Using Machine Learning to Help Vulnerable Tenants in New York City

Teng Ye
tengye@umich.edu
University of Michigan, Ann Arbor

Rebecca Johnson
raj2@princeton.edu
Princeton University

Samantha Fu
samanthacfu@gmail.com
London School of Economics and Political Science

Jerica Copeny
jerica.copeny@gmail.com
DePaul University

Bridgit Donnelly
bridgit.donnelly@gmail.com
Public Engagement Unit, City of New York

Alex Freeman
freemanal@hra.nyc.gov
Public Engagement Unit, City of New York

Mirian Lima
mspencerlima@gmail.com
University of Chicago

Joe Walsh
jtwalsh@protonmail.com
University of Chicago

Rayid Ghani
rayid@uchicago.edu
University of Chicago

ABSTRACT

To keep housing affordable, the City of New York has implemented rent-stabilization policies to restrict the rate at which the rent of certain units can be increased every year. However, some landlords of these rent-stabilized units try to illegally force their tenants out in order to circumvent rent-stabilization laws and greatly increase the rent they can charge. To identify and help tenants who are vulnerable to such landlord harassment, the New York City Public Engagement Unit (NYC PEU) conducts targeted outreach to tenants to inform them of their rights and to assist them with serious housing challenges. In this paper, we[1] collaborated with NYC PEU to develop machine learning models to better prioritize outreach and help to vulnerable tenants. Our best-performing model can potentially help TSU find 59% more buildings where tenants face landlord harassment than the current outreach method using the same resources. The results also highlight the factors that help predict the risk of experiencing tenant harassment, and provide a data-driven and comprehensive approach to improve the city's policy of proactive outreach to vulnerable tenants.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Machine learning approaches.*

KEYWORDS

Machine Learning; Social Good; Public Policy; Resource Allocation; Tenant Harassment

ACM Reference Format:
Teng Ye, Rebecca Johnson, Samantha Fu, Jerica Copeny, Bridgit Donnelly, Alex Freeman, Mirian Lima, Joe Walsh, and Rayid Ghani. 2019. Using Machine Learning to Help Vulnerable Tenants in New York City. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '19), July 3–5, 2019, Accra, Ghana.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3314344.3332484

[1]This work was done as part of the Data Science for Social Good Fellowship at the University of Chicago.

COMPASS '19, July 3–5, 2019, Accra, Ghana

ACM ISBN 978-1-4503-6714-1/19/07…$15.00
https://doi.org/10.1145/3314344.3332484

1 INTRODUCTION

In New York City (NYC), one of the world's most populous and dense cities, housing availability and affordability is a major concern for residents and city government. From 2009 to 2017, rents rose at twice the rate of wages [22], making it more difficult for New York City tenants to afford housing.

To help ensure the long-term existence of affordable housing, the New York State and New York City governments have implemented housing policies, such as rent stabilization, which restricts yearly rent increases, and a voucher program, which subsidizes rent for low-income households. Currently, the city has more than 1 million rent-stabilized housing units [6, 7].

However, the landlords of rent-stabilized units often want to "destabilize" these units [24] by forcing tenants out: that is, they want tenants to move out, voluntarily or through an eviction, to force a larger allowable rent increase that eventually places the unit beyond the purview of rent-stabilization policies. While the overall number of housing units has increased, the number of rent-controlled and rent-stabilized apartments in New York City has decreased by 146,902 units since 1991 [6, 7]. Some of this turnover is the result of landlord harassment, which can take the form of refusal to make essential repairs, illegally locking tenants out of units they have a right to live in, and other tactics aimed at inducing tenant turnover [19].

To help tenants vulnerable to these tactics, in 2015, New York City's Mayor's Office established a Tenant Support Unit (TSU),

a team of outreach specialists from the Mayor's Public Engagement Unit (PEU). TSU specialists proactively canvass door-to-door throughout the city and hold events with local community partners to find tenants in need of assistance with housing challenges. Once they identify a case of harassment or other serious housing challenges, specialists further case-manage tenants to help them access a range of city services, such as emergency repairs, vouchers and free legal assistance.

Canvassing to find tenants in need is a time-sensitive process– TSU's goal is to reach tenants before their problems progress to more serious cases of eviction or other forms of displacement. Currently, TSU identifies buildings that have rent stabilized units in 20 ZIP codes prioritized as part of anti-harassment protection legislation. To locate the buildings, TSU uses an internal address database and canvasses every apartment unit in these buildings. PEU team leads in each borough send specialists to each area until all apartment units have been attempted. Once an area is completed, canvassing begins again in an adjacent area. There are about 150,000 rental units in the 20 ZIP codes where funding is available for TSU to help tenants in need, but TSU specialists only have the resources to knock on an average of 5,000 units a month. Our work is focused on helping TSU prioritize locations where tenants face a high risk of harassment to help TSU specialists better plan their outreach and serve more tenants in need proactively.

In collaboration with TSU, the Data Science for Social Good Fellowship program at University of Chicago deployed machine learning models to help predict which buildings house tenants who face a high risk of harassment by their landlords. By analyzing historical outreach results and building and neighborhood characteristics, we showed that a Gradient Boosting model successfully outperformed the current outreach practice. Specifically, our model increased the precision relative to our baseline — the unit's expert-driven success rate — by 59%, helping TSU better allocate their outreach resources to people most in need and improving their efficiency at helping vulnerable tenants. In addition, we also provided analyses of feature importance, helping the team understand which attributes of buildings and neighborhoods contribute to the likelihood of rental tenant harassment.

In summary, this paper provides the following contributions:

(1) This paper contributes to the prediction of landlord harassment risk by deploying various machine learning models with a direct measure of landlord harassment and well-defined evaluation metrics.
(2) Our model shows significant improvement at identifying buildings at high harassment risk over TSU's current approach.
(3) In addition to yielding risk scores for tenant harassment, this paper also highlights features that can potentially be used as "early warning signs" of future harassment or proxy markers for the presence of harassment.

2 RELATED WORK

2.1 Housing Assistance for Low-income Renters

Social science research documents the negative consequences of housing instability and shows the mixed effects of rent-stabilization and other rental assistance policies on combating this instability. On one hand, such assistance may reduce homelessness [23] and rental burden, increasing financial security (such as to afford health care) among low-income households [16]. On the other hand, suppressing a unit's rent at a level below the rate it would receive on the open market can result in lower-quality housing [12, 21] and creates incentives for landlords to use legal loopholes, such as those that allow landlords to increase the rent each time a tenant moves out, to eventually convert the units to market rate [3].

