Learning from the mundane: enhanced, data driven real estate management

Emma Lejeune (elejeune@stanford.edu), Mariya Markhvida (markhvid@stanford.edu)

Introduction

The status quo in the commercial real estate and property management industry is that the majority of operational data is not gathered in a centralized manner and robust data storage and dissemination methods are atypical. In this project, we examine an exception to this rule. Boxer Property, a property management company, provided a data set collected using Stemmons Enterprise software. This data set is unique because unlike most property management companies, Boxer Property centrally stores **all building operations data**, ranging from calling a locksmith and updating signage to major capital works on all properties.

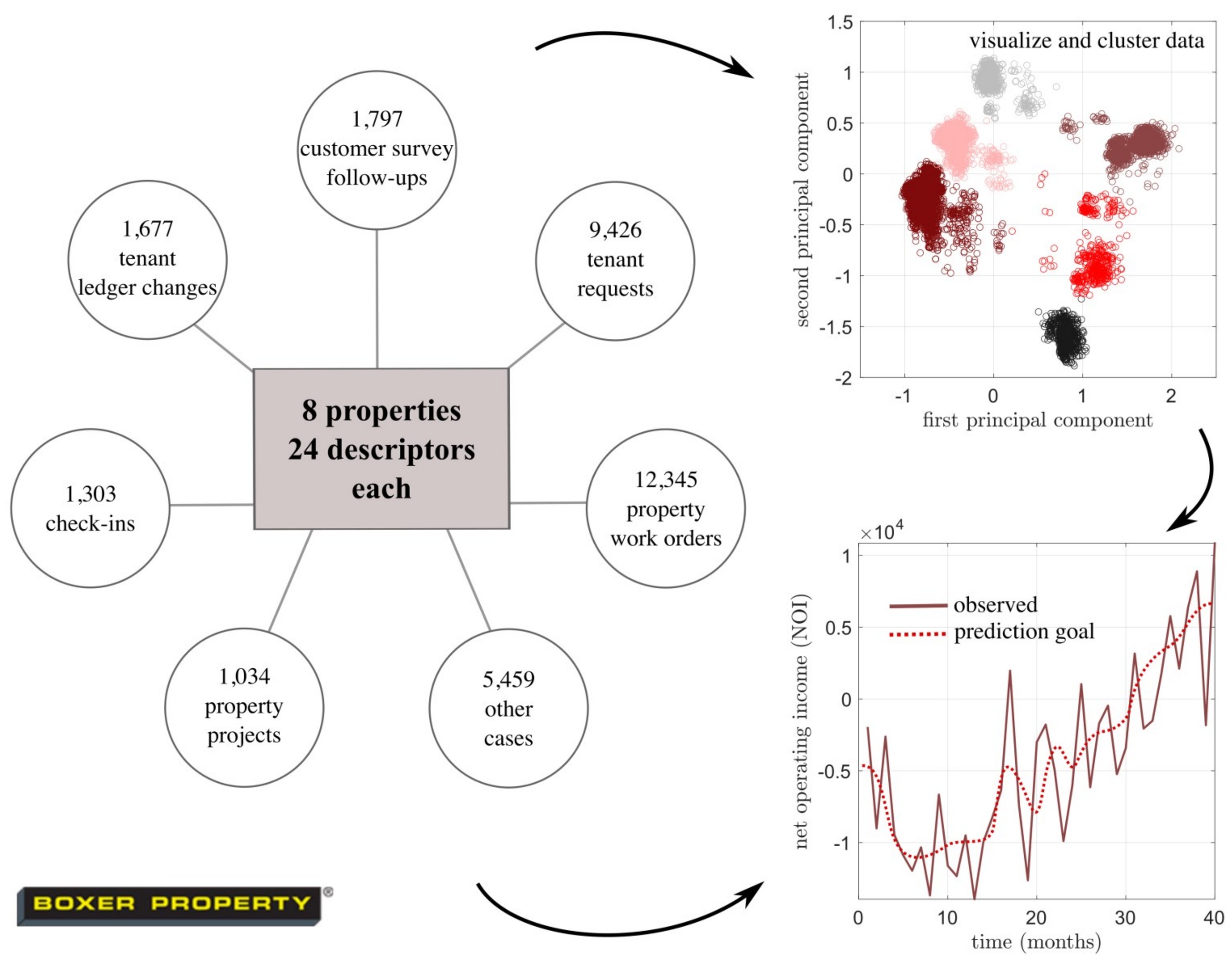


Fig 1. Given a large data set containing multiple data types, our first objective was to explore and visualize the data. Using the insight gained from this process, we constructed features from building operations data as inputs to a model predicting net operating income. The data from 8 properties located in Dallas was used for the analysis.

