Machine Learning Project Progress Report
Evictions in Cook County

Group: Luise San Martin, Liza Soriano, and Jessica Langston
CAPP 30254 | May 21, 2019

In order to supplement the Eviction Lab data and create more features about topics relevant to housing and evictions, we are adding data from the American Community Survey (ACS). We are using the ACS 5-year estimates because this will give us reliable data for the block group level, and since our evictions data is already a little bit old (2012-2016), recency is not the most important criteria. The main tradeoffs between using the ACS 1-year and 5-year estimates are that 1-year data is more current but not available at the smallest geographic levels (like block groups), and only available for areas with larger populations (not a relevant issue for us since Chicago/Cook County has a large population).[1]

To get data from the ACS API, we obtained an API key, reviewed the documentation about how to make API calls, and examined the available variables to find relevant data we could add for more features. As a starting point, we have decided to pull ACS variables on unemployment, income assistance, and tenure (i.e. how long occupants have been living in their residence), as we believe these variables may correlate with evictions and thus add informative value to our model. To join the ACS data with our Eviction Lab data, we concatenate the state, county, tract, and block group fields in the ACS data to form the Census FIPS code, which is called "GEOID" in the Eviction Lab dataset. We create our full dataset by then merging our Pandas dataframes on this column.[2]

Our next step in integrating this data will be to make features from the new columns. To calculate unemployment rate, we will divide the number of unemployed people by the number of people in the labor force. Similarly for income assistance, we'll calculate a rate by dividing number of households receiving income assistance by number of households. For household tenure data, we have variables for how many households moved into their current residence between a range of years, so we can create a column for the rate for each. We can add more features down the line by bringing in more variables from the ACS.

We also wrote code to generate additional features from the Eviction Lab and ACS raw attributes. The code computes a new categorical feature indicating quantile membership and several dummy variables based on this new quantile feature. The desired number of quantile bins (e.g. 5 for quintiles), as well as a list of the raw attributes to transform, are passed to the function. We generate these new features for continuous attributes like poverty rate, median household income, and median gross rent. We will write similar code to generate z-score features and a feature that normalizes the attributes.

[1] https://www.census.gov/programs-surveys/acs/guidance/estimates.html
[2] See these steps in our Jupyter Notebook here:
https://github.com/luisesanmartin/ml-eviction-chicago/blob/master/notebooks/get_acs_data.ipynb

The analysis we conducted consists of (i) exploratory analysis on the Eviction Lab's dataset and (ii) the creation of one-feature classifier models using the original Eviction Lab's dataset features, before applying any transformation to them[3].

The exploratory analysis checked for the existence of missing values, the distribution and occurrence of outliers for each feature, and the correlation between each variable. Each data point analyzed consisted of a block group in Cook County for each year we analyze, from 2012 until 2016. For training the one-feature classifier models, we used features from 2015 and a label from 2016, replicating the temporal holdout specified in our problem setting. We generated the label variable by creating the ratio between the number of evictions and the number of eviction filings by block group and by creating a dummy for the block groups with the highest 10% of that value in 2016. The generation of this dummy had the methodological advantage of transforming our label to a binary variable, converting our prediction problem into a classification. Finally, we created different types of classifiers for each feature in the original dataset, for a total of 140 one-feature classifier models (7 models and 20 features)[4].

The exploratory analysis using the one-feature classifiers shows some clear trends. Firstly, the Random Forest models perform consistently far better than any other kind of model for almost all of the features. Nevertheless, we cannot be sure if this is an effect of using one-feature classifiers or if this is happening because our label is inherently better-predicted by that kind of model. The features with the highest performance metrics are the median household income, the percentage of white population and the percentage of renter-occupied households. In our final project report and as this project progresses, the analysis will be augmented by including new features from ACS, additional variable transformations and temporal holdouts, and more importantly, classifiers with more several features.

[3] For a more detailed description on the exploratory analysis, please see:
https://github.com/luisesanmartin/ml-eviction-chicago/blob/master/notebooks/Data%20exploration%20-%20block%20groups.ipynb
And for a more detailed description on the one-feature classifier exploratory analysis, please see:
https://github.com/luisesanmartin/ml-eviction-chicago/blob/master/notebooks/One-feature%20classifier%20exploration.ipynb

[4] The features used were: population, poverty rate, number of renter-occupied households, percentage of renter-occupied households, median household income, median property value, rent burden, percentage of white population, percentage of african-american population, percentage of hispanic population, percentage of american indian population, percentage of asian population, percentage of native hawaiian or pacific islander population, percentage of population with multiple ethnicities, percentage of population with other ethnicities, number of eviction filings, number of evictions, eviction rate (among all renter-occupied households) and eviction-filing rate (among all renter-occupied households). It is important to note that all of these features were included exactly as they were in the original dataset, and that no variable transformation was applied at this point for analysis.