Thus, policymakers face a dilemma: how can they use policies such as rent stabilization (which sets an upper limit on the rate at which the rent can be increased annually) to promote *access* to affordable housing, while also ensuring that tenants renting in these affordable units live in habitable conditions and do not face landlord harassment aimed at getting them to move out? The bulk of existing research focuses on the former part of the dilemma (the effect of policies on housing access). Less research investigates strategies to ameliorate potential byproducts of rent regulation policies.

Our work, by predicting where tenants in affordable units are likely to experience landlord harassment, fills an important gap. The Mayor's Office of Data Analytics (MODA) [17] also has studied data-driven protection from landlord harassment, and our project builds upon their efforts in several ways. First, through this paper we had a more direct measure of landlord harassment. While MODA [17] defined harassment using a proxy variable (i.e., the number of rent-stabilized units a building lost during a particular time period), TSU's historical canvass data allowed us to use harassment cases tenants reported during outreach. Second, we estimated many different models and evaluated model performance with well-defined metrics. Finally, the different machine learning models we estimated allow us to use significantly more features and to learn complex relationships — i.e., both linear and nonlinear relationships— between these features and a building's observed harassment risk.

2.2 Machine learning for Social Good

In recent years, machine learning has been widely applied to problems of social good and to inform public policies. For example, it has been introduced to forecast issues of criminal justice [4], detect online rumors on social media [25], identify political bias in text [11], map wealth and poverty in given areas [5, 10] and even facilitate medical diagnoses [14].

In particular, government agencies have used machine learning to inform better allocation of resources. For example, random forest and logistic regression have been used to identify students at risk of not graduating, so that school districts can prioritize their limited intervention resources to help these students [15]. Machine learning models can also help government inspectors prioritize inspections to high-risk units. These efforts include using Yelp reviews to help a government agency target hygiene inspections [9, 13] and predicting which buildings face a high fire risk to help the New York City Fire Department narrow its inspection focus [2, 18].

However, far less work has been done to explore how machine learning can inform housing policies, except for making policy recommendations to reduce home abandonment in Mexico [1] and detecting home locations by real life photos on social media [26]

Table 1: Data sources summary

Dataset	Records #	Time Window
(Internal) Knock attempts	100K	2016.4 - 2018.2
(Internal) Case records	8K	2015.6 - 2018.2
(Internal) Case issues	30K	2015.6 - 2018.2
(Internal) Building address	1M	N/A
(External) ACS (tract-level)	2000	2013 to 2016
(External) PLUTO buildings	1M	till 2018.1
(External) HPD violations	4M	till 2018.6
(External) Hous. Court litigation	150K	till 2018.6
(External) Subsidized housing	16K	till 2016

or by tweets [20], as well as MODA's study mentioned in the previous section [17]. In this paper, we highlight a new application by deploying machine learning methods to predict which buildings house tenant(s) facing a high risk of harassment by their landlords.

3 PROBLEM FORMULATION

We formulate the tenant harassment risk prediction as a binary classification problem. For each building, our model produces a risk score for whether there will be at least one harassment case identified if the TSU specialists canvass the building in the next month. Our model answers the question: *Will there be any cases of harassment in a given building in the next month?*

This formulation leads to two further decisions: 1) what time horizon to predict for (e.g., a harassment case within the next week, next month, or next year) and 2) the unit of prediction (e.g., modeling which residential unit faces a high harassment risk versus modeling which buildings contain tenants who face a high harassment risk). Both of these questions need to be answered reflecting the operational and policy constraints of our partner, the Tenant Support Unit at NYC.

For 1), we use a month as the time horizon for our prediction because TSU specialists typically plan their work at the beginning of each month. Monthly prediction thus matches their outreach planning process.

For 2), we focus on each building rather than each tenant for two reasons. First, TSU conducts a building-level outreach process. Out of concern for equity among tenants, TSU specialists believe they should knock on every single unit in a building once they enter. Second, the majority of information in both TSU internal databases and public available datasets describes buildings rather than units. Therefore, it's both more feasible and more important to know the building-level risk of harassment.

4 DATA

To explore variables that can help us predict which buildings may be at risk of harassment, we combined data from multiple sources. Table 1 summarizes the information presented in the data. Details are described in the following sections.

4.1 TSU (Internal) Data

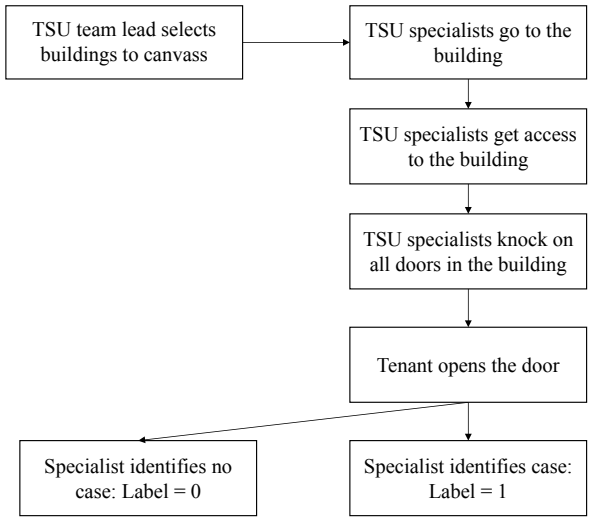
4.1.1 Building address. To locate residential units for canvassing, TSU uses an internal database (which was built using a publicly available dataset) that contains addresses for all the residential buildings in NYC. For each building, the database records the number of units and location information such as address, building identifier number and the tract it belongs to, making it convenient to join with data from other spatial sources.

4.1.2 Knock attempts and case records. During canvassing, TSU specialists knock on every apartment unit in the targeted building(s). If a tenant answers the door, they talk to the person about whether he or she is facing harassment. These activities are recorded at the unit level in *knock attempts* and *case records*, respectively. Each of the records describes the location of the unit, the date it was canvassed, the specialist team that did the canvassing, and the result of the attempt (i.e., knocked, answered, and case identified). The case database also records the source of the case, allowing us to know which cases came from canvassing as opposed to other sources, such as referrals. *Case records* contain information about our outcome variable — whether or not there was at least one case of harassment identified in the building.

4.1.3 Case issues. Once a harassment case is identified, such as a landlord refusing to do essential repairs, the specialists will follow up with the case and separately record each issue related to the housing unit in the *case issues* database. The specialists can then connect the tenants to relevant assistance resources, such as city services or legal support.

Figure 1 shows the TSU specialists' canvassing process and our definition for having a case identified (i.e., the label).

