Rental Insecurity: Predicting evictions in high risk areas

Machine Learning | API 222

Yui Miura
Arianna Salazar Miranda
Mohamed Raouf Seyam

Abstract:

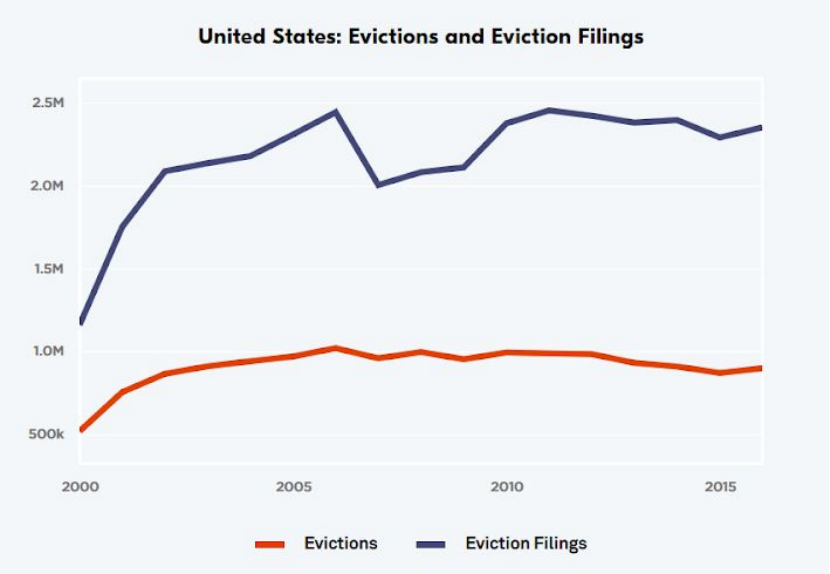
The impacts of evictions are severe and long-lasting. Evictions are a leading cause of homelessness, and research has tied eviction to depression and material hardship outcomes.[1] This project aims to use machine learning to predict the areas with individuals in high risk of eviction. We combine data on evictions gathered by Princeton's Evictions Lab from 2010 to 2016, with demographic and neighborhood characteristics available in the US Census for 2010. Our best performing model is Random Forest, which predicts evictions with a 80% accuracy. Our approach can support policymakers in directing the required resources ahead of time to vulnerable areas to reduce the risk of homelessness for those evicted.

[1] See Desmond, M., & Kimbro, R. T. (2015). Eviction's fallout: Housing, hardship, and health. *Social Forces*, *94*(1), 295–324. https://doi.org/10.1093/sf/sov044

1 | Introduction and Motivation

Evictions in the United States are skyrocketing. In 2000, there were about 500,000 evictions. This figure grew to approximately 900,000 in 2016; out of over 2.35 million eviction court filings (see figure 1).

Figure 1: Evictions in thousands (2000 - 2016)



[2]

Research has linked evictions to multiple severe negative consequences. For low-income tenants, evictions can exacerbate residential instability, increase the chances of job loss[3], and can have long-term psychological effects for both children and adults.[4] Moreover, in the absence of residential stability, it is increasingly difficult for low-income families to invest in their home, social relationships, and community.[5] This is especially true today, when the majority of poor renting families in America devote over half of their income to housing

[2] The Eviction Lab - Princeton University, 2018
[3] Matthew Desmond & Carl Gershenson, Housing and Employment Insecurity among the Working Poor, 63 SOC. PROBLEMS 46, 47, 54–59 (2016).
[4] See, Bartlett, S. (1999). Children's experience of the physical environment in poor urban settlements and the implications for policy, planning and practice. *Environment and Urbanization*, *11*(2), 63–73. https://doi.org/10.1630/095624799101285093
[5] Oishi, Shigehiro. 2010. "The Psychology of Residential Mobility: Implications for the Self, Social Relationships, and Well-Being." Perspectives on Psychological Science 5(1):5–21.

costs.[6] Aside from the adverse effects evictions can have on tenants, they are costly for landlords and can increase economic pressures on homeless shelters.

Given the negative consequences of displacement, it is crucial that governments design strategies to early-identify and prevent involuntary displacement. Moreover, identifying those most at risk of becoming homeless can make it easier to target cost-reducing preventive assistance. It is no surprize that this has been challenging since the U.S. Census Bureau does not track evictions, and the first ever evictions database has been compiled less than a year ago by the Evictions Lab at Princeton. This propelled our motivation to explore this recent dataset to tackle an unaddressed problem. We hope that our work will help decision makers working on evictions in the United States.

