Reducing Evictions by Prioritizing Housing Legal Aid in Chicago Communities

Using Machine Learning to predict next year's eviction rates for Chicago census tracts

Camilo Arias M.
Chi Nguyen
Angélica Valdiviezo I.

Table of contents

Background and goals

There is an eviction crisis going on in the United States. Matthew Desmond's book "Evicted: Poverty and Profit in the American City" shows that eviction causes poverty and also is correlated with other public issues such as mental health or job loss. The lack of legal aid for evicted tenants were one of the main factors that contributed to eviction and exacerbated its impact. We believe that prompt legal support could reduce the eviction rates in low-income communities, by helping tenants exercise their legal rights and, in some cases, avoid eviction.

According to the ACLU[1], many states in the US still lack legal framework to protect tenants from unjust evictions. Under this circumstance, landlords can act with impunity by unfairly and informally threatening eviction. On top of it, in some states including Illinois, personal eviction records can be accessed and could damage future housing applications. Aware of that, tenants may avoid a case by all means and accepted being evicted without exercising their legal rights. In these vulnerable conditions, prompt legal assistance could dramatically improve the outcomes for at risk populations.

Getting free legal aid to those in need might be an effective solution to the eviction crisis. In 2017, the New York City Council passed legislation (the Universal Access law) that made the city the first in the country to commit to provide legal service available for all tenants facing eviction. Though still in early stages, the policy seems promising. For eviction cases from mid 2017 to mid 2018, 84% of the households that received legal assistance from the Office of Civil Justice were able to remain in their houses, "not only saving thousands of tenancies, but also promoting the preservation of affordable housing and neighborhood stability." Bearing that in mind, the need for legal aid plays a fundamental role in tackling eviction.

Our project will provide tenants in neighborhoods with high eviction risk with free legal assistance. According to Matthew Desmond, "eviction is a cause, not just a condition of poverty." Therefore, in reducing evictions, the ultimate policy impact we are aiming for is to limit conditions leading to poverty through empowering individuals and communities. Furthermore, the result of our project, if up to our expectations, will allow advocacy groups to make the case showing the correlation between legal assistance and reduction in evictions, which has already been observed in New York after the Right to Counsel bill was passed in 2017. Therefore, this project will potentially have lasting impact on the state's housing policy.

We have two policy goals to address. Our immediate goal is to identify neighborhoods that faced the highest eviction risk so that we could distribute our legal outreach and assistance most effectively to those most in need. Through working on our immediate goal, we hope to achieve the long-term goal of reducing the absolute number of evictions in Chicago neighborhoods.

[1] Source: Park, Sandra (2017). Unfair Eviction Screening Policies Are Disproportionately Blacklisting Black Women (https://www.aclu.org/blog/womens-rights/violence-against-women/unfair-eviction-screening-policies-are-disproportionately)

Related Work

Sociologists and policy scholars have long been studying the characteristics of neighborhoods having high eviction rate and the impact of eviction on communities and individuals. Harvard professors Matthew Desmond and Carl Geshenson published a study on eviction in 2016 on the Journal of Social Science. In this study, they applied discrete hazard regression models on the Milwaukee Area Renters Study (MARS) dataset to measure the risk factors leading to eviction for individuals and households. They found that having an additional child and being unemployed were both associated with an increase in the probability of getting evicted. Further, neighborhood eviction rate, crime and social network disadvantage were all good predictors of individual evictions.[2]

University of Washington researchers, Timothy Thomas et al, used the Washington State's eviction court orders from 2011 to 2017 to study the characteristics of those being evicted. The study turned court documents into data for the first time in the state and used geocoding to match individual record with neighborhood geo-identifier. The study found that almost half of the eviction cases ended up in a default judgment which resulted from a no-show from tenants on court day. The study also accentuated the evidence from New York City that defendants with legal representation were twice as likely to stay in their homes. The authors further pointed out that black renters were 5 to 6 times more likely to be evicted than white renters.[3]

The study that is the most related to our project was from Charles Solberg, who recently presented at the American Association of Geographers. The author used a random forest model to predict where tenants are most likely being displaced in New York City, to identify the most predictive factors for tenant displacement, and examine their heterogeneous impacts.[4]

Previous studies mainly focused on the characteristics of evicted tenants and primarily used traditional methods such as regression to find predictors of eviction. Very few academic papers had leveraged machine learning to study eviction. We were unable to find any such studies that focused exclusively on eviction issues in Chicago. Our project aims to fill in that gap by using machine learning methods to predict eviction rate in Chicago neighborhoods to inform distribution of legal aid resources. Additionally, we are using a rich eviction dataset containing court record information that was just published by the Lawyers' Committee in May 2019.

