys classified as vacant, identifying 75 percent of them. However, this comes at the cost of almost two false positive predictions for every true positive prediction.

In practice, the cut off should be chosen via a cost benefit analysis. Both types of errors have a cost. If the cost of a false positive error is small relative to the cost of a false negative, we would select a lower cut off.

The relative importance of the predictive variables in the random forest model is displayed in figure 2.[4] As in the logistic regression, complaints, tax delinquency and recent foreclosures are among the greatest contributors to the prediction. Figure 3 displays the distribution of the predicted probabilities of survey vacancy returned by the random forest model. The remarkable contrast with the logistic regression is that the random forest method predicts over 10,000 properties have a zero probability of being survey vacant. With a cut off of .5, the random forest model does not appear to be successful at identifying the vacant properties (see table 4). Only 41 are correctly predicted to be vacant. The share of false positive results are very low, at 203. Considering the skewness of the predicted probability distribution, and the small share of false positive findings, it seems reasonable that one would want to choose a lower cut off.

As visible in figure 4, the gradient boosted model draws predictive power from a wider variety of independent variables. The complaint, foreclosure and tax delinquency variables remain among the most important while the transaction variables move into the top five in the relative importance ranking. The gradient boosted model, even more extensively than the random forest model, declares the majority of the test properties to have zero probability of being surveyed vacant. At the .5 probability cut off, the gradient boosted predictions are only correct for 49 percent of the surveyed vacant homes (see table 5). The false positive predictions are few, as they were for the random forest predictions, suggesting a cut off lower than .5 could be appropriate.

In the parameter rankings and predicted probabilities, we observed dramatic differences between the logistic, random forest and gradient boosted methods. To move forward with application, we need a way to make a direct comparison between the models. To make

[4] Because random forest models are used for prediction or classification, rather than statistical inference regarding specific predictors, users generally do not display parameters or standard errors. The "relative importance" of each predictor is the value usually displayed in the literature. Relative importance can be considered if one is interested in dropping variables to speed computation or reduce data collection efforts.

this comparison, we can examine the receiver operating characteristic (ROC) curves. ROC curves are plots of the true positive rate over the false positive rate. Every point on the curve corresponds to a cut off value between zero and one. As mentioned above, when a model is applied in practice, the user must should select a cut-off based on a cost-benefit analysis. However, the ROC can specify if one model is superior to another without knowing the costs of either type of error. A model with greater area under the ROC curve is generally preferable. Examining the ROC curves in figure 6, we see that the models that appeared disparate in their details appear remarkably similar in their general application. The gradient boosted model is the most effective at delivering true positive values for any given false positive rate, but the differences are small. For example, if we accepted a false positive rate of 0.1, the difference between the best model (gradient boosted) and the worst model (logistic) in terms of the true positive rate is less than .05. The ROC curves of the logistic and random forest models are nearly indistinguishable.

5.2 Property Condition Models

In this section, all the reported results will parallel those reported for the vacancy models, but the dependent variable now has a value equal to one if the property was assessed to have a condition of "F," and zero otherwise. We should recall that 1,531 properties out of the 112,665 properties surveyed were deemed to be in the worst condition category. We are predicting an event that occurs in less than 1.5 percent of cases.

Table 6 presents the coefficients from the logistic regression. As in the vacancy model, recent complaints, tax delinquency, and a tax record designation of poor quality are highly predictive. Two demographic variables, percent Black and ACS-reported block vacancy rate have relatively high Z values.

Moving to the distribution of predicted probabilities from the logistic model, we see in figure 1 that much of the weight of the distribution is below 0.1. The density above a cut off of .5 is not visible at this scale. Table 7 indicates that at a cut off of 0.5, the logistic model predicts 63 percent of the grade F properties correctly. As in the vacancy logistic model, the rate of false positives, 0.74, is very high at the default cut-off.

In figure 8, we can see that three variables contribute extensively to the prediction of grade F by the random forest model. The count and timing of the complaints are most informative. The tax assessor's record of the property condition is the third highest ranked predictor. While it is not surprising that one measure of condition is able to predict another, the strength of the predictive power is interesting. The tax assessor usually does not do an in-person survey to update this record. Rather, homeowners can appeal their assessments, provide evidence that their property is in poor condition, and possibly have their tax assessment reduced. The tax record condition can be raised from poor to good if permits are pulled as part of a rehabilitation, and the post-work inspection confirms that the condition has been improved.[5] The full distribution of tax assessor's property conditions may not be predictive of the full distribution of the surveyed conditions, but the agreement of the measures on the worst condition properties appears to be strong. The tax relief available to owners of properties in very poor condition seems to incentivize them to reveal that property condition and have it recorded in the administrative data.

Beyond the complaint and condition variable, the tax variables are also highly ranked in terms of their contribution to the gradient boosted prediction. This could reflect another strategy frequently used by owners of poor-condition properties. Rather than appealing their property tax assessments, many owners of low-value properties just opt to not pay their property taxes (Whitaker, 2011). They know that the county only has the resources to foreclose on a small fraction of tax-delinquent properties each year, and the county will prioritize higher value properties to maximize collections. The owners of many low-value properties routinely do not pay their taxes, creating a strong relationship between tax delinquency and poor condition.

As with the vacancy model, the random forest predicted probability distribution is heavily weighted toward zero (see figure 9). The actual versus predicted cross tab in table 8 suggests a cut off of .5 is not at all helpful with this classification routine in this context. The

[5] The City of Cleveland offers substantial tax abatement to encourage investment in property rehabilitation, so owners have an incentive to pull permits and have work inspected.

random forest model only predicts a probability above .5 of grade F for 40 properties, and half of those are false positives. Ten times as many properties are assigned a false negative prediction.

When we turn to the gradient boosted model of property conditions, the same three independent variables again have the highest relative importance, namely complaints and the tax assessor's condition measure (see figure 10). Tax delinquency measures contribute, but not to the same extent as seen in the random forest model. The density of the predicted probability is the most skewed of any in this analysis, with almost all density on zero, and no values above .8. In table 9, we see that the predictions are not at all useful if the cut off probability is .5.

Finally we turn to the ROC curves to see if there is evidence to prefer any of the three models of the survey assessed grade F. In contrast to the vacancy ROC curves (figure 6), the property condition ROC curves suggest generally more accurate prediction by having more area under the curves. Among the three models, the random forest model appears to be distinctively disadvantaged. The curves corresponding to the logistics regression and the gradient boosted predictions are very similar, with the logistic predictions holding a slight advantage overall.

6 Conclusion

From this first attempt to model the surveyors' assessments of vacancy and property condition, we have learned that it is possible to predict the assessments out of sample with fairly high accuracy. Using a default cut off of .5 to round the predicted probabilities, the logistic model can predict 75 percent of survey-identified vacancies and 63 percent of survey-identified grade F properties. With a false positive rate of 0.2, logistic, random forest and gradient boosted methods can all predict vacancy with true positive rates between 0.8 and 0.9. With a false positive rate of 0.1, logistic and gradient boosted methods can predict survey grades of F with true positive rates above 0.9.

In their current forms, the random forest and gradient boosted models do not appear to have advantages that outweigh their disadvantages. Their predictive power is not substantially superior to the logistic model for vacancy or property conditions. Logistic regression has the advantage of being available in nearly all statistical software packages, including many platforms available at no cost. Regressions are familiar to professionals in a wide variety of fields, while machine learning techniques are just beginning to be adopted. Logistic regression provides coefficient estimates and enables statistical inference regarding those coefficients, which may be of interest to practitioners in addition to the model's overall predictive power. Logistic estimates can be produced in under a minute, while the random forest and gradient boosted estimates take several hours to produce. The computational intensity discourages expert interactions with the machine learning models because the impact of small changes in the specification take a great deal of time to test.

