Term Project Proposal

Kristen Kwan, Benjamin Jakubowski, Evelina Bakhturina, Rafael Garcia Cano Da Costa

Business Problem

The New York City Department of Housing Preservation and Development (HPD) ensures that building owners comply with the City's Housing Maintenance Code (HMC) and the New York State Multiple Dwelling Law. These regulations provide the minimum housing standards for residential buildings in New York City.

If tenants believe their building or apartment do not meet these standards, they may file a housing complaint with HPD through NYC's 311 Citizen Service Center. After a housing complaint is filed, a housing inspector attempts to contact the building's managing agent to advise them that a complaint has been filed and that a violation may be issued if the condition is not corrected. The housing inspector then attempts to call the tenant who filed the complaint to see if the condition was corrected. If the condition was not corrected or HPD cannot reach the tenant, an inspector is sent to inspect the reported condition and determine whether a violation should be issued.

Housing inspections are labor intensive; in 2015, each inspection team completed an average of only 12.2 inspections per day[1]. Moreover, inspections often do not result in violations being issued. Based on our initial exploratory analysis (sampling the first 1000 records in a filtered subset of the HPD complaints dataset, with the filter passing records in which complaint status updates were entered on or after 08/01/2015), it appears there are six potential statuses for housing complaints:

1. In approximately 49.5% of complaints, a housing inspector conducts an inspection but does not issue any violations.
2. In approximately 35.5% complaints, the housing inspector is unable to access the property.
3. In approximately 9.5% complaints, the housing inspector completes an inspection and issues one or more violation.
4. Approximately 1% of complaints are redundant.
5. Approximately 3% of complaints are resolved when HPD contacts the tenant, and the tenant reports the problem has been resolved.
6. About 1.5% of records are missing complaint statuses.

A tool that can predict the most probable outcome of a housing complaint will aid HPD in effectively allocating human and financial resources and improve more tenants' quality of life and living conditions. Specifically, we envision the following use scenario: A tenant calls the NYC's 311 Citizen Service Center and registers a complaint. The model then assigns probabilities to the possible outcomes of the complaint. Based on the model results, HPD can prioritize inspecting complaints that are likely to result in a violation over complaints that are not likely to result in a violation. Moreover, if the model finds that the probability of successfully accessing the property is low, the housing inspector can make further attempts to coordinate the inspection for when the tenant is available.

[1] http://www1.nyc.gov/assets/operations/downloads/pdf/mmr2015/hpd.pdfa

Data Mining Problem

Our data mining goal is to classify the specific outcome of a complaint filed with the HPD. We will conduct supervised learning using existing datasets (described below) to construct a model that predicts the outcome of a housing complaint (the target variable) as a function of complaint, building, and census tract attributes. A record in our analytic dataset will correspond to a single complaint, and will include (i) the unique HPD complaint ID, (ii) the building's BBL (borough, block, lot) number, which serves as a unique building identifier and will allow us to construct our analytic dataset by merging several disparate datasets, (iii) additional features related to the complaint, building, or immediate neighborhood. More specifically, we plan to explore the following features:

- Complaint type (i.e. location of problems in building, type of problem(s), etc.)
- Building location
- Building age
- Building assessed value and modification history
- Building HPD violation history
- 2000 census-block level demographic and economic characteristics.

Data Details

We are planning to use data from five different datasets. First, our target variable will be constructed by processing "complaint status descriptions" found in the HPD Complaint Problems dataset. This dataset (and many of the others relevant to our problem) is produced by the HPD and made available on NYC Open Data. In addition to our target variable, this dataset includes detailed information about specific problems cited in a tenant complaint.

The remaining datasets will provide additional features, expanding the feature space for model selection and ultimately reducing the uncertainty of our predictions. Most of the supporting datasets are available on NYC Open Data. Ignoring census data (not yet obtained), the size of each of the datasets are described below:

- HPD Complaint Problems: 785814 records, 18 columns, 275.8 MB
- House Maintenance Code Violations (HMCV): 1009996 records, 30 columns, 388.8 MB
- Housing Maintenance Code Complaints: 436288 records, 15 columns, 40.9 MB
- NYC Pluto: 859469 records (split into 5 .csv files, one per borough) 83 columns, 432.9 MB
- Census data

Several of the relevant features are already available in one of the listed datasets. However, a number of features will need to be engineered; for example, we will need to group violations listed in the HMCV dataset by BBL number to construct features that summarize a building's violation history (such as number of violations issued over a defined time period, number of past violations where the managing agent was found non-compliant, etc.). Moreover, the features we anticipate including in our analytic dataset are found across multiple datasets. Thus, in addition to feature engineering, we will need to clean and merge multiple disparate datasets to construct our final analytic dataset.