AUTO THEFT CRIME ANALYSIS

Paula McCree-Bailey, Omar Brandon, Ryan Kaplan, Rohitha Sanka

University of North Carolina at Charlotte

Prof. Nadia Najjar, Prof. Shannon Reid

12/09/2024

Abstract

Keywords: Auto theft, Social disorganization, Violent crimes, Household demographic

Introduction

Every 31 seconds a car is stolen in the United States, which equates to more than 1 million annually (Straughan, 2024). Data published by the FBI unveils that motor vehicle theft (MVT) is one of the most common crimes in the US but one of the least researched topics (Suresh and Tewksbury 2012). The loss of an automobile can pose significant financial and emotional distress on individuals which results in a disrupted lifestyle pattern. The social disorganization theory, introduced by Clifford Shaw and Henry McKay, aligns closely with our topic of auto theft. This theory emphasizes how environmental factors influence delinquency and criminal behavior (Mendez, Spencer, & Stith, 2019). To list some of the variables based on the theory in accordance to auto theft is population size, income, education levels, unemployment rate and housing vacancy. A relevant perspective in auto theft crime can be examined under the circumstance of diminished social control, lack of community cohesion and limited surveillance.

Exploring the relationship between auto theft and social disorganization we see the motivation behind car thieves is emphasized. It was found that many thefts operate on a risk-versus-reward basis. Essentially feeding the concept of an easy to steal vehicle with a higher market value becoming a target (EloGPS 2023). Other factors such as theft for profit, theft to secure transportation, also known as joyriding, were identified as frequent motives committed by adolescents or young adults (Suresh and Tewksbury 2012). Recorded by the National Insurance Crime Bureau (NICB) that Kia and Hyundai sedans were experiencing the highest theft rate in

2023. It was also statistically documented that 85% of vehicles that are reported stolen are recovered by law enforcement, with 34% recovered within the day of the report (NICB 2024).

Throughout this research paper and the technical models built, our study identifies how higher vacant housing rates correlate to auto theft crime rates. The goal of this study targets the main question, how does the presence of vacant housing contribute to an increase in auto theft rates in neighborhoods, and to what extent can social disorganization theory be used to predict this relationship? To accomplish this, the psychology of why and how auto theft occurs is examined, along with the influence of high housing vacancy rates, aiming to bring a solution by emphasizing security measures within a neighborhood. While addressing socio economic challenges and encouraging urban planning, community engagement, and improved law enforcement strategies, the problem can be effectively mitigated. Furthermore, the research and result of the combination of literature and machine learning models emphasizes techniques and prevention of auto theft.

Background

According to Shaw and McKay's work, neighborhoods characterized by weak social cohesion, poor community organization, and high levels of social disorganization are more prone to elevated crime rates, including auto theft. Their pioneering research linked juvenile delinquency and broader criminal behavior to the breakdown of neighborhood social structures. Specifically, the absence of stabilizing institutions, such as strong religious organizations, high-quality schools, and enriching after-school activities, diminishes informal social controls and allows criminal behavior to thrive. This foundational framework helps systematically examine the social and environmental factors that lead to auto theft.

Subsequent studies have validated the relevance of social disorganization theory in explaining modern crime patterns. (Sampson and Groves 1989) expanded this framework, demonstrating that communities marked by low socioeconomic status and family disruption are more likely to experience property crimes, including auto theft. Similarly (Lee et al., 2016) reinforced these findings, illustrating how racial heterogeneity, residential instability, and weakened social ties exacerbate the risk of crime by dismantling the social cohesion necessary to maintain order. Their research emphasized that weakened social ties and disrupted family structures directly undermine collective efficacy, leaving neighborhoods vulnerable to criminal activity. Despite these advances, gaps remain in fully understanding the interplay between socio-economic dynamics and emerging challenges like urban gentrification, technological changes, and variations in housing policies. These areas warrant deeper exploration.

The integration of social disorganization theory into this study reinforces its strength as a guiding framework. It offers an evidence-based foundation for interpreting the relationships between socio-economic factors such as poverty, unemployment, housing vacancy, and auto theft rates. By addressing these structural issues, we aim to identify actionable insights that can inform community-level interventions and policies. Examples of these strategies include neighborhood watch programs that aim to foster community cohesion and collective efficacy, which has been shown to reduce crime. Research by (Bennett, Holloway & Farring, 2006) suggests that active neighborhood watch programs can lead to a reduction in property crimes, including auto theft, as they increase surveillance and informal social control in the community. However, the effectiveness of such programs can vary, particularly in socially fragmented areas, highlighting the importance of tailored approaches.

