Identifying Vacant Lots to Reduce Violent Crime in Dallas, Texas

Laura Lazarescou
Southern Methodist University, llazarescou@mail.smu.edu

Andrew Mejia
Southern Methodist University, mejiaa@smu.edu

Tina Pai
Southern Methodist University, tpai@mail.smu.edu

Sabrina Purvis
Southern Methodist University, spurvis@mail.smu.edu

Robert Slater
SMU, rslater@mail.smu.edu

See next page for additional authors

Follow this and additional works at: https://scholar.smu.edu/datasciencereview

Part of the Data Science Commons

Recommended Citation
Lazarescou, Laura; Mejia, Andrew; Pai, Tina; Purvis, Sabrina; Slater, Robert; and Wilson-Chavez, Owen (2021) "Identifying Vacant Lots to Reduce Violent Crime in Dallas, Texas," SMU Data Science Review: Vol. 5 : No. 2 , Article 6.
Available at: https://scholar.smu.edu/datasciencereview/vol5/iss2/6

This Article is brought to you for free and open access by SMU Scholar. It has been accepted for inclusion in SMU Data Science Review by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
Identifying Vacant Lots to Reduce Violent Crime in Dallas, Texas

Authors
Laura Lazarescou, Andrew Mejia, Tina Pai, Sabrina Purvis, Robert Slater, and Owen Wilson-Chavez

This article is available in SMU Data Science Review: https://scholar.smu.edu/datasciencereview/vol5/iss2/6
Identifying Vacant Lots to Reduce Violent Crime in Dallas, Texas

Laura Lazarescou, Andrew Mejia, Tina Pai, Sabrina Purvis, Robert D. Slater

1 Master of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
2 Child Poverty Action Lab, 1808 S. Good Latimer Expressway Suite 102 Dallas, TX 75226 USA
Laura Lazarescou (llazarescou@mail.smu.edu)
Andrew Mejia (mejiaa@mail.smu.edu)
Tina Pai (tpai@mail.smu.edu)
Sabrina Purvis (spurvis@mail.smu.edu)
Owen Wilson-Chavez (owen@childpovertyactionlab.org)
Robert D. Slater (rslater@smu.edu)

Abstract. Vacant lots have been associated with community violence for many years. Researchers have confirmed a positive correlation between vacant lots and vacant buildings with increased violence in urban and rural geographies. However, identifying vacant lots has been a challenge, and modeling methods were largely manual and time-intensive. This prevented cities and non-profit organizations from acting on the information since it was expensive and high-risk to develop remediation programs without clearly understanding where or how many vacant lots existed.

The primary objective of this study was to provide a predictive model that accelerates and improves the accuracy of prior land classification methods. Labels for 2019 vacant lots from Child Poverty Action Lab (CPAL) were used as the source of truth for model development. Public data from the City of Dallas and Dallas Central Appraisal District (DCAD) were used to determine land value, crime incidents, code violations, and certificates of occupancy. These features were mapped onto lot locations using Geospatial Information System (GIS) modeling techniques.

The study concluded that an XGBoost model with five simple factors: land value, improvement or the building value, total value, land size, and division code offers the best balance of performance and simplicity.

1 Introduction

Vacant lots are a concern for cities in the United States. Past studies have shown that vacant properties experience illegal dumping that may lead to contamination, reduce property values, and offer an unmonitored environment for unlawful behavior. In this study, a vacant lot is defined as a neglected parcel of land with no buildings (US EPA). This is an important distinction as it differs from a lot that is intentionally vacant,
such as a public park or greenbelt. A critical challenge that comes with vacant land is described by “the ‘incivilities’ theory, [which] suggests that physical incivilities, such as abandoned vacant lots, promote weak social ties among residents and encourage crimes, ranging from harassment to homicide” (Branas, 2011, p. 1297).

The City of Philadelphia has experienced high crime. They found that vacant, unkempt lots were associated with substantial gun violence. By applying small interventions over two months, such as clearing debris and laying hydro-soil, they reduced crime by 37% and vandalism by 39% over the next three years where interventions occurred. Furthermore, criminal activity was not simply displaced; it was eliminated in the areas where these interventions were applied (Dengler, 2018).

The City of Dallas recorded more than 200 homicides in 2019, a statistic they have not seen since 2007. In addition, aggravated assaults increased by more than 1,000 between 2018 and 2019. In response to this, and directly following a speech at the funeral of a 9-year-old victim, Dallas Mayor Eric Johnson created the Mayor’s Task Force on Safe Communities (Johnson, 2019). This task force’s goal was to identify evidence-based solutions for reducing urban gun violence. These recommendations must reflect the lived experience from the community, be backed by data and research, and be outside the realm of standard policing (Johnson, 2019). Dallas intends to pursue the same methods for prevention as proven effective in Philadelphia, but they cannot quickly identify where there are genuinely vacant lots.

This research aimed to uncover key attributes which distinguish a vacant piece of land from other parcels. By developing a machine learning classifier, the study may accelerate and improve the quality of vacant lot determination, so that city, county, and state officials may pursue affordable remediation programs.