Figure 1: Canvassing process with our definition of the outcome label.



4.2 Public (External) Data

While the internal canvassing records are critical for understanding where harassment occurs, external data are also important to

capture information about buildings not canvassed by TSU yet or longer term historical data before TSU began their outreach activities. TSU's records focus on violations the agency finds based on outreach that began in 2015. External data provides both an expanded time window — which buildings face high rates of landlord issues documented by agencies that predate TSU's existence? — and a lens into the characteristics of buildings and neighborhoods where TSU has historically detected cases.

4.2.1 American Community Survey (ACS). To gain insight into the demographics of tenants whom TSU specialists conduct outreach to, we collected American Community Survey 5-year estimates from 2013 to 2016 at the census tract level. The ACS data contain demographic information such as racial composition, average income, work hours, age distributions and other demographics of the census tract in which a building is located.

4.2.2 Primary Land Use and Tax Lot Output (PLUTO). The PLUTO records describe attributes of each building, such as its renovation history, its building class (e.g., is it a high-rise or a walk-up apartment?), the number of floors, and its recorded owner. We introduced PLUTO data into our model because we believed building information could shed light upon tenant harassment. For example, landlords often own multiple buildings — if TSU canvassing finds harassment at one of a landlord's buildings, that same landlord might be engaging in harassment in other buildings he or she owns. In addition, if a building has been recently renovated, this could be a signal that the landlord is hoping to displace current tenants and lease the building's units to higher-paying tenants. Therefore, we believe that PLUTO features should improve our predictions of harassment.

4.2.3 Department of Housing Preservation and Development (HPD) violations. The HPD issues violations when, after sending inspectors to a unit in response to a complaint, they find evidence of a Housing Code violation. This database contains recorded housing violations, which range from more minor, non-hazardous violations to severe, immediately hazardous violations (e.g., no heat or hot water, a rodent infestation, lead paint). These housing violations could be indicators of rental harassment since some reflect extreme landlord neglect of living conditions. Mr. Sidibe, a New York resident, is a recent example reported in *The New York Times*. He was first hurt by a broken hot water tap and then was improperly evicted while he was recovering in the hospital [3]. Therefore, we hope to use the HPD violation records to improve the predicted harassment risk of a given building.

4.2.4 Housing court litigation. Similar to the HPD violations, housing court litigation can help the model by integrating historical violations. It shows the cases that city agencies levy against an owner when he or she fails to properly address a violation, such as a case legally compelling an owner to fix the heat and hot water in a unit.

4.2.5 Subsidized housing. This database contains building-level information of 53 different subsidy programs a building might participate in, such as the low-income affordable marketplace program and the HPD mixed income program. The subsidy data complement other building-specific characteristics in the databases described here.

5 METHODS

To predict which *buildings* are likely to house tenants susceptible to experiencing harassment *in the next month*, we experimented with Random Forest (RF), Logistic Regression (LR), Decision Trees (DT), and Gradient Boosting (GB), all of which are implemented with *scikit-learn*. We used the data described above to extract numerous features of buildings: in total we used 92 original features (before further processing such as reformatting to one-hot vectors) generated from our data sources.

5.1 Feature Generation

We generated features based on our discussion with experts at the PEU as well as past research on landlord–tenant issues.

5.1.1 Building-level features. Building-level features mainly included *dynamic features* of what harassment-related behaviors have occurred before and *static features* of basic building characteristics. For *dynamic features*, we first generated behavioral features by aggregating the *canvassing* activities and the results at the building level. To predict harassment risk in the upcoming (*next*) month, for example, we counted the number of knocks, doors opened and case identifications in the current (*this*) month in a given building. We also calculated the number of issues associated with these cases for each type (e.g., repair, legal) separately. Apart from the count, we created binary variables that indicate whether there were any knocks, doors opened, or case identifications in the current (*this*) month. In addition to recording activity in *this* month, we aggregated all the prior historical records (until *this* month) to assess the predictive utility of aggregate measures.

Similarly, we created the *HPD violations* and the *housing court litigation* features. The records are aggregated to indicate the number or existence of violations and litigation, both in *this* month and all the months until now. To further break down the type of violations, we included features that describe the number of violations for each severity class. We also grouped housing court litigation by litigation type (such as heat and hot water litigation versus tenant actions against owners).

For *static features*, ZIP codes and borough information were generated from the internal *building address* database. We also included dummy variables describing each canvassing team to account for potential variation between the individual specialists responsible for given buildings or areas.

We further extracted basic building characteristics from *PLUTO*, such as ownership features like owner name and owner type, as well as building renovation features including the year of each renovation. We also considered the size of the building (indicated by the number of floors and number of residential units), the class of buildings (identifying whether the building was made of brick and whether it has an elevator), and the assessed total value of the building.

Additionally from the *subsidized housing* database, we generated a feature to describe whether the building is included in a subsidy program or not.

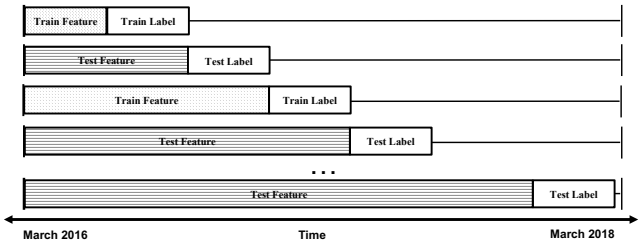
5.1.2 Tract-level features. At the tract level, we generated demographic features by extracting records from the *American Community Survey* database. PEU managers suggested local areas with a certain demographic composition of tenants might contain buildings with more harassment. For example, tracts with a higher percentage of low-income tenants might be more likely to have both a higher concentration of tenants living in rent-stabilized units and a higher concentration of tenants who, due to a lack of awareness of city resources, have unmet needs for help with landlord issues. Our features contain measures of racial demographics, measures of when residents work outside of the home (which affects the tenants' ability to answer the door during the main TSU canvassing hours), and measures of income insecurity, such as receipt of public assistance like Supplemental Security Income (SSI).

We cleaned (i.e., preprocessed such as removing duplicate records) all the data mentioned above to generate the features and match data from different sources by location indicators. We used extrapolation to impute missing data in the features (not the label), such as imputing missing records in 2018.2 with data from 2018.1. We further used the min-max scaler in scikit-learn to normalize continuous features, especially for use in regularized logistic regression models.

