Prioritization of Pro-bono Legal Services to Reduce Evictions in the State of Virginia

By: Felipe Alamos, James Jensen, and Tammy Glazer

Table of contents

Background

Eviction refers to the process by which a landlord forcefully and legally dispossesses a tenant of the property that he/she had been occupying. Each year, over 900,000 evictions occur in the United States, affecting approximately 2.3 million people, many of which are elderly, children, and/or economically disadvantaged. Approximately 6,300 people are evicted each day in the United States. Lack of guaranteed and secure housing sits at the root of a variety of social problems, from poverty and homelessness to educational disparity and health care. Understanding the eviction crisis is critical to effectively addressing these problems and reducing social and economic inequality.[1]

Eviction raises a variety of financial, legal, and ethical concerns. For the tenant, eviction is a cause, not just a condition, of poverty. Eviction can lead to a loss of one's home, possessions, and employment, and can limit a person from obtaining government housing assistance in the future. Evicted individuals often relocate to poor and dangerous neighborhoods and have a higher likelihood of becoming ill and depressed or being forced into homelessness. In fact, for individuals who have experienced eviction, the likelihood of being laid off is roughly 15 percent higher, and the level of material hardship is 20 percent higher, compared to those who have not. Moreover, one in two recently evicted mothers report multiple symptoms of clinical depression. From 2005 to 2010, suicides attributed to foreclosures doubled.[2]

Furthermore, evictions are economically burdensome for the communities within which they most often occur. The Urban Institute estimates that evictions cost US cities from tens to hundreds of millions of dollars. When families cannot consistently pay their mortgages or rent, they are less likely to pay property taxes which feed directly into city budgets. According to Census of Government data, these property taxes account for roughly half of local own-source revenue.[3] Moreover, each individual experiencing homelessness can cost a city tens of thousands of dollars per year. For landlords, evictions not only result in a loss of income but are incredibly expensive and complex to adjudicate.

Finally, at its core, housing is a human rights issue. Human rights advocates would argue that all people should have the right to affordable housing. Homelessness is clear evidence that the government is failing to fulfill its role to protect and provide for its people. In order to ensure the wellbeing of vulnerable and disadvantaged individuals, reduce inequality, save local governments money that can be directly redirected to support constituents, and ensure that people's basic human rights are being met, it is imperative that a data-driven approach be used to prioritize how local governments use limited resources to effectively reduce the number of evictions occurring in cities across the United States.

---

[1] Eviction Lab
[2] Brookings Institute
[3] Harvard University

Goals

Our **project goal** is to identify the Census tracts in the state of Virginia at the highest risk of experiencing an eviction within the next 3 years, in order to prioritize the allocation of limited resources and ultimately reduce evictions. By lowering the number of evictions in the Census tracts at highest risk, **our policy goal** is to ensure the well-being of vulnerable and disadvantaged individuals, reduce inequality, reduce costs at a community-level, and ensure people's basic human rights. A Census tract is an area roughly equivalent to a neighborhood encompassing between 2,500 to 8,000 people, established by the Bureau of the Census for analyzing populations. To accomplish this goal, our objective is to help a nonprofit legal clinic prioritize where to use limited resources to most effectively reduce the number of evictions occurring in Virginia.

Based on a review of the current literature and the results of our analysis, we propose that a nonprofit legal clinic use the following intervention to reduce the rate of evictions in Virginia:

*Expand access to **pro-bono legal consultations with an experienced attorney**, regardless of income level, to those who are facing an eviction in identified high risk locations. Pro-bono legal consultation covers information on rights and responsibilities, entitlement to government benefits, and direct legal support for people representing themselves in court to better navigate the legal system.*

While tenants are often self-represented in eviction cases, the vast majority of landlords -- 85 to 90 percent, according to a 2017 Pew Charitable Trust study -- appear in court with a lawyer.[4] The addition of legal representation can drastically change the result, often without the expense of going to court. One study found that two-thirds of tenants who had legal representation were able to remain in their homes, as compared to only one-third of those who represented themselves.[5] By leveling the playing field between tenant and landlord, increasing access to legal representation is an effective way to reduce eviction rates across Virginia.

