# Using Machine Learning to Target Assistance: Identifying Tenants at Risk of Landlord Harassment

Rebecca A. Johnson[1, 2], Teng Ye[3], Samantha Fu[4], Jerica Copeny[5], Bridgit Donnelly[6], Alex Freeman[7], Mirian Lima[8], Joe Walsh[8], and Rayid Ghani[8]

[1] *Department of Sociology, Princeton University*

[2] *Office of Population Research, Princeton University*

[3] *Department of Information, University of Michigan*

[4] *School of Public Policy, London School of Economics and Political Science*

[5] *Evansville Public Library*

[6] *Public Engagement Unit, City of New York (former)*

[7] *Mayor's Public Engagement Unit, City of New York*

[8] *Center for Data Science and Public Policy, University of Chicago*

March 19th, 2019

### Abstract

Rent-stabilization can combat housing insecurity. Yet landlords use legal loopholes–most notably, the ability to significantly raise the rent each time a tenant moves out until it exceeds the stabilization threshold–to convert these units to market rate. This loophole incentivizes landlords to illegally harass tenants through tactics like neglecting essential repairs or turning off heat to drive tenants out. In this project, we use large-scale administrative and API data to predict tenant harassment in New York City (NYC). We partner with an NYC agency that knocks on the doors of and offers assistance to at-risk tenants. Currently, there is wide variation in the likelihood that a particular knock helps the agency discover harassment. We use machine learning to predict tenant harassment in order to help the agency prioritize outreach to the highest-risk tenants. We discuss preliminary results that show how model-based targeting can mean the same quantity of resources helps more low-income renters.

# Contents

# 1 Introduction

Researchers have leveraged machine learning and big data in a variety of ways to improve our understanding of population processes and social policies (Zagheni et al., 2017; Bansak et al., 2018; Sargsyan et al., 2018; Helsby et al., 2018). One important use in social policy is *prioritization*. Social service agencies hope to target their interventions to certain populations–for instance, those who are the most in need of assistance or those for whom the assistance would provide the largest marginal benefit.[1] Once the agency has agreed

[1] While these two populations–potential recipients with the highest need and potential recipients who would derive the greatest marginal benefit from an intervention–may overlap, the two might also be distinct. For instance, students with the most severe learning disabilities may be the ones most in need of help to

upon a definition of *whom* they hope to prioritize for assistance, the agency then needs to identify the individuals who have the highest need according to this definition. Agencies have long-relied on professional assessments of this need–for instance, a disability examiner reviewing an applicant's case file to determine whether his or her disability is severe enough to prevent work (Bentez-Silva et al., 1999); machine learning can be used to *supplement* (but not replace) these professional assessments by giving those who make prioritization decisions more resources to assess need.

The present paper focuses on a timely application of machine learning for prioritization: prioritizing outreach for New York City rental tenants who are at higher risk of illegal harassment by landlords. We partner with the Mayor's Public Engagement Unit (PEU), from the City of New York, that sends employees to knock on doors of these at-risk tenants. PEU hopes to incorporate additional data in its outreach prioritization on where, if the tenant answers, he or she is most likely to be a tenant in need of help. By targeting the order of door knocks to the individuals who need the agency's help the most, PEU can increase the amount each knock contributes to reducing housing insecurity. While the results we present are subject to external validation by a field trial, the work highlights the role that data science can potentially play in bolstering tenant protection policies.

## 1.1 New York City: high levels of housing insecurity but pioneering access-to-counsel legislation to combat this insecurity

The present paper focuses on New York City, a context with a "perfect storm" of characteristics that, combined, make it a useful case study for two questions: how have cities supplemented federal housing assistance with more local policy solutions? And how can cities use machine learning to target these resources to the most at-risk households?

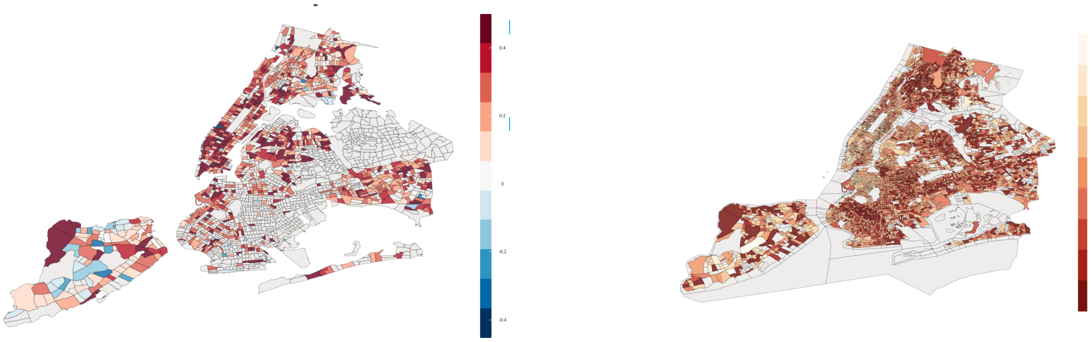
Feature one is a severely limited supply of Section 8 vouchers relative to need. As of March 2017, there were 146,000 households on the waiting list for Section 8 vouchers in New York City. Households who eventually receive a voucher wait an average of 8 years, and the waiting list was last opened up to new applicants over a decade ago in 2007 (Afford-

achieve grade-level performance but may also require substantially more teacher time to derive the same level of benefit as easier-to-educate students.

able Housing Online, 2018). While New York has a variety of city-level voucher programs and income-based housing to supplement the federal Section 8 program, many low-income households remain susceptible to housing instability.

The severe shortage of affordable housing options is both caused by and contributes to feature two: an increasing burden of rent relative to income among New York City households, contributing to housing insecurity. Figure 1 shows that many New York City households qualify as *cost-burdened*, which is generally defined as when a renter allocates more than 30-35% of his or her income to rent. For instance, in 2014, the median rent as a percentage of income for households at $\leq$ 200% of the federal poverty line was 47.8%, an increase of 4.2 percentage points from the 1999 rate of 43.6% (Mironova and Bach, 2018).

Figure 1: *Left panel*: the Figure plots the median rental price of units at the tract level using the American Community Survey (ACS) 5-year estimates between the years 2009 and 2016 ($\frac{rent_{2016}}{rent_{2009}} - 1$). The figure highlights that most tracts have experienced substantial increases in rental prices. *Right panel*: the Figure shows rent as a percentage of household income for the ACS 5-year estimate, with the lightest red showing rent as $<$ 10% of income and the dark red showing rent of $>$ 50% of income. The map shows many tracts where residents qualify as cost-burdened.



One important policy to combat rental cost burdens among the many low-income families who lack vouchers is rent stabilization, which limits the amount by which landlords can increase the rent each year. Tenants benefit from rent-stabilized units not only because the caps on sharp rent increases make housing more affordable long-term; they also are entitled to enhanced legal protections against eviction without cause. These protections mean tenants in rent-stabilized units have a better chance of remaining in a unit that is likely more affordable than market rate ones in the same neighborhood or building (NYU Furman Center, 2012). Rent regulations were originally enacted in New York City in 1943 as part of a wider program of federal wartime price controls. During World War II, a shortage of housing led the federal government to implement rent controls in major metropolitan areas. The federal regulations remained in place after the war. However, in order to create an incentive to continually produce housing, federal rent control laws exempted rental housing that was constructed after February 1, 1947. This created two levels of rental housing: "an older stock subject to regulation and a newer stock not subject to regulation"(Harris and Wagner, 2010).

In 1969, in response to increasing rents and decreasing vacancy rates, the City enacted a new Rent Stabilization Law. Regulation of older housing was preserved, and a new more moderate form of stabilization was imposed on residential buildings with six or more units built after 1946. The New York City Rent Guidelines Board (RGB) was created to design a plan for industry self-regulation, and was given the authority to determine yearly maximum increases of leases for rent-stabilized housing (Harris and Wagner, 2010). During the period from 1971-1974, the significant increase in rents led New York State to pass the Emergency Tenant Protection Act (ETPA) of 1974. This legislation transferred rental apartments in buildings with six units or more constructed between 1969 and 1974 under the newly created rent stabilization laws (Collins, 2018).