5.2 Splitting Data into Training and Testing Sets

To evaluate models with temporal cross-validation, we followed the rule of time-dependent knowledge restriction to temporally split the data into training and testing sets. We needed to ensure that the knowledge in the *future* (i.e., the testing set) does not inform predictions in the *past* (i.e., the training set). For example, in one data split, if we wanted to use data until end of March 2017 (i.e., testing features) to predict the risk of harassment during April 2017 (i.e., testing label), the training set should contain features only until end of February 2017. The training label would then be generated using cases from records during March 2017. Figure 2 shows an example of these training and testing splits, with each row representing one split.

Figure 2: An example illustrating training and testing splits.



5.3 Model Evaluation

5.3.1 Metrics. We used variations of standard metrics to evaluate the model performance: precision and recall at highest predicted risk buildings with a total of k residential units based on the outreach capacity. We select evaluation metrics that have enough flexibility when applied to labels with missing values since many of the buildings we predict at risk will not have been canvassed historically (since the goal of this project is to suggest new buildings to canvass) and we need to evaluate our models in that setting.

To help TSU plan their outreach, at the beginning of each month, we will use the prediction model to recommend a list of buildings with the highest predicted risk of harassment, adding up to k residential units (hereafter denoted as top k, with k limited to TSU's monthly outreach capacity — the number of units they are able to knock on for outreach in a given month).

We want to evaluate the performance of the model according to the true labels of buildings in this prioritized list. Our test set (that we predict on) contains three of types of buildings:

(1) buildings with true positive labels, where TSU knocked and identified case(s)
(2) buildings with true negative labels, where tenant(s) opened the doors when TSU canvassed, but no cases were identified
(3) buildings missing labels, where (i) TSU specialists did not go to the building (no knocks) or (ii) no doors were opened when TSU canvassed the building (knocks but no opens). Traditional precision and recall metrics are not very informative in this case when the true labels of buildings predicted as positive might be missing.

We built upon previous literature [15] focusing on resource allocation in scarce resource settings and used precision and recall at top k as the evaluation metrics. We denote $N_{k,all}$ as total number of buildings in the top k building list, $N_{k,lp}$ as the number of buildings labeled as positive in the top-k list and $N_{k,ln}$ as the number of buildings labeled as negative in the top-k list. $N_{k,u}$ refers to the number of unlabeled buildings. Obviously, $N_{k,all} = N_{k,lp} + N_{k,ln} + N_{k,u}$. As shown by Equation 1, precision at the top k is the proportion of buildings that are labeled as positive (i.e., resulted in true cases) in the top k building list. Recall at the top k represents the proportion of buildings with true positive labels (i.e., with cases identified) that the model captures in the top k list (as shown by Equation 2). While precision measures the efficiency of the model, recall measures model coverage. Figure 3 shows an example of calculating precision and recall at top k.

$$\begin{aligned} \textit{precision at top } k &= \frac{\textit{\# of true positive labels in top } k}{\textit{\# of total labels in top } k} \\ &= \frac{N_{k,lp}}{N_{k,lp} + N_{k,ln}} \end{aligned} \tag{1}$$

$$\begin{aligned} \textit{recall at top } k &= \frac{\textit{\# of true positive labels in top } k}{\textit{\# of true positive labels in testing set}} \\ &= \frac{N_{k,lp}}{\textit{\# of true positive labels in testing set}} \end{aligned} \tag{2}$$

5.3.2 Choices in determining the top k list. First, to determine k, TSU indicated that they would like to keep half of the capacity to their own expert-selected buildings so that TSU specialists could also help residents who lived in buildings outside the top k list. Therefore, each month, we set k as half of TSU's canvassing capacity in a given month (k = 3,000, approximately).

Figure 3: Example of metrics calculation — if a half of TSU capacity $k = 200$, then precision = $\frac{2}{3}$ and recall = $\frac{2}{4}$

Building ID	Prediction Score	# of units	Predicted label	True label
id1112	0.8	153	1	1
id9822	0.79	23	1	1
id9713	0.7	67	1	0
id1751	0.64	11	0	1
id4368	0.48	28	0	0
id4572	0.46	150	0	1

k = 0.5 * TSU capacity

[Top-k list for TSU to canvass]. Second, to suggest a list of the buildings for TSU to canvass, we first rank *all* residential buildings by predicted risk scores and then take the top ones that add up to contain k (apartment) units since the TSU capacity is based on the number of units and we are predicting at the level of buildings. Note that if k is in between two buildings in our list, we include the entire building with at least one unit in the top-k list.

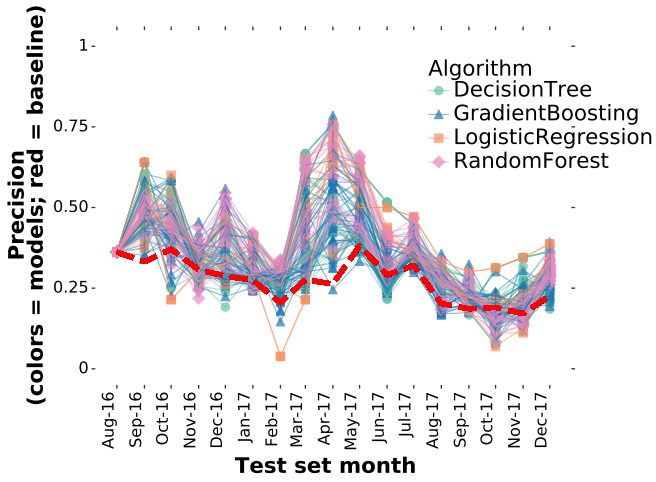
[Top-k list for model performance evaluation]. Third, to evaluate model performance, we generated the top-k list of buildings by only including the *labeled* buildings. We ranked labeled buildings by the predicted risk of harassment, and marked the top-k-units buildings as positive. We didn't deploy the k cut-off on *all* buildings since the top-k list of *all* buildings did not contain enough labeled data to make the precision scores reliable. On average, TSU canvasses about 300 buildings per month out of a total of 6,437 in their outreach area, which covers < 5% of all buildings. It was highly likely that most, if not all, of the (previously canvassed) 300 buildings fell out of the top-k list, leading to few labeled data in top-k list. In fact, about 20% of the top-k lists generated by each model in each test month contained no labeled building, with the rest 80% of models only include a few labeled data. For example, a Random Forest model proposed 19 buildings in the top-k building list, with only one of them observed by TSU. The precision would be 1 if TSU identified case(s) in this building and 0 otherwise. This challenges our confidence in using these precision and recall metrics to represent the model performance. We thus chose to use the labeled data in determining the top k list for model performance evaluation. This is typical in problems with missing labels and we recommend to conduct a field trial with proactive canvassing on the previously not canvassed buildings to further validate the model on both labeled and unlabeled data.