In this project, we combine data on evictions with demographic and neighborhood characteristics to predict the areas with high risk of eviction. In addition, we use our trained model to make predictions out-of-sample for the missing observations in the original dataset.

The remainder of the paper is organized as follows. Section 2 describes the data used for the project, section 3 outlines how we cleaned the data to create our predictors and outcome variable, section 4 overviews the models we implemented and describes in detail our chosen model, section 5 reflects on our conclusions through the lense of fairness and bias, and finally

[6] Desmond, Matthew, and Carl Gershenson. 2015. "Housing and Employment Insecurity among the Working Poor." Working Paper, Harvard University.

section 6 concludes with a discussion of the limitations of this paper and improvements for future research.

2 | Data description

To train our models, we combine two primary sources of data. First, we obtain data from the Eviction Lab at Princeton University released in April 2018, which we use to construct our outcome variable. These data include aggregate counts on evictions filings across the U.S. for the 2010-2016 period.[7] It is important to note that these figures are of evictions that went through court; not informal evictions that were settled outside of court - a reliable dataset is unavailable for those, but they are expected to be much higher.

Our outcome variable is the share of eviction judgments in a given block group. We construct this eviction measure by taking the average of evictions by block for our period of analysis and then dividing it by the total average population over the same period. To ease its interpretation, we dichotomize our outcome as follows: if an observation falls below the 0.5 cut-off, we assign it to a "low risk of eviction" category, and if it falls above the 0.5 cut-off we assign it to the "high-risk of eviction" category.[8] The data used covers approximately 131,288 evictions (after cleaning) and is restricted to include only households that underwent a legal process, not those in which people moved through an informal arrangement. The average eviction rate per census block was 4.75 for the seven year period of analysis. Among

[7] Matthew Desmond, Ashley Gromis, Lavar Edmonds, James Hendrickson, Katie Krywokulski, Lillian Leung, and Adam Porton. Eviction Lab National Database: Version 1.0. Princeton: Princeton University, 2018, www.evictionlab.org

[8] Before dichotomizing the outcome, we considered binning the continuous variable into quartiles. Since the 4 categories were arbitrary, we decided it was more intuitive and natural to implement eviction risk as a binary outcome.

the highest ranked cities in terms of evictions are California, Michigan, Ohio, Virginia, and Texas, which total of 44,229, 43,294, 40,413, 31,762, and 31,728 evictions, respectively.

Second, we combine these data on evictions with demographic and economic indicators obtained from the 2010 Census. We use census block groups as the geographical unit of analysis, which allows us to estimate areas in high risk of eviction at the smallest scale for which the data is available.

3 | Process: Data Cleaning and Feature Engineering

We construct a total of 62 predictors. Table 1 provides an overview of these predictors and their descriptions. As for the data from Eviction Lab, it is natural to assume rent could be one of the most important predictors, but 17% of data is missing. Therefore, we created a new variable (median_rent_imputed) by taking the average median gross rent by block level from 2010 to 2016 --- our period of analysis. We were able to input 130,431 observations out of a total of 261,757 that were missing. We believe this approach is reasonable, given that rents are not likely to vary a lot within a given block group (600 and 3,000 people).

To construct most of our housing and demographic predictors, we used data gathered from the Census. For the housing conditions predictors, we calculated the share of variables, which include the *share of people living in urban areas*, the *share of housing units that have a mortgage*, and the *share of vacant units*, among others. For the demographic characteristics we included the *share of male population* and the *share of population with different levels of*

education (ranging from less than high school to earning doctorate degree), among others. Finally, we replaced the predictor with the median value of that block group, where the value was missing.

