Forecasting Eviction in San Francisco with Spatial Methods

Daniel Basman
Georgia Institute of Technology
Atlanta, Georgia, USA
dbasman2@gatech.edu

Gaurav Sett
Georgia Institute of Technology
Atlanta, Georgia, USA
gauravsett@gatech.edu

Abstract

We employ tax assessor data to forecast evictions in San Francisco. Using the proximate measure of eviction notices filed with the San Francisco Rent Board, we train a machine learning model to predict the likelihood of a filing on a given block in the city. The dataset analyzed includes a wide range of containing property characteristics, ownership information, and tax payment history of land parcels aggregated at the block level We investigate the usefulness of spatial features for this task. We compute averages of our tax data across neighborhoods as well as metrics representing each blocks position in a spatial network. Identifying these factors can help us gain a better understanding of the relationship between property characteristics and eviction likelihood for further research or more effective policy. This in turn can lead to more stable and affordable housing options for residents of San Francisco.

Keywords: machine learning, urban planning, spatial analysis

ACM Reference Format:
Daniel Basman and Gaurav Sett. 2023. Forecasting Eviction in San Francisco with Spatial Methods. In *Proceedings of Fall 2023 (CS 8803: Urban Computing)*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Web Page: http://danielbasman.com/SFEvictionForecasting

1 Introduction

1.1 Motivation

America has a housing affordability problem. With high demand and low supply, housing prices are outpacing stagnant wages, pushing more and more people out of their homes, with San Francisco a prime example of this crisis. The city's skyrocketing rents, limited affordable housing, and increasing income inequality have contributed to a precarious housing situation for many residents. Consequently, eviction has become a pressing issue, with profound implications for social stability, economic mobility, and public health. Understanding the factors that contribute to eviction and developing accurate forecasting models are crucial steps towards mitigating its negative impacts and informing targeted policy interventions.

Traditionally, eviction studies have relied on survey data, court records, and self-reported experiences to analyze the causes and consequences of eviction. Due to the specificity, frequency, and granularity needed to study eviction, such data sources are scarce and often unreliable. Thus, we are limited in our ability to develop a comprehensive understanding of eviction.

Modern data science techniques can provide a complementary quantitative approach to address these challenges. In particular, we believe that machine learning models can be used to identify key predictors of eviction from simple tax assessor data, which includes land value, building value, zoning, land use, property owners and more publicly available values. This data is fairly widely available and standardized across municipalities, which makes it easy to process and utilize. In particular, we believe we can gain insight through a thorough exploration of the spatial aspects of this data such as parcel size and distances. These features are especially relevant because they can be examined in relation to other aspects within the municipality such as transit, jobs, and more. This allows for independent analysis of various areas and a more factors with which we can form our analysis.

Indeed, while urban forecasting models with maximal predictive power will likely incorporate comprehensive, granular, temporal, and spatial data, we will focus on the construction of spatial features to identify their potential. Given the sparsity of relevant data, our paper aims to determine the advantages and disadvantages of a spatial-centric approach.

1.2 Problem Definition

Our project will focus on the city of San Francisco, which has a long history of housing affordability issues and a high eviction rate. We collect tax assessor parcel data and eviction notices records from San Francisco's Open Data project [DataSF 2023] to build a machine learning model that predicts the likelihood of an eviction notice filing for a given

CS 8803: Urban Computing, Atlanta, GA,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

block. In addition to features from the tax assessor data, we build a graph network of blocks in San Francisco to enable the computation of spatial features that may represent key neighborhood characteristics. In our graph, each block is a node, and edges are drawn to neighboring blocks within a distance threshold. For example, we can compute the average land value of surrounding blocks, or the most common zoning classification for the neighborhood.

Our work introduces two hypotheses. First, we hypothesize that the averages of the tax assessor data over the neighborhood of a block will increase the predictive capacity of our model. Second, we hypothesize that the node metrics of a block representing its contribution to the overall network structure.

2 Review

2.1 Background

There are a lot of varying factors that have been determined to play a role in the housing crisis that we are facing in the United States. Many of these factors have been previously discovered such as lowering house supply combined with rising house prices, but with the impact of recent events such as COVID-19 we have seen new factors arise and many previously known factors became more apparent [Schaeffer 2022].