### 2 Literature Review

The study of vacant lots and their impact on communities is not a new topic. This review revealed factors that were commonly associated with vacant lots. Prior vacant lot research, data sources and methods were also considered.

Whether the vacant building is in Philadelphia or Mexico City, the “broken windows theory” suggests that one broken window will lead to more broken windows, and ultimately the decline of a neighborhood. A study of vacant buildings in Mexico City, Mexico, using principal component analysis (PCA) and multilevel random intercept models, demonstrated that social disorder and physical disorder are not independent (Vilalta, 2020). Vacant buildings and alcohol outlets have also been shown to increase gun violence. Alcohol Outlet Density (AOD), in conjunction with empty buildings and vacant lots, contributes to violent crime (Lardier, 2020).

Greening is making a space more environmentally friendly and inviting by adding plants to the area. Many urban renewal projects have involved remediation or removal of waste and debris and planting trees or urban gardens in vacant lots. Vacant lot remediation can improve the mental health of residents (South, 2018). In this Philadelphia-based study, residents who lived near a greened or remediated vacant lot self-reported a 62.8% reduction of poor mental health. Empty lot management can also lower a city’s overall cost of operations. Also, in Philadelphia, another study found that
“… taxpayer and societal returns on investment for the prevention of firearm violence were $5 and $79 for every dollar spent on abandoned building remediation and $26 and $333 for every dollar spent on vacant lot remediation.” (Branas, 2016, page 2158)

The task of identifying vacant parcels of land has been researched in several cities. At the time of this study, no published research was found on identifying vacant lots in the City of Dallas, TX. A typology of urban vacant land was created in Roanoke, Virginia, cataloging five types of vacant land (Kim, 2018). The research was also done in Phoenix, AZ, to identify metropolitan vacant land for greening, cooling, and agriculture (Smith, 2017). Additionally, St. Louis, MO, had a project to understand the health of the neighborhoods, especially as it pertains to the disinvested areas of predominantly Black populations, which had a high concentration of vacant land (Prener, 2020).

These studies used a combination of tax assessor data, imagery, and geographical information system (GIS) techniques to address the task. Kim (2018), Smith (2017), and Prener (2020) used tax assessor data as part of their methods. In the study of urban vacant land in Roanoke, Virginia, Kim looked at both public and private property, used public data sources, municipal statistics, historical records, previous land use, crime, property assessor data, and GIS (Kim, 2018). Smith used agricultural images, cadastral data from county tax assessors, and elevation data from the US Geological survey to derive land classifications and maps (Smith, 2017). Prener used indicators from the city’s parcel ownership and tax database and longitudinal data of building permits (Prener, 2020).

Anderson et al. (1972) described a list of metrics for a useful land use classification system. He asserts that a system should have an accuracy of roughly 90%, and results should be repeatable and generalize on future similar data sets. Land classification systems may be operationalized using hierarchical and heuristic decision tree methods that identify land systems by their most dominant ecologic or socioeconomic components (Ornetsmüller, 2018).

To pivot away from rule-based classification and take advantage of available data and machine learning techniques, three types of algorithms were considered for their application to classifying vacant parcels. While none of these methods have previously been applied explicitly in the scope of Child Poverty Action Lab (CPAL) research, they provide an initial basis of modeling techniques that can be applied to the classification of vacant parcels.

Grippa, Georganos, Zarougui, Bognounou, Diboulo, et al. (2018) propose a workflow for mapping urban land use at the granular street block level using remote sensing and spatial data. This study utilized a data model based on the OLAP framework for GIS. (Pestana, da Silva, and Bedard, 2005)

Wang, Deng, and Wang (2020) demonstrated that gradient boosted techniques are particularly effective when a classification problem is imbalanced. The authors also note the highly scalable nature and broad applications of tree techniques when classifying largely imbalanced classes. Chang, Chang, and Wu (2018) demonstrated boosted tree techniques could outperform self-organizing algorithms and single-stage classification techniques in terms of acceptable metrics such as receiver operator curve (ROC), precision, and recall when attempting to model imbalanced classification problems.
This study hypothesized that tax, crime, and code violations are valuable features in predicting vacant lots.

3 Methods

This analysis was conducted using a machine learning pipeline. Data preparation was extensive, requiring GIS shapefiles to act as the foundation and aggregator of tabular data. Once the data structures were created, a classical machine learning classifier was developed using Random Forest and XGBoost model types. Models were prioritized based on F1 score and model simplicity. To further determine a model’s utility, the false omission rate was reviewed. Statistical importance for features was determined by using binomial logistic regression and analysis of variance Wald Chi-Square for model coefficient techniques.

3.1 Data Sources and Preparation

Data were gathered from three primary sources: CPAL, DCAD, and the City of Dallas. Datasets were prepared according to the structure of the data and the relationship to other features. The account number was the unique identifier that specified unique GIS locations and provided a primary key to join with other DCAD data. The City of Dallas Data was merged using special mapping techniques, then exported to tabular format. Appendix II includes a complete list of software configuration details.