Related Work

Organizations offer pro-bono legal consultations with an experienced attorney to provide tenants with information on their rights and responsibilities as tenants, as well as their entitlement to government benefits. These services may include support for people representing themselves in Housing Court to help them better navigate the legal environment, and help to fulfill a tenant's

[4] Pew Charitable Trust
[5] HLR

right to counsel. Services may also include a legal phone hotline for expedited advice. Washington D.C., Los Angeles, and New York are also implementing a program that will provide city funded legal representation to those facing eviction. We view our work as the connective tissues between outreach organizations on the ground and city governments, ultimately ensuring that the city is assisting those most at risk of eviction. By identifying the top 10% of tracts most at risk of eviction, our model can focus outreach efforts on the areas most at risk, informing them of available services and working with them to navigate the eviction process.

Problem Formulation & Overview of Solution

While evictions are a national concern, the data in the state of Virginia are particularly striking. In 2016, there were 51,821 eviction in the state, and among all large cities in the United States, four of the top six cities that experienced the highest rates of eviction were located in Virginia. Specifically, 11.44% of households in Richmond experienced eviction followed by 10.49% of households in Hampton, as compared to a national average of only 2.34%. Given the gravity of the situation, this analysis specifically focuses on reducing eviction rates in Virginia using a localized methodology and intervention.

Given limited resources to both conduct outreach and provide pro-bono legal representation, the key metric we care about when evaluating our models is precision at 10%. By prioritizing precision at 10%, we will identify tracts that are truly at the highest risk of experiencing eviction (avoid false positives).

Data

Our main data source for this analysis will be a national database of eviction records collected at the tract level from The Eviction Lab at Princeton University, developed in 2018. These datasets are publicly available and can be accessed here: https://data-downloads.evictionlab.org/. Census data has been joined with formal eviction records for each state.

Datasets are available at several geographic levels of aggregation, including states, counties, cities, and blocks. The data contains information on eviction cases reported from 2000 to 2016. For the state of Virginia, the full dataset (considering all different levels of aggregation) contains 135,522 rows of data. We will work specifically with tract level data, which consists of 32,425 rows of data. Through data exploration, we discovered the following data challenges:

- Zeros and nulls appear to be used interchangeably, making it difficult to tell when values are truly 0 (ie. no evictions occur vs. missing data).
- Incomplete data - 14% of records have missing data on eviction rates from 2007-2015
- Socioeconomic indicators are only made available at the tract level once every 5 years, making it difficult to assess trends over time (eg. how has population shifted from 2001 to 2002?)

The following represent an overview of the 27 columns in the data:

- Name of location (either state, city, county, track, or block)
- Socioeconomic variables by area: total population, poverty rate, median household income, racial composition
- Housing variables by area: percentage of occupied housing units that are renter-occupied, median gross rent, median property value, median gross rent as a percentage of household income, count of renter occupied households
- Eviction variables by area: number of eviction judgments in which renters were ordered to leave, ratio of the number of renter-occupied households in an area that received an eviction
- Data collection variables: Imputed: boolean variable indicating whether eviction numbers and renter-occupied-households were imputed, low-flag: Boolean variable indicating whether eviction numbers are estimated to be lower than they really are

This dataset, complemented with Census data, capture the main variables we believe are necessary to predict evictions. In addition, it is sufficiently large. Data exists for 1,797 different tracts, and we have records for each tract over 17 years. Moreover, there are several options for feature creation: the datasource has a reasonable number of columns with information on socioeconomic characteristics, housing, and evictions.

The following statistics provide an overview of the eviction rate variable for 2016 throughout the tracts in Virginia. We observe significant variation across tracts. In addition, we notice that the top 10% of the tracts have an eviction-rate of 10.7%.