Table 1: Predictors

Variables	Description	Source
population	Total population	Eviction Lab
poverty.rate	% of the population with income in the past 12 months below the poverty level	
renter.occupied.households	Interpolated count of renter-occupied households	
pct.renter.occupied	% of occupied housing units that are renter-occupied	
median.gross.rent	Median gross rent	
median.household.income	Median household income	
median.property.value	Median property value	
rent.burden	Median gross rent as a percentage of household income	
pct.white	% population that is White alone and not Hispanic or Latino	
pct.af.am	% population that is Black or African American alone and not Hispanic or Latino	
pct.hispanic	% population that is of Hispanic or Latino origin	
pct.am.ind	% population that is American Indian and Alaska Native alone and not Hispanic or Latino	
pct.asian	% population that is Asian alone and not Hispanic or Latino	
pct.nh.pi	% population that is Native Hawaiian and Other Pacific Islander alone and not Hispanic or Latino	
pct.multiple	% population that is two or more races and not Hispanic or Latino	
pct.other	% population that is other race alone and not Hispanic or Latino	
median_rent_imputed	Our Outcome	
TotalPopulation	Total population	Census
Urban	Total population in urban	
Rural	Total population in rural	
PopulationDensity	Population Density	
MalePopulation	MalePopulation	
AverageHouseholdSize	Average Household Size	
TotalPopulation25Plus	Population 25 years and over	
Pop25Plus_LessHighSchool	Population 25 years and over: Less Than High School	
HighSchoolGraduate	Population 25 years and over: High School Graduate (includes equivalency)	
SomeCollege	Population 25 years and over: Some college	
BachelorDegree	Population 25 years and over: Bachelor's degree	
MasterDegree	Population 25 years and over: Master's degree	
ProfessionalDegree	Population 25 years and over: Professional school degree	
DoctorateDegree	Population 25 years and over: Doctorate degree	
HousingUnits	Housing units	
VacantHousingUnits	Vacant Housing Units	
VacantUnitsForRent	Vacant Housing Units: For rent	
VacantUnitsForSale	Vacant Housing Units: For sale only	
VacantUnitsOther	Vacant Housing Units: Other vacant	
MedianYearStructureBuilt	Median year structure built	
OwnerOccupiedHousingUnits	Specified owner-occupied housing units	
MortgageHousingUnits	Specified owner-occupied housing units: Housing units with a mortgage	
MortgageOrEquity	Specified owner-occupied housing units: Housing units with a mortgage: With either a second mortgage or home equity loan, but not both	
SecondMortgageOnly	Specified owner-occupied housing units: Housing units with a mortgage: With either a second mortgage or home equity loan, but not both: Second mortgage only	
HomeEquityOnly	Specified owner-occupied housing units: Housing units with a mortgage: With either a second mortgage or home equity loan, but not both: Home equity loan only	
BothSecondMortgageEquity	Specified owner-occupied housing units: Housing units with a mortgage: Both second mortgage and home equity loan	
NoSecondMortgageNoEquity	Specified owner-occupied housing units: Housing units with a mortgage: No second mortgage and no home equity loan	
HousingNoMortgage	Specified owner-occupied housing units: Housing units without a mortgage	
share_urban	Urban/TotalPopulation	
share_male	MalePopulation/Total Population	
share_Pop25Plus_LessHighSchool	Pop25Plus_LessHighSchool/TotalPopulation25Plus	
share_HighSchoolGraduate	HighSchoolGraduate/TotalPopulation25Plus	
share_SomeCollege	SomeCollege/TotalPopulation25Plus	
share_BachelorDegree	BachelorDegree/TotalPopulation25Plus	
share_MasterDegree	MasterDegree/TotalPopulation25Plus	
share_ProfessionalDegree	ProfessionalDegree/TotalPopulation25Plus	
share_DoctorateDegree	DoctorateDegree/TotalPopulation25Plus	
HousingUnitsPerCapita	HousingUnits/TotalPopulation	
share_MortgageHousingUnits	MortgageHousingUnits/OwnerOccupiedHousingUnits	
share_MortgageOrEquity	MortgageOrEquity/OwnerOccupiedHousingUnits	
shareSecondMortgageOnly	SecondMortgageOnly/MortgageOrEquity	
shareHomeEquityOnly	HomeEquityOnly/MortgageOrEquity	
shareBothSecondMortgageEquity	BothSecondMortgageEquity /MortgageOrEquity	
shareNoSecondMortgageNoEquity	NoSecondMortgageNoEquity /MortgageOrEquity	
shareHousingNoMortgage	HousingNoMortgage /MortgageOrEquity	

4 | Methods Applied and Model selection

Given the large size of the data, computational power constraints, and the binary classification result we are trying to achieve, we narrowed our scope of analysis to three suitable models: Logistic regression with a penalty, Boosting, and Random Forest. For each model, we first tuned its parameters using cross-validation techniques. Then, we computed its predictive accuracy using the misclassification rate on the randomly generated test subsample that we left a side at the beginning of the procedure. The test subsample included 20% of the observations, and the train sample the remaining 80% of the observations.