Unsurprisingly, many landlords dislike rent-stabilized units for these same reasons and seek to convert rent-stabilized units into market-rate ones. One of the main ways that landlords can move units out of rent-stabilization is through tenant turnover. Each time the landlord leases the same unit to a new resident, the landlord qualifies for a 'vacancy bonus' where he or she can increase the rent by a maximum of 20%. If landlords are able to accumulate these bonuses to increase the rent to $\geq$ \$2700, the unit is then permanently deregulated (Mironova and Bach, 2018). The majority of units that become deregulated become so through this pathway of "High Rent/Vacancy Deregulation"(Meehan, 2017).

Some landlords, armed with an incentive to increase tenant turnover, harass tenants in the hope that they "voluntarily" move out. This harassment, which is illegal,[2] can take the

[2] Mayor Michael Bloomberg passed *IN-627A* in 2008, a statute that made landlord harassment of tenants a Class C (most serious) violation of the city's housing code. The bill defines harassment as "the use of force or threats, repeated interruptions of essential services, the frequent filing of baseless court actions, and other tactics that substantially interfere with or disturb the comfort, repose, peace, or quiet, of any unit's lawful occupant." `https://www1.nyc.gov/office-of-the-mayor/news/087-08/mayor-bloomberg-signs-legislation-establishing-penalties-tenant-harassment`. The statute applies to all rental units in New York, but as the above section discusses, landlords might have stronger

form of long delays in making repairs, intrusive construction, making multiple cash offers for the tenant to move, cutting off heat or hot water, or illegally evicting the tenant (Meehan, 2017; Mironova and Bach, 2018; Barker et al., 2018). Tenants may be unaware that these actions are illegal.

If the landlord escalates the harassment and attempts to evict the tenants–for instance, for purported lease violations called holdovers or because the tenant justly withheld rent due to outstanding repairs–[3] tenants who decide to defend themselves against this eviction face difficult odds. Prior to the passage of recent policies, tenants faced striking imbalances in housing court–in 2013, only 1% of tenants in New York City housing court were represented by attorneys; 99% of landlords had attorney representation (NYC Office of Civil Justice, 2016). And as Desmond (2015) summarizes, randomized trials and quasi-experimental studies show that "tenants with legal counsel are much less likely to be evicted than their unrepresented counterparts, regardless of the merits of their case"(p. 11). He argues that: "establishing publicly funded legal services for low-income families in housing court could prevent the fallout from eviction, decrease homelessness, and help curb discrimination in the eviction decision"(p. 11).

The City of New York, recognizing the vulnerability of these tenants and wanting to decrease the disparities in legal representation, has launched several efforts to expand free legal representation for at-risk tenants. First, in February of 2015, Mayor Bill de Blasio's administration launched the Anti-Harassment Tenant Protection Program (AHTP), which allocated money from the city budget to legal services providers to enable the organizations to offer free legal counsel to more at-risk tenants.[4]

Yet the expanded legal rights are hollow if the low-income tenants who need assistance the most are unaware of their rights. So rather than waiting for tenants to "come to the city"–that is, waiting for tenants to proactively recognize a rights violation by their landlord, decide to report it, and navigate the correct bureaucracies in search of assistance–the Administration created the Tenant Support Unit (TSU) in July of 2015. De Blasio described

incentives to harass tenants living in rent-stabilized units.

[3] Holdover evictions are illegal in rent-stabilized units, but the tenant may not be aware of that protection.

[4] AHTP focuses on target neighborhoods that have been designated for re-zoning and are thus thought to have tenants at greatest risk of harassment or involuntary displacement.

the goals of TSU, housed within the Mayor's Public Engagement Unit (PEU), as one of bringing the city to the doors of tenants:

> When it comes to protecting tenants and affordable housing, we don't wait for a 311 call to come in. We have teams knocking on doors in fast-changing neighborhoods to solve problems then and there. This is a new strategy that's helping us keep New Yorkers in their homes and fight displacement before it happens (Mayor Bill de Blasio, 2016) (PressOffice, 2016).

In February of 2017, the city further bolstered TSU's ability to refer at-risk tenants to free legal counsel when it supplemented AHTP with a pioneering *Civil Gideon* statute that expanded the right to housing counsel to all low-income ($\leq$ 200% of the federal poverty line) New Yorkers;[5] in October of 2017, the Administration increased TSU's funding by $1 million to increase tenants' awareness of these rights.

This proactive outreach–send the city to the doors of tenants–is a promising policy lever to combat housing insecurity. Initial assessments show increases in legal representation among tenants taken to housing court by their landlords: in the targeted zip codes, the legal representation rates for tenants facing eviction jumped from 16.3% in the first quarter of 2016 to 48.0% in the first quarter of 2018, with the largest increases in the Bronx and Brooklyn (NYC Office of Civil Justice, 2017).[6] Residential evictions aided by city marshals have also decreased during this period, declining from 28,849 in 2013 to 21,074 in 2017, though this decrease likely stems from a combination of tenants' increased legal representation and changes in how the city enforces housing laws (NYC Office of Civil Justice, 2017). These changes suggest that low-income New Yorkers have benefited from increased access to legal services.

But sending teams to knock on doors to inform tenants of these services is much more time and resource-intensive than, for instance, staffing a call center that waits for tenants to proactively call. In the next section, we describe how PEU can use machine learning to supplement its existing targeting methods to better identify tenants most in need of services.

[5] The program is being phased in geographically and will be fully implemented citywide by 2022 (HumanResourcesAdministration, 2010)

[6] The 16% before the policy's implementation already constitutes a marked improvement over the low rates documented above, but also has two differences: it is measured in targeted zip codes rather than among all New Yorkers in housing court; it examines legal representation in eviction cases, rather than any housing court case type.

## 1.2 The present paper: using machine learning to better target this assistance

As Mayor de Blasio describes, TSU teams staffed with outreach workers called specialists knock "on doors in fast-changing neighborhoods to solve problems"(PressOffice, 2016). TSU Specialists try to help solve these problems by cutting through the bureaucracy and case-managing tenants with services to address their issue–neglected repairs; an impending eviction–through legal assistance and other free city services. And when knocking on doors, TSU faces a challenge common to many social service agencies: more need than time. While TSU hopes to eventually reach all buildings containing rent-stabilized units in target zip codes, the team wants to further prioritize the order of knocks to direct a valuable resource–specialists' time knocking on doors–to the tenants who are most likely to have problems that TSU can help with.

