Using Machine Learning to Predict Evictions in Cook County

Final Report

Authors:

Jessica Langston
Liza Soriano
Luis Eduardo San Martin

BACKGROUND: Policy Problem

According to our own analysis of data from the Eviction Lab at Princeton University, eviction rates and eviction filing rates across block groups have returned to pre-great recession levels in the City of Chicago. Nevertheless, they disproportionately affect African American population and families living in poverty[1].

According to Matthew Desmond, the author of Pulitzer Prize-winning book "Evicted", evictions are both a symptom and a source of homelessness and poverty in America. In Cook County, many residents are displaced from their homes as rents rise and more high-income earners move into Chicago and the surrounding suburbs.2 Evictions because of property foreclosures, demolitions, or evacuations due to building neglect are also occurring in the county.

This housing crisis in Cook County is disproportionately affecting low-income, minority residents. The vast majority of low-income households in Chicago are rent-burdened (i.e. rent is more than 30% of household income): 84% of households earning 30-50% of Chicago's median income are rent-burdened, while those earning less than 30% of the median income are severely rent-burdened (rent is more than 50% of their household income).3 As a consequence, higher numbers of evictions occur in areas with more poverty, frequently coinciding with locales in the South and West sides of the city where African American and Latino populations traditionally reside. A study shows that the predominantly Latino neighborhood of Pilsen has lost about 10,000 Latino families to gentrification since 2000.4/5 More than 200,000 African Americans have left Chicago during the same time period.6

Lack of affordable housing along with stagnant income rates contribute to conditions that force families to leave their homes. Evictions can further exacerbate this problem because they are put on an individual's court records, which many landlords screen for, thus making it harder for evicted families to secure housing in the future. Therefore, minimizing the adverse impacts of evictions should be a priority for policy makers and organizations tasked with ensuring housing justice for Cook County's residents. Addressing this problem would also be in the interest of entities that want to reduce racial inequalities in the county.

Illinois laws and several Chicago ordinances decently protect tenants' rights in Cook County. These guidelines limit acceptable reasons for evictions, prohibit discriminatory or retaliatory actions by the landlords towards tenants, and in many cases place the burden of following these stringent rules upon the landlords.7 However, when tenants are unaware of their rights or do not have the resources to adequately deal with landlord conflicts, improper and illegal evictions can occur. Free resources provided by the city, along with outreach programs by many housing rights advocates and NGOs therefore try to fill this information gap and assist tenants in dealing with abusive landlords. These services can help tenants stay in their homes or at least lessen the impact of displacement.

[1] See Annex 2 for more information.

INTERVENTIONS: Metropolitan Tenants Organization

The Metropolitan Tenants Organization (MTO) is a coalition of community organizations and individuals located in Chicago, with a mission to "educate, organize, and empower tenants to have a voice in the decisions that affect the affordability and availability of safe, decent and accessible housing."[7] To achieve this mission, MTO provides a wide variety of services: they organize communities, host outreach events, and have free resources for tenants to use in case of any housing-related issues.

MTO sends out community organizers to build relationships with community members and property management. They meet with tenants about issues in their building and help them interact with landlords about properly addressing those problems. They encourage tenants to build tenant organizations to more effectively deal with landlord conflicts. Finally, they involve tenants in social justice campaigns to fight for policies that better protect tenants' rights or improve the affordable housing conditions in the county. MTO has advocated for the successful passage of the Keep Chicago Renting Ordinance (passed in 2013; new owners of foreclosed properties must provide lease renewal or a lease extension to current tenants, or otherwise pay relocation assistance of about $10,000 per family), the SRO Preservation Ordinance (passed in 2015; puts a freeze on closures of single-room occupancy buildings, which are essential to Chicago's affordable housing stock), and an amendment to the county's Human Rights Ordinances (passed 2013; prevents landlord discrimination against Section 8 voucher holders).

Furthermore, MTO hosts leadership development workshops to empower renters to build tenants organizations and fight for their rights. They also engage tenants in public policy forums with elected officials, and distribute educational materials about housing. Lastly, MTO provides a hotline and a phone app that anyone can access for free. The hotline allows people to call in with questions and concerns about landlord conflict and their rights as tenants. MTO can then assist callers in developing strategies to solve their problem or refer them to the relevant agencies. The "Squared Away Chicago" app facilitates communication between landlords and tenants, allowing for faster resolution of conflicts and increased accountability on both sides. The app provides tenants the ability to take photos of issues in the building/unit, specify in detail the problem through a letter template, and report it to their landlord.

The MTO's solutions to addressing the housing crisis in Cook County is unique because they approach issues on multiple levels. They engage individual residents and landlords, but they also work to promote sweeping policy changes that would affect all tenants in the county. More importantly, their aim to educate and empower residents ensures that their mission propagates and that their efforts have longer-lasting effects on the housing conditions in the county.

SUMMARY GOAL: Objective & Problem Formulation

The Metropolitan Tenants Organization is largely a volunteer group that aims to serve the second largest county in the country. Their limited resources could therefore have the most impact if primarily directed towards communities who are in most need of housing-related assistance. The objective of this project then, is to *advise the Metropolitan Tenants Organization about where to focus their efforts in the upcoming year*. More specifically, we would give them a list of block groups in Cook County and recommend that they prioritize these areas for the provision of services. This could mean sending out community organizers to those blocks first, concentrating advertisements about MTO resources (hotline, Squared Away Chicago app, educational materials about rights, etc.) there, or specifically engaging the policymakers and elected officials from those areas about affordable housing and tenants' rights.

We achieve this task by *building a supervised machine learning model to identify the top 10% of Cook County block groups most vulnerable to evictions in the next year*. We use the number of evictions that occured within a block group for one year as the target variable (i.e. indicator of eviction vulnerability), and we focus on the top 10% to take into account MTO's limited resources. Ultimately, the goal is to cause a significant reduction in the number of evictions in Cook County by buttressing MTO's work with data-driven recommendations, and with the primary motivation of mitigating the adverse effects of displacement. Thus, even though we are specifically advising MTO, the models we produce would also be useful to government agencies and any other NGOs working to address Cook County's affordable housing crisis.