6 RESULTS

6.1 Predictive Performance

6.1.1 Baseline: TSU's current outreach method. TSU currently uses a simple approach to plan its outreach in the targeted 20-ZIP-codes areas. TSU specialists systematically go block by block attempting to enter every building where there is at least one rent-stabilized unit. A list of buildings to attempt is assigned via a custom-built canvassing app loaded on an iPad.

Figure 4: Model performance over time (in training stage). X axis represents the time period (month) and Y axis represents the precision scores of models in the month. The figure shows that the baseline has been varying across months and our models generally performed better than the baseline.



6.1.2 The performance of our models. Our final models were trained on data from July 2016 to December 2017 and were tested on outreach records from January 2018. We further split the training data into 17 folds as illustrated in the previous section to conduct temporal validation.

Figure 4 shows the performance of every model on each data split during the training stage. The TSU baseline is represented by the red dashed line. The machine learning models performed better than the baseline by 36% on average.

The figure also shows that the effectiveness of outreach efforts by TSU in terms of found cases of tenant harassment varies over time as well. Therefore, to better interpret how much better our model performed than the baseline in each data split, we calculated the ratio of model precision to baseline precision (hereafter named as precision ratio).

To select the best performing model, we first took the average precision ratio score of all data splits and narrowed down to models that had precision ratios ranked in the top 10. Because we want the model that TSU uses to not only exhibit high *average* precision but also exhibit high *stability* in performance, we incorporated the standard error of the precision ratio scores into the evaluation of a model's performance by calculating: $\frac{\text{precision mean}}{\text{precision std}/\sqrt{\#\text{ of precisions}}}$ [8].

The best model to predict whether there will be at least one case in a building next month was a Gradient Boosting classifier with 100 estimators. In our test month (February, 2018), TSU was able to inspect 312 buildings of 7,374 residential units, covering about 4.85% of all buildings. Therefore, we set $k = 3,687$ units to generate the top k list for evaluation. Table 2 shows how our model performed in terms of false positives, false negatives, true positives and true negatives. Our model was able to identify about 59% more high-risk buildings than the baseline (with a precision score of 0.25 in the test month).

Figure 5: Precision and number of labeled data at each k proportion for the Gradient Boosting model.

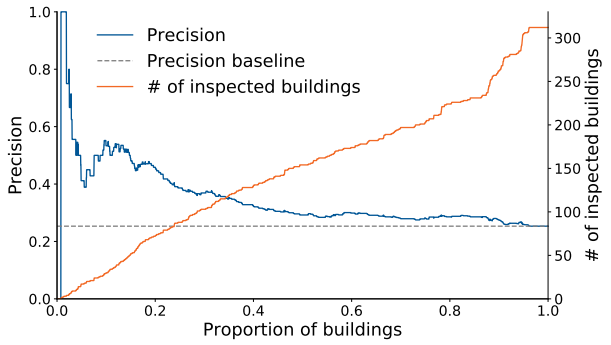
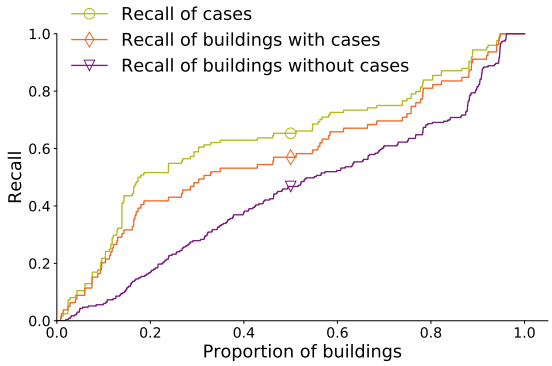


Figure 6: Recall curves at each k proportion for the Gradient Boosting model.



In Figure 5, the precision scores of the building-level prediction at different levels of k is represented by the blue line, with X axis representing the proportion of buildings at k (i.e., $N_{k,all}/Total\ number\ of\ buildings$). We also plotted an orange line to visualize the number of labeled data at each k (i.e., $N_{k,lp} + N_{k,ln}$), which shows the number of (labeled or successfully canvassed) buildings supporting the precision calculations. Since TSU only inspected about 300 buildings per month and left most buildings unlabeled, this supporting number at k helps us assess our confidence in the precision score at k.

In addition to precision, we calculated two measures of recall: recall of the *total* count of cases across buildings and recall of buildings with *any* case. Figure 6 shows that recall of cases (represented by the yellow-green line) was in general higher than the recall of buildings with any case (represented by the orange line), indicating that our model was good at predicting buildings with a larger number of cases rather than buildings with only one case.

Since both precision and recall measures are relying on labeled data, we wanted to understand if our overall list of high risk buildings was good at ranking high risk buildings above low risk buildings. In addition to recall on the labeled positive examples (orange line), we calculated recall on the labeled negative examples (buildings with no cases) using all buildings as the denominator (purple line). The intuition is that a good ranked list will have more (labeled) positive examples than negative examples at the top of the list and vice versa at the bottom of the list. The gap between recall on positive examples and recall on negative examples in Figure 6 allows us to see this was actually the case. The orange line goes up steeply at the beginning (more positive examples) and the purple line goes up steeply at the bottom of the list (more negative examples) giving us confidence in the ranking performance of our model.

Table 2: Confusion matrix of the best performing model.

	Actual True	Actual False
Predicted True	33	50
Predicted False	46	183

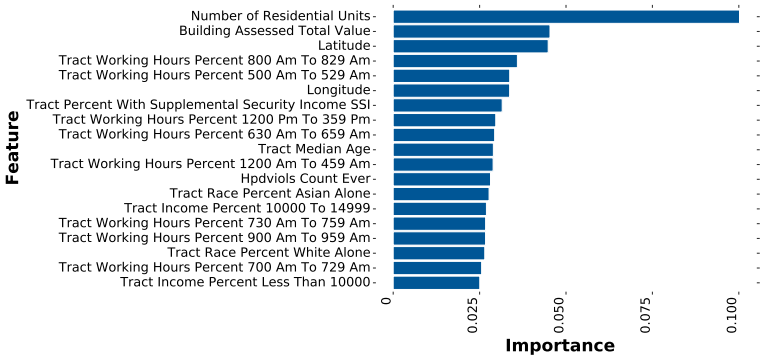
6.2 Interpreting the Models: Feature Interpretation

Figure 7 shows the top 20 features that have the highest feature importances in the best performing Gradient Boosting model.