Machine Learning Problem Formulation and Solution Overview

Our policy goal for this project is to prioritize our limited resources to distribute legal aid services to the neighborhoods tracts that mostly need it. We defined a neighborhood with "the most need" as one where residents are the most at risk of being evicted from their homes. Additionally, since our intervention, educational outreach and legal case filing, will be estimated to take a year to complete, we are going to evaluate the neighborhood's legal aid needs for the next year. Finally, as a non-profit organization, our client's budget is constrained. They have

[2] Source: Desmond, Matthew et.al (2016). Who gets evicted? Assessing individual, neighborhood and network factors. (https://scholar.harvard.edu/files/mdesmond/files/desmondgershenson.ssr_.2016.pdf)
[3] Source: Thomas, Timoty et.al (2019). The State of Evictions: Results from the University of Washington Evictions Project (https://evictions.study/index.html)
[4] Source: Solberg, Charles (2019). Predicting eviction and displacement in New York city (https://aag.secure-abstracts.com/AAG%20Annual%20Meeting%202019/abstracts-gallery/22304)

established that they face almost fixed costs por neighborhood they intervene in. Based on the financial analysis, they can only intervene in 75-80 tracts. Since the City of Chicago has +800 tracts, we are looking to identify the top 10 % of the tracts in need. Based on our definition of "need," our window of intervention, and financial resources, we will narrow down our policy goal to predicting whether a certain census tract will be in the top 10% of areas with the highest eviction rate in the next year. On identifying the tracts being labeled as positive, our client will conduct outreach and legal assistance to residents in those tracts.

Since our goal involves a prediction task and since we have access to a really rich set of data on eviction from the Lawyers' Committee for Better Housing and demographic statistics by tract level from the American Community Survey as the Chicago Open Data Portal, we are going to apply machine learning methods and evaluate to see whether it would offer an optimal solution to help achieve our policy goal. We are going to use five supervised learning classification methods to predict whether a census tract's next year eviction rate will be in the top 10%:

- Decision Tree
- Logistic Regression
- Random Forest
- Bagging
- Gradient Boosting

To solve our prediction problem, we will define and create rows (see Data Description), define and create features and label (see Details of solution). Using preprocessed data, we will vary the parameters for each of the classification models and apply them to the train-test sets created from a temporal hold-out method (see Details of solution). For each model, we are going to calculate the following metrics: precision, recall, f1 and auc-roc. Since our top concerns is maximizing the efficiency of our budget, we are going to pick the best model based on precision at 10%. This strategy will allow us to make sure that every dollar spent on intervention will have the highest likelihood of being used on the right tract (meaning it will end up in the top 10% eviction rate of next year).

Data description

The final DataFrame we used was obtained has 6375 rows. Each row represents a tract in a given year (from the period 2010-2017) and includes 921 features which describe aggregated information about education, evictions cases, demographics and building violations. Features also include information about neighbor tracts, as aggregated by community areas.

This project used three main data sources, described below.

Chicago Data Portal

The Chicago Data Portal is the City of Chicago´s effort to provide public access to relevant information for the city. We identified two datasets that were relevant for the project.

1) The Crime dataset. It presents information from the Chicago Police Department about reported incidents of crime. Each row represents a record, and includes 22 variables,

most of which are information about the location. The dataset also offers information about the date and type of the reported crime, based on the Illinois Uniform Crime Reporting codes. This dataset was included since crime reflects the social fabric of a community, which is the same dynamic that drives eviction. The dataset has 6.9 million rows and is updated daily.

2) The Building Violations dataset. It presents information from the Department of Buildings about violations issued by the latter. Each row represents a violation, and includes 26 variables. These variables describe the violation, its status, the inspection (if any) and the location of the building. We included this dataset because it describes building and landlord behaviors, which could be related to eviction filings. The dataset has 1.7 million rows and is updated daily.

