Syracuse University

Vacant Property Prediction and Pattern Analysis in Syracuse

Group 2-8

Ari Fazlija

Eric Lee

Rajat Rohilla

RIshikesh Thakker

**Project overview**

The objective of this project is to predict vacancy of properties in Syracuse NY as well as identifying the relationships and characteristics that contribute to property vacancy. The data used was available through Syracuse Open Data, which is a verified source that contains public data collected by the city of Syracuse; such as Vacant Property Data, Crime Data, Syracuse Parcel Map and Code Violations. In many former manufacturing hubs in the US, vacant land has become a common issue, particularly in neighborhoods that once thrived. Over the past fifty years, plant closures and job losses have significantly decreased the population, leading to the abandonment of once-bustling residential and commercial properties. (Garvin et al. 413) Our project aims to analyze an ongoing issue of property vacancies which is apparent in various neighborhoods across the city of Syracuse. Property vacancy enables issues such as safety hazards, inhabitant decline and last but not least economic burden to both the local government and inhabitants. residents, especially in low-income neighborhoods, point to these vacancy as primary threats to their health and safety (Branas et al.2946)

In pursuit of combating these issues, the project will be using predictive analytics as an initiative to provide actionable insights regarding vacant housing in the city of Syracuse. Through these insights the local government can allocate resources effectively, as well as identify high risk areas in need of intervention. In order to map out the goals of this project there were two questions asked such as: Can we predict which properties are likely to become vacant, and what patterns or characteristics contribute to vacancy? In order to achieve these goals, the project utilizes random forest trees in order to predict a binary outcome such as vacant or non-vacant. Whereas in order to map out relationships and characteristics that contribute to vacancy the project employs clustering analysis through K-means. The reasons for this approach is that many characteristics predicting vacancy may not be as visible, hence utilizing clustering analysis to group characteristics of properties allows for better pattern recognition.

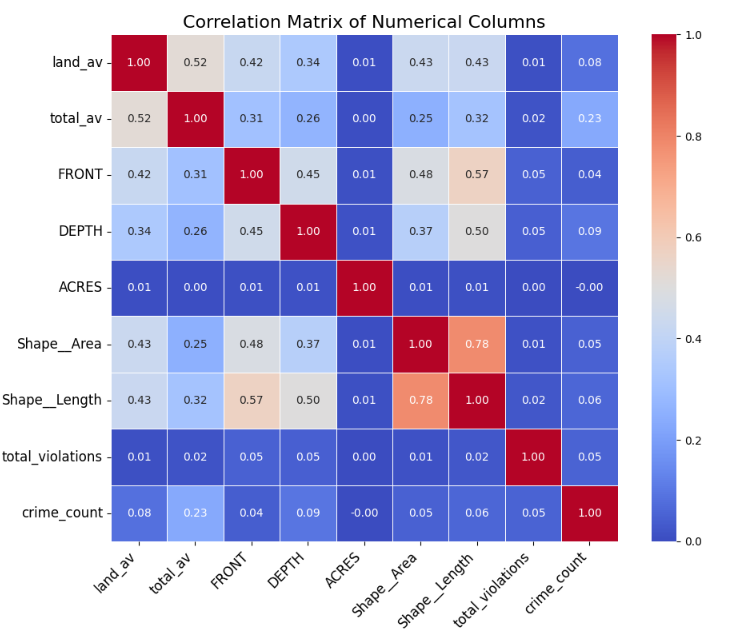
The prediction and inference phase of the project focuses on using supervised and unsupervised machine learning techniques to identify property vacancy as well as characteristics that contribute to vacancy. The first approach involves clustering such as K-means, to identify patterns and relationships associated with vacancy in the city of Syracuse. This will group properties together with similar characteristics, uncovering relationships contributing to vacancy that might not be as visible to the naked eye. The following approach involves a regression such as Random Forest Regression in order to predict weather a property is vacant or not vacant based upon a set of features such as code violations, address, crime rates, and zip code. By analyzing key features the model will be able to provide insight on whether the various properties in Syracuse NY will be vacant or not.

These results can be later used by local government intervene into areas that are associated as high-risk regarding vacancy. These predictive modeling techniques can offer valuable insight for the city of Syracuse. Local government can now properly asses vacancy issues, while targeting areas that are in high risk of becoming vacant in the near future.