The objectives of this project were to (1) develop a method for turning the vast amount of categorical and text data into a usable format for machine learning purposes and to (2) incorporate building operations data in to a model for predicting property performance, in this case monthly Net Operating Income (NOI).

Features

For this data set, feature selection and construction were a major part of the project. For each of the data types shown in Fig. 1, we had to figure out how to turn predominantly categorical and text data in to something usable. Here we illustrate this process for the work orders.

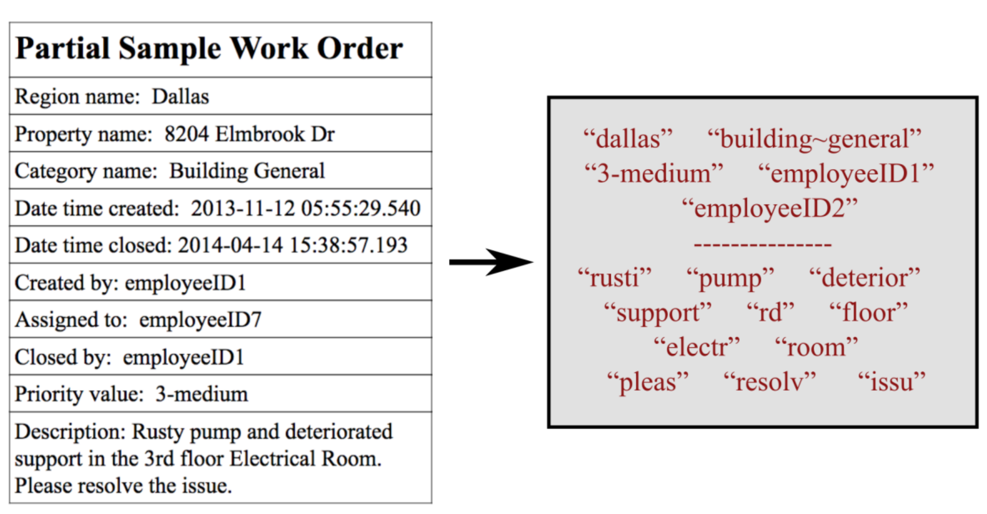
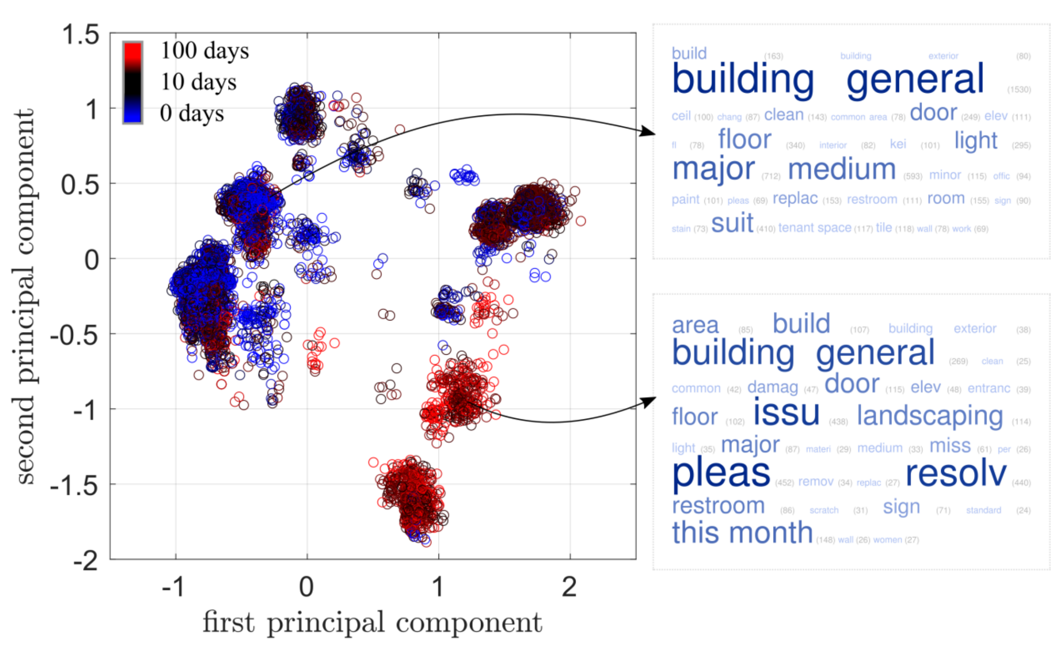