American Community Survey (ACS)

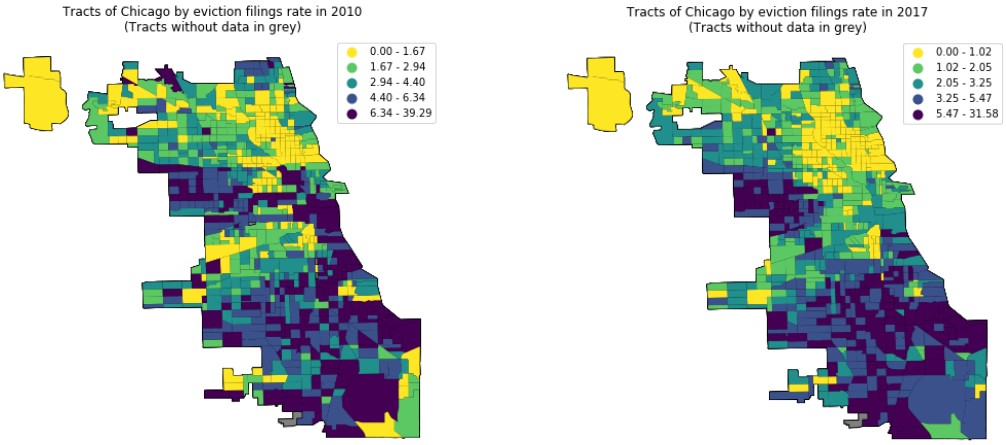
We included information from the ACS to obtain demographic data (education, race, poverty, housing units) of the tracts. We had to use the 5-year estimates over our period of interest, since those estimates are the ones that include the details tables that include the geography we need (tracts). We could only use the 5-year ACS data available at the time of prediction for all of our models. We ended up using the 2010 estimates. Therefore, the data has low variability.

Chicago Evictions Data Portal from the Lawyers' Committee for Better Housing

The Lawyers' Committee for Better Housing (LCBH) is a non-profit law firm that focuses on low and moderate income renters in the Chicago area. They provide free legal and supportive services while advocating for the rights of all renters. In May 2019, they launched the Chicago Evictions Data Portal, which included the first release of data. The data was obtained after LCBH reviewed nearly 300,000 eviction court records for the period 2010-2017. The dataset has almost 40 variables that present aggregated values for tracts. The topics covered include counts for eviction filings (total, rate), if landlords sue for eviction and/or rent, back rent sought, landlord and tenant legal representation, and if the filing ended in an eviction order, among others.

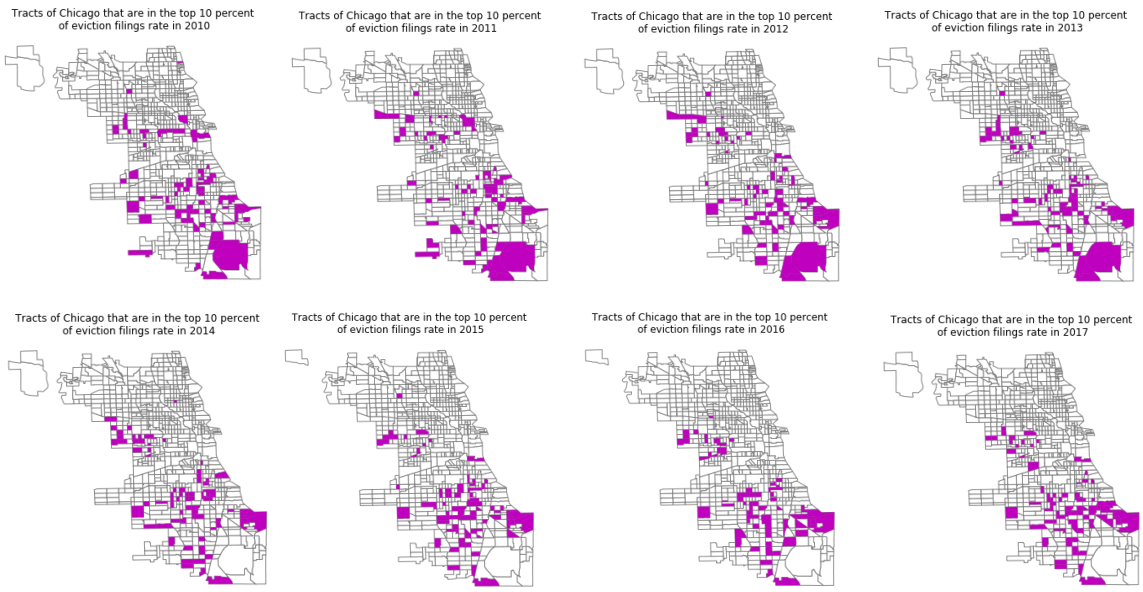
Using geodata, we visualized the tracts with high eviction rates to observe their distribution and change over time. Figure 1 shows that the tracts with the top rates (darkest) were steadily distributed in the South and West Sides of the city, while the North and Center region had tracts in the lowest quantiles.

