Name: Hyomin Seo
Date: 23 Oct 2020
Paper Title: Using Machine Learning to Help Vulnerable Tenants in New York City
Author Names: Teng Yem Rebecca Johnson, and else.
Year Published: 2019

Open questions:

Won't there be a correlation between the econocial status of the region (location of the units) and the risk of harassment? = shouldn't income/ economic state of the borough also be one of the features?

The topic areas covered by the paper are:

The paper introduces and explains the machine learning system that can aid the current New York city tenants. The problem statement is the following : some of the landlords of the city take illegal advantage of 'rent-stabilization' units to maximize their profit out of these units. The machine learning system is employed to first, find more tenants in such abuse of landlords, and secondly to predict the likelihood of a tenant to face such a dilemma beforehand, eventually to a more organized, sysmetican and comprehensive method to successfully improve the city's policy of proactive outreach to vulnerable tenants.

The previous approaches to this problem were:

The city's policy to help the tenants: "rent stabilization", is a policy that restricts yearly increase of rents. Even though there exist over a million of such units, some landlords make aggressive leeways with which they forcefully evict the current tenants under the rent stabilization and replace them with tenants who can provide the maximum rent.

Outline the basic new approach or approaches to this problem:

The paper categorizes each problem statement and its formulation to an algorithm.

1. **Harassment risk prediction** : Numerous of 'harassments' are categorized and organized as data, the machine learning system is fed with such data to predict the risk score of likelihood of harassment taking place in the following months, with units considered as one building, not each tenant.
    a. The category of harassment are the following
    Internal : Canvassing, Knock attempts, Case issues.
    External : American Community survey, Primary land use and tax lot output, Department of housing preservation and development, Housing court litigation, and subsidized housing.

2. **Feature Generation :** Data generated and extracted from various sources were preprocessed and normalized to be used for generating features
    a. **Building Level features** : Active, dynamic harassment feature (knocks and opening of doors). Binary data
    b. **HPD Violation/ Housing Court Litigation feature :** Frequency of violation and categorized litigation issues.
    c. **Static features :** Extracted from internal building address database
    d. **Tract level features :** Racial/ Tenant's workplace/ demographic data

3. **Model Evaluation :** The paper primarily employs evaluation metrics, with three categories of labels (positive, negative and missing). Precision and recall of top k metrics are evaluated.
    a. Method used : Decision Tree/ Gradient Boosting/ LogisticRegression/ Random Forest

4. **Feature Interpretation :**
    a. **Tract-Level demographic feature :** Data that present the socioeconomic status and data relevant to 'Building Level feature' feature
    b. **Building history and value feature** : Data presents how 'external (harassment prediction)' information does contribute to the prediction of the harassment of the consecutive month.
    c. **Building location and size feature** : Data that is used to see if there is correlation between the prediction on harassment and the location of the , and if there is a cluster of these units that are more likely to have higher prediction value.

Critical assumptions made include:

From the features explained, the 'external' feature-data has a very limited probability that the feature will be either reported falsely, unreported or changed without notification. However, the 'internal' data is undoubtedly a subject for high degree of freedom; this data can be reported falsely, unreported, or changed without report, regardless if that is intentional or not. There are layers of other features and considerate data processing done to eliminate any disparincy on this system, however, if the 'flexibility' of the 'internal' feature is unpredictable, the overall accuracy of the system might be in doubt.
Looking the the cluster map of units high risk, it can be assumed that the risk might also have some degree of correlation with overall income of the neighbor,but do not see that being considered

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

Even though the paper employs various types of methods to perform this prediction, it is admitted that there does exist a need for actual field work to validate and enhance such a goal. This is effectively stated under the 'practical implications', where it states that the result of the machine learning system will be a guidance for the field work with result of which the Tenant Support Unit will be able to determine their location of study to maximize their efficiency (number of tenants at risk visited/ time and trip distance). Also it indicates that there might exist bias due to the limited number of labels, and they expect to use the field work to alleviate this bias by categorizing more labels for each feature.

I rate and justify the value of this paper as:

The topic/goal of the paper was not intuitive, because harassment is something considered unplanned and abrupt. However, reading through the paper, it is noticed that there are considerable measures taken to the account and the categorization of the data is very detailed (work time and location, door opening and closing). The paper admits the limitation of how far and accurate this result might be, and successfully addresses that by stating that field work will effectively compensate for such concerns. This is a very recent paper, so we are yet to see the true effect of such technological adaptation of this particular manner.