6.2.1 Tract-level demographic features. From the figure, we see that most features in the top 20 feature list were generated from American Community Survey data, which reflects the demographic characteristics of residents in the tract in which a building is located. For example, measures of income insecurity were important in predicting harassment — these included the proportion of people receiving Supplemental Security Income (SSI) in a given tract (*Tract Percent With Supplemental Security Income SSI* in figure 7) and the percentage of households earning less than $10,000 per year (*Tract Income Percent Less Than 10000* in figure 7). These features might be important because they may reflect unmet need — that is, areas where people are both particularly vulnerable to illegal tactics by landlords and where they also may, prior to TSU's visit, have the most difficulty navigating city services that can help. In addition to the income variables and, more interestingly, 8 of the 20 top features were indicators for the hours that a tract's residents work outside the home. For example, the feature, *Tract Working Hours Percent 800Am to 829Am*, represented the proportion of people who usually leave their apartment to work between 8:00AM and 8:29AM. These features could be important for two reasons — first, they might serve as additional indicators of socioeconomic status (e.g., lower-income individuals might face less standard work schedules); second, they might reflect which tenants are home to answer the door when TSU specialists go canvassing on weekdays and weekends.

6.2.2 Building history and value features. The figure also shows that building-level indicators, such as the total number of HPD violations in a building up until the given month and the total monetary value of a building (generated from PLUTO dataset) are informative in predicting harassment risks. These observations provide support to the idea that external information, including a building's history of violations as well as the physical and economic attributes of a building (i.e., how much is a building worth?), is valuable in predicting whether there will be at least one case of harassment in the building next month.

Figure 7: Feature importance from the gradient boosting model. We plot the 20 most important features to understand the top predictors that help us identify buildings of high risks.



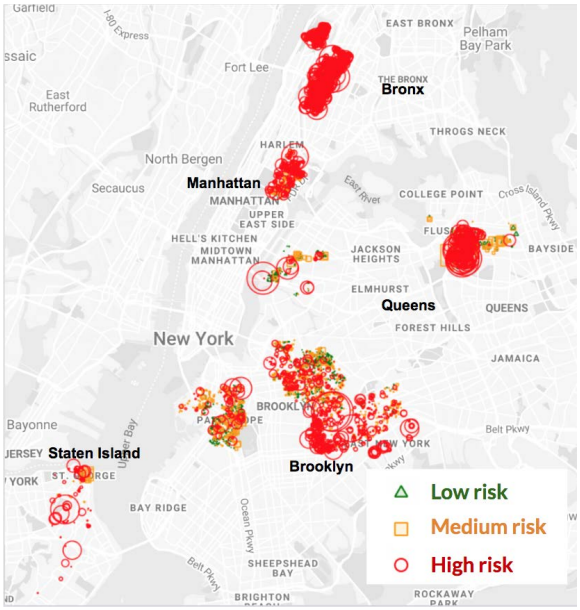
6.2.3 Building location features. In addition, the model identified the longitude and latitude of a building as important features. This indicates that high-risk buildings are perhaps clustered in specific locations. To highlight this clustering feature, we predicted the risk of each building with our best-performing model. We further separated buildings into different levels of risk, with high risk representing buildings with the highest 33.33% risk scores, low risk representing buildings with the lowest 33.33% risk scores and medium risk representing the rest. We plotted each building according to its point location, with high-risk buildings in red, medium-risk buildings in yellow and low-risk buildings in green (see Figure 8). The map highlights clusters of high-risk buildings in Manhattan and the Bronx, with low-risk buildings dispersed throughout Brooklyn, Queens and Staten Island. This finding suggests that in order to balance canvassing efforts across boroughs, we would need to separately rank buildings and provide a high-risk building list for each borough when deploying the model in practice.

6.2.4 Building size. While these features mentioned above indicate that the model took advantage of information in the data, the high importance of the *Number of Residential Units* shows that our problem formulation — predicting *any* case in a building — leads us to identify buildings with many residential units. These buildings have a higher "denominator" of tenants at risk of harassment to generate the label of a single case in a particular month. For the test month (Feb, 2018) in particular,buildings predicted to have high risk of harassment on average contained 70 units per building, which was about 3 times as many as the average size of all buildings in the targeted area. Figure 8 further highlights the correlation between a building being larger and a building being identified as higher risk: buildings with larger numbers of units (indicated by marker size) were more likely to be predicted as buildings of high rental harassment risk (indicated by color of red). Therefore we defined another problem formulation to try and standardize a building's count of cases by the number of tenants who might have a case.

6.3 Reformulation: Predicting case per unit ratio above a threshold

Our reformulated problem uses the label — hereafter called the *any-case* label — defined as follows: $Y \in \{1 = \text{any case}, 0 = \text{no case}\}$ in

Figure 8: Map of buildings predicted as different risk levels. Each point represents a building: high risk (red), medium risk (yellow), low risk (green). Manhattan and the Bronx had most of the high-risk buildings. Low- and medium- risk ones were mainly spread out among Brooklyn, Queens and Staten Island. The marker (i.e., circle, triangle, square) size reflects the # of units in the building.



building i in month m. What we call the threshold label constructed a binary label using a two-step procedure: first, we calculated the ratio of cases in a building per number of units; second, we constructed the binary label as follows: $Y \in \{1 = \text{ratio} \geq \text{threshold}, 0 = \text{ratio} < \text{threshold}\}$ in building i in month m. The results we present focus on buildings with a ratio in the top 10% of the training set.

We used the same procedure as in section 6.1 to select the best performing model for the threshold label. The best-performing model (a Gradient Boosting classifier) identified 14% more buildings

Figure 9: Feature importance from the best-performing threshold model. We plot the 20 most important features to understand the top predictors that help us identify buildings of high risks.

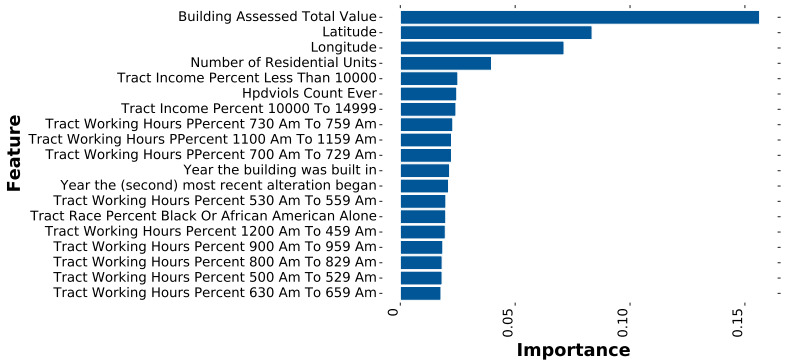


Figure 10: Precision and number of labeled data at each proportion of buildings for the Gradient Boosting model using the threshold label.