Not only are rising costs pushing people out of their homes, but the lack of alternate supply doesn't allow viable options for them to move to. This causes serious problems for those who cannot afford rent and can end up getting evicted. Low income and moderately priced housing grants help persuade housing developers to build cheaper housing that is lower than the current market price. Limits and controls on rent have also been shown to prevent landlords from raising prices and forcing people out in favor of those who can pay [Uhler and Sisney 2016].

This may seem like it but it is not only a house-centered issue, eviction can cause serious long term damage to people which can put them at increased risk for instability and homelessness later in life. Also, it is still undetermined whether eviction is the start of a cycle of problems or a manifestation of this cycle, which makes it near impossible to properly assess [Collinson and Reed 2018].

Eviction is a systematic issue that is difficult to address and fix through policy due to our lack of understanding of its causes. A historical past of public investment into communities is a good opportunity to examine some possible policy choices that have previously been tested. They show that investment in these communities can have good effects, as well as the fact that investments into things that can damage the community like certain transit infrastructure, had the opposite effect leading to higher eviction and gentrification [Zuk et al. 2018]. It was also found that while there was good evidence showing how these societal phenomenons occurred, they were still difficult to interpret and process due to the lack of data. Inferences across sources can make up for some of this, but it still doesn't serve as the best proxy for real data. Gathering data was also a proposed solution which may work within a small land area but has its downsides such as provide a narrow amount of data for a specific region [Easton et al. 2020].

A large part of the housing issues is manifested in the problem of gentrification, which does have benefits such as higher quality housing, reduced crime, and jobs, but also large setbacks to the existing community. The changes brought about are usually exclusionary, with existing residents not benefiting from the addition of these new community infrastructures. Residents can be pushed out through higher prices and evictions, paving the way for landlords to charge more to higher income residents who move in. Longer commutes, substandard housing, and reduced economic opportunity are all problems which displaced residents can end up facing. These can cause problems for anyone especially children who grow up in these conditions and do not face the same opportunities that others have [Dragan et al. 2019].

Data on gentrification is difficult to accumulate for its own variety of reasons. First off, it happens within pockets, and thus to collect this you need people active within the community tracking this issue. This issue is also not an instant occurrence, data needs to be constantly collected over long periods of time before really solid analysis can be run [Easton et al. 2020]. Along with tracking features of a neighborhood, past property renting and ownership data is necessary to understand where people go when pushed out. This data is both confidential and difficult to gather from a single source, making for another larger hurdle that must be dealt with. Put together these encompass some of the largest, but not all of the constraints that are faced when attempting to follow this data.

Recently there have been some revelations in this area that have come in the form of special datasets released by the government or corporations collecting data, as well as creative solutions such as webscraping [Raymond et al. 2021]; ranging from neighborhood amenities on Yelp to eviction filings in court databases [Glaeser et al. 2020]. Even though new sources of data are coming about frequently in this digital age, we still are left with incomplete datasets which are difficult to apply broadly to housing policy in the United States. Combined with this issue, varying policies across states and cities also pose an issue in generalizing this data to make a robust national strategy to counter this problem. There is a need for a better method with which to investigate the topic of eviction in order to properly identify the core cause areas and match them with potential solutions.

2.2 Related Work

There have been a couple recent endeavours into eviction forecasting which have used advanced machine learning and forecasting models in order to predict evictions in various areas. The COVID-19 pandemic also served as a wake up call for this type of work as it became evident how many people nationwide are affected by eviction and its complementary factors. Many of them use demographic or spatial data in order to try and predict which variables will have the highest predictive rate of individuals being evicted.

Other features, such as rent control have been observed as having a direct affect on the likelihood of eviction. Given rent control is designed to assist those most at risk of eviction, this study shows just how much more likely these people are to be evicted from their living place in the future. This validates other demographic features that have been shown to have a correlation with eviction. While the study does not go deeper into what other specific features play the most pronounced role, it serves as a good baseline of understanding what units can be most immediately effected. [Gardner 2022].