Further refinement of the models may reveal advantages not yet seen here. For example, the logistic model relies more heavily on the postal vacancy data to predict surveyed vacancy. If a municipality cannot incur the expense of obtaining and processing that data, then the logistic model may prove inferior in predictive power. Similarly, if one is operating in a county that does not recognize and record a property condition measure, this could change the relative performance of the models. For departments that need to minimize data collection and processing efforts, it is possible that random forest or gradient boosted models outperform logistic models when only a subset of the administrative data is available.

Future work must determine if randomization at the tract level rather than the parcel level can still yield data that can support a predictive model. Currently, we used an 80 percent sample of the data to train the models. This is the approach one would naturally take shortly after completing a comprehensive survey. If a jurisdiction only has funding for a 20 percent sample, it would randomly select tracts for the surveyors to walk rather than selecting individual parcels. Such a sampling strategy might give an advantage to some predictive methods over others.

References

Appel, S. U., Botti, D., Jamison, J., Plant, L., Shyr, J. Y., and Varshney, L. R. (2014). Predictive analytics can facilitate proactive property vacancy policies for cities. *Technological Forecasting and Social Change*, 89:161–173.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Campbell, J. Y., Giglio, S., and Pathak, P. (2011). Forced sales and house prices. *American Economic Review*, 101:2108 – 2132.

Coulton, C. J., Schramm, M., and Hirsch, A. (2010). REO and beyond: The aftermath of the foreclosure crisis in Cuyahoga County, Ohio. Technical report, Federal Reserve Banks of Boston and Cleveland and the Board of Governors.

Ford, F., Hirsh, A., Clover, K., Marks, J. A., Dubin, R., Schramm, M., Chan, T., Lalich, N., Loucky, A., and Cabrera, N. (2013). The role of investors in the one-to-three-family REO market: The case of Cleveland. Technical report, Joint Center for Housing Studies Harvard University.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Groves, J. R. and Rogers, W. H. (2011). Effectiveness of RCA institutions to limit local externalities: Using foreclosure data to test covenant effectiveness. *Land Economics*, 87(4):559 – 581.

Harding, J. P., Rosenblatt, E., and Yao, V. W. (2009). The contagion effect of foreclosed properties. *Journal of Urban Economics*, 66(3):164 – 178.

Hartley, D. (2014). The effect of foreclosures on nearby housing prices: Supply or disamenity?. *Regional Science and Urban Economics*, 49:108 – 117.

Hillier, A. E., Culhane, D. P., Smith, T. E., and Tomlin, C. D. (2003). Predicting housing abandonment with the Philadelphia neighborhood information system. *Journal of Urban Affairs*, 25(1):91–106.

Immergluck, D. and Smith, G. (2006). The external costs of foreclosure: The impact of single-family mortgage foreclosures on property values. *Housing Policy Debate*, 17(1):57 – 79.

Leonard, T. and Murdoch, J. C. (2009). The neighborhood effects of foreclosure. *Journal of Geographic Systems*, 11:317 – 332.

Mikelbank, B. A. (2008). Spatial analysis of the impact of vacant, abandoned and foreclosed properties. http://www.clevelandfed.org/Community_Development/publications/Spatial_Analysis_Impact_Vacant_Abandoned_Foreclosed_Properties.pdf.

Morckel, V. (2014). Predicting abandoned housing: does the operational definition of abandonment matter? *Community Development*, 45(2):121–133.

Naymik, M. (2015). Group begins counting every abandoned property in Cleveland by walking every street in the city. *Plain Dealer*, 15 June 2015. Available: http://www.cleveland.com/naymik/index.ssf/2015/06/group_begins_counting_every_ab.html [Last accessed: 20 December 2016].

Rogers, W. H. (2010). Declining foreclosure neighborhood effects over time. *Housing Policy Debate*, 20(4):687 – 706.

Rogers, W. H. and Winter, W. (2009). The impact of foreclosures on neighboring housing sales. *Journal of Real Estate Research*, 31(4):455 – 479.

Schuetz, J., Been, V., and Ellen, I. G. (2008). Neighborhood effects of concentrated mortgage foreclosures. *Journal of Housing Economics*, 17(4):306 – 319.

Thriving Communities Institute (2016). Cleveland by the numbers: 2015 cleveland property inventory. Technical report. http://www.wrlandconservancy.org/wp-content/uploads/2016/08/ClevelandPropertyInventory_issuu.pdf [Last accessed 22 December, 2016].

Whitaker, S. (2011). Foreclosure-related vacancy rates. *Federal Reserve Bank of Cleveland Economic Commentary*, (2011-12).

Table 1: Sources and descriptions of predictor variables.

Source	**Description**
Cleveland Department of Buildings and Housing	Number of days since open/vandalized property was boarded up
Cleveland Department of Buildings and Housing	Number of times a parcel has had a complaint since 2006
Cleveland Department of Buildings and Housing	Number of violations filed within 1 yr period before survey date
Cleveland Department of Buildings and Housing	Number of violations filed within 2 yr period before survey date
Cleveland Department of Buildings and Housing	Total number of arms length sales since 2006
Cleveland Department of Buildings and Housing	Total number of days since last arms length sales
Cuyahoga County Fiscal Officer (Auditor)	Certified total delinquent taxes owed
Cuyahoga County Fiscal Officer (Auditor)	Total taxes owed
Cuyahoga County Fiscal Officer (Auditor)	Total taxes paid
Cuyahoga County Fiscal Officer (Auditor)	Ratio of tax delinquency to grand total owed
Cuyahoga County Fiscal Officer (Auditor)	Ratio of total paid to grand total owed
Cuyahoga County Fiscal Officer (Auditor)	Grand total property tax balance 2014
Cuyahoga County Fiscal Officer (Recorder)	Total count of foreclosure filings 1 year before of survey date
Cuyahoga County Fiscal Officer (Recorder)	Total count of foreclosure filings 2 years before of survey date
Cuyahoga County Fiscal Officer (Recorder)	Poperty has an active foreclosure filing at survey date
Cuyahoga County Fiscal Officer (Recorder)	Total number of days since last transfer
Cuyahoga County Fiscal Officer (Recorder)	Total number of days since sheriff's deed transfer
Cuyahoga County Fiscal Officer (Recorder)	Total number of transfers since 2006
Cuyahoga County Fiscal Officer (Assessor)	Lot size (square feet)
Cuyahoga County Fiscal Officer (Assessor)	Property condition in tax records as of year 2013 or 2014
Cuyahoga County Fiscal Officer (Assessor)	Square footage, building(s)
Cuyahoga County Fiscal Officer (Assessor)	Tax assessment market value (dollars)
Cuyahoga County Fiscal Officer (Assessor)	Number of buildings on parcel
Cuyahoga County Fiscal Officer (Assessor)	Owner occupied indicator
Cuyahoga County Fiscal Officer (Assessor)	Architectural style: Bungalow, Colonial, Ranch, etc.
Cuyahoga County Fiscal Officer (Assessor)	Year built
Cuyahoga County Fiscal Officer (Assessor)	Property class
Cuyahoga County Fiscal Officer (Assessor)	zip code of parcel
Cuyahoga County Fiscal Officer (Assessor)	Number of units (single family, duplex, etc.)
Cuyahoga County Fiscal Officer (Assessor)	Number of stories
Cuyahoga County Fiscal Officer (Assessor)	Total assessed value of land
Cuyahoga County Fiscal Officer (Assessor)	Total market value of land
Cuyahoga County Fiscal Officer (Assessor)	Total market values per square feet of a parcel
US Postal Service	Counts of months of postal vacancy since April, 2010
Cuyahoga County Land Bank	Land Bank Acquired Indicator
Cuyahoga County Land Bank	Tax foreclosure Indicator
Cleveland Department of Public Safety	Property crimes, rate per 100,000 population, 2014
Cleveland Department of Public Safety	Burglaries, rate per 100,000 population, 2014
Cleveland Department of Public Safety	Part I crimes, rate per 100,000 population, 2014
Cleveland Department of Public Safety	Part II crimes, rate per 100,000 population, 2014
American Community Survey	Vacant housing units, percent, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	Median gross rent, number, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	Persons with bachelors degree or more, percent, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	Poverty rate, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	Median household income, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	White, percent, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	Black, percent, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	Asian/Pac Islander, percent, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	Hispanic, percent, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	Population aged 0-17, percent, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	Population aged 18-64, percent, 2012 5-yr est (ACS 2012 5-year)
American Community Survey	Population aged 65+, percent, 2012 5-yr est (ACS 2012 5-year)