Solution: methods, tools, analysis, models, & features

Feature Generation

We began by generating 113 features from 25 original feature columns. The initial list of features refer to the economic and demographic columns found in the Eviction Lab dataset. Our feature generation expands upon this dataset, adding county-level aggregations by tract (eg. county-level population), binary variables that identify continuous variables over a specified threshold (eg. poverty-rate above X% by tract). For each of the continuous variables we generate features against thresholds of 1%, 5%, 10%, 25%, 50%, and 75%. A full list of our features is attached.

There are several features we were ready to generate in feature_generation.py related to change of variables in time. We finally were not able to use these features (that's why line 30 in feature_generation.py is an empty array) because:

- Census data does not have changes for periods of time less than 5 years.
- Generating features of change compared to 5 years or more in the past would mean having a too big "feature generation gap" (time needed to generate features). In other words, if we wanted to generate these features, we would have been forced not to train our models with data from 2001-2005 (we dont have data from 5 years before that to calculate the features). A future work would be to study the performance of models with less data but with these features, and compare with actual model (which is trained with more data but without these features).

Outcome label

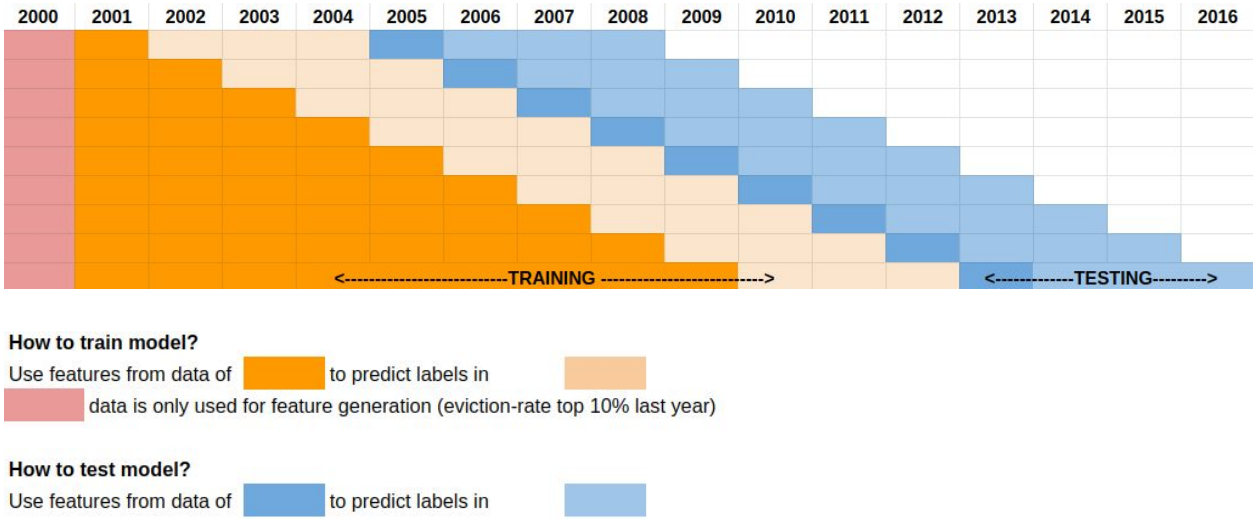
In our initial analysis, given that our last test set was for the year 2016, we aimed to build a model that would predict whether or not a tract would be in the top 10% of eviction rates for the year 2020. We predicted four years in advance because, given the needs of our partner organization, we wanted to build a model that would be immediately usable. However, upon further reflection we refined our model for two reasons: (1) the four year gap could render our model totally useless - data from last 4 years seems determinant for an accurate prediction, and

(2) the implausible fluctuations of yearly eviction rates for a given tract leaves us with an outcome variable that we cannot fully trust. To address both of these issues, we instead built a model that predicts if a tract will be in the top 10% of eviction rates in any of the next three years.