Another study utilized time series regression forecasting in order to predict eviction based on various economic factors. This study was able to get a fairly accurate prediction of the amount of evictions in 2020 and 2021 based on the data from the previous 5 years, starting in 2015. While the forecast that was made was accurate, it also brought about an interesting look into the delayed evictions that were in place to due a moratorium during the COVID pandemic. This was predicted to cause a large uptick in evictions in 2022 due to the delays in prior years. This study reinforces the effect of policy making on evictions and shines light on how it can serve as a predictive mechanism for understanding the areas and people who are most likely to be at mercy of eviction. [Houghton et al. 2021]

More recently, developments at Rutgers University have brought about neural network models which outperform baseline assessments in their predictive accuracy. Past eviction data, labor statistics, and other demographic information was utilized to create a model which could make temporal predictions on evictions in upcoming months. These are solid measures shown by the MARTIAN model doing better than current methods such as XGBoost and SVM, on the data it was trained on, Dallas, TX, as well as other regions, previously unseen by the model. It also makes predictions within the scope of census tracks, which is closer than previous models have done. A model that is both accurate and generalizeable, as well as granular in its representations, is needed in order to understand the leading predictive factors of eviction. This methodology however, does not take into account spatial data, which in itself can serve as a valid feature upon which to make predictions. [Tabar, Jung, Yadav, Chavez, et al. 2022]

Spatial data can serve as a characteristic for determining evictions hotspots, and the specific areas which can be most effected. In combination with demographic data provided by the American Community Survey and satellite imagery, the WARNER model is able to outperform leading baselines by up to 8% on average. Similar to the MARTIAN model, WARNER provides accuracy on its trained data as well as unseen data in a different state. Unfortunately however, the WARNER model depicts results on a county wide granularity, which is not enough to accurately predict eviction within a single city or even subsections of it. [Tabar, Jung, Yadav, Wilson Chavez, et al. 2022]

Current state of the art methods, while performing with good accuracy, utilize primarily demographic data in order to predict eviction and sometimes suffer from being too large in their geographic scope. Our methodology of using spatial features shows promise in finding the parcels and areas most at risk for eviction based on their relation to other parcels in the area. This allows for future predictions of eviction within a city or potentially even specific neighborhoods.

3 Methods

3.1 Data

We collected evictions and property tax data from DataSF, San Francisco's Open Data project [DataSF 2023]. To track evictions, we use the Eviction Notices dataset filed with the San Francisco Rent Board. While a notice of eviction does not necessarily lead to eviction, this measure serves as a strong proxy for the phenomena of displacement we are interested in predicting. The dataset includes basic information like the address of the notices and the filing date. Additionally, it provides the reason the eviction notice was given, such as for non-payment, breach of contract, nuisance, illegal use of unit, access denial, unapproved subtenants, capital improvement, and more. This data is collected from 1997 to 2022.

To track property, we use the Assessor Historical Secured Property Tax Rolls dataset. This dataset includes tax information such as the owner name, assessed property value, assessed fixture value, assessed improvement value. It also provides land use information such as the zoning classification. Moreover, it has information about the buildings on the property such as the number of bedrooms, bathrooms, rooms, stories, as well as the area of the basement and lot. This data is collected from 2007 to 2019.

Parcel geometry changes over time, so we use the more stable spatial feature of blocks to conduct our analysis. For each year, each parcel is grouped with its encompassing block. We take the mean of each tax feature over the parcels in a block. We also create dummy variables for the land use code of the parcels, activated when at least one of the parcels in a block has the respective code. We create a label feature based on if a block contains a parcel with an eviction filing for a given year.

We also standardize the numerical tax features by year to account for exogenous changes in property value such as economic growth or inflation. This means that we take each value, subtract the mean value of that year across all blocks, and divide by the standard deviation of that year.

The features from the tax dataset are found in Table 1.

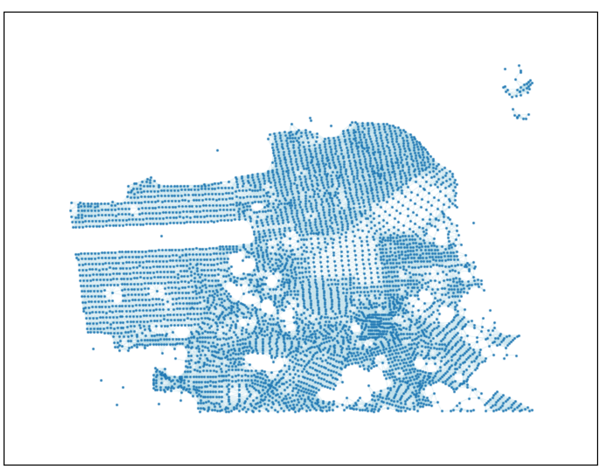
Table 1. List of Tax Features Used in Model

1. *Assessed Land Value:* Assessed value of land.
2. *Assessed Personal Property Value:* Assessed value of personal property.
3. *Assessed Fixtures Value:* Assessed value of fixtures on property.
4. *Assessed Improvement Value:* Assessed value of property improvements.
5. *Number of Units:* Number of units on the property.
6. *Number of Rooms:* Number of rooms on the property.
7. *Number of Bedrooms:* Number of bedrooms on the property.
8. *Number of Bathrooms:* Number of bathrooms on the property.
9. *Property Area:* Square footage of the lot.
10. *Use Code:* Group of property class code (ex. Single Family Residential, Commercial).
11. *Longitude:* Longitude of the block.
12. *Latitude:* Latitude of the block.