Table 2: Logistic regression coefficients. Dependent variable is survey assessed vacancy status. N=78,553 (80-percent training sample from survey data). Data sources: Thriving Communities Institute, Cuyahoga County Fiscal Officer, Cuyahoga Land Bank, City of Cleveland, US Postal Service, and American Community Survey. Significance key: + 0.1, * 0.05, ** 0.01, *** 0.001.

Variables	Coefficients	Standard Error	Z
postal vacancy	0.025***	(0.001)	28.44
# days since complaint	−0.001***	(0.000)	−28.25
poor quality in tax record	1.349***	(0.095)	14.13
# days since foreclosure	−0.000***	(0.000)	−14.08
delq tax paid/taxes owed	1.078***	(0.101)	10.71
# complaints 1yr	0.106***	(0.011)	9.51
# days since sheriff deed transfer	−0.000***	(0.000)	−9.48
# days since transfer	0.005***	(0.001)	8.82
% black	0.005***	(0.001)	7.64
owner occupied	−0.265***	(0.038)	−7.07
total balance	0.000***	(0.000)	6.99
# days since arms length transfer	0.000***	(0.000)	6.80
delq tax balance 13	−0.000***	(0.000)	−6.69
foreclosure 2yr	0.519***	(0.079)	6.53
# complaints 2yr	0.039***	(0.007)	5.90
average quality in tax record	0.382***	(0.086)	4.46
# complaints	0.848***	(0.210)	4.03
land bank acquired	1.385***	(0.347)	4.00
year built	−0.003***	(0.001)	−3.65
total tax balance 14	0.000 **	(0.000)	3.23
foreclosure 1yr	−0.302 **	(0.106)	−2.85
burglary rate	0.000 **	(0.000)	2.76
% with bachelors	0.009 **	(0.003)	2.73
# days since board up	−0.000 **	(0.000)	−2.64
poverty rate	0.005*	(0.002)	2.31
# transfers since 2006	−0.344*	(0.157)	−2.18
price per square foot	−0.005*	(0.003)	−1.97
lot size (sqft)	−0.000+	(0.000)	−1.95
total market value	−0.000+	(0.000)	−1.80
% vacant on block (ACS)	0.004+	(0.002)	1.69
median rent	0.000	(0.000)	−1.53
square footage (house)	0.000	(0.000)	1.37
# stories	0.081	(0.062)	1.31
land bank tax foreclosure	−0.374	(0.295)	−1.27
property crime rate	0.000	(0.000)	1.26
total tax paid	0.000	(0.000)	−1.16
part one crimes	0.000	(0.000)	−1.13
active foreclosure	−0.134	(0.131)	−1.02
median hh income	0.000	(0.000)	0.77
# board up complaints	−1.725	(2.254)	−0.77
% taxes paid	−0.06	(0.100)	−0.61
# arms length transfers	0.192	(0.316)	0.61
land assessed value	0.002	(0.007)	0.32
land market value	−0.001	(0.002)	−0.32
# buildings on parcel	−0.025	(0.115)	−0.22
Constant	−10.408***	(2.328)	−4.47

Table 3: Logistic regression predicted vs. actual for test data, rounding probability at 0.5. Dependent variable is survey assessed vacancy status.

		Predicted		
		Not Vacant	Vacant	Total
Actual	Not Vacant	15,655	2,331	17,986
	Vacant	404	1,249	1,653
Total		16,059	3,580	19,639

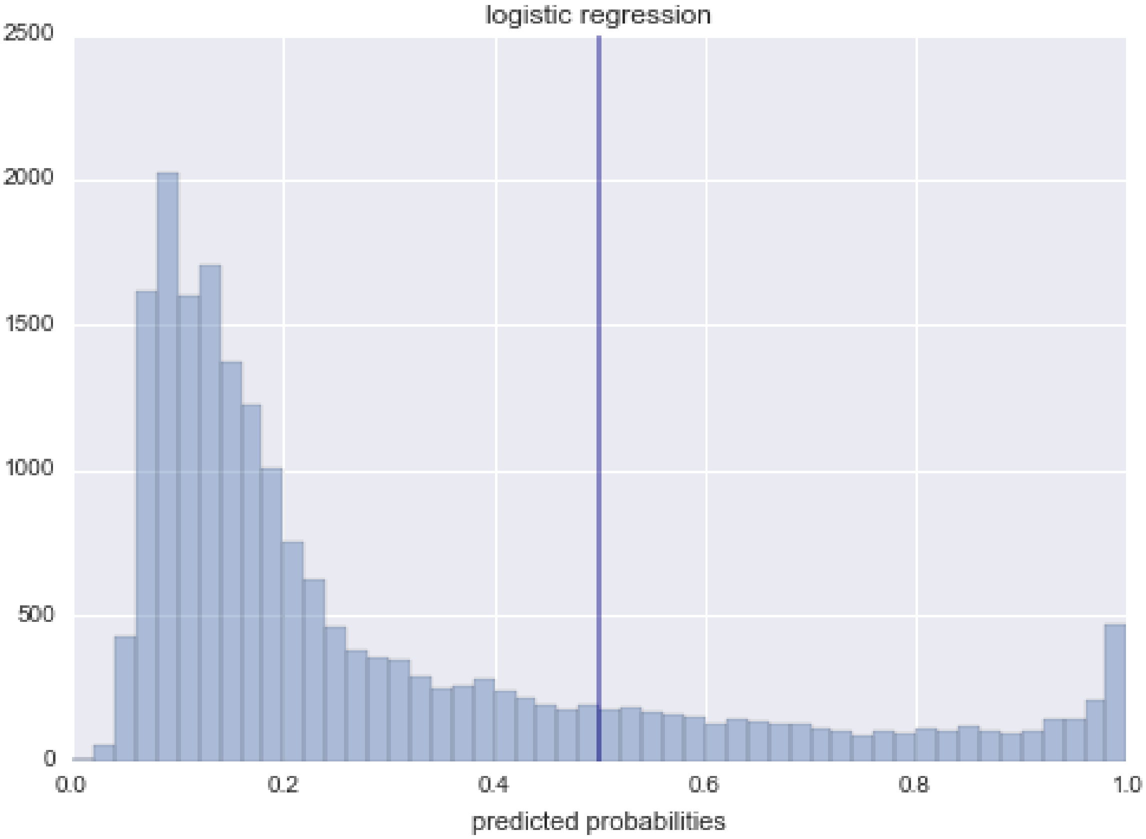

Figure 1: Logistic regression predicted probabilities of survey assessed vacancy status.

