# Assessing the Low-Income Housing Tax Credit Using Machine Learning

## Background

A common criticism of our federal housing policy is that it concentrates poverty and has a negative effect on the neighborhood. Local stakeholders often attempt to block or derail low-income development, as they fear it will detriment their neighborhoods.

The low-income housing tax credit (LIHTC) is a government-sponsored incentive program encouraging the acquisition, rehabilitation or new construction of affordable rental homes. The program is governed by the federal tax code and is administered by state credit allocating agencies. It is estimated to house approximately one million households nationwide. The U.S. Department of Housing and Urban Development (HUD) provides data on LIHTC development.

But do the projects genuinely have a negative impact on the neighborhood? For instance, does the construction of a new LIHTC property depress local income levels, thereby concentrated poverty? Or will it stabilize the neighborhood by bringing an influx of new renter households?

To accomplish this, a linear regression will examine the relationship between property characteristics and change in median income, within the relevant census tract.

Understanding which property characteristics affect the neighborhood's income growth could provide key insights into how to best optimize the program and disperse the funding effectively.

Each states' government controls the location of projects, and all have the intention of fair housing practices – meaning giving people equal access to housing. Running the model across different states is an important final step in the assessment of the model.

## Problem Statement and Hypothesis

Using the data provided by the U.S. Housing and Urban Development, can we predict a change in income within a census tract, based on the characteristics of affordable housing development?

I hypothesize that the construction of large LIHTC developments (large, as percentage of total households in the census tract) will result in reduced income growth relative. Other factors will contribute to the change in income, such as the tenancy (1 bedroom v 2 bedroom, ect.) of households targeted, and whether it's a new-construction or rehabilitated property.

## Methods

I plan on using a linear regression model to determine which parameters contribute to the change in income. In doing so, we want to minimize r-squared and mean-square error, with a goal of generalizing the model effectively.

### Simple Outline

1. Research data sets and build data dictionary.
2. Check for outliers.
3. Check for covariates.

4. Access API for relevant income data and append to property data.
    a. Build custom Requests function to pull in data from Census API.
5. Drop duplicates, null, and incomplete rows.
6. Look for places where categorical data makes more sense than continuous data. For instance, breaking up number of units into blocks (i.e. properties with 0-50 units).
7. Create dummy variables (i.e. rural development deal – y/n to 0/1).
8. Train OLS model.
9. Run predictive dataset.
10. Cross validate model.
11. Visualize results.
12. Graph predictive model against Universe of census tracts without properties in them.
13. Iterate model over states, if necessary
14. Interpret results.

Upon an initial review of the data, much of the time will be spent cleaning and removing incomplete data.

Data from the American Community Survey API will be scraped and appended to each individual census tract containing a LIHTC property. There are over 38,000 properties, so this may take time and has the potential to be incomplete.

Below lists the libraries and languages used to run the model.

| Language/Library | Documentation |
|---|---|
| Python | https://www.python.org/doc/ |
| Pandas | http://pandas.pydata.org/ |
| Jupyter notebook | http://jupyter.org/ |
| Scikit learn – linear regression | http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html |
| Scikit learn – lasso | http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html |
| Requests | http://docs.python-requests.org/en/master/ |
| Seaborn data visualization | https://seaborn.pydata.org/ \| https://seaborn.pydata.org/tutorial/regression.html#regression-tutorial |

## Data

See below for an outline of data. The website listed below provides a data dictionary for the variables needed.

| | Website | Specific Variable | Geography | Time Series? |
|---|---|---|---|---|
| LIHTC Database | https://lihtc.huduser.gov/ | N/A | Point data, with Lat/Longs | Each project's "Placed in Service Date" |
| American Community Survey 5-Year Estimates API | https://www.census.gov/data/developers/data-sets/acs-5year.html | DP0362 – Median Household Income | Census tract polygons (provided by the Census Bureau) | Each Year from 2010-2015 |

The LIHTC database contains significant amount of columns. See the data dictionary in the website link.

## Questions, Assumptions, Risks, and Benchmarks

One assumption that I have, is that low-income development will *improve value* neighborhoods that are poor, while *decreasing value* in neighborhoods that are wealthier. This may require the best model to contain polynomials.

The problem statement faces a number of risks, the largest being incomplete data for the time series for which Income data in available. Properties in this program have been built starting in 1987, however the time series of income is only available from 2010-2015.

It may be necessary to compare the growth rates of census tracts with no affordable development as a relevant benchmark. This could be crucial in assessing the model and visualizing the effects of introducing a subsidized housing project. A classification may not be necessary, but being able to see the expected growth rates between the two types of census tracts is appealing.

There is also the potential that the model does not generalize well across the country. It may be necessary to run a model on each state – which would be a helpful comparison regardless. This is a valuable benchmark.

Missing demographic data is a potential risk. Filtering out incomplete data will present a risk, since it will lower our sample size.

## Goals

My goal is to create accurate prediction model that describes which variables contribute to a change in median income in census tracts that contain LIHTC properties. Crucial to the project is determining if the overall slope is negative or positive (for a simple linear regression) or where the slope changes (for a polynomial regression).

$$Predicted\ Change\ in\ Rent = Variable1_{Cofficient} + Variable2_{Cofficient} + ect..$$

Interpreting the model is an equally important goal. Building a model that makes sense and can be "digested" to by industry professionals is imperative.

In a larger context, the goal is to optimize the locations of LIHTC buildings. By identifying predictive features, stakeholders will be able to understand the effect that LIHTC projects have on the neighborhoods in which they are sited.