DATA DESCRIPTION

We used data from the Eviction Lab at Princeton University (EL). It includes demographic and housing spatial information at the block group level for the Cook County. The timespan of this dataset ranges from 2000 to 2016. We decided to focus on the five most recent years, 2012 to 2016. This decision was motivated by two reasons: (i) Cook County data from before 2007 was inconsistent, and some of the years had a very small number of evictions in the whole County; and (ii) we wanted to keep some validity over time of the data we were analyzing, specifically by not having a large range of years in our data.

The first data exploration we conducted on the EL dataset showed it was plausible to use it for the purpose of our project. Each data point in this dataset corresponds to a unique block group / year combination, and it includes a column for the number of evictions reported in every data point. Importantly, the dataset also contains two columns flagging the data points where some kind of data imputation was conducted and where the data was contradictory with other data sources. None of the data points we include in our analysis fall in those cases.

The dataset contained a total of 3,993 block groups for each of the years of analysis. None of them has missing values for the column number of evictions evictions during 2012-2016. Nonetheless, some of the potential predictor features of this dataset present missing values for this time frame, like rent burden (missing 18% of data points), median property value (missing 5%), median household income (missing 3%) and median gross rent (missing 20%). A comprehensive of the features included in the EL dataset can be found in Annex 1.

Our data exploration also helped illustrating some of the ideas mentioned in the problem definition. In 2016, the eviction filing rate across all Chicago was 3.4%, while in block groups where the African American population was higher than 50% it reached 6.7% and in block groups where the poverty rate was higher than 30% it was 5.1%. Eviction rates have decreased during the last five years. For more details on the data exploration we conducted, please see Annex 2.

Table 1: Eviction and eviction filing rates in the Cook County, 2012-2016

Year	Eviction rate	Eviction filing rate
2012	1.6%	3.9%
2013	1.4%	3.6%
2014	1.2%	3.5%
2015	1.0%	3.2%
2016	0.9%	3.8%

Data source: Eviction Lab at Princeton University

For our final dataset generation, we augmented the EL dataset with 12 labor market, housing and social assistance features from the American Community Survey (ACS) using the ACS API. The ACS API includes five-years aggregated data at the block group level, available from 2013 onward. In the feature generation section, we will explain why the ACS data availability from 2013 (and not from 2012) did not pose a problem for conducting our 2012-2016 analysis. The complete list of the features added from the ACS API is included in Annex 3.

METHODS AND ANALYSIS

1. Feature generation

The augmented dataset we generated contained 20 EL features and 12 ACS features. We used the number of evictions and the number of eviction filings to generate one more feature we called "eviction effectiveness", and it was defined by the ratio between the number of evictions and eviction filings for a given block group and year.

The way we defined our objective - to use previous-year EL features to predict a future outcome - means that, for analyzing and obtaining the results of a given year, we should use the previous-year EL features. For this reason, all the EL features we used were always included with a one-year lag. The ACS dataset, on the other hand, was already available for 2017 (the time of our final prediction outcome) when our analysis was conducted. Thus, it was possible to use those features for the same year of analysis. That is why we used ACS data from 2013 onward, even though our analysis considered EL data from 2012-2016.

Table 2: Year of features included in the analysis

Features	**Year included in the analysis**			
	2013	2014	2015	2016
EL features from...	2012	2013	2014	2015
ACS features from...	2013	2014	2015	2016

The 33 features we originally included in our dataset were continuous. Each of these variables were discretized into quartiles by year, and then we created a dummy for every quartile for every variable. 4 of these 33 variables also had missing values, so we also created dummies for those cases. The total numbers of features we used for our analysis was 136, 88 from the EL dataset and 48 from the ACS dataset.

Lastly, we generated our predicted label by creating a dummy variable indicating if a certain block group was in the upper ten percent of number of evictions in its year. This is the label we used for our Machine Learning analysis.

2. Training and testing datasets

Following our problem definition, we constructed our training and testing datasets following a year-wise temporal holdout approach. We trained our models with a dataset from a given year and tested its results with data for the following year. The following table summarizes this procedure.

Table 3: Temporal holdouts used for the training and testing datasets

Set	Year of analysis			
	2013	2014	2015	2016
1	Train	Test		
2		Train	Test	
3			Train	Test

A total of three different datasets were built with this approach, spanning EL data from 2012-2015 and ACS data from 2013-2016[2].

3. Methods

We used nine different kinds of classifiers from the scikit-learn Python library in our analysis. Each of these classifiers was deployed with a number of different parameters, generating unique models. In total, the number of distinct models we used in this analysis was 866.

[2] Recall that every year of analysis uses last-year EL features and current-year ACS features. 2016 EL features and 2017 ACS features were only used for the final prediction outcome, as will be further explained in the final prediction results section.

Table 4: Number of models used for each classifier

N°	Classifier	Number of different parameters combinations (models) used	Parameters varied for each classifier
1	Gradient boosting	540	subsample, max depth, n_estimators, max_features, learning_rate
2	Adaptive boosting	54	base_estimator, n_estimators, learning_rate
3	Bagging	80	base_estimator, n_estimators, max_samples, max_features
4	Random forest	120	n_estimators, max_depth, criterion, max_features
5	Support vector classifier	10	C, kernel
6	Linear support vector classifier	10	C, penalty
7	Logistic regression	10	C, penalty
8	Decision tree	40	max_depth, criterion, min_samples_split
9	Nearest neighbors	2	n_neighbors
Total		**866**	

Each of these models was iterated on each of the three datasets we built, generating a total of 2,598 model/dataset iterations. Each iteration fitted the model on the corresponding training set and also tested that result on the testing set of each dataset used, generating a series of performance metrics for each model and dataset. The performance metrics we generated were the following:

- A baseline indicating the proportion of block groups in a dataset with a value of one on our prediction label[3].
- Precision and recall for the data points with the upper 1% of predicted probabilities.
- Precision and recall for the data points with the upper 2% of predicted probabilities.
- Precision and recall for the data points with the upper 5% of predicted probabilities.
- Precision and recall for the data points with the upper 10% of predicted probabilities.
- Precision and recall for the data points with the upper 20% of predicted probabilities.
- Precision and recall for the data points with the upper 30% of predicted probabilities.