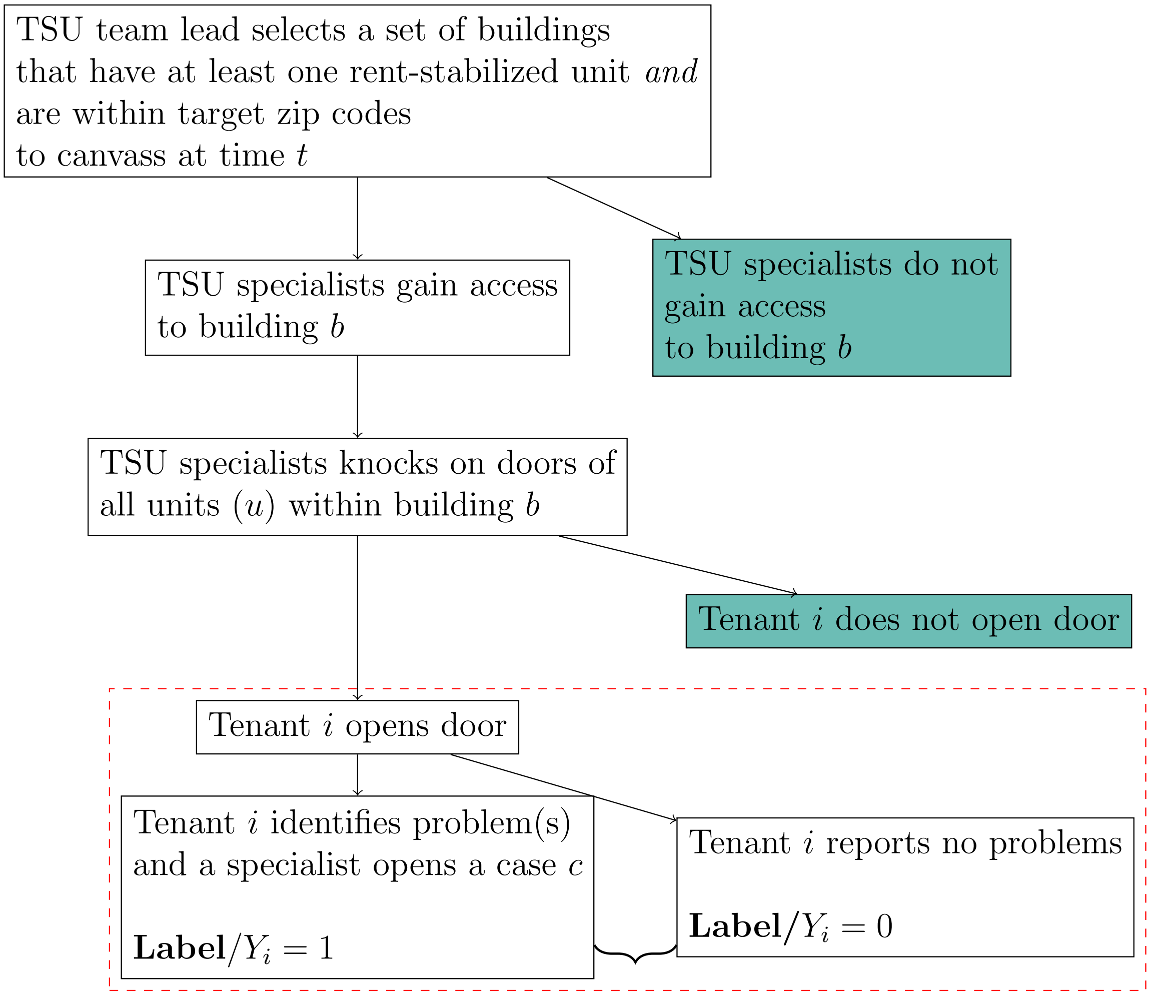
Figure 2 illustrates the process behind each door knock. Currently, TSU uses human judgment in the first step: team leads give specialists a list of doors to knock on using two criteria for prioritization.[7] First, the lists are composed of buildings in TSU's target zip codes, which began with 5 at the team's founding and expanded to the 20 depicted in Appendix Figure A2 by May of 2017, with most added by March of that year. Second, most team leads further filter the list to buildings that contain at least one rent-stabilized unit.[8] Finally, team leads use software to draw polygons ("cutting turf") around buildings using judgment about specialist capacity and high-risk areas. Team leads give specialists a list of buildings falling within that polygon to visit during the week; specialists pull up the list on `ipads` when they go canvassing.[9]

[7] New teams have been added during the course of TSU's operations; by the end of the analysis, there were two Manhattan teams (Inwood and East Harlem), three Brooklyn teams (East New York; Bushwick; Gowanus), two Queens teams (Flushing and Long Island City), and one team each in the Bronx and Staten Island. Team leads usually generate a month's worth of doors to knock on, though vary in their approach to the frequency with which they generate lists.

[8] Rent-stabilization status, both for TSU team leads and in our data, is defined at the building–does this building have any rent-stabilized units?—rather than the unit level. See discussion in Data Sources and (RentGuidelinesBoard, 2019)

[9] The remainder of specialist time is spent on case management and raising awareness at community events.

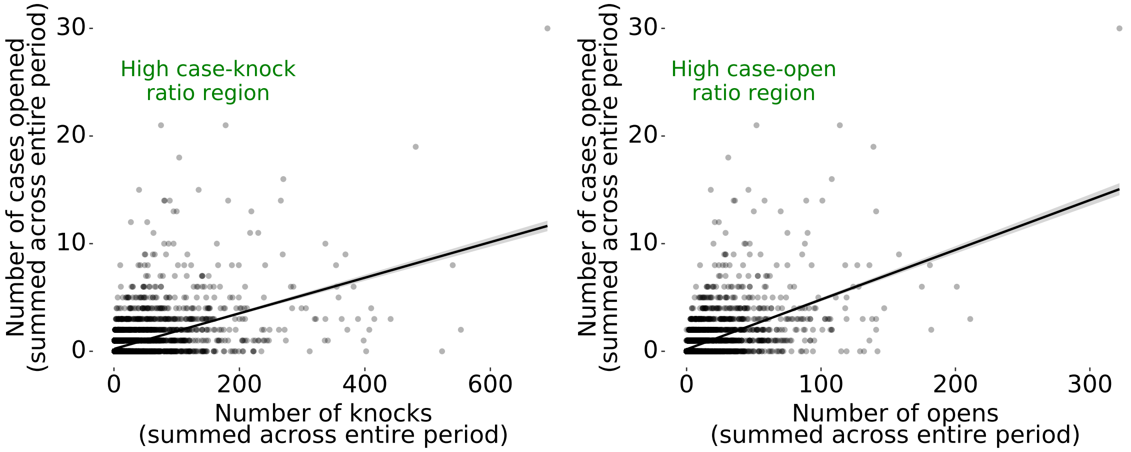
Figure 2: **Process for identifying tenants facing landlord harassment** Blue boxes indicate ways that buildings end up with a "missing label" (e.g., if TSU does not knock on a door, we do not know what the tenant *would* have reported had she or he answered; likewise with tenants who do not answer the door either because they are not at home or ignore the specialist's knock). The red dashed box indicates the outcome measure we focus on in the present analysis: conditional on a door knock *and* a door open in month $m$, do tenants at building $b$ in month $m$ report any cases of landlord harassment?


TSU team lead selects a set of buildings that have at least one rent-stabilized unit *and* are within target zip codes to canvass at time $t$
TSU specialists gain access to building $b$
TSU specialists do not gain access to building $b$
TSU specialists knocks on doors of all units ($u$) within building $b$
Tenant $i$ does not open door
Tenant $i$ opens door
Tenant $i$ identifies problem(s) and a specialist opens a case $c$
**Label**/$Y_i = 1$
Tenant $i$ reports no problems
**Label**/$Y_i = 0$

Two forms of aggregation in analytic sample:

1. Aggregate tenants $i$ to buildings $b$
2. Aggregate counts from specific time $t$ to months $m$

Figure 3: **Variation across buildings in the likelihood that TSU specialists find issues while knocking** Each dot represents one building, with the sample subsetted to the 6049 buildings with at least one knock during the canvassing period. *Left panel*: The x axis represents a building's cumulative count of knocks during the canvassing period and the y axis represents the cumulative count of cases opened at that building, and a linear fit depicting the average relationship across buildings. The figure shows substantial variability around the mean rate. *Right panel*: the x axis plots the cumulative count of opens at a particular building, with the y axis depicting the cumulative count of cases (same as left panel). Similarly, there is high variability around the mean rate. Each figure highlights the potential for TSU to target knocks to buildings in the upper left corner where either each knock or each door open is more likely to yield a case of landlord harassment.



Even employing these two filters, TSU team leads face a challenging task in deciding which buildings to include on the list they give specialists to knock on. There are roughly 147,000 units in rent-stabilized buildings in TSU's target zip codes. Faced with constraints on specialists time, how should team leads prioritize which buildings to include on the list for specialists to knock on?