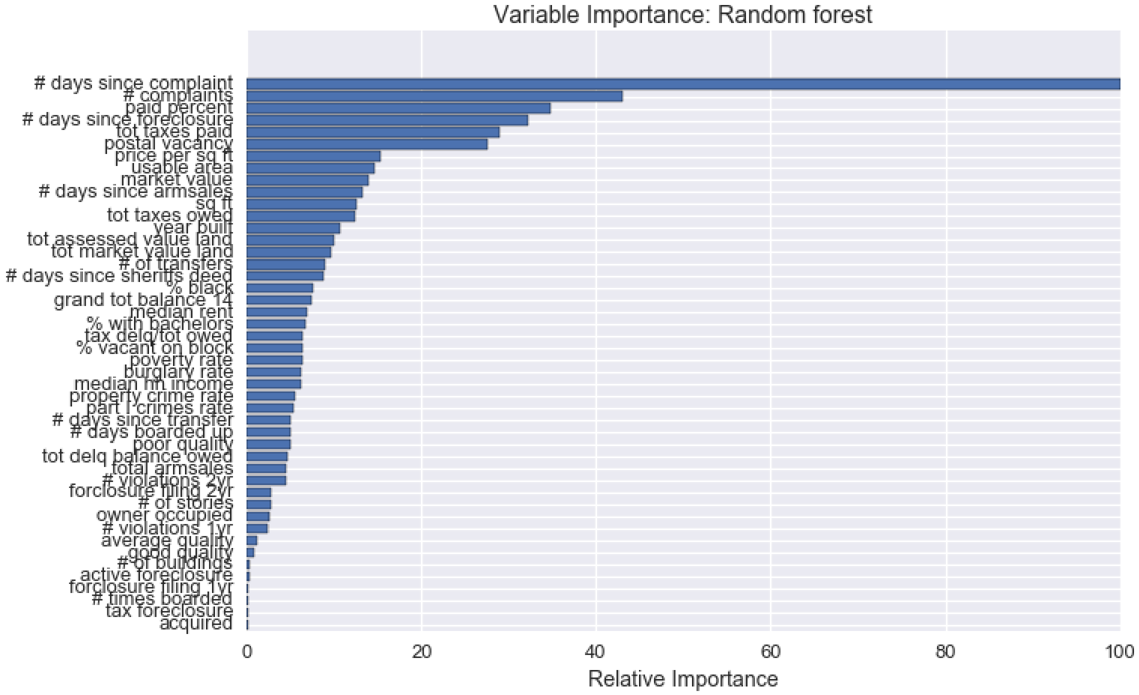

Figure 2: Random forest independent variables relative importance in predicting survey assessed vacancy status.

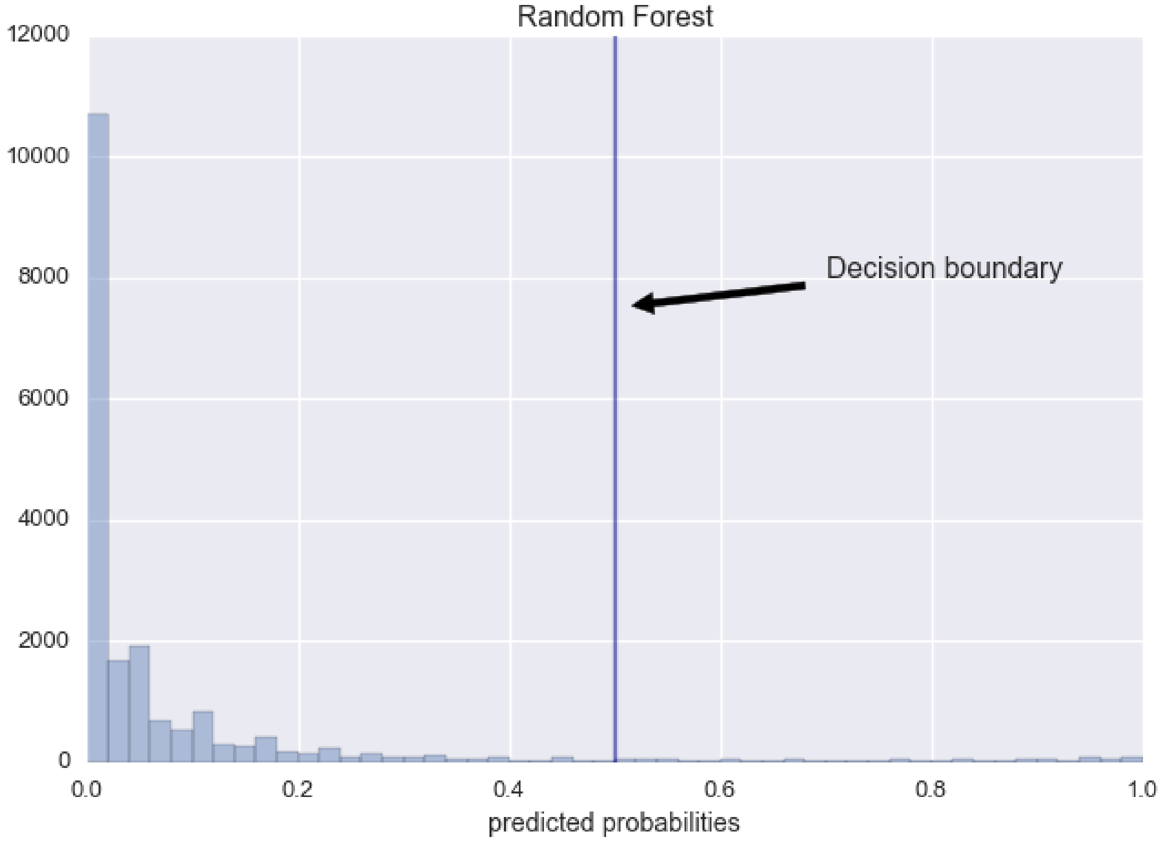

Figure 3: Random forest predicted probability of survey assessed vacancy.

Table 4: Random forest predicted vs. actual for test data, rounding probability at .5. Dependent variable is survey assessed vacancy status.

		Predicted		
		Not Vacant	Vacant	Total
Actual	Not Vacant	17,783	203	17,986
	Vacant	966	687	1,653
Total		18,749	890	19,639