CPAL Data

CPAL provided a list of Account numbers or lot identifiers for lots that were classified as vacant in the “Mayor’s Task Force on Safe Communities” (Johnson, 2019). Lots that were not present in this list were classified as not vacant. All lots were within the boundaries of the City of Dallas and Dallas County. This data served as the annotations that were used to train and validate the machine learning models. Figure 1 provides a map of vacant lots based on CPAL annotations. Restricting data to the area within the City of Dallas minimized information “leakage”. This is the propensity for some of the features within the feature space to “leak” information about the classification of the target class during the training of the machine learning model, which may not be available during prediction and the production deployment of the model (Coron, Naccache and Kocher 2004).
Figure 1. Vacant Lots within the City of Dallas and Dallas County, CPAL 2019

Dallas Central Appraisal District (DCAD) Data

DCAD is the taxing authority for Dallas County. Therefore, the dataset was developed to describe taxable entities, their owners, and exemptions. Not all features available from (DCAD GIS 2019) and (DCAD Data 2019) were relevant to classifying parcels of land. The study focused on features that CPAL and the researchers believed directly related to vacant lots and their classification. A data dictionary of all features may be found in Appendix I.

GIS shapefiles provided a GIS description of all parcels in Dallas County. Dallas County extends beyond the City of Dallas, and the City of Dallas is present in five counties (Maghelal, P., Andrew, S., Arlikatti, S., & Jang, H. S. 2014.), so GIS shape
files were trimmed to focus on the City of Dallas and Dallas County locations exclusively. DCAD data included:

- Account information for the taxed individual or enterprise.
- Appraised entities, which may include land, improvements (buildings), and business personal property.
- Expanded detail for Account, Land, Improvements, Residential and Commercial entities.

**City of Dallas Data**

The City of Dallas Open Data initiative was the third primary source of data (City of Dallas Data, 2019). Beyond land features, the study sought to understand if social parameters might indicate whether a lot is vacant or not. Prior studies had revealed a relationship between vacant lots and increased crime, but would the presence of increased crime or code violations point to the presence of vacant lots? These data were collected from the City of Dallas:

- **311**: Service requests that were created by citizens calling into the 311-contact center and reporting neighborhood concerns. This data represented suspected code violations.
- **Building Permits**: This data represents permits requested by developers or property owners if they wished to demolish, build, or modify a building. The study hypothesized that recent or numerous permit requests might indicate the presence of a building, which would classify a parcel as not vacant.
- **Crime Reports**: 2019 crime data was a mix of all types of crime reports. Some exploratory data analysis was performed to see if a violent crime was more significant than all crimes combined. The final models were developed using all crime reports.

### 3.2 Data Preparation

The annotated GIS data originated from CPAL, which based their research on DCAD tax assessments of parcels of land within Dallas County. CPAL utilized a rule-based approach to assess vacant lots within Dallas County. Incomplete data collection and subjective determination of “vacant” by field assessors are concerns that may affect the accuracy of the annotations (Maghelal et al., 2014).

The data required three GIS coordinate systems to be used: North American Datum 1983 (EPSG:2276 – NAD83), EPSG:6584 - NAD83(2011), and EPSG:4269 - NAD83. It is recommended to keep all geometries within the same coordinate system, however
that was not possible since different remote sensing systems collected different data, and each used a separate coordinate system. This affected the reference geometry and joining of the points to the base reference layer. To overcome this problem, the GIS software reprojected the GIS layers to a common coordinate system WGS84 (World Geodetic System 84), which became agnostic once the features were blended and serves no further significance in the analysis.

DCAD GIS shapefiles provided the foundational location references. (ESRI, 2021). ACCOUNT_NUM was the primary identifier of a taxable entity or account, and it was present on the DCAD GIS shapefiles. ACCOUNT_NUM tabulated the 311 calls, police reports, and permit requests, and new features were created to represent the number of 311, crime, or permit occurrences for each ACCOUNT_NUM.

ACCOUNT_NUM was the primary key for all locations in the GIS shapefiles. This value represents both the account or owner of the land and the location of the land. In some cases, one account owned many properties, so this was not a unique value. In other cases, a single location was associated with multiple accounts. There was no complimentary key in the dataset to resolve the relationship between account and location, so the study had to filter all observations where the account was not unique. This semantically translated to limiting the study to lots that were owned by a single account. It eliminated lots that were owned by multiple owners as well as Accounts that owned numerous lots.

To join these new features to physical locations, the GIS parcels were represented by polygon shapes. The 311 calls, crime, and permit counts per location were mapped onto the GIS feature space using a point with a radial buffer of 10 feet. It is recommended to allow these buffers to be applied dynamically (Miguel et al., 2010). However, this research used a buffer of 10 feet to be consistent, and the study may be reproduced. This buffer provided enough distance to allow the points to be joined to the relevant location shapes without overlapping multiple irrelevant shape polygons.