Figure 1: Eviction Filings Rate per quantile in 2010 and 2017



Additionally, as Figure 2 shows, the tracts within the top decile of evictions filings rate tended to be in the same areas, the South and West Sides.

Figure 2: Tracts on the Top Decile of Evictions Filings Rate by Year, 2010-2017



Finally, when we look at the tracts that have been in the top decile of the eviction filings rate each year, we can notice that there are some tracts that are consistently in the period of interest. We found that there are 47 tracts with that are on the top decile for at least four years in the period 2010-2017. If we increase the number of years, we found 11 tracts with that are on the top decile for at least six of the eight years in the period.

Details of solution

The eviction data at the tract level was merged with the demographic tract level estimates from the American Community Survey. Since we only downloaded 2010 estimates, all the ACS

features are constant within tracts. In addition, we also merged to the master dataset the information about crimes and building violations. To do it, we first aggregated the building and crime counting the number of events by tract and by year. For example, we obtained the number of crimes and the number of crimes by type of crime for each tract and year. Since we had year to year variation in these two datasets, we calculated the year to year percentage change to capture the dynamics of the neighborhoods. Then, we merged the eviction data with these two DataFrames.

The resulting data frame contained information by tract level and year. For the missing values, we imputed them using the median since median value is less vulnerable to outliers. Then, we identified the community area of each tract. Thinking that eviction may be strongly correlated spatially, we decided to include in the analysis the average value of all the variables in its community area. For the feature creation, we discretized every variable in three categories: low, medium and high by splitting the range of data into three bins.

In terms of the outcome label, it was key to consider that at the moment of prediction we only have data from the previous year. Thus, our outcome label was 1 if a certain census tract will be among the tracts with the 10% highest eviction rate and 0 otherwise. For example, we use a tract's eviction rate from 2012 to construct the label for 2011.

We built five types of classifier models: Decision Trees, Logistic Regression, Random Forests, Bagging and Gradient Boosting, each one with different parameters. We used a moving window to test and train out models. We trained the first model using 2010 to 2012 data, and we tested with 2013 data (which had the label from 2014). Then, we kept moving the test data one year, which resulted in four training and testing, that can be seen in the additional material file. To convert the scores of out models into a predicted label, we sorted the scores from highest to lowest and made a cut off so that we classified exactly 10% of the tracts with the highest predicted score as 1 (positive).

To be able to evaluate out precision, we created two different baselines: a random baseline and a baseline using previous data. For the former one, we analyzed the scenario of predicting the highest tracts at random. Given that we want to predict the highest 10%, the precision of the random baseline is consistently of 10% at any threshold. The second baseline was the sensible exercise of picking the same tracts that were among the top 10% this year as our prediction for next year. Since eviction is a fairly consistent phenomena, this baseline had a substantially higher precision than the random baseline and remained in the range of 45-50% over the years.