3.2 Feature Engineering

To build our spatial network, we create a graph where each block is a node. Then, we draw an edge between all nodes within 2% of the maximum distance between blocks. Our mean node had 20.4 neighbors. The 25th, 50th, and 75th percentiles were 16, 20, and 25 respectively. Few nodes have 0, and they are quite isolated on the map. Nodes in downtown have up to 53 neighbors.

Figure 1. Spatial graph for the City of San Francisco



To assess our first hypothesis, we computed features of neighborhood averages for our tax data. For features 1-9 in Table 1, we compute the mean value across each vertex's neighbors to create our neighborhood features.

Related to our first hypothesis is the claim that our tax features have spatial autocorrelation. In our data exploration, we applied Moran's I statistic for evaluation. We found that the features assessed land value, assessed improvement value, number of units, and property area are spatially autocorrelated with significance $p < 0.05$.

To assess our second hypothesis, we computed features based on each block's contribution to the network structure of our graph using the NetworkX library [Hagberg et al. 2008]. These found in Table 2.

Table 2. List of Graph Features Used in Model

1. *Degree:* Number of edges incident on the node.
2. *Clustering:* The fraction of possible triangles through that node that exist.
3. *Betweenness Centrality:* The sum of the fraction of all-pairs shortest paths that pass through the node.
4. *Closeness Centrality:* The reciprocal of the average shortest path distance to the node over all other reachable nodes.
5. *Eigenvector Centrality:* The centrality based on the principle eigenvector of the network's adjacency matrix, also considering the centrality of neighbors.
6. *Page Rank:* The importance of a node based on the importance of nodes it links to, computed by the convergence of random walks.
7. *Constraint:* The extent to which a node is invested in those nodes that are themselves invested in the neighbors of the node.

3.3 Model

The XGBoost (eXtreme Gradient Boosting) architecture is machine learning algorithm designed for solving classification and regression problems [T. Chen et al. 2015]. It is widely used for similar tasks due to its efficiency, effectiveness, and interpretability. We believe this model's simplicity and power will enable us to best illustrate the general capacity of the features we are interested in evaluating.

It is an advanced implementation of the gradient boosting framework, which combines the strengths of multiple weak learners (usually decision trees) to create a robust and accurate model. XGBoost has gained immense popularity among data scientists and researchers due to its exceptional performance, scalability, and ability to handle a wide range of feature types. The core idea behind gradient boosting is to iteratively add weak learners to the model while minimizing a loss function. At each iteration, the algorithm computes the gradient of the loss function with respect to the current model's predictions and fits a new weak learner that best aligns with the negative gradient. The weak learner's output is then combined with the existing model through a weighted sum to update the model. This process is repeated, and new weak learners are sequentially added until a predefined stopping criterion is reached, such as a maximum number of iterations or a convergence threshold. XGBoost improves upon the traditional gradient boosting framework by introducing several key enhancements including online learning, regularization, sparse learning, and more.

Our XGBoost classifier is initialized with 100 estimators, a maximum depth of 10, and a learning rate of 0.05. These hyperparameters were chosen through grid search.

We sequentially train our model over from 2007 to 2019. We treat each year as a batch. Specifically, we use our features at time step t to predict whether the same block at step $t+1$ will have an eviction filing. At each step t, we evaluate our model with $k = 5$ fold validation, using the last fold for the next step's training.

4 Results

In this section, we conduct an evaluation of our model and qualitative analysis of its outputs.

4.1 Model Evaluation

For our prediction task, we tested our model with four different sets of features: tax features found in Table 1, neighborhood features computed by averaging values in Table 1 from each node's neighborhood, graph features found in Table 2, and all features. The performance of each model measured by accuracy, precision, recall, and f1 are found in Table 3.