To account for the inconsistency in how Null values are handled in the data and the potential unreliability of single year eviction rates, our outcome variable is whether a tract has an eviction rate in the top 10% at any point within a 3-year period. If so, this tract receives a label of 1. For example, if x_train covers 2000-2003, y_train covers 2004-2006. If a tract is in the top 10% of evictions during 2004, 2005, or 2006, its outcome label = 1.

Temporal holdouts methodology

We used temporal holdouts methodology to split our dataset into 9 pairs of train and test sets. We present a diagram of how each train/test is composed.



Models

We used a variety of classification methods to make predictions: random forests, decision trees, logistic regression, gradient boosting, bagging, ada-boosting, and extend trees. Each of them was run with a set of different parameters.

Metric of focus

Our key metric of evaluation is precision at 10%. This is due to the fact that the intervention comes at a high cost, so we must be sure to identify tracts that are truly at the highest risk of experiencing eviction (avoid false positives). We also calculate recall, AUC, and F1, for a variety of different thresholds.

Baseline model

Our analysis contains two baselines against which we compare our models. (1) A simple dummy baseline that classifies all outcomes as 1. (2) A baseline that uses the prior year's outcome classification (whether or not the tracts was in the top 10% of eviction rates) to assign the outcome variable in a given year. By using the previous year's outcome as our baseline, we aim to determine whether or not a machine learning approach adds any value. That is to say, can we just focus on trend of outcome variable recent history to predict the near future, or do we need a deeper understanding of different characteristics among census tracts, and how those characteristics vary over time, to determine changes in our outcome variable.

Results

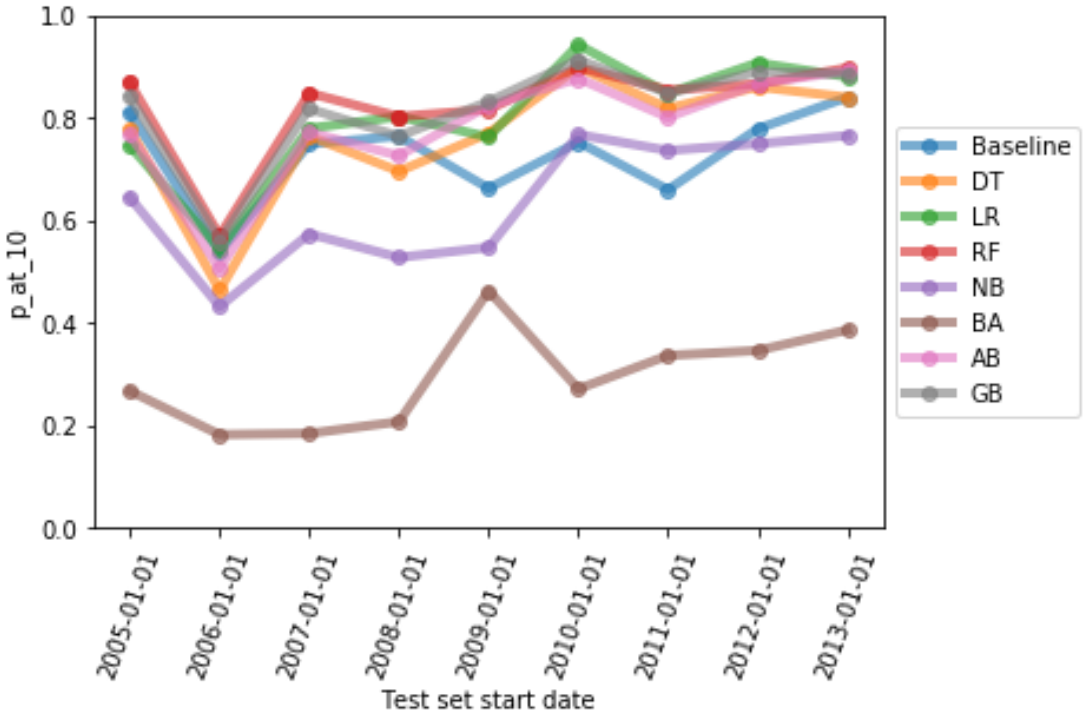
File predicted_tracts.csv contains our predictions of the tracts with the highest risk of being in the top 10% for any of the following 3 years.