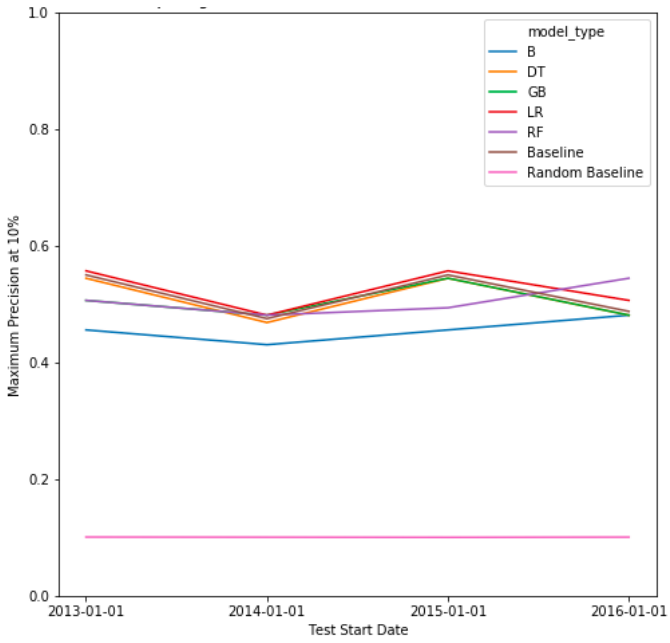
Evaluation

For each model that we built, we stored four metrics for evaluation purposes: precision, recall, f1 and area under curve. Because the client organization is constraint on budget, we prioritized precision among the four metrics because it measures the effectiveness of each dollar our client spends on intervention. Since the budget will be enough to cover intervention effort in the top 10% of census tracts in Chicago, we would classify the 10% of the tracts with the highest predicted scores as positive (to intervene). Given our goal and constraints, the best performing

model is the one with the highest precision at 10%. Since we are performing cross validation on train and test sets split on different times, among the top 10 best performing models based on precision at 10%, we would choose the one that performed the most consistently and whose performance improved over time.

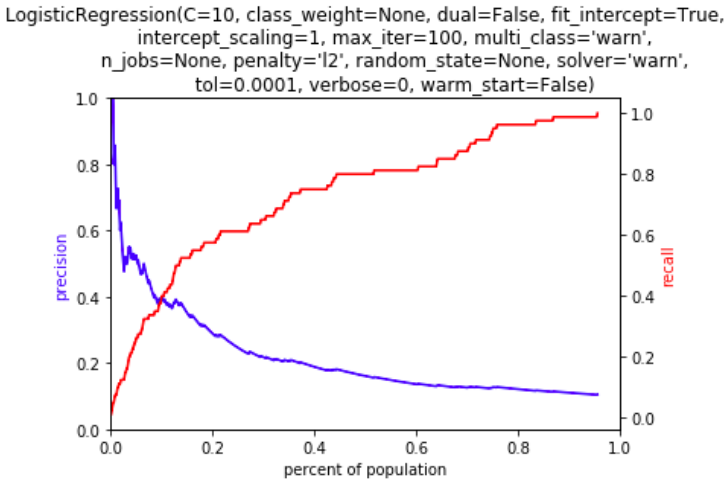
For each classifier, we calculate the maximum value of precision at 10% across all variation of the parameters for that classifier and visualize its performance over time. Figure 3 shows this comparison. All models performed better than the random baseline, which is 10%. We also constructed a heuristic baseline based on the tracts in the top 10% highest eviction rate from last year. Comparing to this heuristic baseline, only logistic regression's best models and random forest (trained on 2010-2015 data) performed better.

Figure 3: Precision performance of models and baselines



As we can see in Figure 3, When the models were trained on the data in 2010 to 2012, 2010-2013 and 2010-2014, Logistics Regression consistently outperformed the other models in terms of precision. Interestingly, when trained on the data in 2010-2015, Random Forest returned the highest precision despite having lower precision in the preceding years. Overall, however, logistics regression still had the most consistently high precision records across time.

Figure 4: Precision and Recall Curves for the Best Model



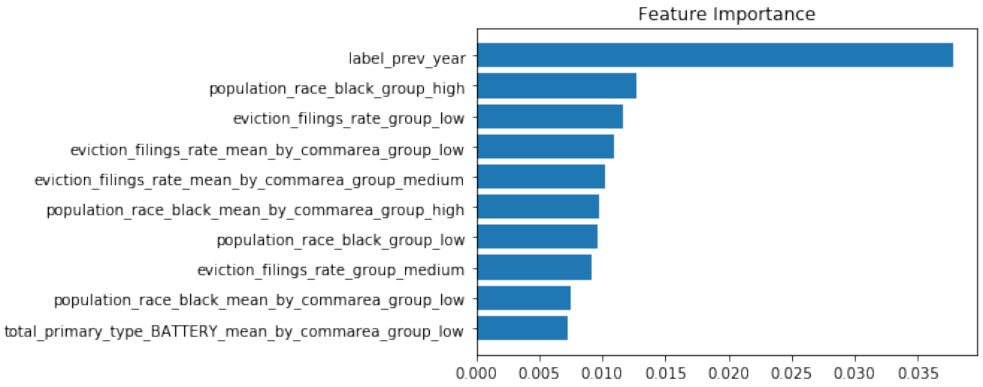
We trained 652 models in total. We ranked all the models based on precision at 10% in descending order and found that our best performing model was Logistic Regression with parameters C = 10 and penalty level l2. The model's precision-recall curve can be seen in Figure 4:

Discussion and interpretation

Our best model's precision score is 55%. This means that our of all tracts predicted to be among the top 10% highest eviction rate next year only 55% were correct. In other words, for each $100 spent on intervention, we should expect around $55 to go to the correct neighborhoods that are, by our definition, the most in need of legal assistance. For the same model, recall score is also 55%. This means that our intervention covered more than half of the tracts that needed intervention.