We began by considering the LDA and QDA models. Although both of these methods are more interpretable than trees, we disregarded the models for two main reasons: first, they both assume that $Pr(X = x \mid Y = k)$ is normally distributed to achieve low bias in the model. Second, these models will suffer from high variance when the dimension of the data is high. In addition, we also considered the SVM model. However, upon trial, SVM proved to be very computationally expensive - it is also prone to overfitting, which further discouraged us from using it. Similarly, kNN proved to be computationally challenging when we tried to find the optimal k. This came as no surprise given the size of our data. Moreover, we feared the risk of overfitting if we were to select an arbitrary value for k, so we decided to prioritize other approaches.

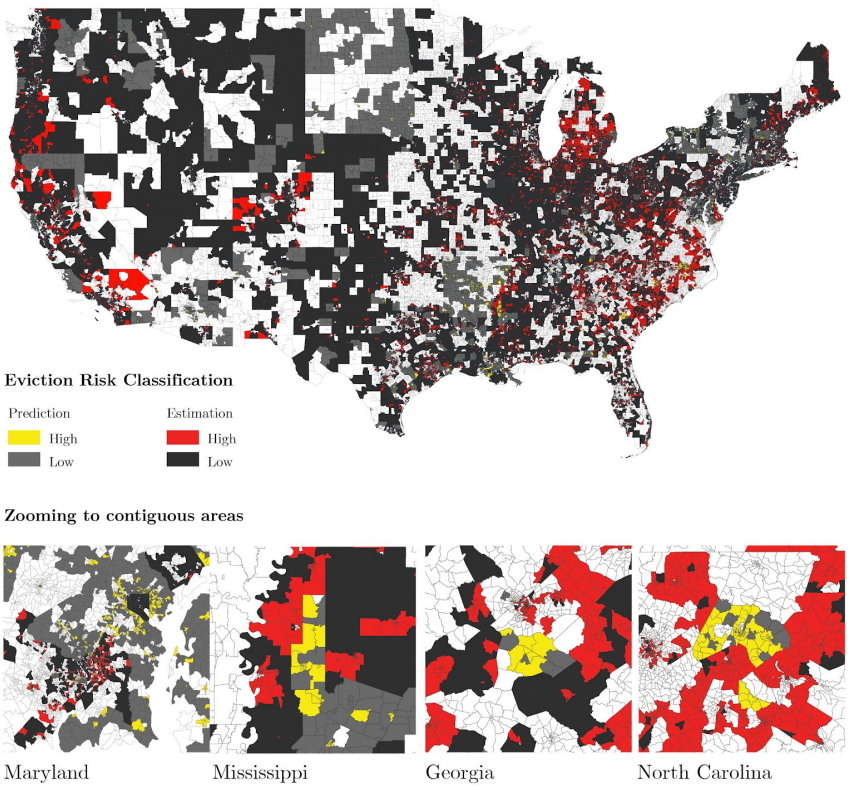
Since we have multiple variables, our first approach was to use a **penalized logistic regression**, which imposes a penalty to the logistic model for having too many variables. The optimal penalty, which we found using cross-validation, is 0.0031. Overall, the model selects 37 predictors and has an accuracy of 79%. Among the most important predictors are the share of individuals older than 18 with some college, the share of vacant units for sale, the average household size, and the share of housing with mortgage or equity.

For our second method, **Boosting**, we ran a slow learner model (lambda = 0.01) with 5 fold cross validation and a maximum tree depth of 4 on 2000 trees. The model's accuracy turned out to be lower than that of Random Forest, and it took a longer time to converge. We could improve its accuracy by increasing the number of trees, but that requires additional computing power; and we also reasoned that Random Forest might be better suited as it de-correlates the trees.

Last, we implemented **Random Forest**: an algorithm that instead of using the full set of predictors (p = 64) at each splitting point, considers only a randomly selected subset of predictors (M). In a broader sense, the random forest generates training data sets by boosting and then uses this data sets to train individual decision trees. To tune the algorithm, we set the number of subset predictors (M) as 7 and set the number of trees as 500. The model has an accuracy of 81%, which is the result of taking the average across all the trees. This model ranked the highest concerning predictive accuracy among all the implemented models, and as a result, we selected it as our best model.