The answer depends on what we assume about the *distribution of harassment risk* across buildings. If we assume that this risk is uniformly distributed–at each building, TSU specialists have an equal probability, upon knocking on the door, of speaking with a tenant who reports harassment that leads the specialist to open a case–then it might make sense for team leads to choose buildings randomly. Yet anecdotally, uniform harassment risk seems unlikely–for instance, private equity investors that purchase buildings hoping to convert the units to more lucrative condos/co-ops may have stronger incentives to harass tenants than other landlords (Barker et al., 2018). And empirically, Figure 3 shows substantial variation in risk. While there is an average of 8 cases for every 100 TSU knocks, and an average of 11 cases for every 100 tenants who open the door,[10] there is substantial variability around these averages. The upper left regions of each graph highlight buildings with a high risk of harassment–for instance, a building in the Bronx with 75 knocks, 52 answers, and 21 cases opened. Meanwhile, the lower right region shows buildings with a low risk of harassment–for instance, a building in Flushings, Queens with 523 knocks, 115 doors answered, and 0 cases opened.

The wide between-building variation in how likely TSU specialists are to find problems when a tenant answers the door suggests the potential to use data to find buildings in the promising upper left-hand corner of Figure 3. To do so, we use machine learning (Section 3) to model the step in the TSU process highlighted in red in Figure 2. When TSU knocks on the door and the tenant opens, does the tenant report a housing issue that falls within TSU's purview?[11]

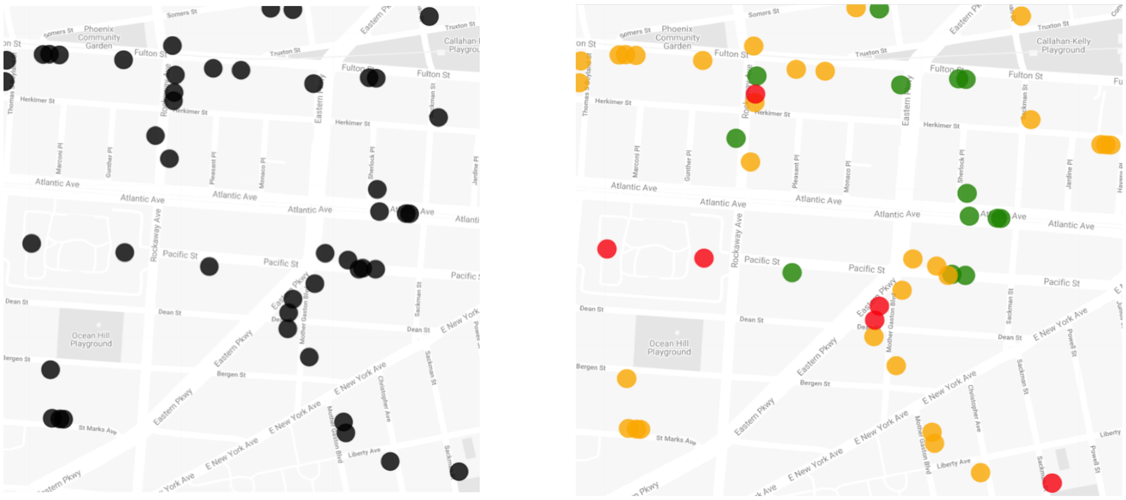
By using *historical* TSU data on where TSU has found cases in the past, and validating our predictions with a held-out test set,[12] the goal is to move from team leads looking at a map of buildings undifferentiated by risk levels to team leads looking at a map of buildings with information about the unequal harassment risk across buildings (Figure 4).

[10] As described later, this rate increases to closer to 25% once we aggregate to 'any case' over 'any open' at a particular month.

[11] As we discuss in more detail in Section 3, all data are aggregated to the building and month level. So more precisely, we model: when TSU knocks on at least one door in building $b$ in month $m$, and at least one tenant opens the door, what predicts the likelihood that a tenant reports a case to a TSU specialist?

[12] More precisely, we use multiple held-out test sets and average the results to find the best-performing model(s) across test sets.

Figure 4: **Modeling goal:** the map zooms in on blocks in one of TSU's target zip codes in Brooklyn and each dot represents one building. The goal of the methods we cover in Section 3 is to develop scores for each building so that team leads move from the current approach (*left hand side*) of unknown risk of buildings when they generate the list to having information about the building's risk on hand when generating the list (*right hand side, where red dots represent high-risk buildings*).



## 2 Data Description

Table 1 describes the data sources used for the "features" (covariates) and the "labels" (the outcome variable).

The first set of data ('Internal') come from records PEU keeps of TSU's canvassing activities, as well as the master roster of addresses that TSU draws from when the team leads "cut turf" and choose the buildings that specialists will canvass (the first step in Figure 2).

We augmented these internal data sources with external data of three types. First is more information on building characteristics, such as who owns the building and its property value (PLUTO). Second is more information on violations identified by city agencies other than TSU. For instance, the Department of Housing Preservation and Development (HPD), an agency that pre-dates TSU by over three decades (HPD was established in 1978), has data on serious violations that may identify problem buildings on which TSU can focus. Third, we use American Community Survey (ACS) 5-year estimates at the tract level to provide demographic information on the surrounding neighborhoods.

Table 1: **Data sources**. The table shows how we linked internal data sources that record knock and case records at buildings to a variety of external data sources indicating features like who owns a building (since there may be "high violation" landlords who own multiple buildings), neighborhood demographics, and violations identified by city agencies other than TSU.

| **Type** | **Name** | **Description** | **Approx. N** |
|---|---|---|---|
| Internal | Address | Addresses of all buildings containing rental units in NYC; roster of buildings that TSU then filters to ones based in target zip codes with at least one rent-stabilized unit | 1M buildings |
| Internal | Canvass | TSU knock attempts from April 2016 to March 2018 | 100K attempts |
| Internal | Case | TSU cases from inception (June 2015) to March 2018 | 8K distinct cases |
| Internal | Contact | Attributes of buildings where TSU opens cases | 8K distinct cases |
| Internal | Followups | TSU follow-up attempts for each case | 100K follow-ups |
| Internal | Issue | Issues related to each TSU case (cases can have multiple issues) | 30K issues |
| Internal | Zip_region | List of TSU target zip codes and dates added | 21 zip codes |
| External | ACS | 2013 to 2016 American Community Survey 5-year estimates at the tract level containing race, income, work hours, and other demographics | 2000 tracts x 4 years |
| External | PLUTO | Primary Land Use and Tax Lot Data indicating building ownership and renovation data | 1M buildings |
| External | HPD violations | Department of Housing Preservation and Development-confirmed violations | 4M violations |
| External | Housing Court litigation | Owner-directed litigation in Housing Court (e.g., legally compelling an owner to restore heat or hot water) | 150K cases |
| External | Subsidized Housing | Buildings contain units under subsidized housing programs | 16K buildings with any subsidized units |

# 3 Methods

## 3.1 Problem formulation: binary prediction of any case at a building in month $m$

We set up the prediction task as a binary classification problem focused on the red box in Figure 2. Conditional upon a TSU specialist knocking on any doors at building $b$ in month $m$, and conditional on at least one tenant opening a door, is there any reported landlord harassment/housing challenge? Features used to predict the outcome are measured at $m$,

$m-1, m-n$. All outcomes occur at month $m+1$ (in the next month). Put more formally, we define the label as follows, where $b$ indexes a building, $k$ indicates a knock at that building, $o$ indicates a door opened by a tenant to talk to an outreach specialist, and $c$ indicates the count of cases:

$$y_{bm} = \begin{cases} 0 & if\ k_{bm} \geq 1, o_{bm} \geq 1, \\ & c_{bm} = 0 \\ 1 & if\ k_{bm} \geq 1, o_{bm} \geq 1, \\ & c_{bm} \geq 1 \\ NA & \text{otherwise} \end{cases} \tag{1}$$

### 3.1.1 Additional label: cases/residential units > threshold

As we highlight in the results section, the binary label defined in Equation 1 flags larger buildings as higher risk. In particular, imagine two buildings:

1. **Building 1**: 1 case; 100 residential units in the building
2. **Building 2**: 1 case; 10 residential units in the building

The binary label, in predicting "any case", treats these outcomes as equivalent. However, we might think that building 2 has a higher harassment risk because there was a case despite there being fewer opportunities to find harassment among its tenants.