We select based model based on precision at 10% as already explained. According to this metric, a Random Forest model, with parameters {'max_depth': 5, 'max_features': 'sqrt', 'min_samples_split': 2, 'n_estimators': 1000, 'n_jobs': -1}, performs the best for the most recent test set (2013 to 2016 data). Its precision at 10% is of 0.89, this means that 89% of the predictions of tracts being in their top 10% for any of the next 3 years were correct. We can also observe that this model performs relatively well throughout the different test sets starting points, so it's a reliable model.

Interestingly, Logistic Regression performs very well for the most recent test sets, but its performance is generally not that stable compared to others (particularly noticeable for 2009 test set.

Finally, we can compare the performance of the ML models with the baseline. Lets remember the baseline prediction consists of predicting that the tracts in risk for the following years are those that in the previous year where in their top 10%. Interestingly, the baseline behaves relatively well compared to other ML models. As a matter of fact, its precision at 10% for the last test set is very similar to Decision Tree. Nevertheless, we can see that our selected model, Random Forest, perform significantly better than the baseline and hece justify the ML approach to this problem.

Models performance on precision at 10% through time

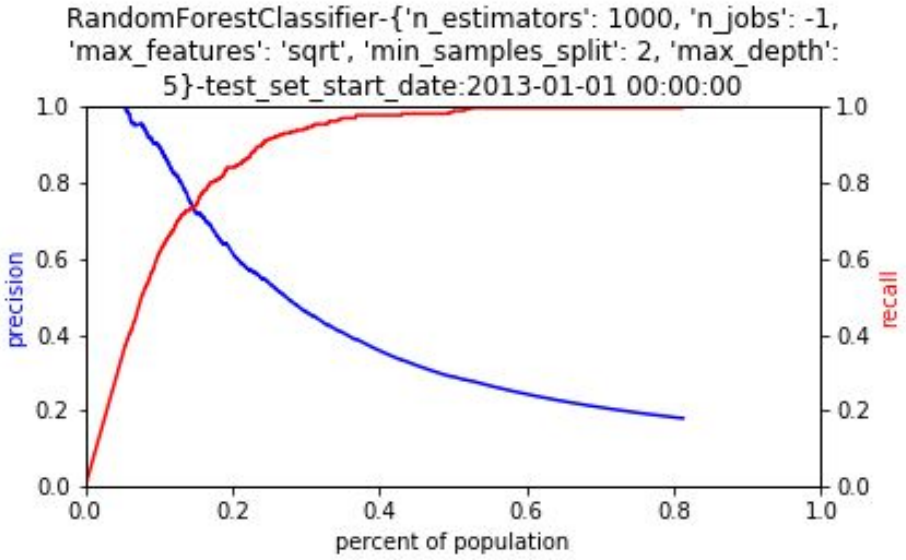


Best model according to precision at 10% for the different test sets

model_name	parameters	test_set_start_date	baseline	p_at_10	r_at_10	f1_at_10
RF	{'max_features': 'sqrt', 'n_jobs': -1, 'max_de...	2005-01-01	0.140202	0.867021	0.617424	0.721239
RF	{'max_features': 'sqrt', 'n_jobs': -1, 'max_de...	2005-01-01	0.140202	0.867021	0.617424	0.721239
LR	{'C': 0.1, 'penalty': 'l1'}	2006-01-01	0.111524	0.579787	0.519048	0.547739
RF	{'max_features': 'sqrt', 'n_jobs': -1, 'max_de...	2007-01-01	0.154198	0.847328	0.549505	0.666667
RF	{'max_features': 'sqrt', 'n_jobs': -1, 'max_de...	2007-01-01	0.154198	0.847328	0.549505	0.666667
LR	{'C': 10, 'penalty': 'l1'}	2008-01-01	0.145802	0.816794	0.560209	0.664596
AB	{'n_estimators': 100, 'algorithm': 'SAMME'}	2009-01-01	0.146154	0.823077	0.563158	0.668750
GB	{'n_estimators': 100, 'learning_rate': 0.001}	2009-01-01	0.146154	0.823077	0.563158	0.668750
LR	{'C': 10, 'penalty': 'l2'}	2010-01-01	0.170803	0.941606	0.551282	0.695418
RF	{'max_features': 'sqrt', 'n_jobs': -1, 'max_de...	2011-01-01	0.148040	0.852941	0.573123	0.685579
LR	{'C': 0.1, 'penalty': 'l2'}	2012-01-01	0.149533	0.906433	0.605469	0.725995
RF	{'max_features': 'sqrt', 'n_jobs': -1, 'max_de...	2013-01-01	0.146552	0.896552	0.611765	0.727273