[3] Since our label was a dummy indicating if a block group belonged to the upper 10% of number of evictions in a given year, the baseline always was 10% or 10.01% across our three datasets. We presume this discrepancy is due to the number of block groups by year (3,993) not being an exact multiple of ten.

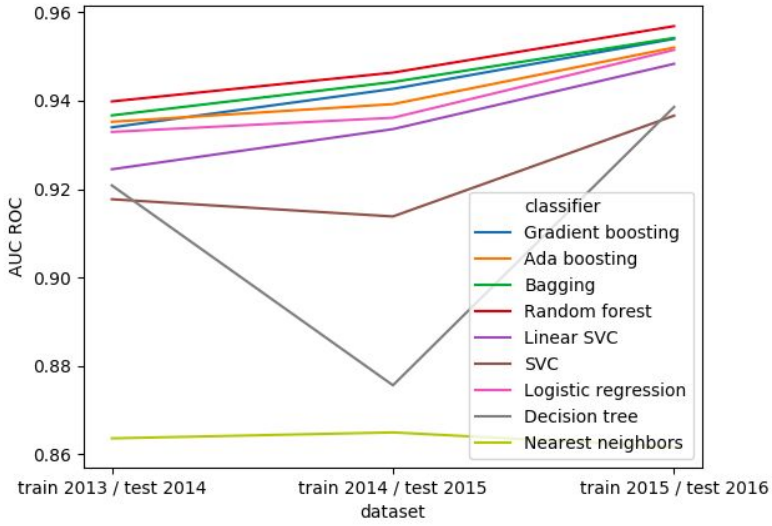
- Precision and recall for the data points with the upper 50% of predicted probabilities.
- Area under the curve (AUC-ROC).

EVALUATION OF RESULTS

We used some of these performance metrics to evaluate each model we generated across our three datasets. Each graph we present contains the model with the highest average metric (across our three datasets) for every classifier.

1. AUC-ROC

Graph 1: AUC-ROC for selected models



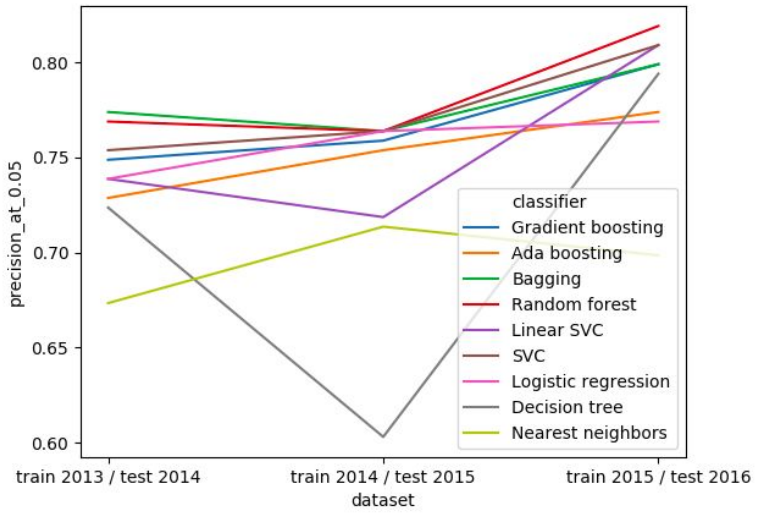
Note: the graph presents the models with the highest average AUC-ROC for each classifier, among the three datasets evaluated.

In general, all the classifiers evaluated achieved a high AUC-ROC, possibly due to our highly imbalanced label (10% of the data points are actually positive). The classifier with the better performance in this metric is the Random Forest with an AUC-ROC of 0.956 for our latest dataset.

2. Precision at different levels

We will compare the precision at different levels of our models with our baseline of 10%.

Graph 2: Precision at 5% for selected models

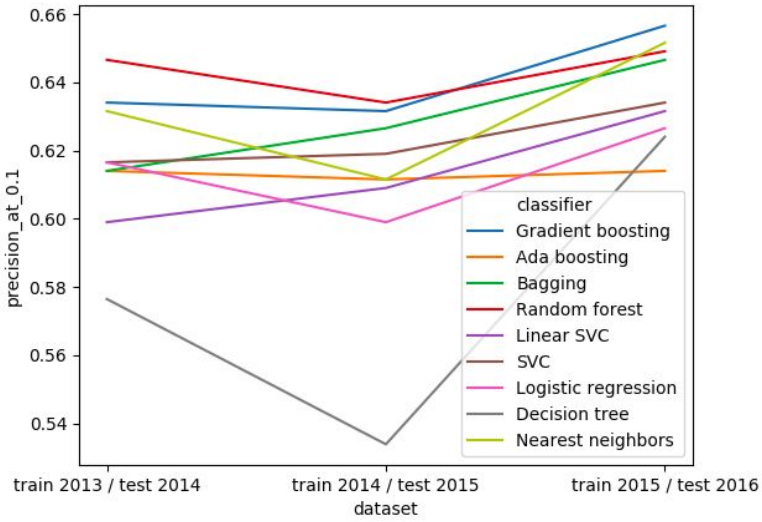


Note: the graph presents the models with the highest average precision at 5% for each classifier, among the three datasets evaluated.

The classifier with the highest average precision at 5% was the Random Forest. It had a precision at 5% of 83.4% for our latest dataset.

3. Precision at 10%

Graph 3: Precision at 10% for selected models



Note: the graph presents the models with the highest average precision at 10% for each classifier, among the three datasets evaluated.

Once again, the classifier with the highest average precision at 10% was the Random Forest. It achieved a precision at 10% of 64.9% in our latest dataset.

DISCUSSION AND INTERPRETATION OF RESULTS

Given that our problem definition consists on finding the 10% of block at the most risk of having the highest number of evictions, we will exclusively focus on precision at 10% ten percent to select our final model. Then, we will create a straightforward method to select our final model: the one with the highest average precision at 10% across the three datasets we use.

As the image 3 illustrated, this model is a Random Forest classifier. The exact parameters of this final model are summarized in the following table. Any parameters not specified in this table followed the default scikit learn library configuration for Random Forest classifiers[4].