Table 3. Performance of Models by Features

	all	tax	neighborhood	graph
accuracy	0.8833	0.8787	0.8763	0.8783
precision	0.4580	0.3848	0.3060	0.3606
recall	0.1096	0.0979	0.0689	0.0844
f1	0.1760	0.1545	0.1116	0.1353

All models performed relatively similarly. Each had very high accuracy, able to predict nearly 90% of the time which blocks in the following year would have an eviction filing. However, the poor precision scores indicate the proportion of true positives of all predicted positives is low. Indeed, the poor recall scores indicate the model is failing to identify most of the positives.

Eviction forecasting is a difficult task, depending of highly local and individual phenomena that is difficult for a simple model with such limited data to predict. However, it appears that the marginal increase in f1 between the all features model and the tax data only model indicates moderate evidence for both of our hypotheses.

To better understand how the features we computed contribute to our model's predictive capacity, we can look into the weight given to each feature by our model. Given the architecture of the XGBoost classification model, we are defining weight as the number of times features are used to split the data across trees.

In our all features model, we can immediately see the cause of our high accuracy compared to the low f1. The primary feature is the number of units on a block. Our model understands most evictions will happen where most residential units are. Next is the eviction filing count of neighboring blocks, followed by eigenvector centrality. Generally, our all features model gives greater weight to the neighborhood features than the block level features. This may be an indication that the model can infer trends from patterns in the surrounding neighborhood. Further details can be found in Figure 3.

Our tax features model gives the greatest weight to the number of units as well, in addition to the land value and improvement value. This is followed by information about the structure such as number of rooms, property area, and number of bedrooms.

In contrast, our neighborhood features model gives the greatest weight to the eviction count of neighbors, followed by features indicating commercial properties in the neighborhood.

Lastly, our graph features model gives greatest weight eigenvector centrality, longitude, and latitude, as shown in Figure 4.

4.2 Qualitative Analysis

It appears that this approach fails to produce features that enable precise forecasting of eviction filings given the low f1 score. However, qualitative analysis shows that this model has produced several interesting insights. Our forecast using data from 2019 to predict eviction filings in 2020 is shown in Figure 2.

One common trend across all models is that the features with greatest weight are correlated with the block being downtown where most units. Indeed, this is where we would expect to see most evictions. Interestingly, each model, with different sets of features, were able to reconstruct this spatial pattern and product predictions with relatively similarly accuracy. For the tax features, it was the number of units and land value. For the neighborhood features, it was the commercial properties in the neighborhood. For the graph network, it was longitude and latitude.

Additionally, the neighborhood features can help smooth out coarse datasets. This can especially be seen with the eviction filing counts neighborhood feature. Each eviction is a highly individualized phenomenon, but aggregation over a spatial graph can not only help predictive models but also enable broader understanding.

While spatial graph features may not initially seem like they offer much insight into an economic problem, our model has made clear that we can reconstruct much of the economic data in our tax dataset by the nature of the distribution of blocks, evidenced by the similar graph feature model performance.

The eigenvector feature given high weight in the all features model and the graph features model appears to be an interesting quirk. The measure is concentrated in the southern neighborhood of Bernal Heights which appears to continually have few eviction filings.

Our clustering and constraint features appears to roughly represent blocks on the edges of separated areas within the city. These separations may be due to a park, a waterfront, hills, or a road. They are also activated by large blocks such as those south of Market Street (SoMa), a largely commercial space with many offices.

Visual inspection of the forecasts indicate that while the model often fails to identify the exact block eviction filings will occur on, it is generally able to identify broader neighborhoods where evictions are likely. This holds even outside of Downtown in areas like the Richmond District, Mission District, and Bayview.

5 Discussion

This paper illustrates the usefulness of the creation of spatial features for urban analysis. We take the example of eviction forecasting to show how averaging values over a neighborhood and metrics representing the contribution of a node to the structure of the graph.

Our model is accurate in its year by year predictions of eviction filings, trading off against precision and recall. The models learn to identify parts of town where housing is prevalent. However, visual inspection shows the model is close in its spatial proximity between blocks where it expects eviction filings and blocks where they occur across the city.

The approach outlined in this paper may prove useful for enabling more complex models to grasp spatial features of urban trends that may combine well to empower greater prediction.

The ability for each model to learn where housing is concentrated despite having different representations indicates that the approach may be useful in situations where data is sparse or information about individual neighborhoods is not know. For example, if one were to conduct a study about housing within a broad region with differing local data collection, the graph features may be able to reconstruct missing information with only the location of units of land.

