Deliverable 1 - City of Cambridge Evictions Study

Team Members:
Tiancheng Zhu
Kevin Peters
Drew Abram
Syahrial Dahler

The data for this project has been collected and preliminary analysis has been performed. One of the strategic questions from the client is: How does tenant representation by attorney affect outcomes? To examine this, data was scraped from the Massachusetts court dockets from 2017 to the present. A sample of this data was studied to yield a preliminary answer to this strategic question.

The Massachusetts court dockets consists of data with the following categories: case number, status, plaintiff, prosecutor, defendant, defendant attorney, property, docket recording, court date, final judgment, judgment method, judgment payment, and final payment. First, relevant data was selected from this dataset. Namely, the judgment type would serve as the classification for the outcome of the case. The primary negative outcome was “judgment for plaintiff for possession and rent,” which led to a payment by the defendant. The outcome classification was binarized into plaintiff wins and plaintiff non-wins (which could fall under one of several categories, including notices of dismissal and resolutions between the plaintiff and defendant).

Other attributes selected for preliminary analysis include plaintiff, lawyer representation, judgment method, and judgment payment. Each of these attributes, aside from judgment payment, could be binarized. In many of these cases, the plaintiff was the Cambridge Housing Authority; the plaintiff data was categorized into Cambridge Housing Authority and other. Lawyer representation is easily binarized into representation versus no representation. Judgment method is the method of resolution and includes agreement between parties, agreement after mediation, and, most notably, dismissal after the defendant failed to appear in court. Judgment method was binarized into cases in which the defendant appeared in court and did not appear in court.

Each of these categories was correlated by a Pearson correlation coefficient. Table 1 summarizes the correlations between attributes.

CORRELATION	Plaintiff	Attorney	Judgment	Appear	Payment
Plaintiff	1	-0.02229	0.014265	0.017534	0.015351
Attorney	-0.02229	1	0.149076	-0.16381	0.058454
Judgment	0.014265	0.149076	1	-0.40033	0.170677
Appear	0.017534	-0.16381	-0.400331	1	-0.13411
Payment	0.015351	0.058454	0.170677	-0.13411	1

Table 1: Correlations Between Attributes, Court Cases

Pearson correlation coefficients range from -1 (perfect negative correlation) to +1 (perfect positive correlation). Values near zero reflect a lack of correlation between attributes.

Most notably, representation by an attorney did not have the highest correlation with judgment outcome (0.149); instead, whether the defendant appeared in court (-0.400) had the strongest correlation with the judgment outcome. In future iterations of data analysis, perhaps reasons for why the defendant did not appear in court could be examined; reasons for not appearing in court could include not having a means of transportation, not being able to take leave from work, or other personal factors that could be a result of socioeconomic status. Additionally, the strategic question of whether lawyer representation impacts judgment outcome could be rephrased to reflect two distinct subsets of data: defendants who appeared in court (and, therefore, may or may not have representation) and defendants who did not appear in court (which appears to be a dominating factor in outcome). These subsets rely on conditional relations between attributes and may yield clearer trends in data than examining the data as an unfiltered set.

Additionally, in our preliminary analysis, we ran K-means and GMM algorithms for the aforementioned attributes, using the binarized judgment attribute categories as a baseline for classification. The data could be classified into one of two categories, namely plaintiff win and plaintiff non-win. The classifications by K-means and GMM were then compared with the actual outcomes of the court cases. Percent classification as a positive case () and percent error relative to the actual outcomes are listed in Table 2. The number of attributes examined was varied (either all four attributes or just attorney representation and court appearance status). Additionally, even though the data is classified into two different categories (plaintiff win versus plaintiff non-win), the data could be readily divided into three categories; in future analyses,

perhaps these three categories could be examined to check whether they reflect the aforementioned subsets of defendant appearance versus defendant absence, in addition to plaintiff win versus plaintiff non-win.

	2 Clusters		3 Clusters	
	% Positive Classification, 2 Attributes	% Positive Classification, 4 Attributes	% Positive Classification, 2 Attributes	% Positive Classification, 4 Attributes
K-means	59.0	59.0	61.2	61.2
GMM	59.0	59.0	61.2	61.2

Table 2: Results for Preliminary K-Means and GMM Classification Methods

The K-means algorithm and GMM algorithm yielded the same results regardless of variations in number of attributes. However, dividing the data into multiple clusters yields a higher percent positive classification. This could support the idea that the data may need to be divided into subsets, or perhaps the binarized data needs to be categorized into more than two categories.

The actual positive classification percent from the judgment classification is approximately 68%. In future analyses, the positive classifications generated by each classification algorithm can be compared with the positive classifications of the judgment outcomes in more detail.

The code for the preceding analysis can be found in our team's GitHub repository.