After the new features were mapped onto the feature space, the tabular merge of location (ACCOUNT_NUM) with associated counts (311, crime, permits) were exported into comma-separated value files. All data was evaluated for extreme, null, and negative values. Negative values were erroneous and were imputed with zeros. Null values were converted to categorical labels or set equal to zero if numeric. Features with outliers were log-transformed to improve the scale and distribution. The land size was converted to square feet since the original data was a mix of acres and square feet. All numeric features were scaled using min/max scaling and normalized between zero and one. These features provided input for the exploratory data analysis phase of the pipeline.

3.3 Exploratory Data Analysis

The distinction between residential and commercial property was an essential factor. Division code represents the type of taxable entity. This categorical factor includes residential, commercial, and business personal property values. Business personal
property items had chairs, printers, and boats that may belong to businesses, but they would never be vacant lots. They were retained within the dataset because there were so few (n=1), and the study philosophy was to let the model determine vacantly or not vacant. 87\% of all lots were residential properties, as seen in Figure 2.

The distribution of 311 call counts, permit counts, and crime counts were a key question in this study. The study hypothesized that the City of Dallas had varying concentrations of these elements.

Figure 2. Commercial and Residential Property Breakdown
Figure 3: Parcels Associated with Crime Reports
Figure 4: Parcels Associated with 311 Calls
When the study examined these features individually, it identified more meaningful patterns. The City of Dallas records all calls made through the 311-customer service help desk. This program is the primary method of reporting code violations, specifically bulky trash violations in this instance. Each recorded report describes the specific address and also the latitude and longitude coordinates. As shown in Figure 6, concentrations can be identified easily, suggesting that areas of Dallas are more unkempt than others.
Police Incident data reports all police reports made, citing location, description of the event, and zip code. The study was focused on identifying vacant, blighted properties, and the underlying criminal activity prescribes where best to apply these findings.

Narrowing the focus by identifying only incident reports where a weapon was reported as used, Dallas’ activity map shifts to show widening areas affected by weapon-related crimes. Figure 7 demonstrates how there is not a single area that would easily be targeted to reduce gun violence in Dallas based on activity reports alone.

![Figure 6. 311 Reports of Bulky Trash](image)
Improvement value is a feature that describes the value of an office building, home, or shed that may be present on a property. From Figure 8, there is evidence of approximately 90% of the parcels having an improvement value greater than $1000.

Figure 7. Crimes with Weapons

Figure 8. Parcels with Improvement Value Less than $1000
3.5 Model Design and Implementation

The overall approach of the analysis was to design a tree-based classification method using the Random Forest classifier as the baseline classification model. Then the baseline model was compared to hyper-tuned RF and gradient boosted tree classifiers with 10 fold cross-validation.

This was a binary classification analysis where the target class was not balanced with respect to all possible parcels within the City of Dallas. The study used Random Forest (Breiman, 2001) and gradient boosted trees or XGBoost (Zhang et al., 2020).

The analysis loosely followed the framework proposed by Grippa, Georganos, Zarougui, Bognounou, Diboulo, et al. (2018). The approach in this analysis differed mainly from how the geospatial data was processed initially.

The study was performed in R because it was most compatible with CPAL’s current business processes. The researchers wanted to produce models that may be useful to CPAL in the future. QGIS was chosen over the R package RGDAL because the size of the data and the meta-data that accompanies joining points to a polygon shape is computationally prohibitive and exceeds R’s memory constraints.

Appendix II includes a complete list of software that was used for this study.

3.6 Evaluation Criteria

The machine learning models in this analysis were evaluated using the F1 score. The choice of F1 was based on the desire to account for both false positives and false negatives. F1 is a blend of these two metrics, and it offered a more balanced approach than Accuracy or Precision. The results were further analyzed using the false omission rate (FOR). The impact of incorrectly predicting a non-vacant lot when the truth is a vacant lot (type II error) was potentially more damaging than the inclusion of predicting a vacant lot when the lot was not (type I error). The act of excluding a vacant lot may present a cost to the community if it is not considered for remediation. However, the study did not seek to prioritize false positives because the potential price of site-surveying lots that are not vacant might also place an unnecessary cost and time burden on the city or non-profit.

4 Results

4.1 Model Selection

The final model was chosen for similar performance and greater simplicity when compared to other models. Other models that used additional features included data for building permits, certificates of occupancy, 311 calls, crime records, and the SPTD code that indicated the property tax classifications. These more complex models were
found to have only a small increase in performance compared to the simplest model, so the simple model was chosen as the final, in favor of the ease of generating the dataset without extensive preprocessing. Although many models were created using various algorithms and feature combinations, only one—the full XGBoost model—will be highlighted for comparison against the final model, as it contained all the features and was the highest performing.