Technological interventions such as anti-theft devices and vehicle tracking systems also play a critical role. Devices like immobilizers, vehicle alarms, and GPS tracking systems have been demonstrated to lower theft rates, as they make vehicles more difficult to steal and increase the likelihood of recovery (Prasetyo & Hidayat, 2021). The National Insurance Crime Bureau (NICB) found that cars equipped with GPS systems have a recovery rate of over 90% when stolen, compared to a general recovery rate of about 50% for all stolen vehicles. Additionally, the introduction of electronic immobilizers in vehicles has contributed to a significant decrease in auto theft rates over the years (NHTSA, 2012). Future research could explore how newer technologies, such as advanced AI-enabled surveillance and connected vehicle systems, further influence these trends. The NICB recommends the use of layered security measures, combining common-sense practices (such as locking doors) with advanced technologies such as kill switches and vehicle recovery systems. A report by the NICB highlights that cars equipped with these technologies are significantly less likely to be stolen.

Furthermore, Crime Prevention Through Environmental Design (CPTED) strategies have gained traction in urban areas across the U.S., emphasizing creating environments that reduce opportunities for crime, such as installing better street lighting, increasing natural surveillance, and reducing isolated areas (Welsh & Farrington, 2008). Studies conducted in U.S. cities like Philadelphia have demonstrated that environmental modifications, particularly the introduction of street lighting and security cameras, can reduce property crimes, including auto theft, by as much as 30% (Welsh & Farrington, 2008). While these solutions have proven effective, neighborhoods experiencing severe social disorganization face unique challenges in implementing them due to limited community engagement and resource allocation. Addressing the root causes of crime, such as poverty, unemployment, and a lack of social infrastructure, is

critical. A multifaceted approach that tackles both the causes and the symptoms of crime is necessary to effectively mitigate auto theft.

Dataset Description

The Communities and Crime dataset contains 2,215 observations and 147 features which combined socio-economic data from the 1990 U.S. census, law enforcement data from the 1990 U.S. Law Enforcement Management and Administrative Statistics (LEMAS) survey and the 1995 FBI Uniform Crime Report (UCR) crime data. The range of features include percentage of urban, median income, race per capita, percentage population under poverty, marital status, information on police officers in the community and various types of criminal offenses, provide a lot of information for algorithm predictions.

This dataset offers a baseline for understanding the relationships between variables, despite being more than thirty years old. Our target variable is auto theft, while the independent variables were selected based on the results from our exploratory analysis. To assist in feature selection, we utilized heatmaps to identify relationships that could influence model performance. The goal was to find relationships that could affect the performance of future models and to preselect features to be included in the initial models. The original data set contained missing values. After the preprocessing, the dataset was reduced to 2,212 observations and 20 features.

This section presents descriptive statistics for key variables related to auto theft and socio-economic factors in urban areas. Understanding these statistics helps in analyzing patterns and correlations that may influence crime rates, specifically auto theft rates. These results are summarized in Table 1.

Table 1

Descriptive Statistics

Features	Mean	Median	Std Deviation	Minimum	Maximum
autoTheft	516.69	75.00	3258.16	1.00	112464.00
ctPopUnderPov	11.61	9.33	8.60	0.64	58.00
HousVacant	1748.09	557.50	6508.12	36.00	172768.00
medIncome	33991.57	31444.50	13428.21	8866.00	123625.00
PctBSorMore	23.07	19.66	12.69	1.63	79.18
PctEmploy	62.03	62.45	8.31	24.82	84.67
PctNotHSGrad	22.28	21.38	10.97	1.46	73.66
PctUnemployed	6.04	5.44	2.89	1.32	31.23
pctUrban	70.43	100.00	44.10	0.00	100.00
PctUsePubTrans	3.04	1.22	4.92	0.00	54.33
PctVacMore6Mos	34.77	34.10	13.92	3.12	82.13
RentMedian	428.53	397.00	170.79	120.00	1001.00

The Communities and Crime dataset reveals various socio-economic and crime-related patterns. Auto thefts show a positively skewed distribution, indicating that some regions have extremely high rates. The percentage of the population under poverty varies significantly, with substantial disparities in housing vacancies and median income across different areas. Educational attainment levels range widely, while employment conditions appear relatively healthy overall. Public transportation usage is generally low, although some regions rely on it more heavily.

Social disorganization theory suggests that crime rates are influenced by the breakdown of social institutions and community structures, often due to factors like poverty, residential instability, and ethnic diversity. By examining these twenty societal factors from the dataset such as median income, house vacancy, and percentage urban, it presents the opportunity to look at how these elements contribute to the social fabric and potentially influence crime rates, including

automobile theft. This approach can help identify key predictors and possibly inform policy decisions aimed at reducing crime.

To gain further insight, a heatmap was created to illustrate the strength and direction to the relationships between the features and the target variable. Overall, most of the correlations are nearly zero, which suggests that there is no meaningful relationship between these two variables. For clarity, the number of features shown in the results (Figure 2) was reduced to thirteen, revealing several noteworthy correlations. The heatmap shows the strongest positive correlation of 0.90 between auto theft and house vacancy, suggesting that as house vacancies increase, the number of auto thefts also rises. Additionally, there is a moderate positive correlation of 0.34 between public transportation and auto theft. Other relationships include Auto Theft and Percentage Boarded Vacant houses with 0.18 correlation , Auto Theft and Percentage Unemployed with 0.12 correlation, and Auto theft and Percent Poverty with 0.11 correlation. All these correlations are positive, suggesting that as one feature increases, the other is expected to increase as well.