Figure 2 maps the resulting eviction risk classification from the random forest model. The top panel overlays both the estimation evictions and the predicted evictions. Areas shown in yellow or red are places with high-risk prediction or estimation, respectively. The bottom panel zooms into four selected areas: Maryland, Mississippi, Georgia, and North Carolina, where the number of contiguous high-risk areas where the most agglomerated. Overall, the high risk eviction areas are located in the States of California, Texas, Michigan, Ohio, and New York, which total 6,729, 5,475, 4,750, and 4478, and 4085 evictions, respectively. Conversely, the areas with the lowest predicted high risk eviction rates are located in the States of Wyoming, Vermont, Hawaii, District of Columbia, and Montana, which total 46, 47, 60, 137, and 141 evictions, respectively.

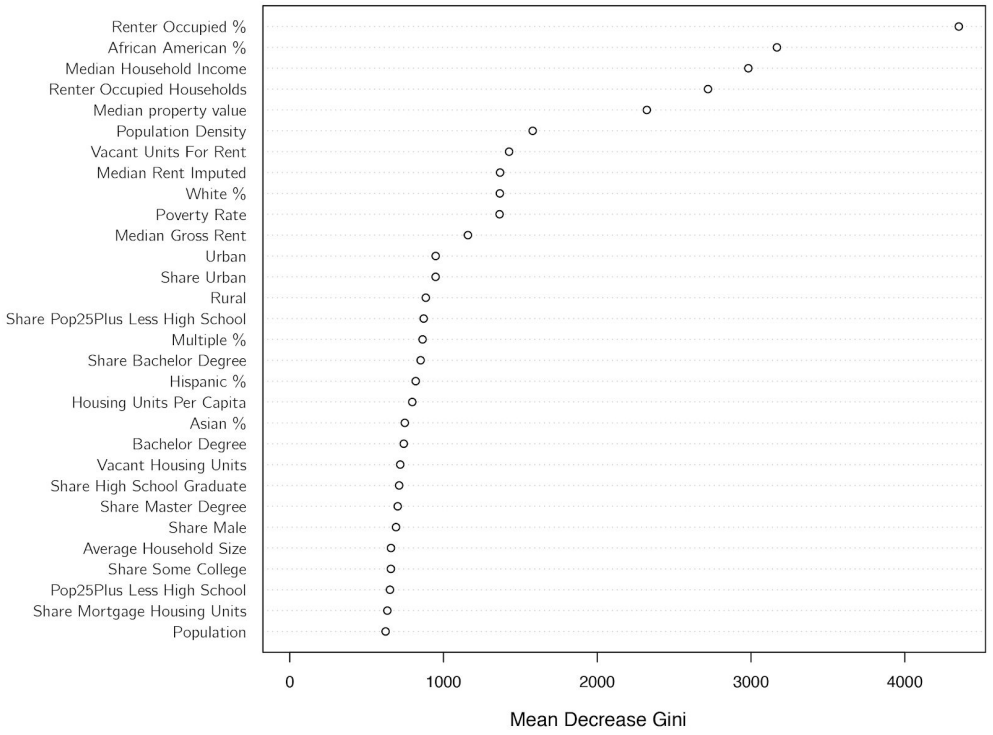
Figure 2: Eviction classification map



How your model works

Ensemble methods such as random forest models are a useful approach to discover complex relationships in the data. Although the Random Forest model had the best predictive performance, it also suffers from low interpretability. In other words, gaining a full understanding of the decision process and the variables chosen in each tree is unfeasible. To mitigate this, we use variable importance metrics that show the most important variables (based on predictive performance) used for splitting. Figure 3 plots the Mean Decrease Gini Index, which gives a rough estimate of the loss in prediction performance for each variable included in our model.

Figure 3: Random Forest Variable Importance



Fairness and Bias

Fairness and bias are two fundamental considerations when implementing machine learning models. Potential sources of biases and fairness in our model relate directly to issues of representation and discrimination, respectively.

On the one hand, our model might suffer from bias due to *who is* and *who is not* represented in the data. For instance, it might be the case that due to limited economic resources many tenants who are being forcefully displaced by their landlords might not be able to afford to file a court case. In this case, our model might be underestimating the number of evictions because we only observe the subset of individuals who can file a claim in court.

On the other hand, the data used in our model might also be missing one underlying mechanism by which evictions happen: the discrimination against people with different ethnicities or incomes. Although this paper addresses a prediction problem rather than aiming to understand the determinants of eviction, a shortcoming of the dataset we use is that it does not include illegal evictions. This means that high levels of discrimination in the residential market could be problematic because we might be underrepresenting the most vulnerable populations: those that are evicted and have no economic resources to take the case to court. In fact, estimates suggest that the number of acts of actual discrimination in the rental market may exceed four million each year and is higher than reported.[9] Given that people with different demographics are likely to be affected differently, our results should be interpreted as descriptive and exploratory.