To evaluate the performance of the models, confusion matrices and performance metrics were generated. The chart of performance metrics in Table 1 compares the full model to the simple model. It reveals that both models demonstrate high performance in all categories, including both sensitivity and specificity. The F1 score for the full model was 0.9907, and the F1 score for the simple model was 0.9889, and the FOR was 0.00502 for the full model and 0.00598 for the simple model. Ideally, the FOR should be as small as possible.

<table>
<thead>
<tr>
<th></th>
<th>Simple Random Forest</th>
<th>Full XGB</th>
<th>Simple XGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>0.986</td>
<td>0.991</td>
<td>0.989</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.938</td>
<td>0.954</td>
<td>0.945</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.978</td>
<td>0.986</td>
<td>0.984</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.974</td>
<td>0.983</td>
<td>0.980</td>
</tr>
<tr>
<td>False Omission Rate</td>
<td>0.00680</td>
<td>0.00502</td>
<td>0.00598</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>XG-BOOST Confusion Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAC</td>
</tr>
<tr>
<td>NVAC</td>
</tr>
<tr>
<td>Prediction</td>
</tr>
<tr>
<td>VAC</td>
</tr>
<tr>
<td>NVAC</td>
</tr>
</tbody>
</table>

Figure 9. Simple XGB Confusion Matrix
4.2 Final Model Performance

The ROC curve in Figure 11 presents evidence of the overall model performance. With an AUC (area under the curve) of over 99%, the initial presumption is that the model is effective.
Diving deeper into the model performance, a look at the confusion matrix in Figure 9. Simple XGB Confusion Matrix Figure 9 for the model's predictions shows that the vast majority of forecasts are correctly classified. Considering the imbalance between target classes in the data—approximately 90% of the parcels were non-vacant—unsurprisingly, sensitivity is slightly lower than specificity. But at .95 sensitivity, the model is still quite effective at identifying positives.

### 4.3 Final Model Features

The final model was an XGBoost model with five features: log of improvement value, division code, log of land value, log of area square footage, and log of the total value. As shown in Figure 12, features found most impactful centered around the parcel's value, size, and categorization.
Figure 12. XGBoost Feature Importance Plot

Improvement value is the most important feature in the model, which is not surprising considering that a piece of land with a valuable improvement built on it is unlikely to be vacant. Figure 13 shows the distributions of log improvement values for vacant vs. non-vacant classes. The vacant parcels are clearly separated from the non-vacant parcels on log improvement value—the vacant parcels almost all have 0 or close to 0 improvement value, whereas the non-vacant parcels almost all have \( >e^7 \) (approximately $1000) improvement value.
The remaining features do not show such a clear separation between the vacant and non-vacant classes (see Figure 13, Figure 14, Figure 15, and Figure 16). Still, in combination with improvement value, it is evident that they are valuable to separating the non-vacant observations.

**Figure 13.** Kernel Density of Vacant and Non-Vacant Parcels on Log Improvement Value

**Figure 14.** Kernel Density of Vacant and Non-Vacant Parcels on Division Code
Figure 15. Kernel Density of Vacant and Non-Vacant Parcels on Log Land Value

Figure 16. Kernel Density of Vacant and Non-Vacant Parcels on Log Area Square Footage
Figure 17. Kernel Density of Vacant and Non-Vacant Parcels on Log Total Value

Figure 18 shows a scatterplot of log improvement value by division code, with points slightly jittered to allow for better visibility. Observations with 0 improvement values for division codes 1 and 2 primarily make up the vacant class. In comparison, most non-vacant parcels have a greater than 0 improvement value or a division code of 0. This makes sense because division code 0 is Business Personal Property, which are items belonging to a business such as chairs, and would never be plots of vacant land.

Figure 18: Improvement Value by Division Code
Similarly, log land value also helps separate those points with 0 improvement value between vacant and non-vacant. Figure 19 shows that points with 0 improvement value and a land value greater than $150 tend to be vacant, while others tend to be non-vacant.

![Figure 19. Improvement Value by Log Land Value of Vacant and Non-Vacant Parcels](image)

### 4.4 Statistical Significance of Features

The statistical significance of the features was examined through logistic regression and a Wald chi-squared test for model coefficients. Table 2 that four of the top five final features have statistical significance in separating the means of vacant and non-vacant populations, p-value < 2e^{-16}.

Features of interest relating to crime and community issues—count of crime and count of 311 calls—show that crime has a significantly different mean between vacant and non-vacant populations, but 311 call count does not. Despite not making it into the final model feature set, the crime feature’s Wald test results support the hypothesis that there is an increased level of crime in vacant lots than in well-maintained properties, p-value < 0.019. The p-value for 311 call count is far from significant (p=.85), which is unintuitive considering that the calls are for bulky trash. It is reasonable to think that the count of bulky trash reports would be a strong indicator of vacancy. A possible explanation for this will be discussed in the discussion section.