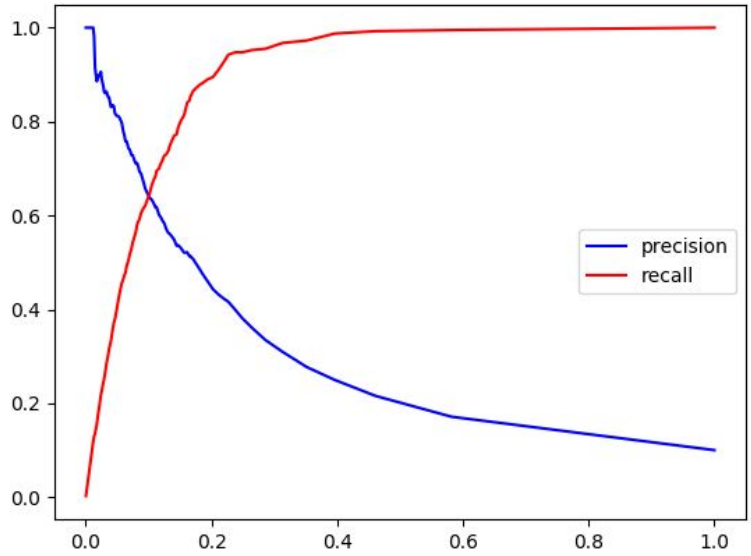
Table 5: Parameters of our final model

Parameter	Value
Classifier	Random forest
Max depth	20
Max features	0.1
Split criterion	Entropy
Number of estimators	10,000

Our final model is also the best performing in AUC-ROC and precision at the top 5%. As mentioned previously, it correctly classifies 64.9% of the data points at the highest 10% predicted probabilities.

[4] This includes: bootstrap=True, class_weight=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_jobs=10, oob_score=False, random_state=False, verbose=0, warm_start=False. Notice that the parameters “n_jobs” and “random_state” did not use their default values and were configured to improve computation time and to ensure replicability of results, respectively.

Graph 4: Precision and recall curves for the final model and the latest dataset available



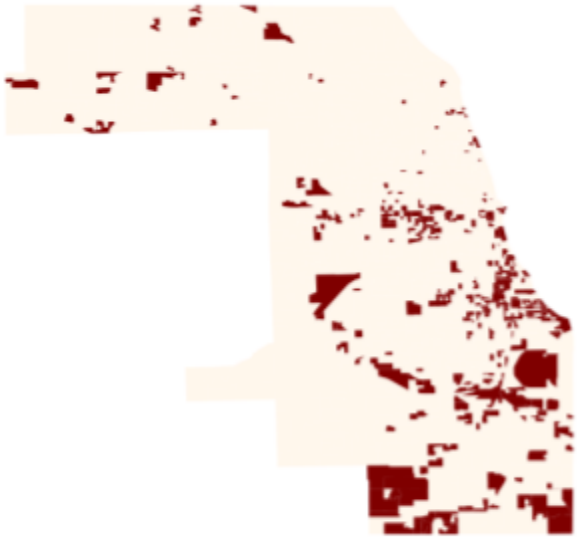
Y-axis: precision or recall
X-axis: upper rate of the population after sorting by predicted score

The precision and recall curves of our final model are consistent with the results we presented previously. The precision curve shows a value around 0.65 (y axis) for the upper 0.1 rate of population (x axis). It also displays a value close to 0.1 for the upper 1.0 population, which is equal to our baseline.

We fitted this final model with 2015 EL and 2016 ACS features and then used 2016 EL and 2017 ACS parameters to produce our final outcome: a final list of predicted probabilities whose upper 10% represents the block groups we suggest intervening on.

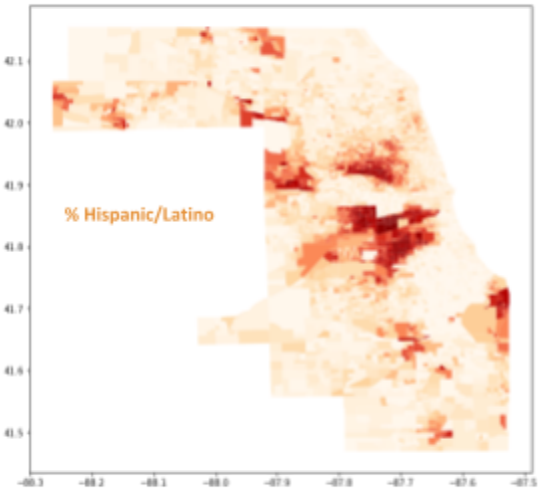
POLICY RECOMMENDATIONS

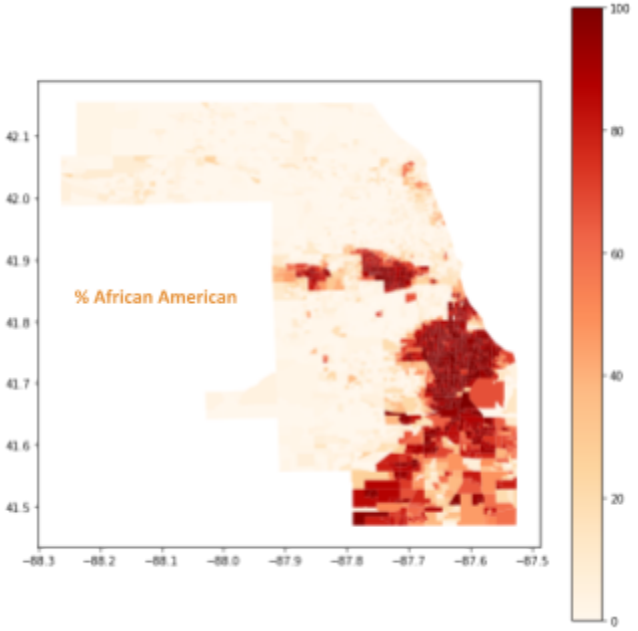
Our chosen model (built from a Random Forest classifier) gave us a list of predictions of each Cook County block group's likelihood of being in the top 10% for the number of evictions in 2017. Given these predictions, we grabbed the top 10th of the block groups having the highest likelihood. The map to the right plots these block groups. According to our analysis, we predict that these areas are in the top 10% for being most vulnerable to evictions in 2017. We therefore would recommend that the MTO focus their services on the geographic areas pictured.