In terms of the feature importance of most of our models, the most important feature we had was the indicator of being among the top 10% of eviction this year. For the City of Chicago, this means that eviction is a fairly stable event, and that the most affected tracts do not change much within years. Since our best model was Logistic Regression, the importance of each feature is not calculated as straightforward as for the other classifiers. However, if we see the feature importance Figure 5 presents of the model that we identified as second best, it can be seen how the previous year label was very relevant.

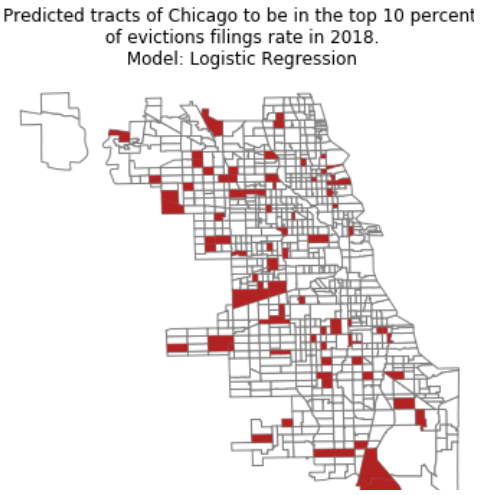
Figure 5: Feature Importance of the Top Ten Features of the Best Model



Policy Recommendation

Our model provides a list of tracts we predict will be in the top 10% of evictions filings ratings in 2018, which is one year after our dataset finished. Figure 3 presents a map of the predicted tracts. The full list is available in the folder of the project.

Figure 3: Predicted tracts



In the above sections, we have compared the model's performance to the random baseline and, most importantly, to the heuristic baseline . This analysis is relevant for our client, as we have to evaluate how a Machine Learning approach to the problem could be useful for them in the long run and, if so, what recommendations we could provide to implement it and work further for the policy goals.

As we have seen previously, our selected model performs just slightly better than the heuristic baseline, as measured by precision at 10%. Our recommendation is for the organization to test the Machine Learning approach for at least two years and see how it compares to the same-as-last-year baseline. The reason for this is that we do not believe a slightly more precise Machine Learning model is reason enough to recommend the full implementation of this approach. In our opinion, the gap between the heuristic baseline and the Machine Learning approach is not large enough to discard either approach.

Through those two years, we could recommend to follow either approach's prediction of the tracts to intervene to. We would suggest recording high quality data of the free legal aid interventions, as to build a stronger design to potentially test the causal relation between legal aid and eviction reduction, as well as the results of a Machine Learning approach in predicting the most at risk tracts. Additionally, as discussed in the section Limitations, caveats and future work, we suggest a few options to explore to look to improve the performance of a Machine Learning solution.