To give the models a better chance at flagging building 2 as higher risk than building 1, we supplemented the primary *any case* label with what we call the *threshold* label. For each building, we calculated the count of cases per number of residential units, a ratio that would flag building 2 as higher risk than building 1. For the results we present, we then coded a building as 1 = yes harassment if its ratio of cases per residential units was in the top 10% of a particular training set.[13] Equation 2 describes more formally, with $\tau$ indicating the threshold and $i_b$ indicating the # of residential units at building $b$:

[13] The quantiles differ across training sets because different training sets have different numbers of months represented so have slight differences in the distribution of risk.

$$y_{bm} = \begin{cases} 0 & if\ k_{bm} \geq 1, o_{bm} \geq 1, \\ & \frac{c_{bm}}{i_b} < \tau \\ 1 & if\ k_{bm} \geq 1, o_{bm} \geq 1, \\ & \frac{c_{bm}}{i_b} \geq \tau \\ NA & \text{otherwise} \end{cases} \tag{2}$$

## 3.2 Analytic sample for training set

A building needs to satisfy three criteria in a given month $m$ to be included in the training set:

1. *Contains at least one rent-stabilized unit:* as we discuss in Section 1, TSU targets tenants living in rent-stabilized units both because these tenants are at higher risk of harassment and because, when they *are* harassed, these tenants have more legal rights against eviction. Because rent-stabilization status is only available at the building level–does this building contain *any* rent-stabilized units–rather than at the unit level–is this particular unit rent-stabilized or not–we subset the training set to buildings with any rent-stabilized unit.

2. *Is located in a TSU target zip code:* similarly, TSU's outreach areas (Appendix Figure A2) are zip codes that contain areas that the city has both slated for re-zoning and where the city has bolstered legal aid to prevent displacement during this re-zoning.

3. *Has a non-missing label (so has at least one knock and at least one door open in that month):* Figure 2 highlights two ways in which a building can satisfy the two prior inclusion criteria–contains at least one-rent-stabilized unit; is located in a TSU target zip code–but has a missing label that prevents inclusion in the training set. These are:

    1. *TSU does not knock on any doors in the building in month m*

    2. *TSU knocks on at least one door in the building but no tenants open the door in month m*

The vast majority of missing labels come from source one (TSU does not knock at that building in a particular month) rather than source two (TSU knocks but no one opens). In the Discussion, we discuss how these labels may not be missing at random and steps to address that issue.

Buildings with missing labels *cannot* be used to estimate the risk of harassment, since there is an unknown relationship between the covariates and the outcomes. However, because we still have feature information from these buildings, they are included in the test set (so we can still generate risk predictions based on the parameter estimates/model objects from the training set estimation).

## 3.3 Features/covariates

Appendix Table A1 highlights the features used to predict each label (the *any case* label and the *threshold* label). In total, there were approximately 118 features after pre-processing that included generating dummy indicators for levels of categorical variables with > 10 buildings–for instance, landlords who owned more than 10 buildings received a binary indicator feature; those who owned less than 10 were grouped together under "Other."

## 3.4 Models

All models were estimated using `sklearn` in `Python` after standard pre-processing (feature imputation; normalization; generating dummy indicators for the most frequent categories in categorical features)(Pedregosa et al., 2011). Appendix A2 summarizes the hyperparameters we varied within each model class. We focused on three main classes of models:

1. **Tree-based methods:** Decision Tree (DT) and Random Forest (RF), each with varying depths (for DT and RF) and number of trees estimated (for RF)

2. **Logistic regression (LR):** the logistic regression models varied the penalty term (L1 regularization/lasso; L2 regularization/ridge) and the cost parameter (C), with a smaller C representing stronger regularization (more sparsity and/or smaller magnitude coefficients) and a larger C representing weaker regularization

3. **Gradient Boosting (GB), an ensemble classifier**: GB is an ensemble classifier that, rather than estimating a single deep tree with many splits (DT) or a simultaneously estimating a forest of trees seeded with different values (RF), follows a sequential procedure that (roughly):

    - Estimates a shallow tree
    - Takes the residuals from step one and upweights poorly predicted observations
    - Estimates another shallow tree on those re-weighted training observations
    - Repeats...

The approach we took to choosing classifiers and estimating models was relatively theory-agnostic. Rather than assuming that we can *a priori* identify the single best algorithm + hyperparameters or model group (e.g., Random Forest) that will outperform the others, we take a data-driven approach to choosing the best classifier for modeling building risk. That is, we fit a number of classifiers with a range of hyperparameters, store the complete set of results in a `SQL` database, and then perform analysis on the tables in this database to select the best-performing model(s) using the evaluation methodology we describe in the next section.

## 3.5 Model evaluation

Model evaluation was carried out using temporal cross-validation. The model was trained using data up to the first of every month and then tested against data for that month. For example, to make predictions for June 2017, we would train the model on data from March 1, 2016, through May 31, 2017, then use the model to make predictions for June. Appendix Figure A1 provides a general example of how splits were generated. More precisely, the process is as follows:

1. *Train on data up to month m*
2. *Test on data in month m+1*: for the results that follow, the test set month was the one immediately following the end of the training set month.

3. *Evaluate using the metrics we describe below*

4. *Repeat with the next temporal split*: for instance, in the above example, after training on 04.2016-05.2017, and predicting for 06.2017, the model would train on 04.2016-06.2017, and predict for 07.2017. The earliest split was in July of 2016, which means the model was trained on four months worth of data. The latest split was in December of 2017, which means the model was trained on twenty one months worth of data.

5. *Compare performance across splits and choose best-performing model as the one with the highest performance across all temporal splits*[14]

We used temporal cross-validation–rather than an a-temporal train/test split–for two reasons. First is that although the model aims to *predict* which buildings have the highest risk of having a case when specialists knock on the door, rather than *identify* the causal effect of specific covariates–for instance, a building's owner–on this risk, it is important that all covariates are temporally prior to the outcome. Second, temporal cross-validation most closely mirrors how the agency may use the model; the model's predictions will be used to generate a list based on data up to month *m* for specialists to use in month *m+1*, so testing how well the model performs on a variety of "m+1" months in different test sets is important.

### 3.5.1 Metrics

We use variations of standard resource-constrained machine learning metrics: precision and recall. In the present case, the constraint–$k$–represents the number of residential units that TSU can conduct outreach to in a given month. The variations we use are due to missing labels: we do not know whether a case would have been found if TSU had knocked on doors in that building in that month. When all observations have labels and we have a fixed $k$, maximizing precision is equivalent to maximizing recall. That guaranteed relationship does not necessarily hold when there are missing labels

*Precision in the top k* is the proportion of the k highest risk buildings (as identified by

[14] In the preliminary results that follow, we present model performance separately for each split date. Prior to the meeting, we will highlight models that perform well across an aggregation of split dates.

the model) that resulted in cases. It measures the efficiency of the model predictions. *Recall in the top k* is the proportion of buildings with problems that the model puts in the top k. It is a measure of coverage. There is typically a tradeoff between precision and recall. If TSU were to only knock on buildings where it was very confident a problem exists, it would have high precision (most buildings knocked have a case) but low recall (lots of buildings with problems would not be knocked). And if TSU were to knock on almost all buildings, regardless of its confidence, it would have low precision (lots of wasted knocks on buildings that do not have problems) but high recall (most buildings with problems receive knocks).