Recommended Areas for Focused Interventions

The map of recommendations shows a large number of block groups in the south side of the county (areas both in the south suburbs and in the city), some masses on the west side, and a few blocks scattered along the north side. The general locations of these selected block groups makes sense considering poverty rates and where historically disadvantaged populations traditionally reside. The maps below show the percentage of block groups' 2016 populations that are Hispanic/Latino (left) or African American (right)--the darker the red, the more dense in population the given race is for the area. One can observe that these darker red locales generally correspond with the south and





west sides of the county, therefore consistent with what we have previously mentioned regarding the geographic distribution of these minority groups.

The recommended block groups are also reasonable given the spread of eviction filing rates for 2016. The map to the right shows this distribution. While eviction filings seem to be happening all over the county, the number of filings given the number of renter occupied households (i.e. eviction filing rate) are higher once again within clusters in the south and west sides. However, perhaps the more telling rate in terms of the recommended block groups that lie in the middle of the county is the poverty rate.

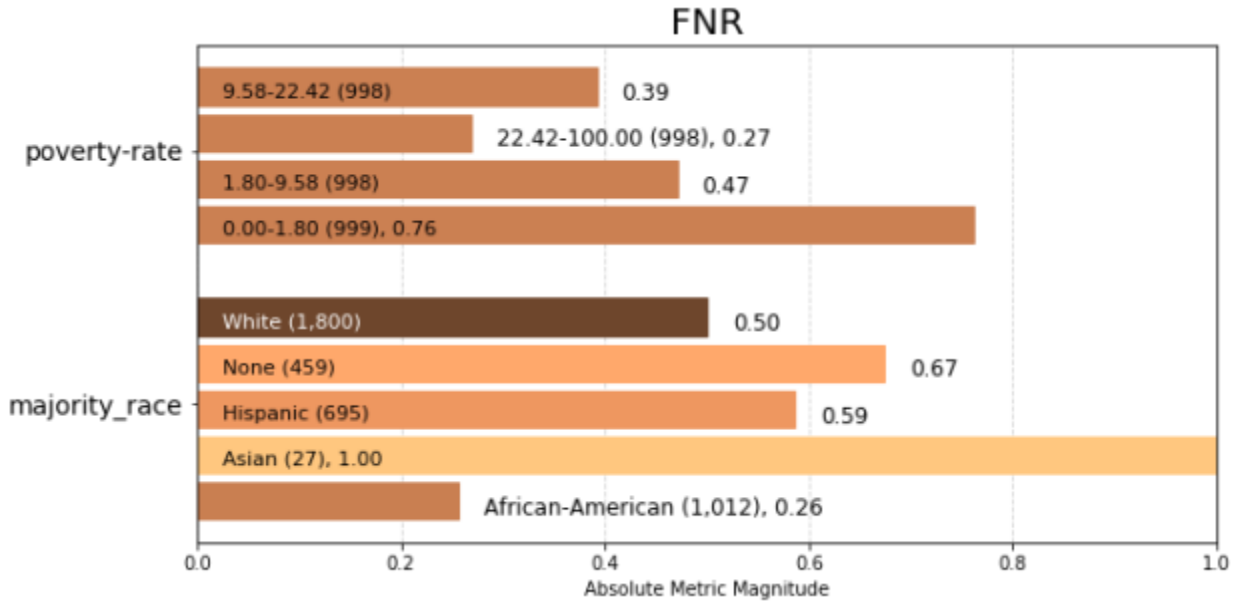
Although the recommended block groups seem to exist in clusters, they are nevertheless spread throughout different parts of the county. Perhaps MTO would like to focus their services on these block groups using different strategies. Similar to what we have suggested in the problem formulation, their actions given our recommended list could be to send out community organizers in some blocks first and increase advertisements in others. For instance, they could send out community organizers to the many clusters of block groups in the south and west sides of the county early in the year. Since these block groups are close in proximity with each other, it might be easier to establish a presence in order to gain momentum with organizing efforts there. On the other hand, they could increase advertising of their resources on the scattered block groups in the north and middle parts of the county to gain community recognition without expending the limited number of volunteer or staff organizers. In this way, MTO will still be able to reach vulnerable populations and provide assistance from a distance.

BIAS AND FAIRNESS ANALYSIS

For our analysis, we care about accurately extracting the most vulnerable block groups in Cook County. While our selected model's precision at our population target of 10% is pretty good, this precision might not necessarily extend to different demographic groups in our sample. We therefore utilized Aequitas to perform a bias and fairness analysis of predictions given by our chosen machine learning model.

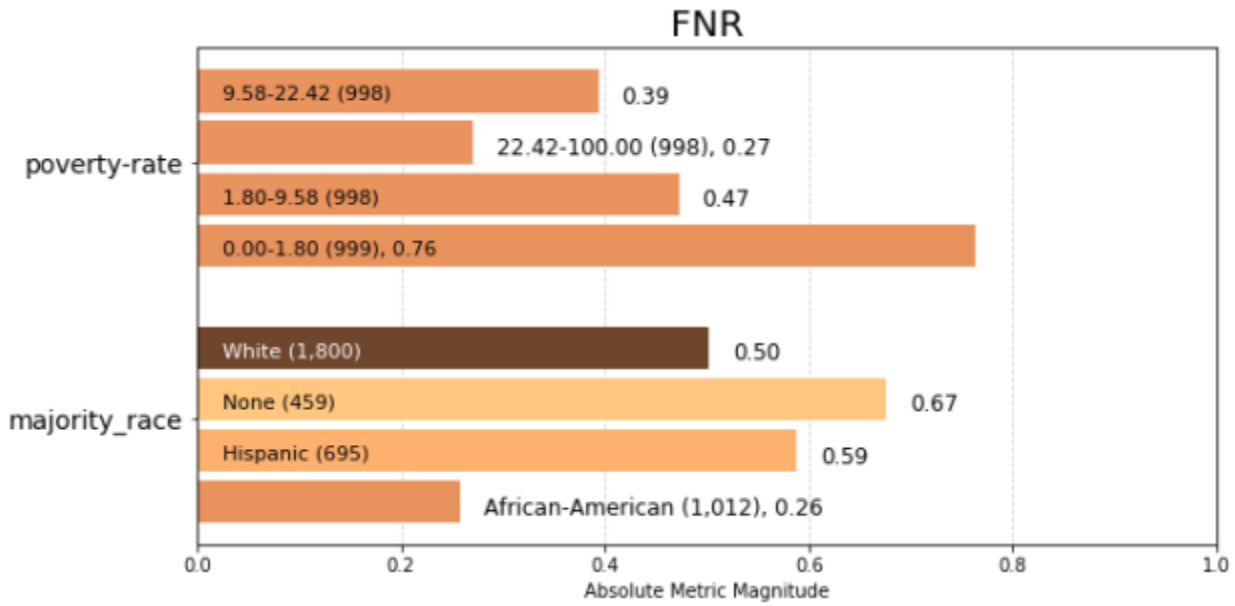
We decided to look at the following attributes for this analysis: poverty rate and majority race. Poverty rate is kept a continuous variable, which in pre-processing Aequitas will discretize into quartiles. Majority race is a categorical variable defined as the race that makes up more than 50% of the block group's population; if no singular racial group is defined as such, then we passed the value 'None'. We used 2015 data to train our chosen model, and predicted on 2016 outcomes. We then passed a preprocessed data frame into Aequitas.