Future work may develop the temporal aspects of forecasting to better understand how the spatiotemporal data can help with our prediction task. It can also investigate other localities to determine the extent to which these studies generalize. Furthermore, future researchers may want to incorporate census data to see the interaction between tax data and demographic information. Lastly, the specific prediction task could be changed to determine the usefulness of these features for general urban analysis.

6 References

Tianqi Chen et al.. 2015. “Xgboost: extreme gradient boosting.” *R package version 0.4-2*, 1, 4, 1–4.

Robert Collinson and Davin Reed. 2018. “The effects of evictions on low-income households.” *Unpublished Manuscript.[Google Scholar]*, 1–82.

DataSF. 2023. *Eviction Notices.* (2023). https://data.sfgov.org/Housing-and-Buildings/Eviction-Notices/5cei-gny5.

Kacie Dragan, Ingrid Ellen, and Sherry A Glied. 2019. *Does gentrification displace poor children? New evidence from New York City Medicaid data.* Tech. rep. National Bureau of Economic Research.

Sue Easton, Loretta Lees, Phil Hubbard, and Nicholas Tate. 2020. “Measuring and mapping displacement: The problem of quantification in the battle against gentrification.” *Urban studies*, 57, 2, 286–306.

Max Gardner. 2022. “The effect of rent control status on eviction filing rates: Causal evidence from San Francisco.” *Housing Policy Debate*, 1–24.

Edward L Glaeser, Michael Luca, and Erica Moszkowski. 2020. *Gentrification and Neighborhood Change: Evidence from Yelp.* Tech. rep. National Bureau of Economic Research.

Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. *Exploring network structure, dynamics, and function using NetworkX.* Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

James Houghton, Hugh Keenan, Nandini Swami, Alexandra Weber, and Sachin Mathew. 2021. “Forecasting Evictions to Determine Areas at High Risk of Housing Insecurity.” In: *2021 IEEE MIT Undergraduate Research Technology Conference (URTC).* IEEE, 1–6.

EL Raymond, S Stein, and V Haley. 2021. *Atlanta Region Eviction Tracker.* Tech. rep. Georgia Institute of Technology.

Katherine Schaeffer. 2022. “Key facts about housing affordability in the US.”

Maryam Tabar, Wooyong Jung, Amulya Yadav, Owen Wilson Chavez, Ashley Flores, and Dongwon Lee. 2022. “Forecasting the Number of Tenants At-Risk of Formal Eviction: A Machine Learning Approach to Inform Public Policy.” In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence.*

Maryam Tabar, Wooyong Jung, Amulya Yadav, Owen Wilson Chavez, Ashley Flores, and Dongwon Lee. 2022. “WARNER: Weakly-Supervised Neural Network to Identify Eviction Filing Hotspots in the Absence of Court Records.” In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3514–3523.

Brian Uhler and Jason Sisney. 2016. *Perspectives on Helping Low-income Californians Afford Housing.* Legislative Analyst’s Office.

Miriam Zuk, Ariel H Bierbaum, Karen Chapple, Karolina Gorska, and Anastasia Loukaitou-Sideris. 2018. “Gentrification, displacement, and the role of public investment.” *Journal of Planning Literature*, 33, 1, 31–44.

7 Appendix

Figure 2. 2020 Eviction Filing Forecast
Legend represents model predicted probability of filing, red markers show true filings.