[9] Report by the U.S. Department of Housing and Urban Development

Out-of-sample performance

We expect to observe moderate out of sample performance. On the one hand, we are predicting missing eviction rates during the same period and geographical context as in-sample. Since we are not extrapolating much, our model should perform adequately. On the other hand, however, we cannot rule out concerns regarding the external validity of the model because the places for which we are predicting out of sample (those that do not have evictions data) might not be random. In other words, it might be the case that places with no data also have a different relationship between the explanatory variables and eviction, which would compromise our prediction.

Finally, it is worth noting that a model with a good fit does not necessarily lead to a good out of sample forecast, and vice-versa. For example, overfit models will typically have small in-sample errors but will render poor results when forecasting. Our best diagnostic to mitigate overfitting is to use cross-validation, which can tell us what to expect when we predict out of sample. In this sense, it is reassuring that the error measures of the model in the validation period are similar to those of the estimation period.

5 | Conclusion:

Housing is one of the most basic and necessary elements in our life, which makes the eviction crisis a serious and pressing problem to tackle. Due to limited access to data on evictions our comprehension of the problem has been limited. To address this situation, in April 2018, the

Eviction Lab at Princeton University gathered data on evictions spanning all of the U.S, which reveal the top evicting cities. However, looking at the data in detail, we find that about 20% of the evictions for 2016 are missing. To tackle this, we first we construct a training data set by imputing missing data and combining historical and census data. Despite our efforts to clean the data, around 10% of eviction data was still missing. Therefore, we use tree algorithms to predict these missing values out of the sample. More specifically, we tested penalized logistic regression, Boosting, and Random Forest algorithms, and select Random Forest since its performance is the best (lowest error rate).

Based on our estimation and prediction, we created the eviction risk map shown in figure 2. This map shows which states, counties, and block groups have higher eviction risk, with the hopes that such visualization can be a first step towards enabling policy makers to design more efficient policies that mitigate the negative consequences of eviction. For example, governments can give priority to such high risk areas when allocating the budget to construct public housing. Despite our prediction efforts, it is worth noting that the data used is gathered from eviction judgments and does not contain the full eviction population. Therefore, if the data is systematically misrepresenting the evicted population, our prediction might be biased.

6 | Discussion

There is no one-fits-all policy in place to prevent evictions in the United States. Eviction laws are implemented locally at the state level as part of the residential renting regulation and vary by jurisdiction, which means that there is considerable variance in how each State addresses the eviction issue. What worsens the situation is that the US Census Bureau has not collected data on evictions, which has made limited access to high-quality, reliable data. Hopefully, access to new sources of data (such as the one provided by the Evictions Lab) can help researchers and practitioners examine the problem further to propose innovative policy solutions. The goal of this project was to give more clarity to decision makers, by predicting areas with high eviction rates with reasonable accuracy.

7 | Recommendations

The model developed in this project is useful to predict census blocks with high eviction rates. We constructed our model by considering both robustness of prediction and simplicity of interpretation. One direct application is for policymakers, activists, non-profits, and citizens, which can use the model's output to inform their decisions and direct their limited resources to affect positive change. Such applications range from improved funding allocation to targeted free legal counseling efforts.

For example, a targeted application would be to predict blocks that are likely to witness higher evictions in the coming year. Consider for example predicting future evictions that will take place in Virginia - a state that we have signaled as having amongst the highest

eviction rates. As a result, the state's budget office could allocate funding to those areas more efficiently, and work in tandem with local organizations to offer services that can mitigate the negative consequences of displacement.

Implementing these types of models will appeal to policymakers and local organizations alike. Policymakers will benefit because preempting the problem of evictions can reduce the cost and burden on the legal system to process thousands of evictions cases. Similarly, local organizations will benefit because the model will help them build a case-for-support to request funding for their initiatives.

Limitations:

Most prominently, this model is built solely based on evictions cases that ended up in court. Undocumented evictions are thought to outnumber documented ones greatly. As such, our prediction may not be an accurate representation of the reality on the ground. It is important to keep in mind the limitations inherent to Random Forest (mentioned throughout the paper) when evaluating the results of this model. Finally, it should be noted that our model is predicting an outcome, and as such it makes no claims regarding causation.