We cared mostly about False Negative Rates (FNR) for this analysis, although we looked at other metrics as well. We wanted to see how often we are missing certain groups in our predictions and at what rate compared to some base group. An initial look at absolute FNR (with no comparison to some base) is shown in the chart below. The numbers in parentheses show how many block groups in the sample fall into the given group (this characteristic is also portrayed by the darkness of the bars), while the length of each bar represents the FNR.



The above chart suggests that there's a very unfair rate of false negatives for block groups that are majority Asian (100% FNR!). However, when we looked at those block groups, we found that this high FNR rate is due to 1 block group being incorrectly labeled by the model. While we would ideally not want to miss this block group in our predictions, incorrectly labeling one block may not be the worst or most obviously biased outcome. On the other hand, missing the one majority Asian block group in high need of housing assistance may arguably be very detrimental and discriminatory against Asian communities.

The above chart shows all attribute groups regardless of their share of the sample. As seen above, there are only 27 block groups with Asian majorities--less than 1% of the total number of block groups in Cook County. Filtering out attribute groups that make up less than 5% of the sample, produces a similar chart below.



This chart only eliminated the majority Asian block groups, and otherwise displays the same information as the previous chart. There is a much higher rate of false negatives for blocks with majority Hispanics or Whites compared to African American-majority blocks. The rate is the highest for more mixed-race blocks ('None'). Similarly, the false negative rate for blocks at the lowest rates of poverty have a much higher FNR than the rest (76%). These might pose unfairness problems in that blocks without strong signals that the models may be using (mixed-race blocks or blocks where impoverished households live within the same geographies as more privileged homes) can be missed, adversely affecting populations who may be in dire need of housing assistance.

We generated a table of disparities compared to a majority group, automatically chosen as White majority race and the lowest quartile for poverty rate (shown below).

	attribute_name	attribute_value	ppr_disparity	pprev_disparity	precision_disparity	fdr_disparity	for_disparity	fpr_disparity	fnr_disparity	tpr_disparity	tnr_disparity	npv_disparity
0	majority_race	African-American	8.131579	14.463283	1.391586	0.608414	9.761773	12.057072	0.512111	1.487889	0.879470	0.904490
1	majority_race	Asian	0.000000	0.000000	NaN	NaN	3.434698	0.000000	2.000000	0.000000	1.010901	0.973460
2	majority_race	Hispanic	0.578947	1.499432	1.090909	0.909091	2.342535	1.392445	1.172414	0.827586	0.995722	0.985365
3	majority_race	None	0.815789	3.199174	0.903226	1.096774	6.283571	3.789727	1.348837	0.651163	0.969590	0.942406
4	majority_race	White	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
5	poverty-rate	0.00-1.80	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
6	poverty-rate	1.80-9.58	3.750000	3.753758	1.440000	0.685714	1.553515	2.655604	0.617647	2.223529	0.988065	0.990879
7	poverty-rate	22.42-100.00	21.166667	21.187876	1.596850	0.573678	5.140625	15.483330	0.352273	3.072727	0.895589	0.931771
8	poverty-rate	9.58-22.42	7.416667	7.424098	1.591011	0.577849	2.578795	4.651974	0.514175	2.554639	0.973673	0.973985

Positive Predictive Ratio Disparity

- We are positively predicting on blocks that are majority African American 8 times more than blocks that are majority White, while we are positively predicting on majority Hispanic blocks only about half the time.
- We are positively predicting on blocks with the highest poverty rates 21 times more than blocks at the lowest poverty rate.

False Omission Rate Disparity: Number of blocks incorrectly predicted negative (FN) / Number of blocks predicted negative (FN + TN)

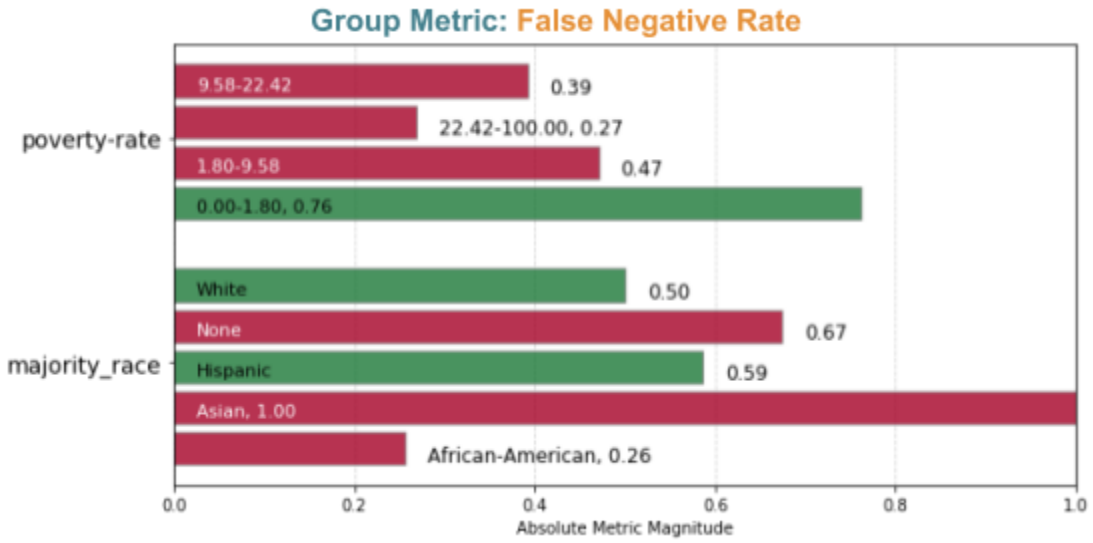
- Compared to majority White blocks and given our analyses' negative predictions, we are about 10 times more likely to incorrectly omit blocks that are majority African American, and 6 times more likely to omit blocks without a race majority.