Precision Recall Curve

For the selected and recommended model Random Forest we present the precision-recall curve. This curve tells us, for a given percent of population we want to intervene, what the precision is (of all tracts we predicted will be in the top 10% for any of the next 3 years, what fraction was effectively in that top 10%), and what the recall is (of all the tracts that were effectively in the top10% in any of the next 3 years, what fraction our model was able to correctly identify). We can observe that precision is relatively high for approximately the top 20%, but then rapidly decreases to the baseline value. Different values of recall can also be observed.



File predicted_tracts.csv contains our predictions of the tracts with the highest risk of being in the top 10% for any of the following 3 years.

Feature importance

Feature importance for the selected model can be found in feature_importances.csv. Most important ones listed here:

	Importance
evictions	0.1774633249
eviction-rate	0.1059220771
eviction-filing-rate_above_10	0.0791829379
eviction-filing-rate_above_5	0.0590898186
eviction-filings	0.0550825888
pct-other	0.0406758537
county_average_median-gross-rent	0.0301942632
pct-white	0.0276614664
county_average_evictions_y	0.0266628306
pct-af-am_above_25	0.0200995581
county_average_eviction-filing-rate_above_5	0.0196550675

Policy Recommendations

As mentioned, we recommend to expand access to pro-bono legal consultations, regardless of income level, to those who are facing an eviction in identified high risk locations. Pro-bono legal consultation covers information on rights and responsibilities, entitlement to government benefits, and direct legal support for people representing themselves in court to better navigate the legal system.

To identify the high risk cases, we recommend using a Machine Learning approach to prioritize the pro-bono legal consultations. We specifically suggest using a Random Forest classifier with 1000 estimators and max_depth 5. This will have a precision of almost 90%, which is significantly higher than a baseline approach that does not use machine learning techniques.

Attached (predicted_tracts.csv) is the list of tracts recommended to intervene.

Ethical Issues, Bias, and Fairness

Conducting a bias and fairness analysis using Aequitas highlights specific metrics for which the suggested model is imposing bias on given attribute groups, which must be considered as proposed policy interventions are evaluated. In this case, a few specific

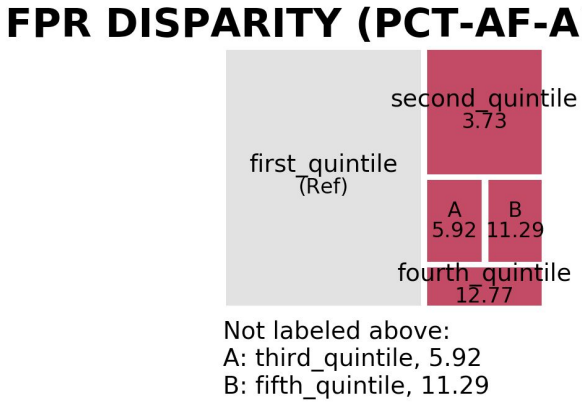
groups stand out as having high predicted positive ratios, indicating that they are predicted to be positive significantly more than other groups. This applies to the second quintile for median gross rent (low rent), the first quintile for median household income (low income), and the first quartile for percent white (low percent white), with 29% of tracts in this category predicted at risk. Because we are particularly interested in false positive rate, it is interesting to note that the fourth quintile for poverty (high poverty) had a higher than average false positive rate of 17%.