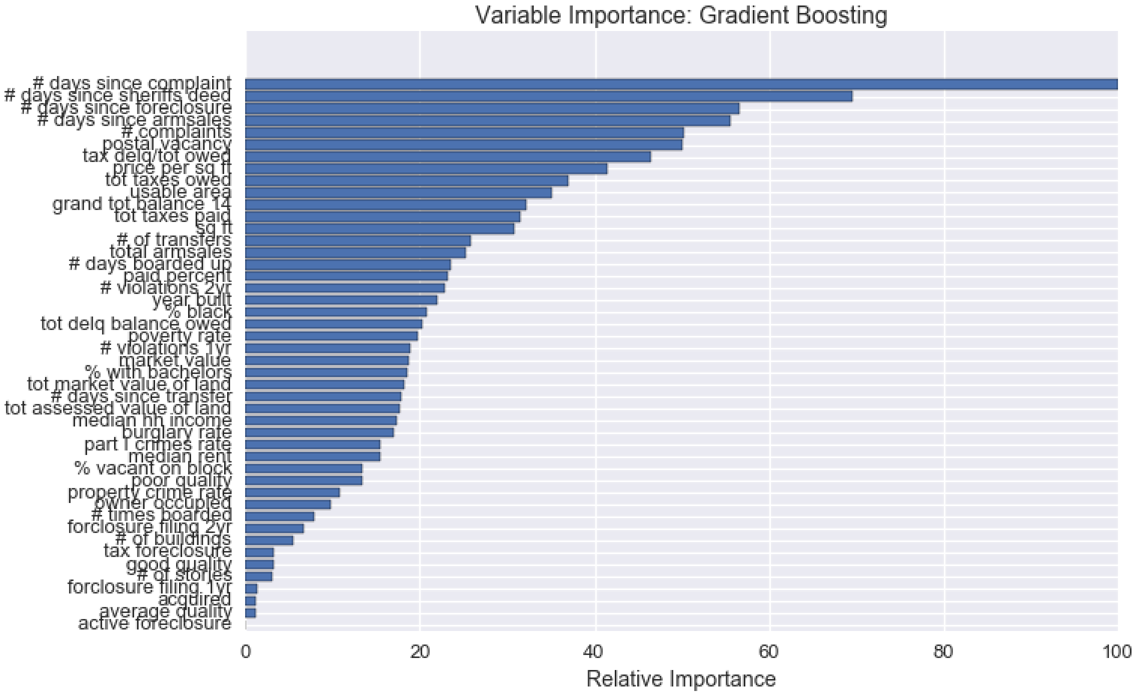

Figure 4: Gradient boosted independent variables relative importance in predicting survey assessed vacancy status.

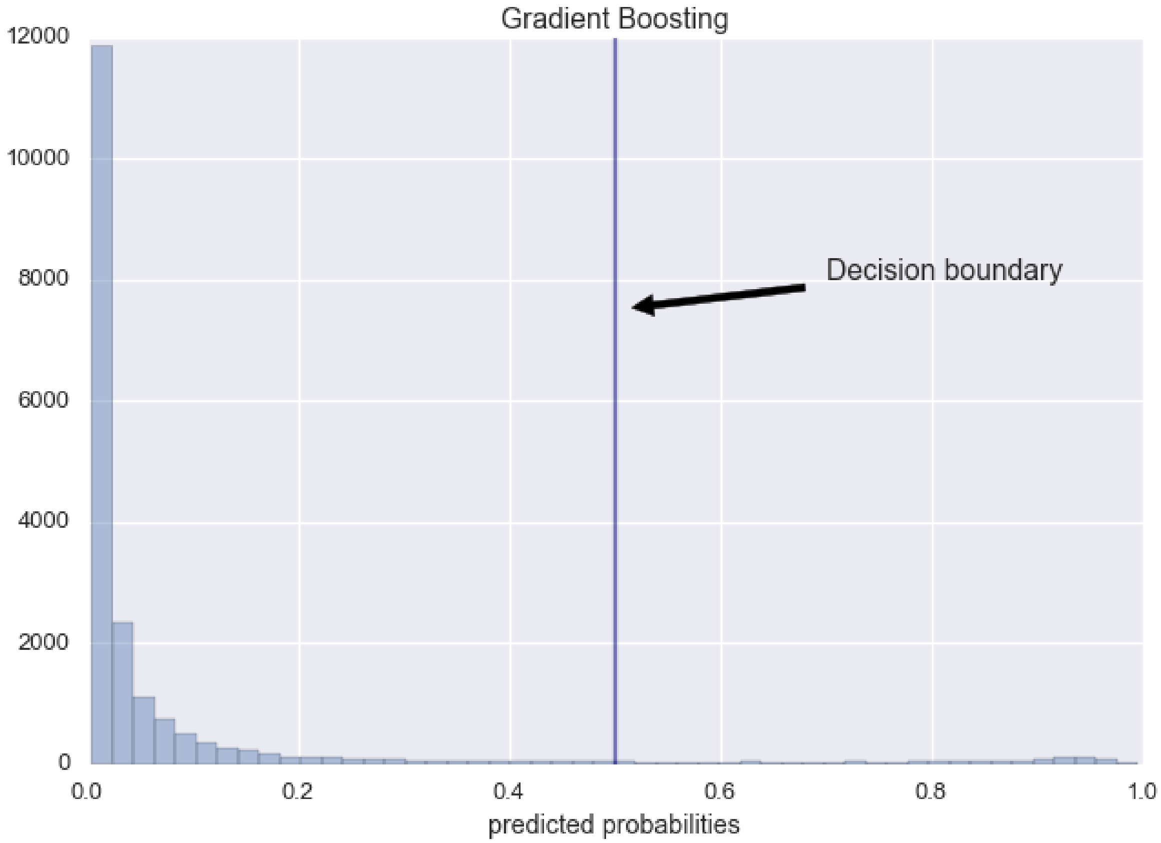

Figure 5: Gradient boosted predicted probability of survey assessed vacancy.

Table 5: Gradient boosted predicted vs. actual for test data, rounding probability at .5. Dependent variable is survey assessed vacancy status.

		Predicted		
		Not Vacant	Vacant	Total
Actual	Not Vacant	17,728	258	17,986
	Vacant	849	804	1,653
Total		18,577	1,062	19,639

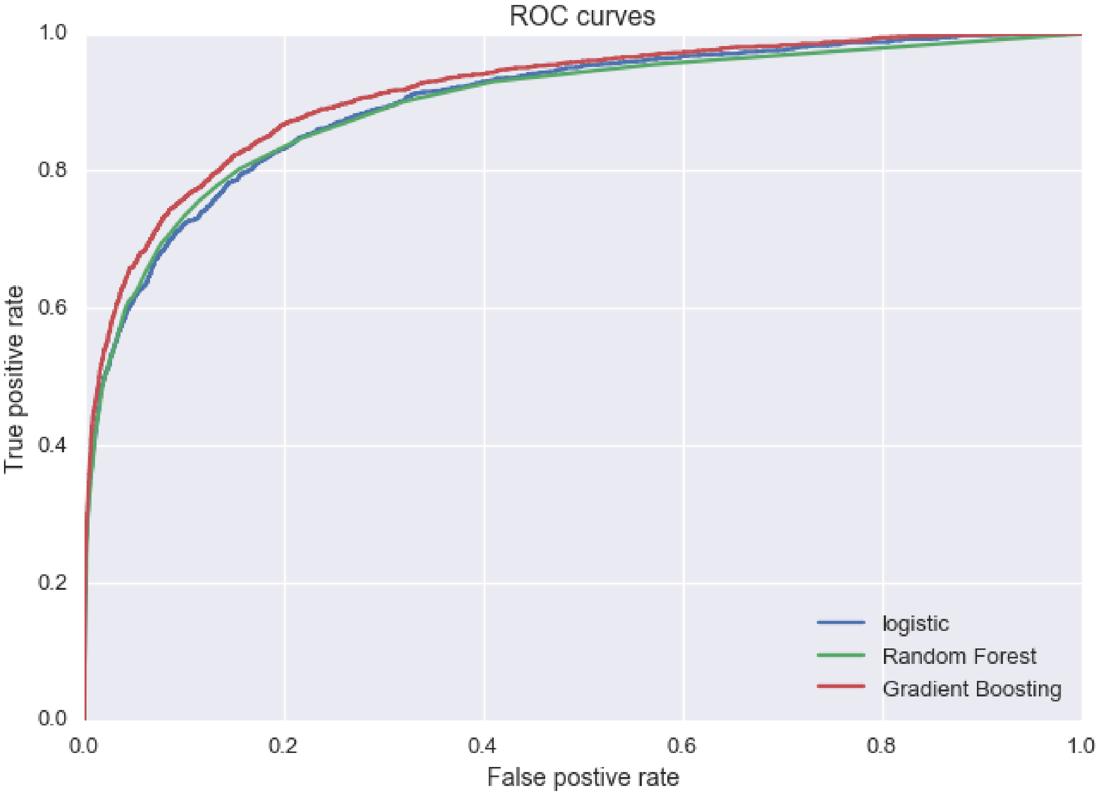

Figure 6: Receiver Operating Characteristic curves for models of survey assessed vacancy status.