**Data Exploration**

Our data exploration focused on understanding the key characteristics of the datasets and identifying patterns related to vacant properties in Syracuse. Below are the major insights obtained:

Correlation Matrix

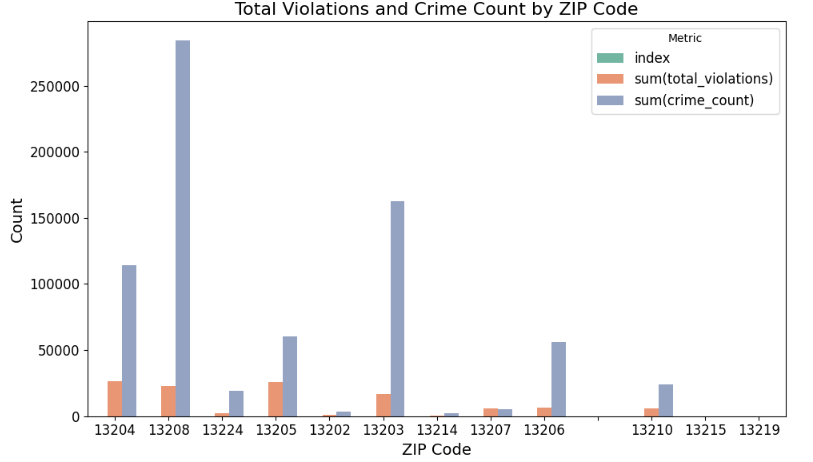


We visualized the relationships between numeric variables such as land area value, total area value, year built, number of residential units, and acreage using a correlation heatmap. This visualization helped uncover significant relationships:

* Land Area and Total Area Value showed a strong positive correlation, indicating that larger properties tend to have higher assessed values.
* Interestingly, Year Built had a weaker correlation with property vacancy than anticipated, suggesting that older properties are not necessarily more likely to be vacant.

This analysis provided a foundation for selecting predictive features for our model.

Crime Type Distribution



We examined the distribution of the most frequent crime types across the dataset. The analysis revealed:

* Vandalism and burglary were the most common crimes in neighborhoods with high vacancy rates.
* Crimes such as trespassing and illegal dumping were also notably frequent, possibly reflecting neglect and lack of property oversight in these areas.

This understanding of crime patterns provides insights into community-level issues that may influence property vacancies.

**Interesting/Surprising Results:**

**Surprising Correlation Patterns**

One of the surprising findings from the correlation matrix was the relatively weak relationship between the number of residential units and vacancy rates. This was counterintuitive, as we expected multi-unit properties to exhibit different vacancy dynamics compared to single-family homes.

**ZIP Code Clusters**

Certain ZIP codes displayed a notable clustering of high violations and crimes, reinforcing the connection between community safety and property neglect. However, some areas with high violations had lower crime rates, suggesting other factors, such as economic conditions or municipal policies, might be influencing vacancies.

Also, areas with low violations and high vacancies pointed to potential issues with market demand rather than safety.

**Unexpected Trends in Property Characteristics**

Box plots revealed that some older properties with high maintenance costs remained occupied despite economic challenges. This was surprising as newer properties were expected to have lower vacancy rates but showed mixed trends, potentially due to location and market demand.

**Influence of External Factors**

The analysis highlighted external factors such as proximity to schools, public transportation, and commercial areas influencing vacancy rates. Surprisingly, some neighborhoods with good infrastructure still had high vacancies, suggesting deeper socio-economic issues.

These results show that property vacancies are influenced by a mix of economic, social, and safety factors. They also provide important insights for creating better models and solutions.

**Methods used to Solve Vacancy problem**

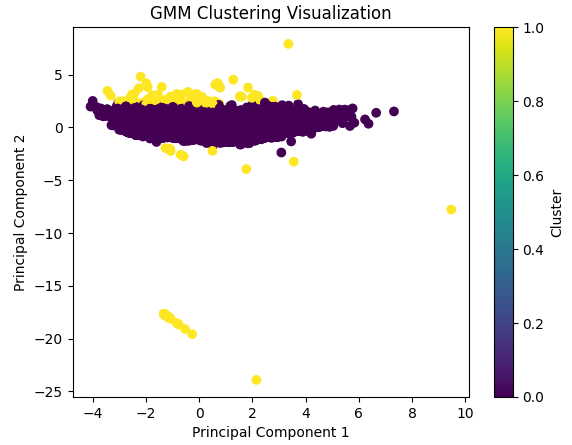
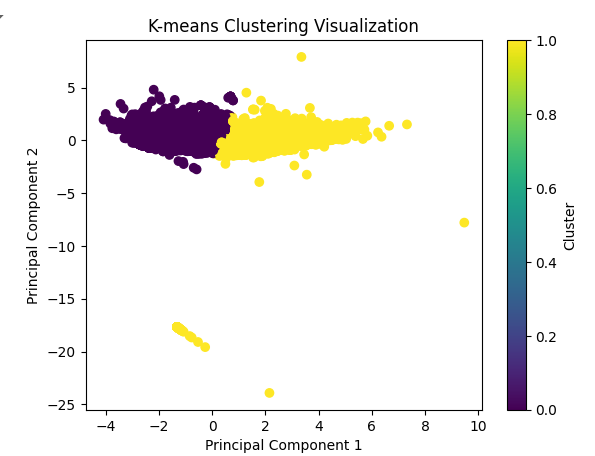
Preprocessing Step

After merging all four datasets into one big dataset and completing the exploratory data analysis (EDA), we enhanced the dataset with additional features such as a binary variable for the western region (west\_region), a overgrowth violations (overgrowth\_violation), and insights from crime and code violation data, where characteristics of vacancies were frequently observed. These engineered features provide valuable context for understanding factors influencing vacancy rates.