In addition to recall of true positive labels, we measure *recall of number of cases in the top k*, which refers to the proportion of cases that the model puts in the top k. This is a measure of coverage of cases. This helps us understand whether the model finds as many cases as possible. A model with high precision and recall may flag large buildings that do not have systemic problems while not capturing as many cases. For example, buildings with a large number of units would have a high probability to have at least one case but might have a low case per door open ratio.

*k* can vary from list to list. TSU is unit-constrained, as specialists can only knock on so many doors in a month, rather than building constrained. If there were a high-risk building with 10,000 units, TSU could spend all month knocking on doors there and nowhere else and still not reach all units in the building. We rank all buildings by risk and choose *k* using the following process:

- Examine how many residential units TSU specialists knocks on in a particular month (ranges from 3096 units earlier in the agency's outreach to 7374 units once they expanded their capacity)

- Assume that model-guided predictions will only be used for half those units $\frac{k}{2}$– for shorthand, we call this the *outreach constraint*

- Evaluate buildings ranked in order from highest to lowest predicted risk using that *outreach constraint*

In turn, there are two types of buildings, each of which is used for different aspects of the modeling process:

- *Buildings with an observed harassment label in month $m$:* these are buildings that TSU visits and either finds harassment $y_{bm} = 1$ or finds no harassment $y_{bm} = 0$. These are used for:
  - Training the model (since we know the relationship between the covariates and risk)
  - We predict harassment for these buildings ($\hat{y}_{bm}$) so they may appear below the *outreach constraint* if they are predicted high risk
  - We can use these buildings to evaluate the model's performance because we observe both $\hat{y}_{bm}$ and $y_{bm}$
- *Buildings with a missing harassment label in month $m$:* these are buildings that TSU does not visit. Because they do not visit the building and talk to the tenants, we do not know whether they would have found harassment or found no harassment.[15] These are used for:
  - We predict harassment for these buildings ($\hat{y}_{bm}$) so they may appear below the *capacity threshold* if they are predicted high risk

Put differently, the top-ranked buildings–buildings below the *outreach constraint*–can include buildings that lack labels. The metrics should not.[16] We calculate the precision and recall scores as follows:

1. Precision at top $k$ units:

$$\frac{\text{\# of true positive labels in top k list}}{\text{\# of labels in top k list}} \tag{3}$$

[15] Missing labels also stem from buildings they visit but where no tenants answer the door. This is significantly rarer.

[16] Assuming the missing labels are actually 0's implies TSU is nearly perfect at identifying problem buildings: it gets all the problem buildings (and some non-problem buildings) in the top $\frac{k}{2}$ buildings. Assuming the missing labels are actually 1's implies TSU performs poorly at identifying problem buildings: it gets all the non-problem buildings (and some problem buildings) in the top k. The truth is likely somewhere in the middle, but we dont have data to tell. However, as we discuss in the conclusion, we can use these assumptions to form bounds on our evaluation metrics.

Table 2: **Hypothetical example to illustrate how evaluation metrics would be calculated**

| Address ID | Risk Score | # of Units | Pred. Label | True Label | # of Cases |
|---|---|---|---|---|---|
| a5 | 0.81 | 153 | 1 | 1 | 34 |
| a7 | 0.68 | 23 | 1 | | |
| a8 | 0.62 | 77 | 1 | 1 | 12 |
| *Total units* | | *253* | | | |
| | | | | | |
| a4 | 0.48 | 300 | 0 | | |
| a3 | 0.46 | 100 | 0 | 0 | 0 |
| a1 | 0.4 | 25 | 0 | | |
| a6 | 0.3 | 83 | 0 | 1 | 14 |
| a9 | 0.2 | 110 | 0 | 0 | 0 |
| a2 | 0.11 | 60 | 0 | 1 | 9 |
| a10 | 0.1 | 65 | 0 | 0 | 0 |

2. Recall of any label at top $k$ units:

$$\frac{\text{\# of true positive labels in top k list}}{\text{\# of true positive labels in test set}} \tag{4}$$

3. Recall of total cases at top $k$ units:

$$\frac{\text{\# of cases identified in top k list}}{\text{\# of cases in test set}} \tag{5}$$

Table 2 illustrates these metrics with a hypothetical example. Suppose that TSU faces a 250-unit knock constraint,[17] we choose the three highest ranked buildings (with a total of 253 units) for knocks (predicting they have a problem). Then precision is 1 (two true positives out of two labels), recall of true labels is 0.5 (two true positives out of four positive labels) and recall of cases is 0.67 (forty-six cases out of sixty-nine cases ). We can also calculate recall of 0s at k, which in this case would be 0 (zero true negatives out of three true negative labels). Ideally, precision and recall are high and recall of 0s is low at the top of the list.

In the present results, we focus on one metric–precision at *k/outreach capacity.*

[17] This represents about half of their observed capacity in that month of $\sim 600$.

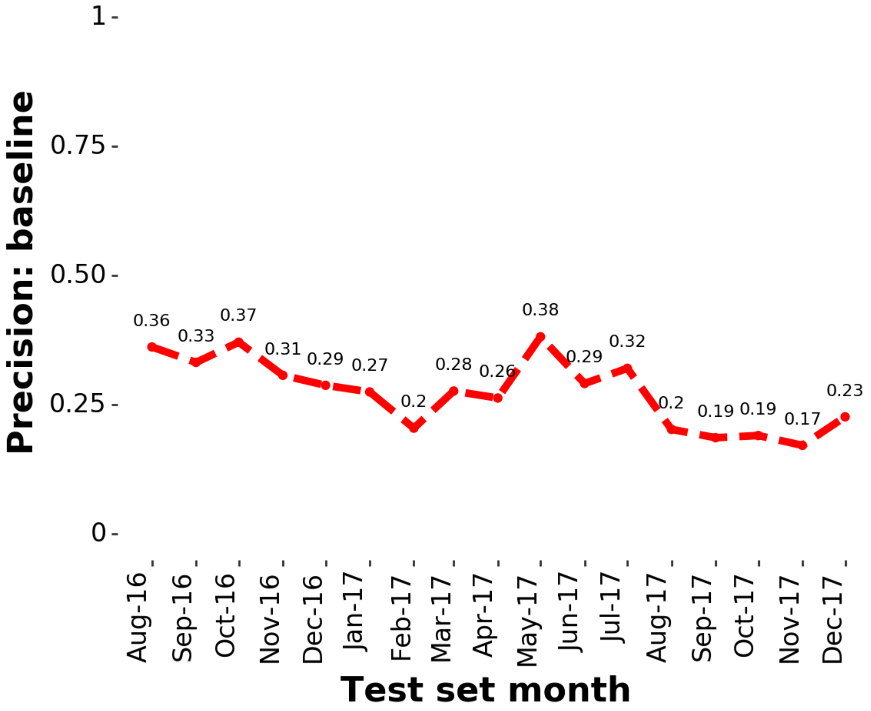
### 3.5.2 Baseline to which to compare performance

The above metrics give us measures of how well the model performs–for instance, precision at $k$ of 0.4. But in order to know if this performance is "good"–that is, if specialists' use of the model's predictions would result in meaningful improvements in finding cases relative to their current practice–we need to compare the model's performance to a reasonable baseline.

We use TSU's observed success rate at finding cases with each door open in a test set month as the baseline measure of performance. This roughly corresponds to the counterfactual–if TSU team leads continue what they're already doing and do not use the model, at what rate would they find cases? This baseline is *time-varying*–in different test set months, the TSU specialists have different ratios at which they find cases conditional on opens.[18] We see that on average, for every 4 tenants who open the door for a TSU specialist, 1 case is opened. In addition, we see substantial variation over time. In particular, TSU expanded its set of target zip codes in March of 2017, and following the expansion, we see a sharp increase in the ratio–perhaps due to specialists targeting the "low hanging fruit" of known problem buildings in the new zip codes. For each test set month, the goal is for the models' ratio to out-perform the baseline.