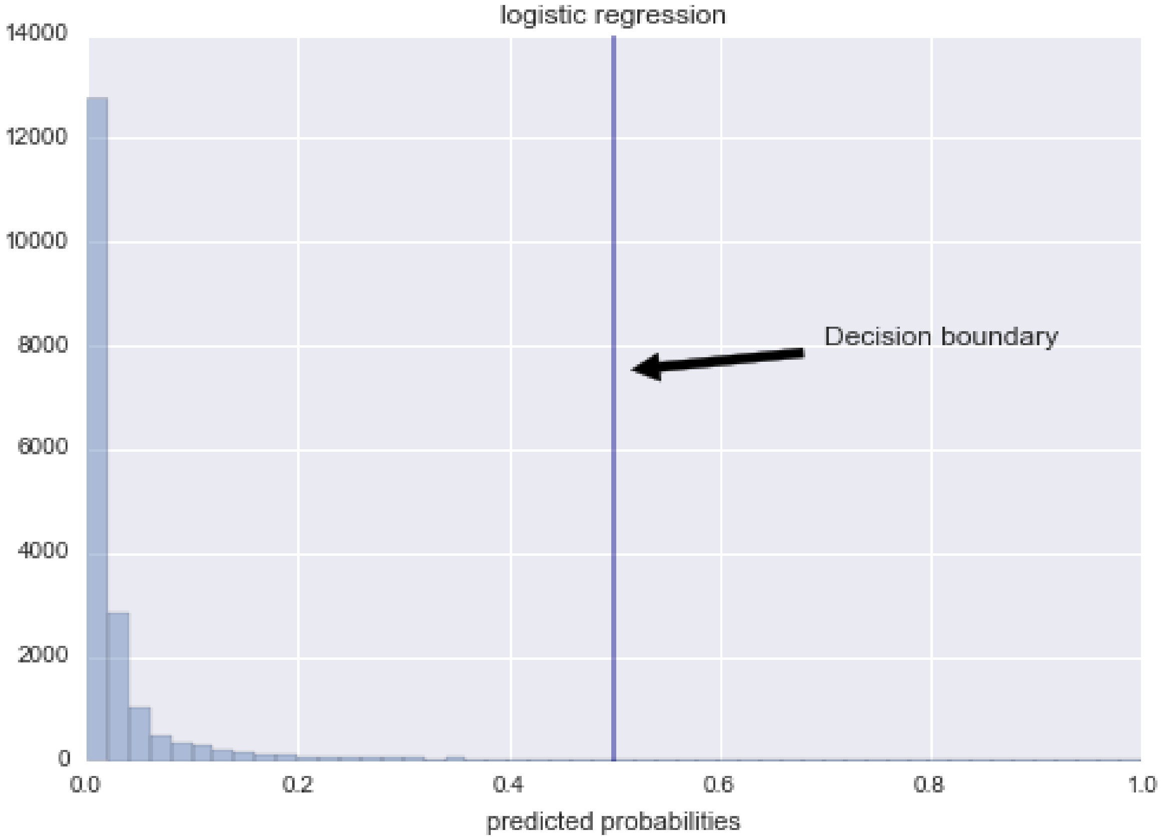

Figure 7: Logistic regression predicted probability of survey assessed condition letter grade "F".

Table 6: Logistic regression coefficients. Dependent variable is survey assessed property condition grade “F”. N=78,553 (80-percent training sample from survey data). Data sources: Thriving Communities Institute, Cuyahoga County Fiscal Officer, Cuyahoga Land Bank, City of Cleveland, US Postal Service, and American Community Survey. Significance key: + 0.1, * 0.05, ** 0.01, *** 0.001.

Variables	Coefficients	Standard Error	Z
# days since complaint	−0.001***	(0.000)	−14.24
% black	0.020***	(0.002)	9.09
# days since transfer	0.009***	(0.001)	7.91
delq tax paid/taxes owed	1.301***	(0.227)	5.74
poor quality in tax record	2.619***	(0.512)	5.12
% vacant on block (ACS)	0.026***	(0.006)	4.31
price per square foot	−0.040***	(0.011)	−3.79
# complaints 2yr	0.031**	(0.010)	3.21
# days since arms length transfer	0.000**	(0.000)	3.20
# complaints 1yr	0.049**	(0.015)	3.18
average quality in tax record	1.490**	(0.511)	2.92
part one crimes	−0.000**	(0.000)	−2.70
# days since foreclosure	−0.000**	(0.000)	−2.65
property crime rate	0.000*	(0.000)	2.53
total tax paid	−0.000*	(0.000)	−2.50
owner occupied	−0.199*	(0.087)	−2.30
poverty rate	0.011*	(0.005)	2.28
total balance	0.000*	(0.000)	2.24
land bank acquired	1.543*	(0.704)	2.19
year built	−0.003*	(0.001)	−2.08
postal vacancy	0.003*	(0.002)	2.03
% with bachelors	−0.020+	(0.010)	−1.90
square footage (house)	0.000+	(0.000)	1.77
delq tax balance 13	−0.000+	(0.000)	−1.66
lot size (sqft)	0.000	(0.000)	−1.54
foreclosure 2yr	−0.289	(0.203)	−1.43
# buildings on parcel	0.178	(0.156)	1.14
median hh income	0.000	(0.000)	−1.08
# transfers since 2006	0.353	(0.361)	0.98
land assessed value	−0.014	(0.015)	−0.94
land market value	0.005	(0.005)	0.94
land bank tax foreclosure	−0.452	(0.517)	−0.87
total tax balance 14	0.000	(0.000)	0.83
% taxes paid	0.167	(0.261)	0.64
burglary rate	0.000	(0.000)	−0.50
median rent	0.000	(0.000)	−0.49
# days since sheriff deed transfer	0.000	(0.000)	−0.48
# complaints	−0.102	(0.241)	−0.42
# stories	0.058	(0.154)	0.38
foreclosure 1yr	0.085	(0.236)	0.36
# board up complaints	0.615	(2.061)	0.30
# arms length transfers	0.163	(0.629)	0.26
total market value	0.000	(0.000)	−0.21
# days since board up	0.000	(0.000)	0.21
active foreclosure	0.016	(0.268)	0.06
Constant	−33.455***	(4.707)	−7.11

Table 7: Logistic regression predicted vs. actual for test data, rounding probability at 0.5. Dependent variable is an indicator of the survey assessed property condition letter grade “F”.