| Feature            | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------------|----------|------------|---------|---------|
| Division Code      | -0.71717 | 0.07274    | -9.859  | <2e^{-16} |
| num_division_cd    |          |            |         |         |
Table 3. Wald Chi Square Test for model Coefficients

| Feature                        | Pr(>|X^2|)|
|--------------------------------|----------|
| 311 Counts count of 311 scaled | 0.85     |
| Permit Counts count permits scaled | 0.019   |
| Crime Counts count of crime scaled | 0.00015 |
| 311 Counts Permit Counts Crime Counts | 6.7e-5 |

Practical significance is another consideration in addition to statistical significance. Considering a large number of observations in the dataset (downsampled to approximately 40,000), a small difference between the means of the vacant and non-vacant populations would be detected as significant. Figure 20 and Figure 21 show the kernel densities of the 311 and crime counts by vacant/non-vacant, and they show no visible separation between the two classes. Figure 22 shows the entire XGBoost model top twelve features, which have crime count near the bottom at 11th place and 311 counts not even making a list. The visual inspection and the low rank in feature importance from the XGBoost model suggest that there is little to no practical significance in the counts of crime and 311 calls for identifying vacant lots.
Figure 20. Kernel Density of Count of Crime Scaled

Figure 21. Kernel Density of Count of 311 Scaled Data
Discussion

This study combined GIS mapping with a machine learning modeling approach to provide practical solutions for the future classification of vacant lots. CPAL had pursued a manually intensive approach, and it was extremely time-consuming. By using R and GIS mapping techniques, the methodology was more repeatable and easily consumed by CPAL and other researchers. The magnitude of the data and system requirements were also a factor in choosing QGIS and shared compute resources.

One of the inherent challenges of vacant lots is that they may constantly be changing. The lot may become a dumping ground, or it may become developed or remediated. It is challenging to ensure that each of those changes is captured in databases designed for taxation purposes. The study would have benefited from image analysis; however,
it was essential to align the data to CPAL’s annotations that were developed in 2019, and no reliable image data existed for that period. In the future, it may be useful to establish an image database that can be updated and analyzed on an ongoing basis.

Another major challenge in this study was the lack of clarity in the DCAD data model. With the given information, the study could not uniquely identify all Accounts in the total population. ACCOUNT_NUM was the primary key for all locations. This value represents both the account or owner of the land and the location of the land. In some cases, one account owns many properties, so this is not a unique value. In other cases, a single site is associated with multiple Accounts. There was no complimentary key in the dataset to resolve the relationship between account and location, so the study filtered all observations where the account was not unique. This semantically translates to limiting the study to lots that are owned by a single Account. It also eliminated all accounts that owned multiple lots. The latter restriction is most unfortunate since the City of Dallas holds multiple lots, so those lots were not part of this study.

The choice of buffer length or radius from which points were joined to GIS locations is an important contributor to these results. In this study, several techniques were attempted: variable length buffers and buffers of different lengths. This is an important element that may yield different results in feature selection. An intentionally longer buffer length would allow researchers to quantify the number of 311 calls counts or crime counts in a region or neighborhood instead of on a per-lot basis. This is an opportunity for future studies.

The Wald Chi-Square test for model coefficients tested whether there is an effect on a lot’s vacancy status for 311 call counts. The p-value was insignificant, at .85. This low significance level may reflect an issue with the data categorization of the 311 calls. The 311 calls in the data set consisted entirely of bulky trash violations in 2019, with no other request type listed. However, it would be expected that there are various types of 311 calls, not all of which are bulky trash reports.

Additionally, it is reasonable to think that the count of bulky trash reports would be a strong indicator of vacancy. Still, not all 311 calls may be for issues that would be indicators of vacancy. If all types of calls are lumped together, it is sensible that they would dilute the signal and be only weakly significant in distinguishing vacant lots from non-vacant. Another consideration is when all three features were tested together, 311 counts, CO counts, and crime counts, the p-value was $6.7e^{-7}$, suggesting a possible interaction among all three factors. This potential feature should be further investigated in future research.

Future research may also be limited by the need for local subject matter expertise. In this example, CPAL and the City of Dallas guided to help the team evaluate the practical significance of certain features. Without these advisors, it would be difficult for another study to provide valuable conclusions. For example, one feature labeled neighborhood seemed to have statistical significance; however, when discussing it with CPAL, the team learned that the value did not relate to a neighborhood as a citizen might understand it. It was a code that had meaning for DCAD, but it did not help describe the location of a lot.

Finally, future researchers will be challenged by the lack of standardized data and validated annotations. Without a national standard that clearly describes lot size, lot
value, improvement value, and lot type (residential or commercial), this study cannot be replicated. Validated annotations are necessary to build new models or confirm this model, so that is another challenge that must be overcome in future research.

One possible replacement for CPAL’s annotations was the State of Texas’ State Property Tax Division Code or SPTD code included in the DCAD dataset. This code is a factor of more than twenty codes that may be consolidated to indicate vacant or not vacant. Compared with CPAL’s annotations, less than ten percent of the SPTD codes were different from CPAL’s labels. This may provide an alternative source of truth for other cities within the State of Texas. No known dataset classifies vacant lots on a national or global level.

