Name: Hyomin Seo
Date: 6 Nov 2020
Paper Title: Mining Administrative Data to Spur Urban Revitalization
Author Names: Ben Green, Alejandra Caro, and else.
Year Published: 2015

Open questions:
How can we avoid model misspecification in general?

The topic areas covered by the paper are:
The paper introduces and explains the machine learning system that can aid the systematic revitalization of urban towns, specifically - Memphis, TN. The problem statement is the following: the majority of the urban neighbors are frequently overlooked by the government and dismissed from upkeep with the growth of the city, and the government must acquire a practical, systematic way to plan the improved maintenance of such urban areas. (key features - population, housing stock, area/taxes) The main goal of the implemented system is to calculate and predict the neighbors that need rehabilitation and identify the best optimization strategy of restoration that can positively contribute not only to the specific area about the more prominent community/ town.

The previous approaches to this problem were:
The government attempted urban community reinvestment, but due to the several external impacts, including the 2008 financial meltdown, the reinvest rarely achieved its revitalization; in most cases, such reinvest resulted as either uneven or unyielding. The challenging aspect of the reinvestment is redeveloping a traditional core neighbor and abandoned properties. The city of Memphis is selected as the objective of the study because it embedded the core issues listed before.

Outline the basic new approach or approaches to this problem:
The paper categorizes the data they have collected to conduct the study and feed it to the system.

1. **Administrative Data**: The majority of the data were extracted straight from the 'City internal database of every parcel in Memphis.'The data not only provides the numerical measure of the city's property (buildings, systems) but also states the all disconnected, abandoned property. There is another data category within the administrative data - 'foreclose,' indicating the distressed properties financially-struggling homeowners. The authority of the data is all verified since the platforms (City of Memphis administration, US census, and American Community Survey) are nationally certified.
2. **Neighborhood Surveys:** Community survey conducted by The Center for Community Building and Neighborhood Action -2008 ~2010. The description of the data is relatively straightforward; however, this data is less reliable due to the nature of the inevitable error created during human (non-professional) to the human survey, leaving considerable room for inconsistent classification.

3. **Data, Features, and model:** 30 input features were implemented, mostly the features that concern identifying the distressed, underfunded system and aspect of the city.
    a. **Model:** The system primarily employs random forest to predict each residential property in Menships that will be distressed in the next couple of years.

b. ; the risk score was calculated. This is a direct quote from the paper to simply explain the model.

'*if a property is labeled as distressed with a probability of 0.75, for example, 75% of the most similar properties in the training set were labeled as distressed while the other 25% were labeled as not dis- tressed. We define a property's class probability for being distressed as its risk score.'*

c. **Classifier Validation:** The paper approached two methods to validate such classification, Cross-validation, and Manual - qualitative.
The 'Manual- qualitative' is an interesting approach. The team visited the site with different risk scores to manually check the consistency between the risk score and the physical property. The resulting example is presented below. It is shown that the stuff with a higher risk score is visible ' not - well maintained' than the one with the small risk score, on the left.

Figure 2: There is a visible difference between homes with a low risk score (0.368, left) and a high risk score (0.807, right).

The team also employed five-fold cross-validation, and the calculated result was consistently accurate.

d. **Implementation:** The model estimation is featured as a color map of Memphis that clearly shows which region has a higher risk of being distressed. The community groups and city officials were provided with such a plan, enabling them to work undoubtedly efficiently by prioritizing (investing the most in) the regions according to its risk score without conducting a massive and unreliable survey over the entire city.

4. Revitalization strategy evaluation: This is when the paper studies the relationship between the real estate value of the distressed home and other ones in their vicinity to find a relatively simple, expected result. *more severely-distressed properties and their immediate neighbors are valued below dis- tressed homes further away in their neighborhoods".*
Next, the paper studies the rehabilitated homes as those that were labeled as distressed in the prior survey. The identification for the rehabilitated property is- *received non-demolition building permits valued over $10,000 between 2008 and 2013``.Also,* to find an expected, ideal result that the value of rehabilitated homes and its neighbor are higher than those of the demolished homes.

5. Future works and evaluation: The last few sections of the paper concern with the tax appraisals (simulated the effect on tax on such rehabilitated region) and governmental data science challenge (to expand the study adopted at Memphis to more part of the country, with the aid of more massive software data)

Critical assumptions made include:

Even though the paper employs various methods to perform this prediction, it is admitted that there is a need for actual fieldwork to validate and enhance such a goal. Some percentage of such conditions is covered by employing the 'manual - qualitative research' on their result, manually cross-checking the property, and according to the risk score.

There might be an underlying, fundamental factor that is causing the entire city to be stagnant. The paper primarily focuses on the beach, particular property.

The performance of the techniques discussed in the paper was measured in what manner:

The data selection for the study was mainly from governmental- approved platforms, so even though man is liable to man surveys, overall, the data is reliable. The model evaluation was pretty flawless and verified, though the technical aspect of the assessment itself was not too detailed. Similar to the paper discussed previously (New york city tenant help with machine learning), this system also provides an efficient aid to the current dilemma, expediting the process of improving a defect.

New background techniques are used in the paper:

Proximity Matrix (proximity matching)
Economical Matrix
Model misspecification

I rate and justify the value of this paper as:

I lived in TN, visited Memphis a few times during my residence in TN. There do exit numerous distressed properties in the city of Memphis but perhaps all over TN. I consider that this system/ paper provides a tentative solution to improve such a dilemma faced in Memphis. Still, it may exclude a more significant, broader perspective to malate such stagnant growth of these regions. Instead of revitalizing each building or a section of a city, encouraging the overall development of the town and living might be, admittedly slower, but steadier and sustainable e solution.