		Predicted		
		Not Grade F	Grade F	Total
Actual	Not Grade F	19,005	407	19,412
	Grade F	84	143	227
Total		19,089	550	19,639

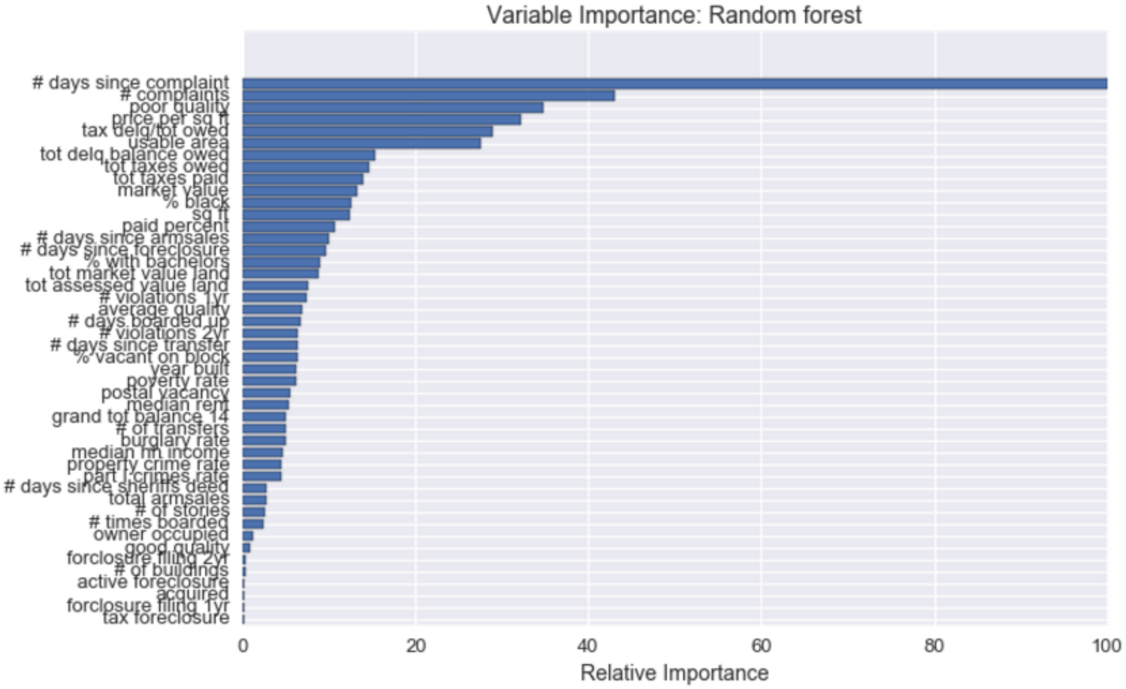

Figure 8: Random forest independent variables relative importance in predicting survey assessed condition letter grade “F”.

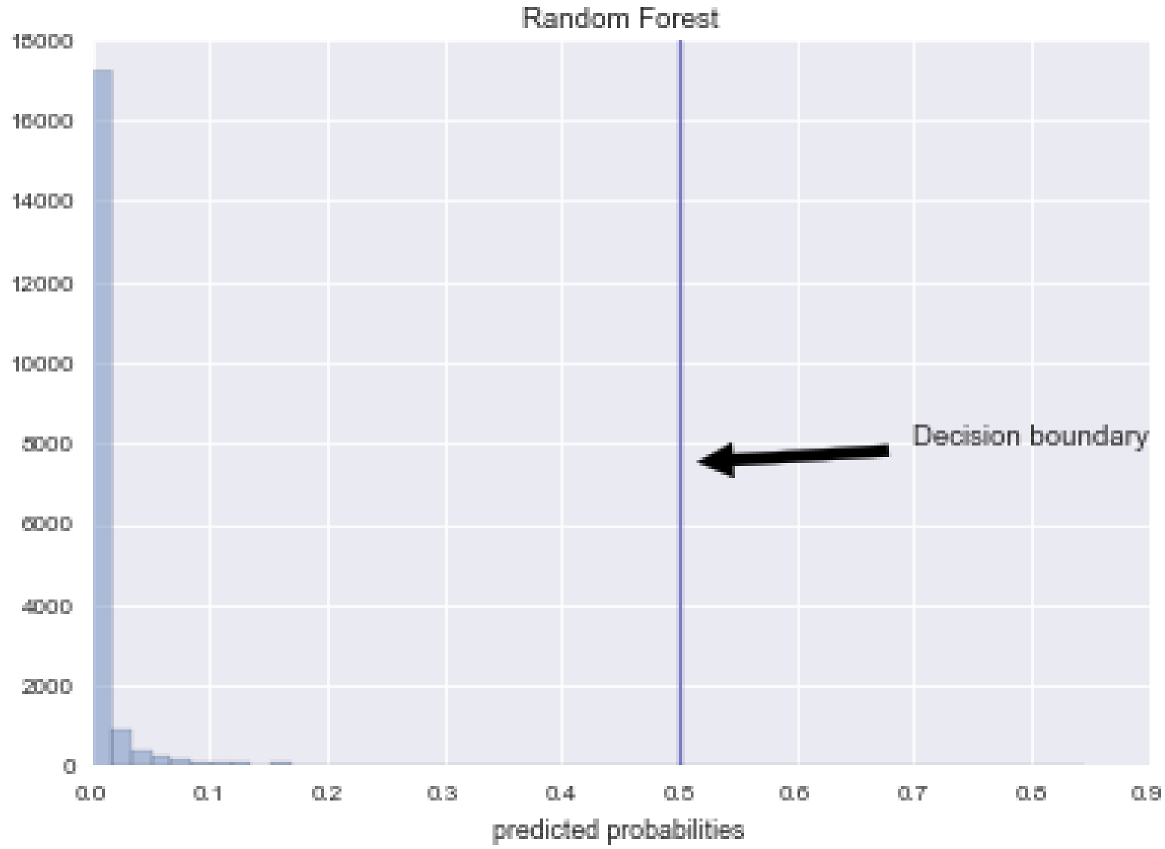

Figure 9: Random forest predicted probability of survey assessed condition letter grade “F”.

Table 8: Random forest predicted vs. actual for test data, rounding probability at .5. Dependent variable is an indicator of the survey assessed property condition letter grade "F".

		Predicted		
		Not Grade F	Grade F	Total
Actual	Not Grade F	19,391	21	19,412
	Grade F	208	19	227
Total		19,599	40	19,639

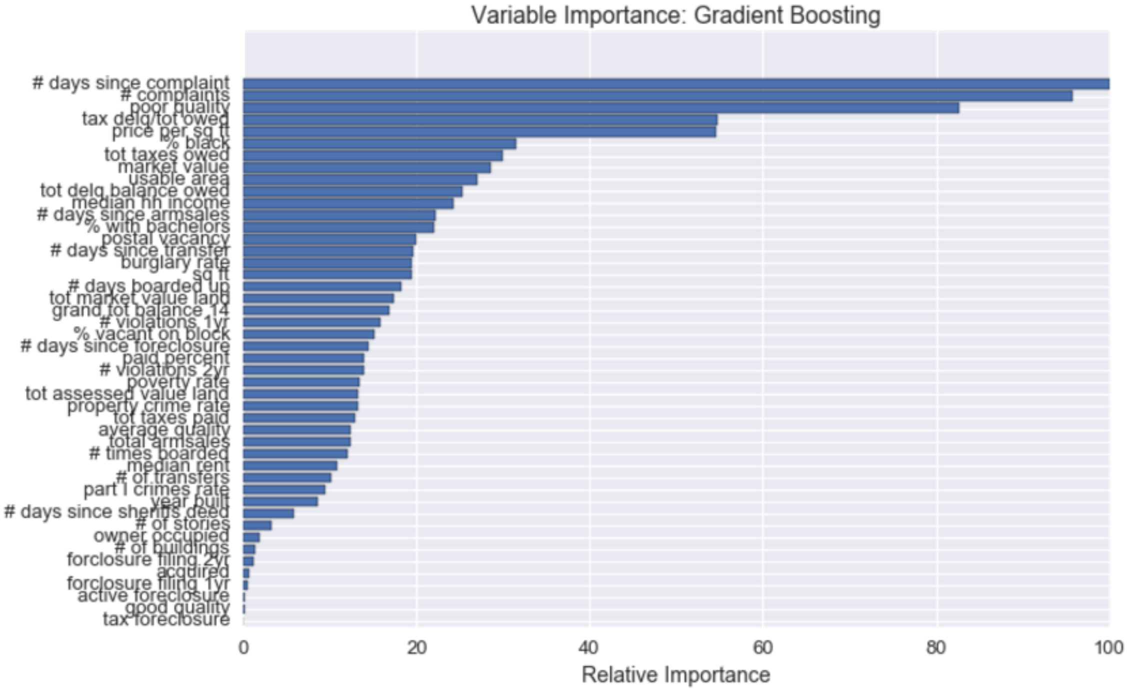

Figure 10: Gradient boosted independent variables relative importance in predicting survey condition letter grade "F".