Despite these challenges, the study produced a highly accurate model. It confirmed that simple features are highly effective in classifying vacant lots. The City of Dallas within Dallas County disproved the hypothesis that crime and 311 calls may indicate vacant lots.

The strength of this approach is the ease of use and repeatability for other researchers. Should the City of Dallas or DCAD desire to re-evaluate vacant lots based on 2020 or 2021 data, they may apply the XGBoost model with the final features, producing a list of vacant lots in minutes. The model can be embedded in an application that can be incorporated into their regular business processes. This will free up time to evaluate which lots are the highest priority for remediation.

Beyond this direct application, this study may also encourage city, county, and state agencies to explore machine learning as a tool to accelerate and improve the accuracy of other classification problems.

**Ethical Considerations**

This study was conducted with full consideration of ethical implications. A disproportionate number of vacant lots exist in socioeconomically disadvantaged neighborhoods. Researchers and government officials must pay careful attention to the desires of residents in communities where remediation programs may occur. The use of machine learning and artificial intelligence is a sensitive topic and may have good and bad consequences. In this study, the researchers took care to use anonymized data that was publicly available. All citizens may access this same data, so there is no lack of transparency.

A potential misuse for this research would be for policymakers to decide to use this classification model as part of a larger policy for either revenue generation via fines or undue increase of police patrols in a community that may be better served by plot remediation rather than enforcement. That said, remediation program leaders should take extra care to engage communities and capture their qualitative input before proceeding with remediation programs.
6 Conclusion

Simple features can yield beneficial models that identify vacant lots. This study confirmed the original hypothesis that a machine learning model may be developed to classify vacant lots. An unexpected conclusion was that more complex features like crime data, 311 calls, and permit data were not significant contributors to the final model. However, the good news was that the most valuable features, like land and building value, land size, and land type, were the simplest data that are likely to be published by taxing agencies.

Additional areas of research may extend to modeling lots with abandoned buildings. CPAL has been working to identify lots with abandoned buildings using manual techniques. This is a more challenging objective since the value of a building may or may not indicate its state of disrepair or abandonment.

Another extension of this research is to develop a method of prioritization for vacant lots. One approach may be to increase the GIS buffer, define a neighborhood or regional area, and then re-assess 311 and police report counts for larger spaces. This may help prioritize lot remediation by associating the lots with communities that experience more crime and disrepair. Remediation efforts may be more impactful if conducted on a regional basis vs. an individual lot basis.

It is difficult to determine if these results are generalizable to all cities because there is no city-wide source of annotations to build different models. As we discussed in the State of Texas, there exists an SPTD code. This code was shown to be comparable to CPAL’s annotations, so it may be helpful for training models for other cities in the State of Texas.

Simple models that are easy to understand have positive socioeconomic implications. Many city and county officials may not consider remediation programs due to their perceived cost and complexity. If machine learning can provide insights, this reduces the barrier of uncertainty and encourages more remediation investments that may ultimately improve the quality of life for more citizens.

A national land database would be beneficial for future research. Tax information is collected inconsistently across the country, and not all counties provide data to the public. The availability of land and improvement data would allow the Federal government to quantify the magnitude of vacant urban lots across the United States and their association with violent crime. Federal infrastructure programs focusing on digital equity, urban agriculture, and economic development may benefit from converting a vacant lot’s liability into a gem of a community.

Acknowledgments

We want to thank Professor Jacquelyn Cheun, Ph.D., for her guidance and encouragement throughout this study. We are extremely grateful to Owen Wilson-Chavez from Child Poverty Action Lab for sharing data and the background of CPAL’s earlier work to determine vacant lots. He has also provided constant feedback as we worked to develop a useful tool for future research and community programs.
References