Fig 2. In it's raw format, a single work order is a 1x456 spreadsheet entry. Prior to analysis, meaningful features such as the ones shown in the partial sample work order here are extracted. Then, each work order is converted to a "bag-of-words" including both fixed tokens such as employee IDs and stemmed freely typed text, stemmed with the Porter stemming algorithm and stop words. The result is a matrix X.

Fig 3. We perform Principal Component Analysis (PCA) and compute the principal eigenvectors of $\Sigma = \frac{1}{m}xx^T$. Ten thousand work orders are shown in principal component space and obvious clusters emerge in the data. The color gradient indicates the duration of the work orders, which was not an input feature. The word clouds show which tokens (out of 3,358) were most prevalent in the cluster with shortest average duration and the cluster with the longest average duration, excluding the employee IDs.



After performing PCA, we clustered the work orders using k-means clustering (results shown in Fig. 1). As a result, for each property, for each month, we had 3 features per cluster: the number of "work order type created", "work order type is active" and "work order type closed". Similar features were created for other operations data shown in Fig.1. These features are referred to as the "time-variant" features.

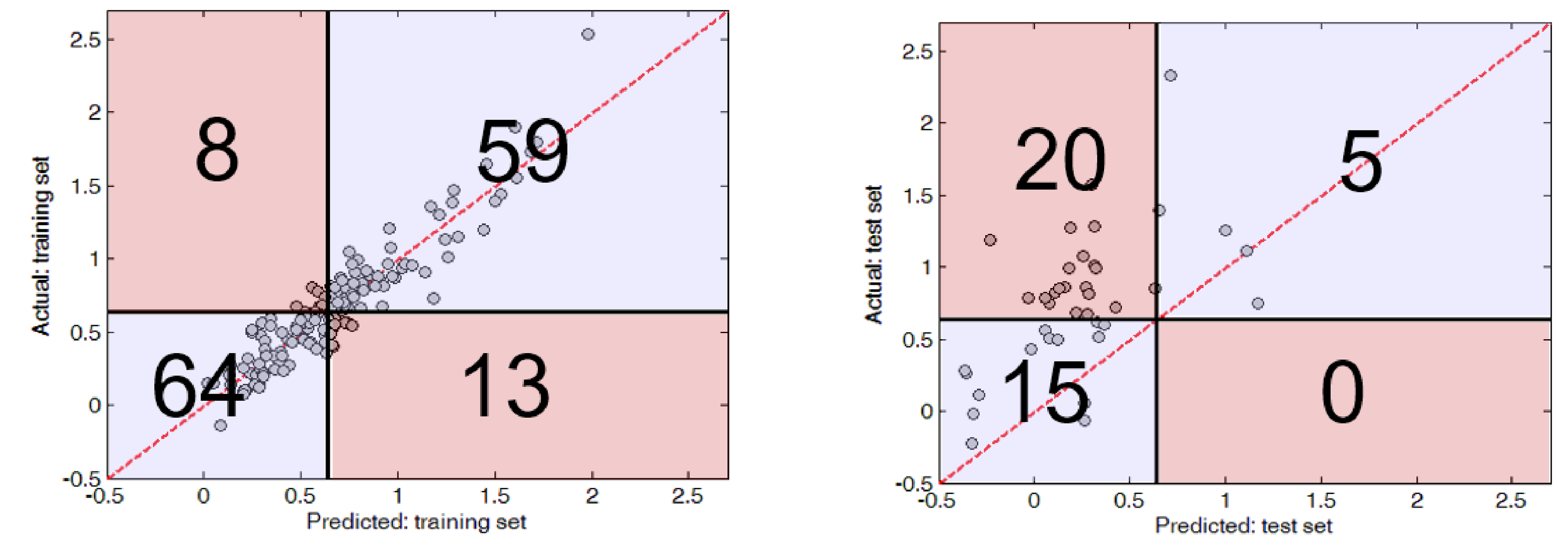
Predicting Property Performance

In order to predict monthly NOI (per ft^2), we used both the time-variant features based on operations data created using PCA and K-means clustering, and static property specific features such as average rent, gross area, year of acquisition, etc. We predicted classifications **"profitable"** (above median) and **"non-profitable"** on a monthly basis.

Panel Analysis:

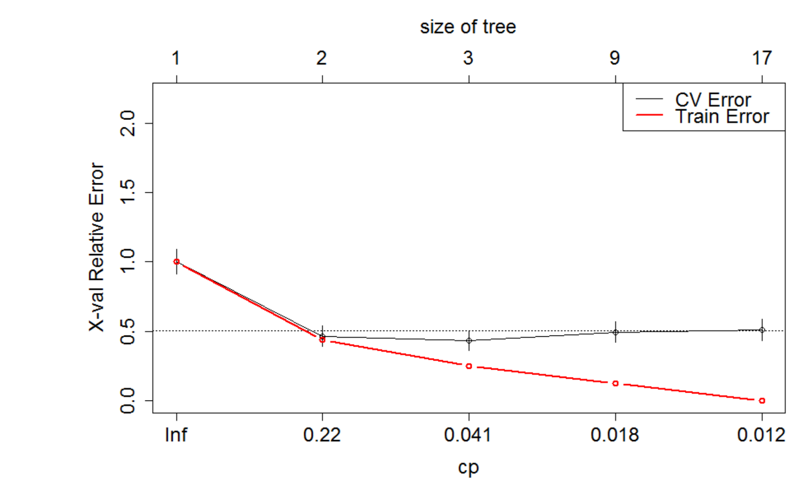
Used random effects model to capture property heterogeneity.

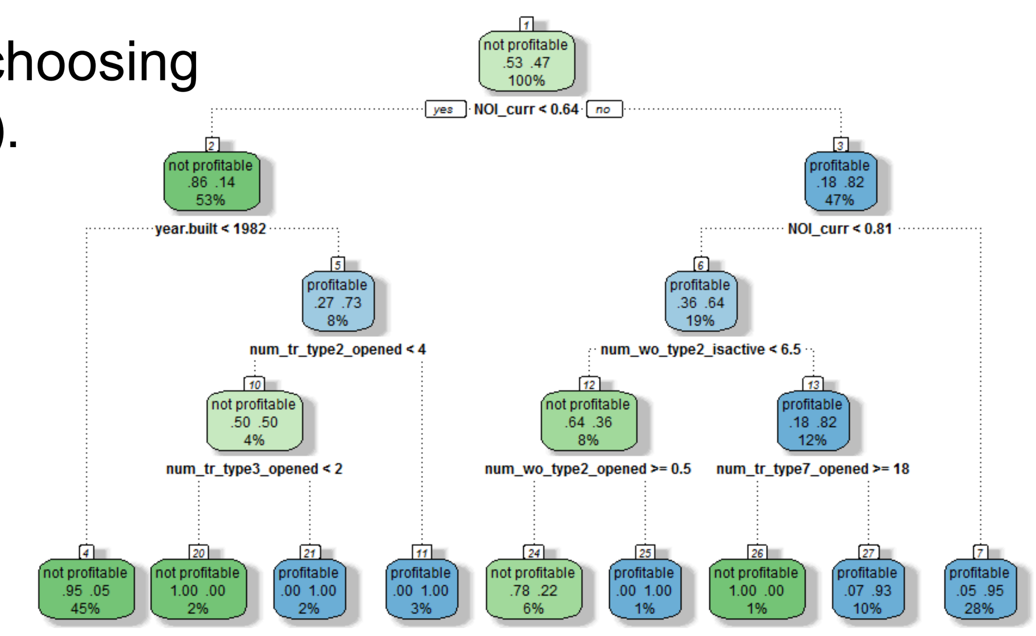
$$y_{it} = \alpha + \beta^T x_{it} + \mu_i + \epsilon_{it}$$