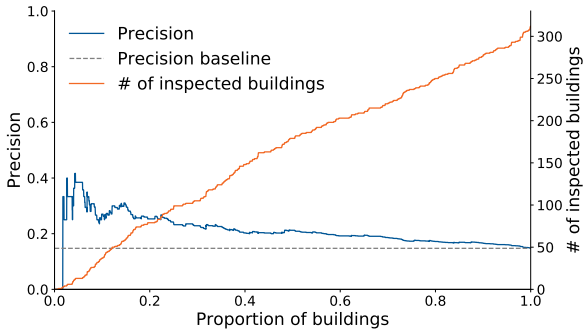
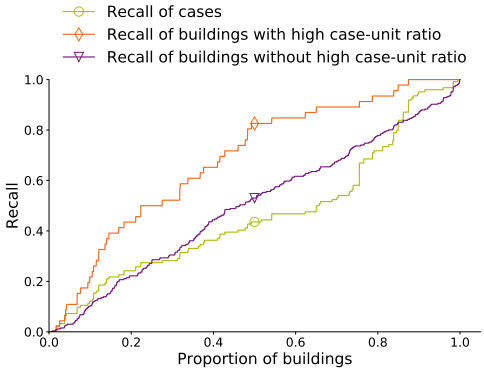


Figure 11: Recall curves at each proportion of buildings for the Gradient Boosting model using the threshold label.



of high case-per-unit ratio than the baseline (with *precision* = 0.15 in the test month). Figure 10 plots its precision scores and the supporting number of buildings at each level of k. We found that the threshold model successfully prioritized buildings with a higher proportion of cases than the any-case model (with case-per-unit ratio = 0.94 and 0.30, respectively), which means about 213% more cases could be identified by the threshold model than the any-case model, holding the number of units canvassed constant.

Figure 9 plots the 20 most important features from this model. Comparing to the model using the any-case label, the best-performing model using the threshold label put more weights on features such as the assessed building value, the year the building was built in, the year the building was recently renovated (i.e., *Year the (second) most recent alteration began*), and number of violations HPD had ever recorded, while it was less informed by features such as percentage of households receiving SSI in the tract and number of units.

In fact, the model using the *threshold label* prioritized buildings with smaller sizes (average number of units = 11) than the model using the *any-case label* (average number of units = 70). This may also account for the phenomenon in Figure 11: Recall of buildings with high case-per-unit ratio was higher than average (*slope* > 1); recall of cases was not as high.

To further understand how model of any-case label over optimized large buildings, we additionally compared the two models in two ways. First, we ranked all 6,437 buildings according to the predicted risk score using both models and found that they were somewhat uncorrelated. Second, for each model, we plotted the buildings in the top-k list the model suggested TSU to canvass, respectively (see Figure 12). Each point refers to a building with the size of the point representing the number of units in the building. This map shows that predictions using the threshold label (represented by orange square) top ranked more small-size and geographically distributed buildings than predictions using the any-case label (represented by green circle).

These findings support our assertion that if we only predict whether there would be any case next month, the model would be more likely to provide a list of large buildings as opposed to buildings of high case-per-unit ratio. Depending on their goals, policymakers and canvassing specialists might prefer one or the other — larger buildings might allow for more efficient canvassing to knock on doors that are more geographically co-located (supporting the *any-case label*). On the other hand, the threshold label gives tenants living in smaller buildings more of an opportunity to receive outreach and results in a possibly more equitable outreach process.

Figure 12: Comparing of any-case label and threshold label suggestions. Each point is a building with size representing the # of units. Predictions using any-case label prioritize large size buildings and were more geographically clustered.

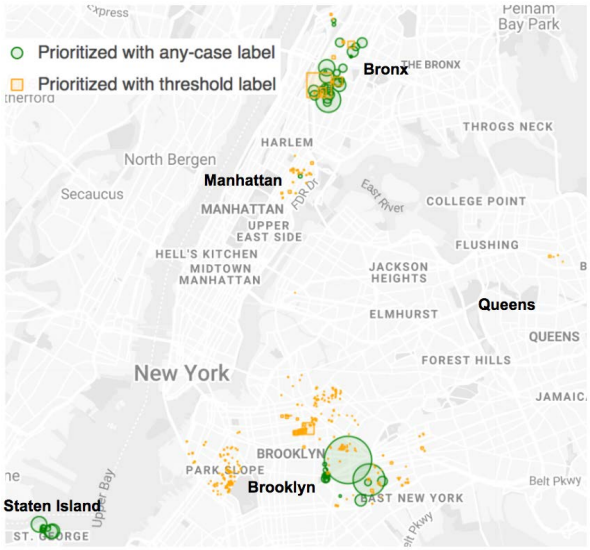
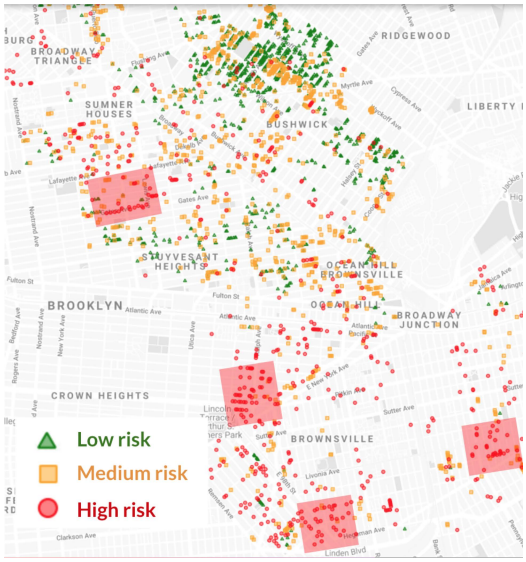


Figure 13: Example of post-model implementation with high-risk buildings clustered.



7 PRACTICAL IMPLICATIONS AND NEXT STEPS PRIOR TO IMPLEMENTATION

The Tenant Support Unit hopes to efficiently find more individuals in need of their help with fewer outreach knocks by generating a list of buildings where tenants are most likely to experience harassment. At the beginning of each month (when TSU team leads typically decide which areas to visit next in the upcoming month), the model will generate a list of buildings for each borough where tenants face high risks of harassment. Prior to the results being used to inform TSU's process, the agency should conduct a field trial to validate the predictions of our model as well as run a thorough bias and fairness analysis. This field trial can better inform whether buildings that the model flags as high risk are more likely to yield cases than buildings that the model flags as low risk. If the model is able to successfully differentiate buildings in this way, next steps should include efforts to use the list more efficiently — that is, to not waste time travelling across the city to canvass buildings in exact order of high to low risk. TSU could determine clusters of buildings that have a high enough density of units in high-risk buildings to canvass in one or over multiple days (see Figure 13). Once these cluster areas are created, specialists can canvass every target building in the cluster area without unnecessary travel among exclusively high-risk buildings across boroughs or neighborhoods.