False Positive Rate: Number of blocks incorrectly predicted positive (FP) / Number of blocks that are actually negative (FP + TN)

- We are 15 times more likely to incorrectly predict positive on blocks that are at the highest poverty rate compared to the blocks at the lowest poverty rate.
- We are 12 times more likely to incorrectly predict on blocks that are majority African American compared to blocks that are majority White.

The above statistics suggest that we are overpredicting on blocks with high poverty and blocks that are majority African American, yet we are at the same time omitting more majority African American blocks given our negative predictions. All this may simply demonstrate what we have seen before--that block groups most vulnerable to eviction are more likely to be high poverty or mostly composed of African Americans.

The final Aequitas output summary on the fairness of our model is shown in the chart below. For the FNR metric, it shows that our model is predicting "fairly" only on block groups that are majority White or Hispanic and on block groups that are in the lowest quartile in poverty rates.



CAVEATS AND LIMITATIONS

We only had data from the Eviction Lab until 2016, so the final model we built made predictions of which Cook County blocks were most susceptible to evictions as if it were the year 2017, the most recent year we can predict with the data limitations we face. Nonetheless, we built our pipeline in a generalized way such that we could easily apply our analysis to more recent data if it suddenly became available.

The eviction rate data we are using corresponds to official evictions and it comes with some limitations in that it does not always paint the full picture of what happens when someone is evicted or when an eviction request is filed against a tenant. For example, it does not count scenarios where tenants agree to move out before facing a judiciary process, which is what happens before an eviction officially occurs. Then, cases such as when the tenant might find alternate housing quickly, when she/he might accept cash from the landlord in exchange for the house keys, or when she/he might not show up at mandatory arbitration at courthouse are not included in the number of evictions variable we use. Therefore, the dataset we use and our results are likely to overlook such manifestations of the affordable housing crisis across Cook County.

Our analysis uses a relatively small number of datasets (three) to train and test the models we generate. This was mainly due to three reasons:

1. We explicitly wanted to keep some validity over time of the data we were analyzing. Specifically, we chose not to have a large range of years in our datasets.
2. The number of evictions for Cook County block groups before 2007 was inconsistent and some years presented unstable and inexplicably low numbers for this feature. Though this variable seemed to stabilize from 2007 onward, we considered it was a good practice to leave a gap of some years before start using the data points available. In our specific case, this gap was 5 years (2007-2011).
3. ACS block-group-level data was not available for 2012 and before that year. While using instead tract or zip code level data would have also been feasible to overcome this restriction, we preferred to maintain a higher resolution for the features we were using.

Though any of these three would have been a serious issue for the validity of our analysis and results, having a small number of datasets might be an issue for the consistency of the final model we chose. An analysis including more datasets, as soon as more data is available, is needed to further ensure the selected model is the best possible.

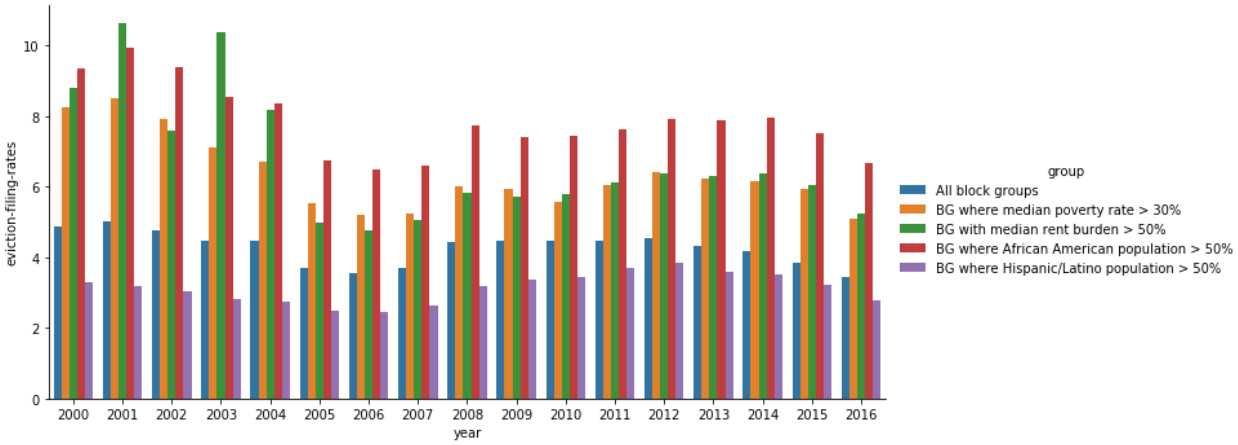
ANNEX 1

Features included in the Eviction Lab dataset:

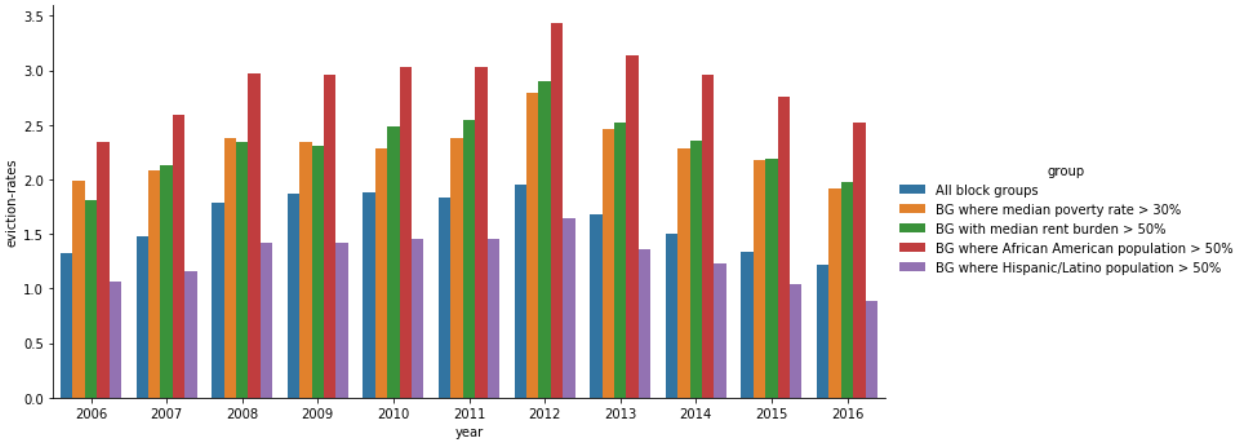
- GEOID: Census FIPS code
- name: Census location name
- parent-location: Parent location of given geography. Counties for tracts and block groups
- population: Total population
- poverty-rate: % of the population with income in the past 12 months below the poverty level
- pct-renter-occupied: % of occupied 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, max is 50% representing >= 50%
- 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-other: % population that is other race alone and not Hispanic or Latino
- pct-multiple: % population that is two or more races and not Hispanic or Latino
- renter-occupied-households: Interpolated count of renter-occupied households
- evictions: Number of eviction judgments in which renters were ordered to leave in a given area and year
- eviction-filings: All eviction cases filed in an area, including multiple cases filed against the same address in the same year
- eviction-rate: Ratio of the number of renter-occupied households in an area that received an eviction judgement in which renters were ordered to leave
- eviction-filing-rate: Ratio of the number of evictions filed in an area over the number of renter-occupied homes in that area
- imputed: Boolean variable indicating whether eviction numbers and renter-occupied-households were imputed
- subbed: Boolean variable indicating whether eviction numbers and renter-occupied-households were pulled from another source other than Eviction Lab sources
- low-flag: Boolean variable indicating whether the eviction numbers are estimated to be lower than they are in reality.

ANNEX 2

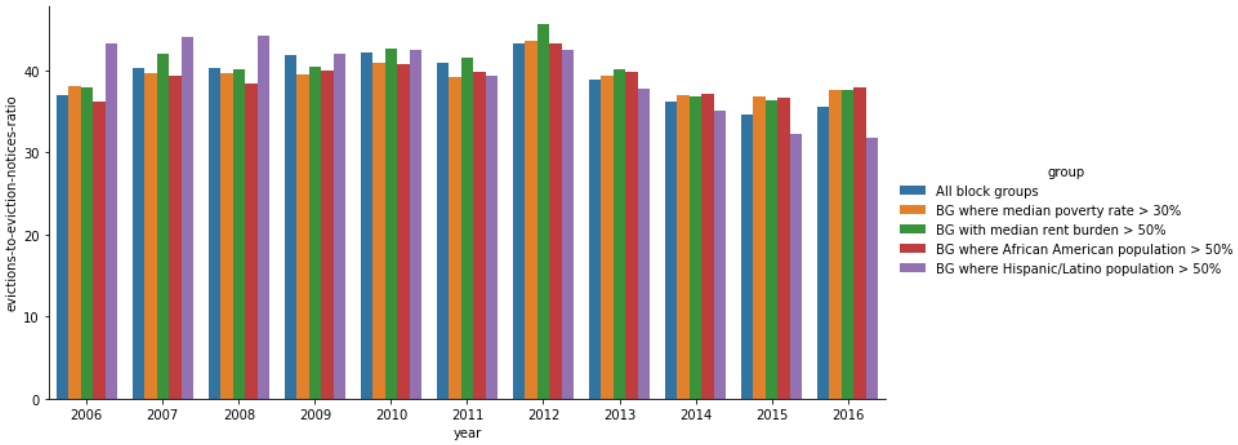
Graph A1: Eviction filing rates (percentages) across years, for selected block groups



Graph A2: Eviction rates (percentages) across years, for selected block groups



Graph A3: Eviction to eviction notice ratio (percentage) across years, for selected block groups



ANNEX 3

Table A1: Features added from the ACS API

N°	ACS API code	Name in our dataset	Variable description	Level
1	B23025_002E	estimate_total_in_labor_force	Total population in labor force	Block group
2	B23025_005E	estimate_civilian_unemployed	Total population in labor force unemployed	Block group
3	B19057_001E	total_for_public_assistance_income	Total households	Block group
4	B19057_002E	with_public_assistance_income	Total households with public assistance income	Block group
5	B25038_001E	total_for_householder_tenure	Total occupied housing units	Block group
6	B25038_009E	renter_occupied	Total renter-occupied housing units	Block group
7	B25038_010E	renter_moved_2015/2010_later	Total renter-occupied housing units with renters who moved during 2015/2010 or later	Block group
8	B25038_011E	renter_moved_2010-2014/2000-2009[5]	Total renter-occupied housing units with renters who moved during 2010-2014/2000-2009	Block group
9	B25038_012E	renter_moved_2000-2009/1990-1999	Total renter-occupied housing units with renters who moved during 2000-2009/1990-1999	Block group
10	B25038_013E	renter_moved_1990-1999/1980-1989	Total renter-occupied housing units with renters who moved during 1990-1999/1980-1989	Block group
11	B25038_014E	renter_moved_1980-1989/1970-1979	Total renter-occupied housing units with renters who moved during 1980-1989/1970-1979	Block group
12	B25038_015E	renter_moved_1979/1969_earlier	Total renter-occupied housing units with renters who moved during 1979/1969 or earlier	Block group

[5] Features from N°8 to N°12 vary in their description depending on the year of the data point. For example, the description of renter_moved_2015/2010_later is "Total number of renter-occupied housing units with renters who moved during 2015 or later" for data points from 2015 onward, while from data points from 2013-2014 it is "Total number of renter-occupied housing units with renters who moved during 2010 or later". An analogous explanation applies for the rest of these features.