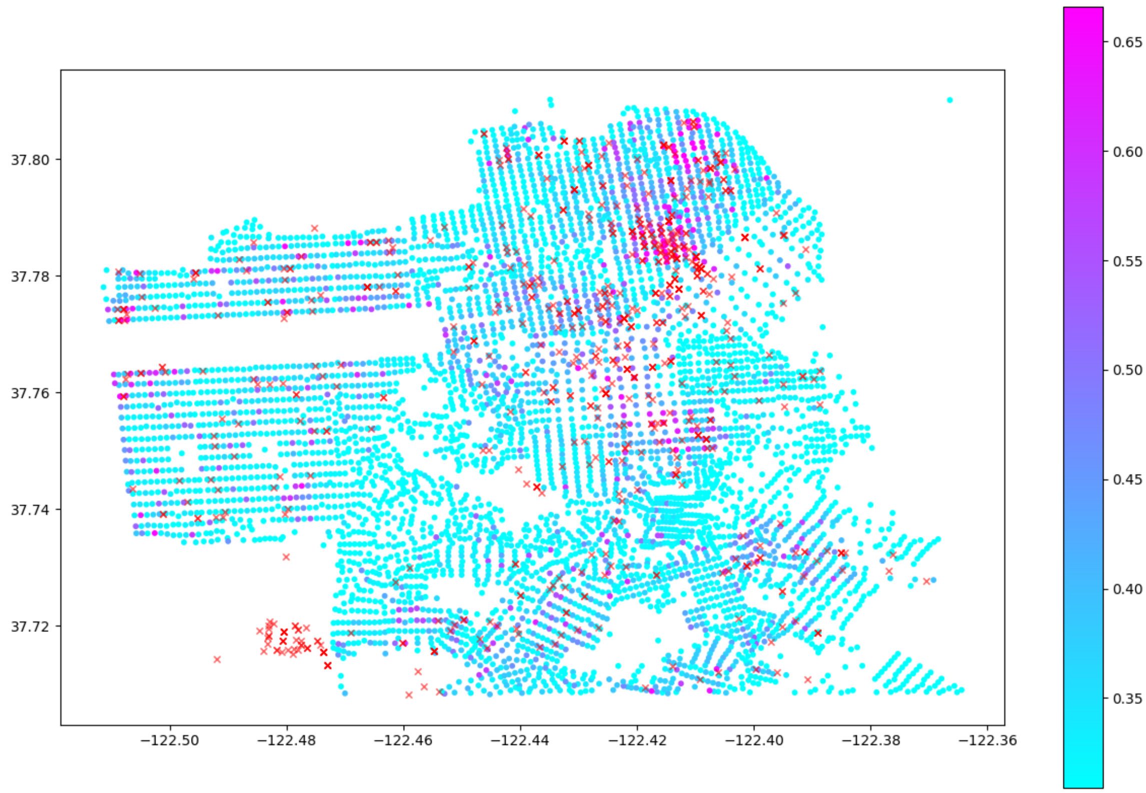


Figure 3. All Feature Model Importances

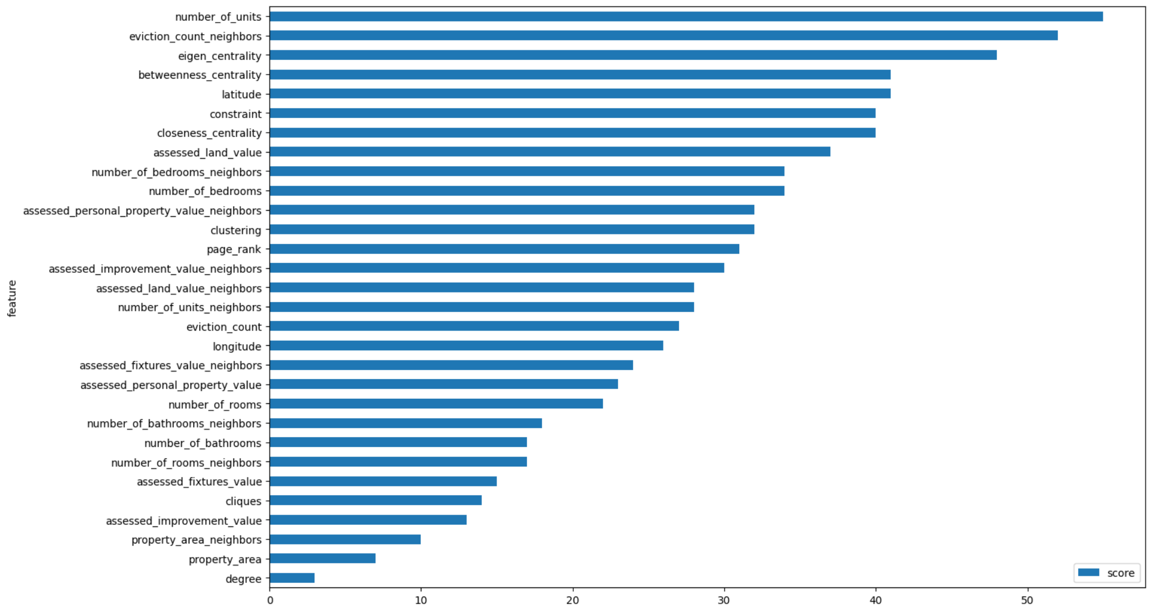


Figure 4. Graph Feature Model Importances