Decision Trees:

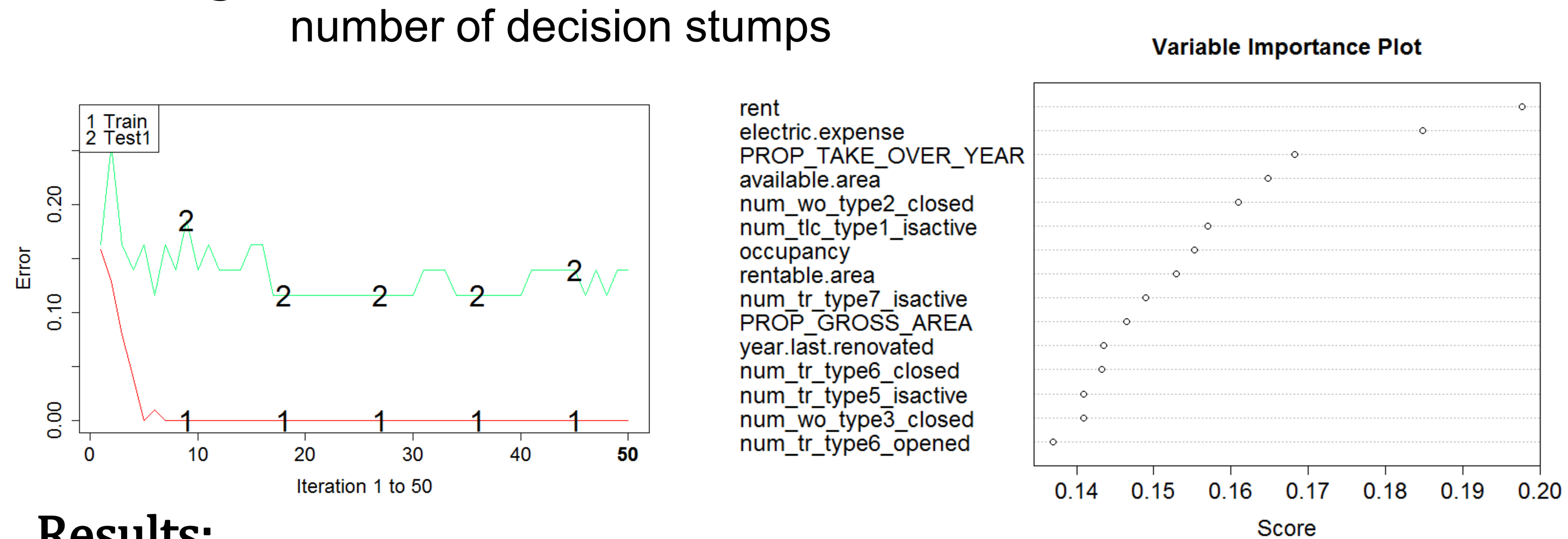
Used 10-fold cross validation for choosing optimal complexity parameter (cp).





Boosting:

hold-out cross validation (70/30) was used to select the number of decision stumps



Results:

	Training Error	Test Error	No. Training Samples	No. Test Samples
Panel Analysis	14.6%	50%	144	40
Decision Trees	5.6%	22.5%		
Boosting	0%	20%		

Discussion

One of the most interesting outcomes of our analysis was the clusters that emerged from the property work order and tenant request data. By superimposing work order duration on the data in principal component space, we could identify clusters of work orders that were taking much longer than average to fulfill. We also saw that using building operations data was important in predicting NOI, where the best model achieved 20% test error. **Future work** is required to better interpret these results. The bottleneck for this project was the amount of data we could handle given our hardware limitations. Therefore, the next steps would be to upgrade hardware and explore the data from regions all over the country. Furthermore, we believe that our general approach could be used to predict and see patterns in property performance and other aspects of real estate management, where online learning can be implemented to quantitatively predict future performance. This exploratory project shows the potential of machine learning on well aggregated real estate data.

References: Croissant, Yves, and Giovanni Millo. "Panel data econometrics in R: The plm package." *Journal of Statistical Software* 27, no. 2 (2008): 1-43.