One area of future work we want to explore is to deal with selection bias in our labels and actively collect new labels. Since we only have labels from buildings canvassed by TSU, and there is some bias in how they select buildings to canvass, our model is trained only on that data and will most likely be only confident on predictions made on similar buildings. We want to use the field trials to understand this bias and use the TSU team to also help improve the model by canvassing new buildings to provide more representative labels to train our model.

8 CONCLUSIONS

We used a machine learning approach to help NYC identify buildings where tenants might face landlord harassment. Our model significantly outperforms the current outreach method. The predicted risk scores can help the agency more accurately prioritize areas of high rental harassment and better allocate their building canvassing resources to help more tenants in need in an equitable manner.

In addition, our model provides insights into the important correlates of harassment that might be useful for researchers without access to agency data confirming harassment, complementing efforts to look at harassment using proxies like a loss in rent-stabilized units [17]. Our feature importance results not only find the relevance of building-specific attributes — for instance, the building's history of code violations — but also the utility of local demographic data to highlight where tenants might face housing issues.

By comparing different formulations of the prediction problem, with different prediction labels, we also showed that although formulating the harassment prediction as a binary classification — whether there will be any case next month — significantly increased the precision, it might be biased towards buildings with many units. Finally, we discussed how our model can better facilitate canvass planning and resource allocation by clustering the high-risk buildings for efficient deployment and outreach.

9 ACKNOWLEDGMENTS

We greatly thank the reviewers for their time and insightful feedback. We also thank our partner, the New York City Public Engagement Unit for their support.

REFERENCES

[1] Klaus Ackermann, Eduardo Blancas Reyes, Sue He, Thomas Anderson Keller, Paul van der Boor, Romana Khan, Rayid Ghani, and José Carlos González. 2016. Designing policy recommendations to reduce home abandonment in Mexico. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 13–20.

[2] Susan Athey. 2017. Beyond prediction: Using big data for policy problems. *Science* 355, 6324 (2017), 483–485.

[3] Kim Barker, Jessica Silver-Greenberg, Grace Ashford, and Cohen Sarah. 2018. The Eviction Machine Churning Through New York City. *The New York Times* (2018). https://www.nytimes.com/interactive/2018/05/20/nyregion/nyc-affordable-housing.html

[4] Richard Berk. 2012. *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

[5] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350, 6264 (2015), 1073–1076.

[6] NYC Housing Preservation Development and United States Census Bureau. 2018. New York City Housing and Vacancy Survey. *NYC Housing Preservation Development* (May 2018). https://www1.nyc.gov/site/hpd/about/nychvs.page

[7] NYC Housing Preservation Development and United States Census Bureau. 2018. New York City Housing and Vacancy Survey. *United States Census Bureau* (May 2018). https://www.census.gov/programs-surveys/nychvs.html

[8] Drew Fudenberg and Annie Liang. 2018. Predicting and Understanding Initial Play. (2018).

[9] Edward L Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review* 106, 5 (2016), 114–18.

[10] Edward L Glaeser, Scott Duke Kominers, Michael Luca, and Nikhil Naik. 2018. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry* 56, 1 (2018), 114–137.

[11] Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21, 3 (2013), 267–297.

[12] Joseph Gyourko and Peter Linneman. 1990. Rent controls and rental housing quality: A note on the effects of New York City's old controls. *Journal of Urban Economics* 27, 3 (1990), 398–409.

[13] Jun Seok Kang, Polina Kuznetsova, Michael Luca, and Yejin Choi. 2013. Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1443–1448.

[14] Igor Kononenko. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine* 23, 1 (2001), 89–109.

[15] Himabindu Lakkaraju, Everaldo Aguiar, Carl Shan, David Miller, Nasir Bhanpuri, Rayid Ghani, and Kecia L Addison. 2015. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1909–1918.

[16] Alan Meyers, Diana Cutts, Deborah A Frank, Suzette Levenson, Anne Skalicky, Timothy Heeren, John Cook, Carol Berkowitz, Maureen Black, Patrick Casey, et al. 2005. Subsidized housing and children's nutritional status: data from a multisite surveillance study. *Archives of Pediatrics & Adolescent Medicine* 159, 6 (2005), 551–556.

[17] Mayor's Office of Data Analytics. 2018. Tenant Harassment Project. (2018). https://github.com/MODA-NYC/Project_TenantHarassment

[18] Jesse Roman. 2014. In Pursuit of Smart. *National Fire Protection Association Journal* (2014). https://www.nfpa.org/News-and-Research/Publications/NFPA-Journal/2014/November-December-2014/Features/In-Pursuit-of-Smart

[19] Eric T. Schneiderman. 2018. NYS Attorney General Tenant's Right Guide. *NYC Government* (2018). https://www1.nyc.gov/assets/buildings/pdf/tenants_rights.pdf

[20] Dan Tasse, Alex Sciuto, and Jason I Hong. 2016. Our House, in the Middle of Our Tweets.. In *ICWSM*. 691–694.

[21] Gregg G Van Ryzin and Thomas Kamber. 2002. Subtenures and housing outcomes for low income renters in New York City. *Journal of Urban Affairs* 24, 2 (2002), 197–218.

[22] Ameena Walker. 2017. In New York, Rents are Increasing Twice as Fast as Wages. *Curbed New York* (2017). https://ny.curbed.com/2017/8/16/16154956/nyc-rent-prices-wage-increase-comparison

[23] Michelle Wood, Jennifer Turnham, and Gregory Mills. 2008. Housing affordability and family well-being: Results from the housing voucher evaluation. *Housing Policy Debate* 19, 2 (2008), 367–412.

[24] Elvin Wyly, Kathe Newman, Alex Schafran, and Elizabeth Lee. 2010. Displacing New York. *Environment and Planning A* 42, 11 (2010), 2602–2623.

[25] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1395–1405.

[26] Danning Zheng, Tianran Hu, Quanzeng You, Henry A Kautz, and Jiebo Luo. 2015. Towards Lifestyle Understanding: Predicting Home and Vacation Locations from User's Online Photo Collections.. In *ICWSM*. 553–561.