[18] The case:open ratio is substantially higher than the case:knock ratio depicted previously, and the former is the ratio we use as the baseline.

Figure 5: **TSU baseline against which we compare performance**



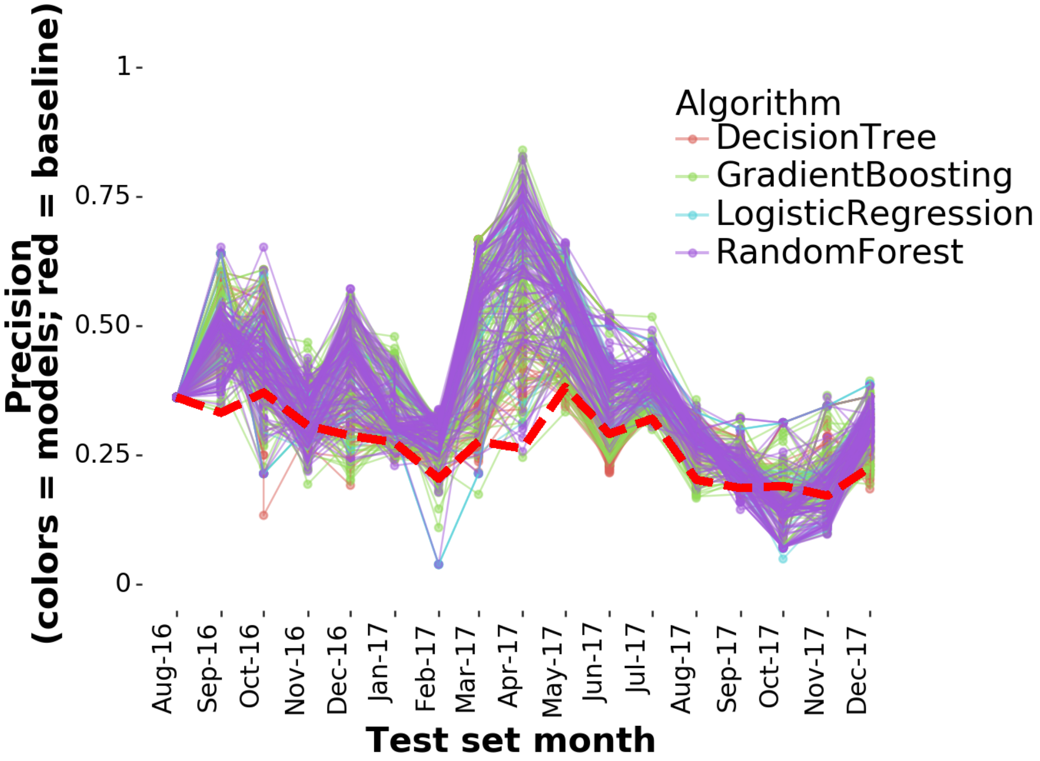
# 4 Findings

## 4.1 Performance across all models

How well do the models we discuss in Section 3 perform at the goal of finding more cases per each door opened? Figure 6 depicts the model performance for different dates at which a TSU team lead would receive a ranked list of model predictions. This ranked list includes both buildings with a non-missing harassment label (either yes or no) and buildings with a missing label (those that TSU did not knock on), with performance then evaluated only among the subset of buildings in the top k that have a non-missing label. The colored lines depict the *precision at k* of each model, discussed in Section 3.5.1. The red line of the Figure depicts the same baseline as Figure 5.

For the binary label, the Figure highlights that the best-performing model classes were Random Forest and Gradient Boosting. The models out-performed TSU's baseline by a median of 40% across all models and all test set months and a maximum out-performance of 85%. While subject to confirmation in a field trial, these initial results indicate that the majority of models outperform TSU's own internal process for finding cases with which to assist tenants.

Figure 6: **Model performance compared to TSU baseline on buildings that they canvassed** The Figure highlights the results of approximately $\sim 800$ models estimated across 4 model classes. The x axis depicts one test set month. The y axis depicts the model's performance. Each color represents a different algorithm type (e.g., logistic regression versus decision tree), with each separate line within the color then representing a different set of hyperparameters within that algorithm (e.g., within logistic regression, the type of penalty term and the strength of regularization). The red line shows the same baseline depicted in Figure 5 of TSU's current outreach practice. The Figure highlights that most models substantially out-performed TSU's outreach process in most months.



# Appendix

Table A1: **Features used in models for which we show preliminary results**. The table shows the raw feature names, which have the general syntax: *variable source_variable content_timeperiod* (if applicable). For instance, the first set of variables come from the PLUTO data we describe in Table 1; others come from internal data sources and are aggregated in different ways (e.g., knocks this month versus knocks ever)

| | |
|---|---|
| pluto_ownertype_static | internal_peu_team_queens_static |
| pluto_ownername_static | internal_peu_subteam_bk_bushwick_static |
| pluto_unitsres_static | internal_peu_team_manhattan_static |
| pluto_numbldgs_static | internal_borough_manhattan_static |
| pluto_numfloors_static | internal_peu_team_bk_eny_static |
| pluto_bldgclass_static | internal_peu_team_bronx_static |
| pluto_assesstot_static | internal_borough_queens_static |
| pluto_yearbuilt_static | internal_latitude_static |
| pluto_yearalter1_static | internal_knocks_count_this_month |
| pluto_yearalter2_static | internal_knocks_any_this_month |
| pluto_yearbuilt_years_static | internal_opens_count_this_month |
| pluto_yearalter1_years_static | internal_opens_any_this_month |
| pluto_yearalter2_years_static | internal_cases_opened_count_this_month |
| hpdviols_count_this_month | internal_cases_opened_any_this_month |
| hpdviols_count_classa_this_month | internal_issue_legal_cases_opened_count_this_m... |
| hpdviols_count_classb_this_month | internal_issue_repair_cases_opened_count_this_.... |
| hpdviols_count_classc_this_month | internal_issue_service_access_cases_opened_cou... |
| hpdviols_count_classi_this_month | internal_issue_other_cases_opened_count_this_m... |
| hpdviols_any_this_month | internal_cases_closed_count_this_month |
| hpdviols_any_classa_this_month | internal_cases_closed_any_this_month |
| hpdviols_any_classb_this_month | internal_nya_cases_closed_count_this_month |
| hpdviols_any_classc_this_month | subsidized_housing_flag_static |
| hpdviols_any_classi_this_month | internal_cases_opened_count_ever |
| hpdviols_count_ever | acs_tract_median_age_all |
| housinglitig_count_this_month | acs_tract_percent_white_alone |
| housinglitig_tenantaction_count_this_month | acs_tract_percent_black_or_african_american_alone |
| housinglitig_heatwater_count_this_month | acs_tract_percent_american_indian_and_alaska_n... |
| housinglitig_any_this_month | acs_tract_percent_asian_alone |
| housinglitig_tenantaction_any_this_month | acs_tract_percent_1200_am_to_459_am |
| housinglitig_heatwater_any_this_month | acs_tract_percent_500_am_to_529_am |
| housinglitig_count_ever | acs_tract_percent_530_am_to_559_am |
| internal_zip_static | acs_tract_percent_600_am_to_629_am |
| internal_peu_team_statenisland_static | acs_tract_percent_630_am_to_659_am |
| internal_peu_team_bk_bushwick_static | acs_tract_percent_700_am_to_729_am |
| internal_peu_subteam_bronx_static | acs_tract_percent_730_am_to_759_am |
| internal_peu_subteam_statenisland_static | acs_tract_percent_800_am_to_829_am |
| internal_peu_subteam_bk_gowanus_static | acs_tract_percent_830_am_to_859_am |
| internal_borough_bronx_static | acs_tract_percent_900_am_to_959_am |
| internal_peu_subteam_manhattan_eh_static | acs_tract_percent_1000_am_to_1059_am |
| internal_borough_statenisland_static | acs_tract_percent_1100_am_to_1159_am |
| internal_peu_subteam_bk_eny_static | acs_tract_percent_1200_pm_to_359_pm |
| internal_peu_subteam_queens_lic_static | acs_tract_percent_400_pm_to_1159_pm |
| internal_peu_subteam_queens_flushing_static | acs_tract_percent_living_in_household_with_sup... |
| internal_longitude_static | acs_tract_percent_less_than_10000 |
| internal_peu_subteam_manhattan_inwood_static | acs_tract_percent_10000_to_14999 |

Figure A1: **Conceptual diagram illustrating temporal cross-validation**. The $x$ axis represents time. The orange bar shows the time span of features in the training set, with the first gray box showing the label span (using features up to month $m$ to predict labels in month $m+1$). The navy bar shows the time span of features in the test set. It is important that the model is tested on the results of knocks that are *not in* the training set, to reduce the likelihood of over-fitting.