Disparities between groups are calculated as a ratio of a metric for a group of interest compared to a base group. In this model, it is very notable that tracts with the fourth and fifth highest quintiles for percent African American are falsely identified as being at a high risk of eviction at over 10 times the rate of tracts in the first quintile (low percentage).

Aequitas has an important get_disparity_major_group() method, which allows researchers to quickly understand how much more or less often groups are falsely or correctly identified as high risk in relation to the group we have the most data on. Using this metric, a few interesting elements in this data stand out. The analysis reaffirms the finding mentioned above regarding tracts with a high percentage of African Americans. Additionally, tracts in the lowest quintile for percent white are over 5 times more likely to be falsely identified as high risk as compared to tracts in the highest quintile for percent white. Tracts in the fourth quintile for rate of poverty are over 10 times more likely to be falsely identified as high risk as compared to tracts in the lowest quintile for rate of poverty.

Finally, based on a fairness analysis which leverages these parity determinations, it is important to note that false positive rate disparities between tracts in lower and higher racial group quintiles largely do not fall into the “fair” range. Relative to the base groups, the predictions do not provide supervised fairness to groups based on income, poverty rate, rent, or racial composition. These findings mark an inherent trade-off between false positive and false negative fairness, which is present in any decision system where base rates are not equal.

Figure: False Positive Rate Disparity – % African American



Limitations and Caveats

Several assumptions and analysis limitations must be considered in the evaluation and application of these results. First, it is important to note that our analysis is based on the number of evictions that have been formally recorded. This does not include evictions that occur informally, without a formal adjudication process. Therefore, our eviction rate predictions will likely be less than the true number of future evictions in a given area since the model will be trained on a subset of the true values. Additionally, we assume that households deemed “evicted” were in fact forced to dispossess their homes. This assumes that all judgements of eviction were effectively carried out. Because our interventions intend to help those at the highest risk of homelessness and being forced from their homes, it is imperative that our model is trained based on data for families that have actually been forced to dispossess their property.

Furthermore, our analysis assumes that all tracts are equally receptive to the proposed interventions. In other words, we assume interventions are equally effective in different locations. We must assume this in order to prioritize areas based solely on a predicted eviction rate. By contrast, it could be the case that areas with high predicted eviction rates are not actually receptive to the suggested interventions, making it more effective to intervene in other areas with lower eviction rates and higher receptivity. We do not have data on receptivity to the proposed interventions, and hence, we need to assume receptiveness is constant.

Finally, we assume that we should target tracts with the highest risk of eviction, rather than locations at a lower risk that may benefit more from minor interventions than areas experiencing a myriad of setbacks.

Our dataset is limited to evictions that have occurred between 2000 and 2016. This is important to note, because we are using this dataset to draw conclusions about evictions that will occur at least three years in the future (2019). Therefore, we are assuming that no significant housing policy or socioeconomic changes have occurred since 2016 that would shift the results of our analysis. Given additional time and resources, it would be interesting to expand this analysis by leveraging data from more recent years. In addition, as was mentioned above, a future

improvement would be to better understand when a value of zero actually represents an eviction rate of zero instead of null. To do this, we could have conversations with the people collecting the data to understand and verify their methods.

Future Work

We know there are several limitations with our current approach, the first being that the most recent year in our dataset is 2016. To further improve upon our work, we wish to: update the model with the most recent data (2017-2019) when it is released; conduct and compare the same analysis on other high risk states; Consider alternate interventions, including tenant educational programs, affordable housing, counseling, and public policy changes; consider alternative evaluation metrics based on the selected intervention; further investigate the use of 0s and Nulls to capture when 0 actually indicates 0 and when it does not, and it impacts on our predictions; leverage additional ACS features for further exploration; review how we can improve the model and make changes to our proposed intervention based on the results from aequitas.