Appendix I: Data Dictionary

<table>
<thead>
<tr>
<th>Logical Name</th>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>311 Counts</td>
<td>count_of_311_scaled</td>
<td>Number of calls to 311, City of Dallas Customer Service.</td>
</tr>
<tr>
<td>Account</td>
<td>ACCOUNT_NUM or Acct</td>
<td>Unique identifier for the lot as well as the entity responsible for paying assessed taxes.</td>
</tr>
<tr>
<td>CO Age</td>
<td>days_since_issue_scaled</td>
<td>Number of days between the issuance of a CO and January 1, 2021.</td>
</tr>
<tr>
<td>CO Approval to Issue</td>
<td>days_from_CO_appr_o_to_issue_scaled</td>
<td>Difference in time between the approval of a Certificate of Occupancy (CO) and the Issuance of a CO.</td>
</tr>
<tr>
<td>CO Area</td>
<td>CO_sqft_scaled</td>
<td>Building size in square feet.</td>
</tr>
<tr>
<td>CO Counts</td>
<td>count_COs_scaled</td>
<td>Number of Certificates of Occupancy related to a specific lot</td>
</tr>
<tr>
<td>CO Type</td>
<td>CO_type</td>
<td>Type of CO</td>
</tr>
<tr>
<td>Crime Counts</td>
<td>count_of_crime_scaled</td>
<td>Number of crime reports per lot.</td>
</tr>
<tr>
<td>Days Since Permit</td>
<td>days_since_permit_scaled</td>
<td>Number of days between the most recent permit request and January 1, 2021.</td>
</tr>
<tr>
<td>Division Code</td>
<td>num_division_cd</td>
<td>Taxable entity type: 0=Business Personal Property, 1=Commercial, 2=Residential</td>
</tr>
<tr>
<td>Improvement Value</td>
<td>log_impr_val</td>
<td>The value of a building or immovable structure on the lot.</td>
</tr>
<tr>
<td>Land Area</td>
<td>log_area_sqft</td>
<td>Lot area in square feet, logged.</td>
</tr>
<tr>
<td>Land Value</td>
<td>log_land_val</td>
<td>Value of the parcel, not including any improvements. The value was logged to normalize.</td>
</tr>
<tr>
<td>Neighborhood Code</td>
<td>num_nbhd_cd</td>
<td>Code used by Dallas County to assess taxes. Not a valid indicator of location.</td>
</tr>
<tr>
<td>Permit Counts</td>
<td>count_permits_scaled</td>
<td>Number of permit requests for each parcel.</td>
</tr>
<tr>
<td>Permit Type</td>
<td>permit_type</td>
<td>Type of permit requested from the City of Dallas.</td>
</tr>
<tr>
<td>SPTD Code</td>
<td>num_sptd</td>
<td>Numeric translation of the State of Texas Property Tax Division code</td>
</tr>
<tr>
<td>Total Value</td>
<td>log_tot_val</td>
<td>Total value of a parcel including improvements and any business personal property assigned to the given account.</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Vacant</td>
<td>vac_par</td>
<td>Classification of the lot. 1 = vacant, 0 = not vacant</td>
</tr>
<tr>
<td>Zoning</td>
<td>zoning_buckets</td>
<td>Zoning classification of the lot.</td>
</tr>
</tbody>
</table>

**Appendix II: Software Configuration**

<table>
<thead>
<tr>
<th>Software</th>
<th>Package</th>
<th>Version</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>qgis</td>
<td></td>
<td>3.18.3</td>
<td>Processing GIS data</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td>4.0.2</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>dplyr</td>
<td>1.0.6</td>
<td>General data processing</td>
</tr>
<tr>
<td>R</td>
<td>tidyverse</td>
<td>1.3.1</td>
<td>General data processing</td>
</tr>
<tr>
<td>R</td>
<td>superml</td>
<td>0.5.3</td>
<td>General data processing</td>
</tr>
<tr>
<td>R</td>
<td>Bbmisc</td>
<td>1.11</td>
<td>String Processing</td>
</tr>
<tr>
<td>R</td>
<td>caret</td>
<td>6.8-86</td>
<td>Machine Learning data Structures and pre-processing</td>
</tr>
<tr>
<td>R</td>
<td>parallel</td>
<td>4.0.2</td>
<td>Parallel processing for algorithms</td>
</tr>
<tr>
<td>R</td>
<td>e1071</td>
<td>1.7.4</td>
<td>Support Vector Machines algorithm</td>
</tr>
<tr>
<td>R</td>
<td>ggplot2</td>
<td>3.3.4</td>
<td>General visualization Plotting</td>
</tr>
<tr>
<td>R</td>
<td>ROCR</td>
<td>1.0.11</td>
<td>Receiver Operator Curve (ROC) plotting</td>
</tr>
<tr>
<td>R</td>
<td>mlr</td>
<td>2.19.0</td>
<td>Machine learning data structures processing</td>
</tr>
<tr>
<td>R</td>
<td>ggmmosiac</td>
<td>0.3.3</td>
<td>General visualization plotting</td>
</tr>
<tr>
<td>R</td>
<td>parallelmap</td>
<td>1.5.0</td>
<td>Parallel processing for algorithms</td>
</tr>
<tr>
<td>R</td>
<td>xgboost</td>
<td>1.3.0.1</td>
<td>Extreme Gradient (XG) boost classification algorithm</td>
</tr>
<tr>
<td>R</td>
<td>randomForest</td>
<td>4.6-14</td>
<td>Random Forest classification algorithm</td>
</tr>
<tr>
<td>R</td>
<td>randomForest</td>
<td>4.6-14</td>
<td>Random Forest classification algorithm</td>
</tr>
<tr>
<td>R</td>
<td>randomForest</td>
<td>4.6-14</td>
<td>Random Forest classification algorithm</td>
</tr>
<tr>
<td>R</td>
<td>randomForest</td>
<td>4.6-14</td>
<td>Random Forest classification algorithm</td>
</tr>
<tr>
<td>R</td>
<td>randomForest</td>
<td>4.6-14</td>
<td>Random Forest classification algorithm</td>
</tr>
<tr>
<td>R</td>
<td>aod</td>
<td>1.3.1</td>
<td>Wald Chi Square Test</td>
</tr>
</tbody>
</table>