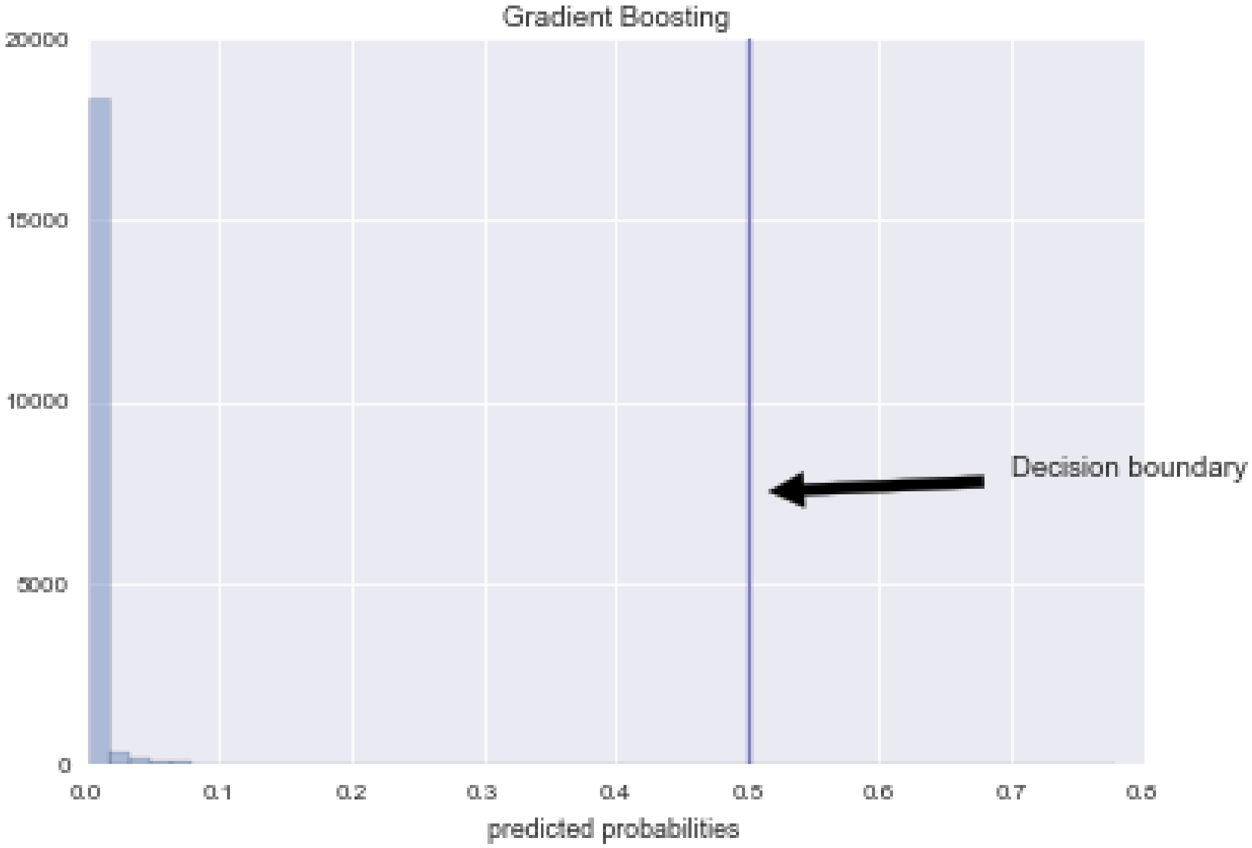

Figure 11: Gradient boosted predicted probability of survey assessed condition letter grade "F".

Table 9: gradient boosted predicted vs. actual for test data, rounding probability at .5. Dependent variable is an indicator of the survey assessed property condition letter grade "F".

		Predicted		
		Not Grade F	Grade F	Total
Actual	Not Grade F	19,396	16	19,412
	Grade F	215	12	227
Total		19,611	28	19,639

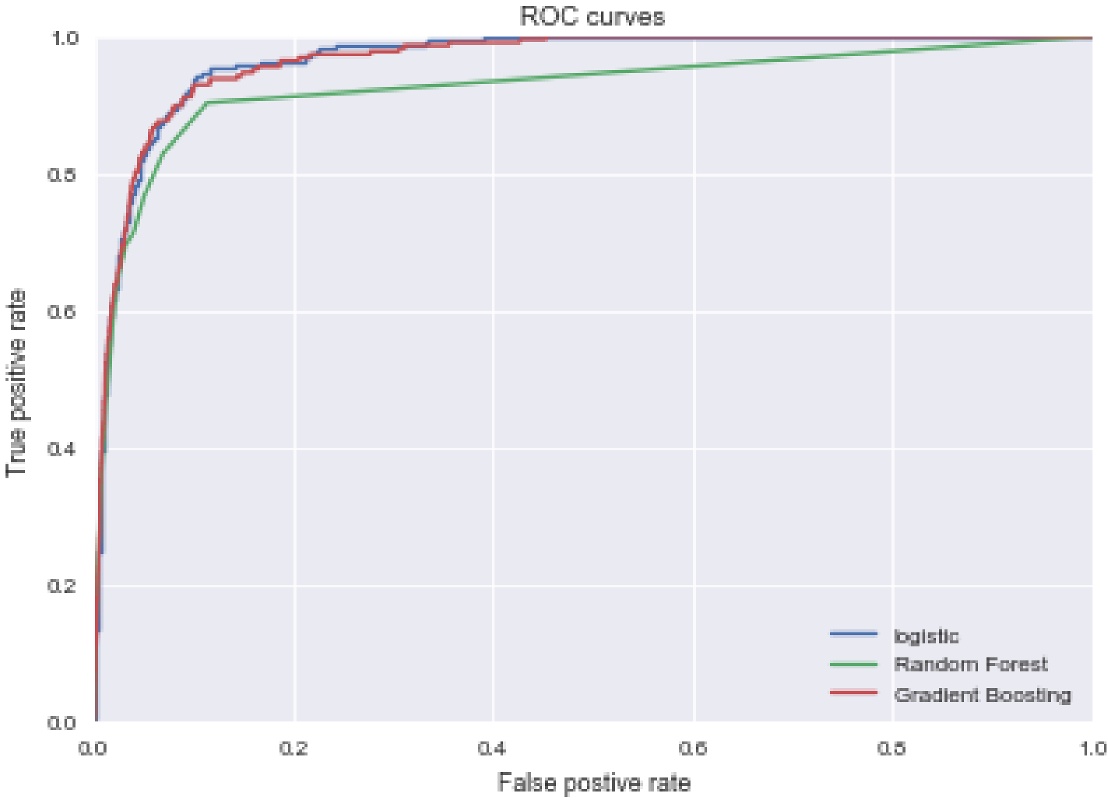

Figure 12: Receiver Operating Characteristic curves for models of survey assessed property condition letter grade “F”.