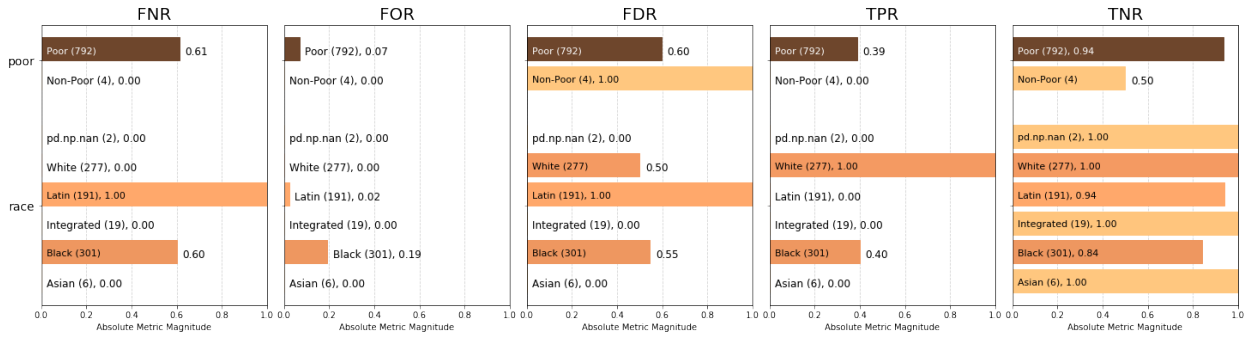
Ethics

Since we are working with many demographic features, many of which come from the American Community Survey, we would like to conduct a bias and fairness analysis to understand our model's limitations and make sure that it will be applied in the right context. Our team used an open source package, Aequitas, developed by a team of researchers at the Center of Data Science and Public Policy at the University of Chicago. Among the metrics offered by the package, we would focus on False Negative Rate and False Omission Rate. We want to focus on False Negative Rate because we want to be aware of the percentage of census tracts with

high eviction rates that our model predictions missed. Second, we want to focus on False Omission Rate because this metric informs us of the percentage of tracts that should have been intervened among the tracts marked as not intervene. Overall, these metrics will allow us to evaluate our models on the vulnerable neighborhoods that will not be covered by our predictions. A high measurement for either metrics will be a call for concern.

Below is the bias metrics produced by Aequitas for our best model, Logistic Regression. A positive neighborhood is one that has eviction rate next year in the top 10% highest in Chicago. In terms of poverty, the model produced predictions that missed 61% of the true positive poor neighborhoods. In terms of race, the model produced predictions that missed 100% of the true positive Latin neighborhoods and missed 60% of the true positive black neighborhoods. This measurement does not necessarily mean that our model was biased against poor and black tracts. The False Negative Rate was higher for these neighborhoods because the group of true positive tracts tended to contain a lot more poor and black tracts than others, and there were not enough nuances in our features to distinguish one black or poor tract from another yet. Looking at the False Omission Rate, 7% of poor tracts and 19% of black tracts were missed by our models. For both poor and black tracts, the fact that false negative rate is bigger than false omission rate suggest that the number of true positive tracts is smaller than the number of predicted positive tracts that is produced by our model.

Figure 6: Bias Analysis for the Best Model



Limitation, caveats and future work

In this section, we discuss the limitations and caveats of our project and the data it uses. Our main concern is that the goal is this project is to reduce evictions in Chicago. Data and, thus, the analysis is only considering *formal* eviction, by aiming to predict eviction filings whose records were complete. This could leave behind informal evictions that do not go through a court record. The main concern is that this omission could potentially impact worse vulnerable populations who are already more affected by the eviction crisis.

A second concern is that the assumption that free legal aid and educational programs have a causal relation with evictions. The objective of the prediction is to be able to provide our client, the Chicago Housing Legal Aid Association, a list of tracts that, based on our prediction, whose evictions filings rate will be in the top decile. Then, the organization can focus their efforts in the

most needed populations and that will have a negative impact on evictions. This cannot be affirmed unless evidence supports it.

In terms of the data caveats, for the Crimes dataset of the Chicago Data Portal, we identified two potential concerns. The first, which concerns us the least, is that not all reported crimes necessarily represent a crime. The second, of higher relevance, is the dark figure of crime. Moreover, we suspect this figure is not random, which can introduce bias to the data. On the other hand for the eviction data from the Lawyers' Committee for Better Housing, the first concern is for replicability. This dataset is the first release, but there is no clarity of when further releases will happen. Additionally, the dataset is already preprocessed and has no missing values. Their methodology did not specify how data were imputed. Finally, for the ACS, we are aware of the limited variability of it and how that affects the ability of the models to learn from it.

In terms of the analysis, the main concern is that we do not have formal methods to break ties between tracts. This means that we can have a large group of the tracts with the same predicted scores and include randomly a few of them, as we can only label the top 10%.

Among our ideas of future work to continue improving the project, the most relevant are the following:

- Improve the label. Our goal is to reduce the absolute number of evictions, so we want to produce a label that takes into account both eviction rate and the population size of the tract. Alternatively, we can do an analysis post prediction that normalizes by the population size.
- Improving features. As discussed, ACS statistics have limited variability, so we could use a predictive model to calculate demographic data for each year based on historical ACS statistics.
- Smaller size. The project could be analyzed from a smaller geography, such as block, and see if the model improves.