In PySpark, we are processing the dataset by handling both categorical and numerical features, converting them into a single feature vector. Foremost process categorical columns, replacing missing values or empty values with “Unknown” using the withcolumn method. StringIndexer is applied to each categorical column to convert string labels into numerical indices. OneHotEncoder is then applied to the indexed columns to create binary vectors representing each category. Using VectorAssembler, we are combing the numerical and one-hot encoded categorical features into a single vector which will be stored in a new column named features.

Implement the StandardScaler to normalize the features in the features column to ensure that all features have mean 0 and a standard deviation of 1 and added it to a new column called scaled features. This step is relatively of utter most importance since we are adopting machine learning which are sensitive to the scale of the features.

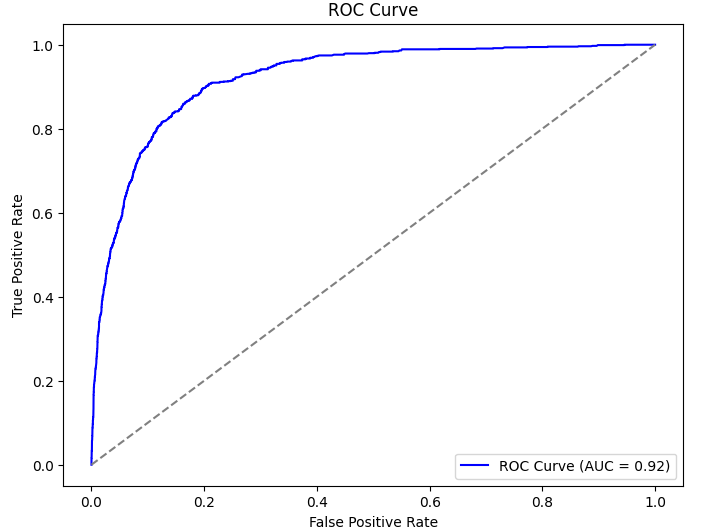
Clustering



We applied K-means clustering to identify distinct patterns in the property data, using features like building and neighborhood characteristics to group properties into clusters, potentially separating vacant and non-vacant properties. K-means was selected for its scalability and ability to combine categorical and numerical features, but challenges such as cluster interpretability, linear separability, and sensitivity to initialization were noted. The initial silhouette score was 0.114, reflecting poor cluster separation.

To enhance performance, we incorporated a new feature derived from the insights gained during the EDA process. Additionally, we combined K-means with PCA, which addressed the "curse of dimensionality," reduced noise, and enabled better visualization and improved cluster separation. (Jordi Soria-Comas and Josep Domingo-Ferrer, 350). This combination significantly improved the silhouette score from 0.114 to 0.602, demonstrating better clustering results. Compared to the Gaussian Mixture Model (GMM) used alongside, K-means with PCA delivered significantly better clustering performance, which is why we opted to use K-means for this project.

Regression



We also implemented a Random Forest Regressor predict properties that are not currently vacant but have a high likelihood of becoming vacant in the future. Random Forest was chosen for its reliability, ability to handle a mix of categorical and numerical features, and capability to capture complex relationships within the data (Zhang et al.714).

To ensure the model paid appropriate attention to the underrepresented vacant properties, we introduced class weighting by assigning higher weights to vacant properties based on their proportion in the dataset. Additionally, to prevent overfitting and mitigate issues such as class duplication during training, we employed cross-validation. This approach ensured the model's reliability and better detection of vacant properties.

The Random Forest delivered a strong performance, achieving a Test AUC of 0.92, Precision of 0.42, Recall of 0.79, and an F1-Score of 0.55. This high recall score demonstrates the model's effectiveness in identifying properties at risk of becoming vacant. By combining class weighting and dimensionality reduction, the model achieved a balanced focus on accuracy and recall, making it a valuable tool for identifying potential vacancies.