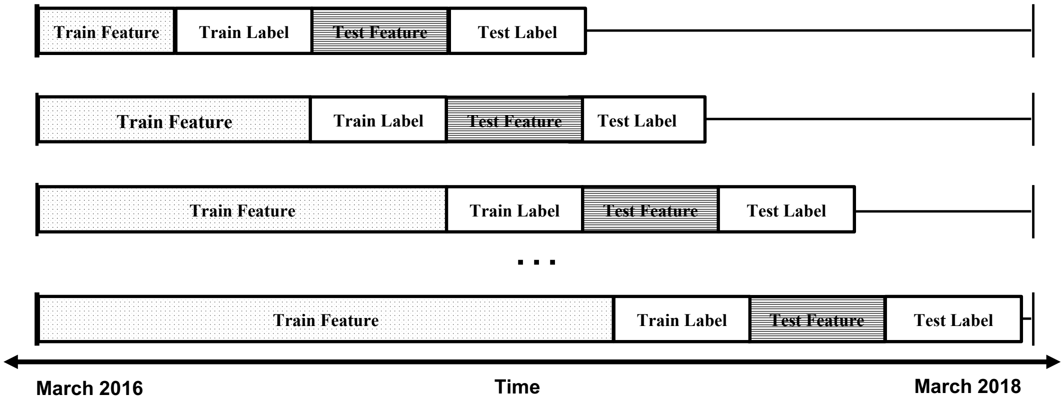


Table A2: Hyperparameters varied for each model class

| **Model class** | **Hyperparameters we varied** |
|---|---|
| Decision Tree (DT) | Tree depth; criterion for measuring split quality (Gini impurity v. information gain) |
| Random Forest (RF) | Tree depth; number of trees; # of features to consider for split |
| Logistic Regression (LR) | Type of regularization (L1 v L2 penalty); cost parameter |
| Gradient Boosting (GB) | Number of trees; tree depth; criterion for evaluating split |

# References

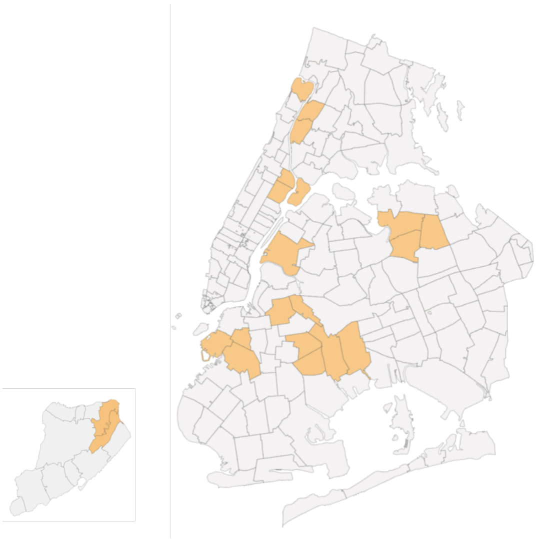
Affordable Housing Online (2018). New York Section 8 Waiting Lists Open Now.

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., and Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science*, 359(6373):325–329.

Barker, K., Silver-Greenberg, J., Cohen, S., and Ashford, G. (2018). The Eviction Machine Churning Through New York City. *The New York Times*.

Bentez-Silva, H., Buchinsky, M., Chan, H. M., Rust, J., and Sheidvasser, S. (1999). An empirical analysis of the social security disability application, appeal, and award process. *Labour Economics*, 6(2):147–178.

Figure A2: **TSU target zip codes**. *Right*: the map shows the full set of TSU target zip codes by the end of the modeling period (with knocks in new zip codes (how we define zip code expansions) occurring in June of 2016, March of 2017, and May of 2017). Target zip codes are chosen based on the city's AHTP target zip codes and which zip codes are the first to have the universal access to housing counsel legislation phased in. They represent outreach areas and TSU also takes cases from tenants outside the target zip codes who attend community events or otherwise hear about their services. *Left*: the table provides the borough and neighborhood names (informal) of the target zip codes where applicable.



| **Borough** | **Zip code** | **Neighborhood** |
|---|---|---|
| Bronx | 10452 | Jerome Avenue |
| Bronx | 10453 | Morris Heights |
| Brooklyn | 11217 | Gowanus |
| Brooklyn | 11215 | Gowanus |
| Brooklyn | 11207 | East New York |
| Brooklyn | 11208 | East New York |
| Brooklyn | 11231 | Gowanus |
| Brooklyn | 11212 | East New York |
| Brooklyn | 11233 | Bushwick |
| Brooklyn | 11206 | Bushwick |
| Brooklyn | 11237 | Bushwick |
| Manhattan | 10035 | East Harlem |
| Manhattan | 10034 | Inwood |
| Manhattan | 10029 | East Harlem |
| Queens | 11355 | Flushing |
| Queens | 11101 | Long Island City |
| Queens | 11354 | Flushing |
| Queens | 11358 | Flushing |
| Staten Island | 10304 | NA |
| Staten Island | 10301 | NA |

Collins, T. (2018). History: Rent regulation and the rgb.

Harris, W. and Wagner, C. (2010). Rent regulation: Beyond the rhetoric.

Helsby, J., Carton, S., Joseph, K., Mahmud, A., Park, Y., Navarrete, A., Ackermann, K., Walsh, J., Haynes, L., Cody, C., Patterson, M. E., and Ghani, R. (2018). Early Intervention Systems: Predicting Adverse Interactions Between Police and the Public. *Criminal Justice Policy Review*, 29(2):190–209.

HumanResourcesAdministration, N. Y. C. (2010). Legal services for tenants: Universal access to legal services.

Meehan, S. (2017). Zoning, tenant harassment, and the property contradiction: Lessons from the special clinton district. *CUNY Law Review*.

Mironova, O. and Bach, V. (2018). Tenants at the edge: Rising insecurity among renters in new york city. *Community Service Society of New York*.

NYC Office of Civil Justice (2016). Annual Report.

NYC Office of Civil Justice (2017). Annual Report.

NYU Furman Center (2012). Rent stabilization in new york city.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, . (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

PressOffice, N. Y. C. (2016). Protecting tenants and affordable housing: Mayor de blasio's tenant support unit helps 1,000 tenants fight harassment, secure repairs.

RentGuidelinesBoard, N. Y. C. (2019). Buildings that contain rent stabilized units.

Sargsyan, A., Karapetyan, A., Woon, W. L., and Alshamsi, A. (2018). No Fragile Family Left Behind - Targeted Indicators of Academic Performance. *arXiv:1806.02615 [cs]*. arXiv: 1806.02615.

Zagheni, E., Weber, I., and Gummadi, K. (2017). Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review*, 43(4):721–734.