**Problems Encountered**

Throughout the project, we faced several challenges that required careful consideration and problem-solving. Below are the main issues encountered and how we addressed them:

1) Imbalanced Data

* Problem: Our dataset was highly imbalanced, with 39,753 non-vacant properties and only 1,344 vacant ones, making it a challenging machine learning problem. Despite the models achieving high accuracy, they were misleading, as they mostly predicted properties as non-vacant and failed to identify vacant ones accurately.
* Solution: We applied oversampling techniques to increase the representation of vacant properties, tripling their data size by sampling with replacement. Additionally, class weighting was applied to assign higher importance to vacant properties based on their dataset proportion in our regression model.
* Outcome: The recall score in regression model improved significantly from 0.21 to 0.79, highlighting the model's enhanced ability to correctly identify vacant properties. Recall is vital for predicting potential vacancies, enabling proactive decisions. The improvement from 0.21 to 0.79 shows how oversampling and class-weighting effectively addressed class imbalance, enhancing vacant property identification.

2) Data Accuracy and Completeness

* Problem: Both the crime dataset and the building dataset presented challenges due to inaccuracies in location data. The crime dataset used approximate addresses for privacy reasons, and the building dataset contained latitude and longitude values that did not accurately correspond to the actual locations.
* Solution: Using an alternative dataset, we created a heatmap of vacant properties in Syracuse. Additionally, merging the crime dataset with matching addresses led to inaccuracies, adding crime data to addresses where no crimes had occurred but were in the same area.
* Outcome: Successfully visualized the vacant property situation in Syracuse on a map and analyzed crime trends through EDA. However, due to potential inaccuracies in crime data leading to prediction and classification errors, crime data was excluded from the modeling process.

3) Model overfitting

* Problem: By oversampling the vacant property data, the risk of overfitting increased, making it challenging to preserve the model's generalization capabilities
* Solution: To address the increased risk of overfitting due to oversampling, cross-validation was implemented to ensure the model's robustness and maintain its generalization capabilities.
* Outcome: By implementing cross-validation, we mitigated data bias and enhanced the model's predictive reliability, ultimately improving the accuracy and stability of vacant property predictions.

Conclusion

This project used PySpark machine learning techniques to identify vacancy patterns and predict properties at risk of becoming vacant in Syracuse. To address the imbalanced dataset, we applied oversampling and class weighting, improving the model's detection of underrepresented vacant properties. Feature engineering, EDA, and PCA enhanced the performance of both clustering and regression models.

The K-means clustering analysis revealed distinct patterns in the dataset, grouping properties based on average characteristics such as building size, violations, and neighborhood conditions. For example, vacant properties tended to have higher overgrowth violation rates and were more commonly located in specific regions of Syracuse. The lower code violation rates and relatively higher land values revealed by our model demonstrated characteristics that differed from the expected traits of vacant properties.

The Random Forest model achieved a robust AUC of 0.92, demonstrating its effectiveness in distinguishing between vacant and non-vacant properties. Additionally, the Recall improvement from 0.21 to 0.79 highlights the success of our approach in identifying vacant properties, a critical metric for this project. The relatively high recall score demonstrated the model's potential to effectively identify vacant properties, providing a practical solution for accurate vacancy prediction and proactive urban planning.

While the project achieved significant results, challenges such as inaccurate geospatial data and imbalanced representation of vacant properties underscore areas for future improvement. Addressing the class imbalance more effectively in future iterations could further enhance model accuracy. Despite these limitations, the clustering insights and predictive models provide valuable guidance for city planners and local government to take proactive measures in addressing potential vacancies, ensuring efficient resource allocation and community well-being.

Reference

Branas, Charles C., et al. “Citywide Cluster Randomized Trial to Restore Blighted Vacant Land and Its Effects on Violence, Crime, and Fear.” *Proceedings of the National Academy of Sciences*, vol. 115, no. 12, Feb. 2018, pp. 2946–51, https://doi.org/10.1073/pnas.1718503115.

Garvin, Eugenia, et al. “More than Just an Eyesore: Local Insights and Solutions on Vacant Land and Urban Health.” *Journal of Urban Health*, vol. 90, no. 3, Nov. 2012, pp. 412–26, https://doi.org/10.1007/s11524-012-9782-7.

Jordi Soria-Comas, and Josep Domingo-Ferrer. “Mitigating the Curse of Dimensionality in Data Anonymization.” *Lecture Notes in Computer Science*, vol. 11676, Springer Science+Business Media, Jan. 2019, pp. 346–55, https://doi.org/10.1007/978-3-030-26773-5\_30. Accessed 14 Dec. 2024.

Zhang, Junfei, et al. “Modelling Uniaxial Compressive Strength of Lightweight Self-Compacting Concrete Using Random Forest Regression.” *Construction and Building Materials*, vol. 210, Jan. 2019, https://doi.org/10.1016/j.conbuildmat.2019.03.189. Accessed 4 Mar. 2023.