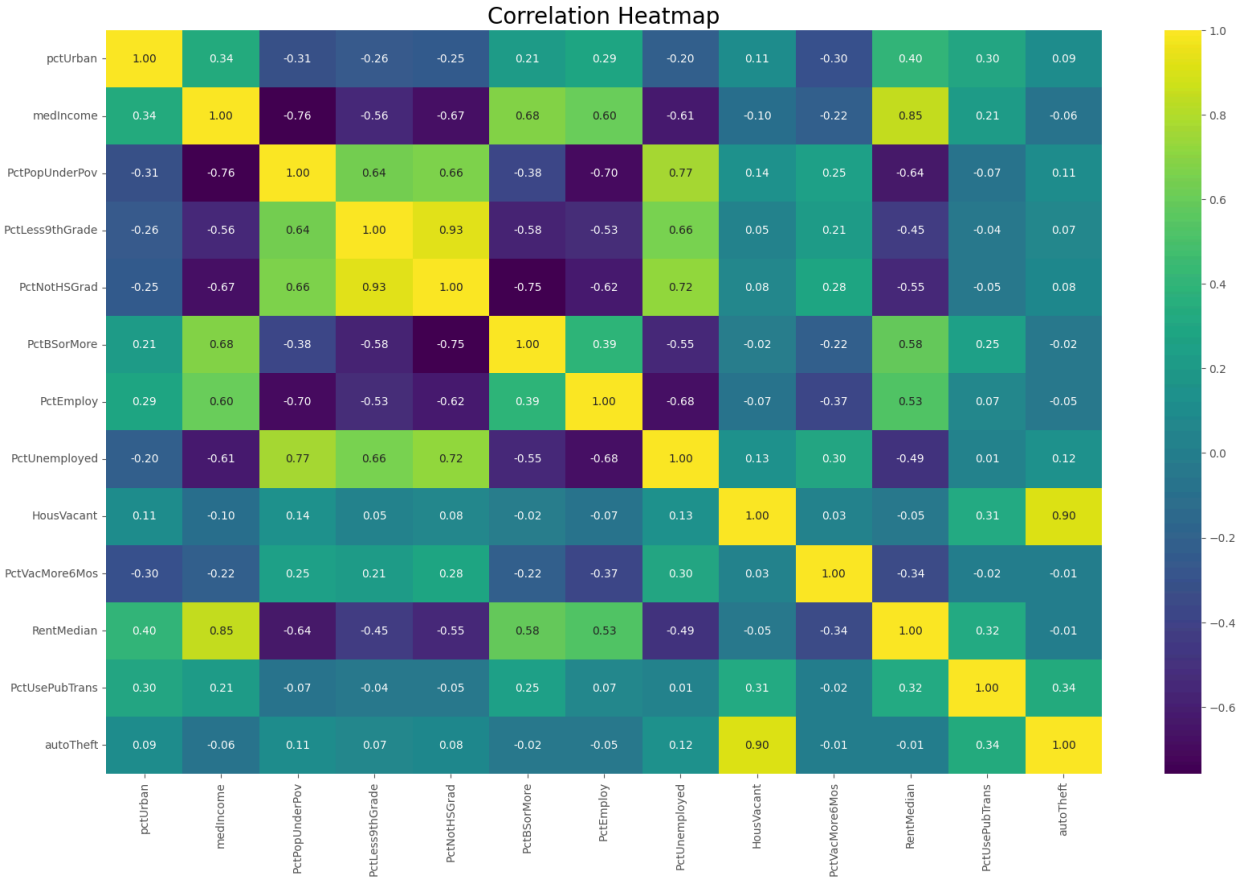

Figure 1

Methodology

The dataset required significant pre-processing to prepare it for analysis, particularly in terms of handling missing values. Some of the features contained missing values, which were represented by the symbol "?". Rather than attempting to impute missing data, which could introduce bias or reduce the integrity of our analysis, we opted to remove all rows containing missing values. After this entire process, our dataset was left with 2,212 observations and 20 features that we would be looking at. This decision ensured that the remaining dataset was complete and that our analysis could be conducted without concerns about incomplete data.

Although this alteration to the dataset reduced the number of usable observations, it simplified the analysis and ensured consistency across all variables. The removal of incomplete rows reduced noise and prevented issues that could arise from inaccurate imputation, such as skewed relationships between variables. We then examined the dataset for outliers and inconsistencies. A few outliers were identified, but none were significant enough to warrant removal or transformation. Additionally, all fields were appropriately structured for analysis, and no further alterations to the dataset were necessary in terms of reformatting or variable mapping.

We dropped features that were not relevant to the research focus, simplifying the dataset to only include socio-economic factors that could potentially influence auto theft rates. In terms of data preparation, we applied binning to some of the continuous variables, such as median household income and education level. This was done by defining specific ranges for each category (e.g., low, medium, and high) and using conditional statements in the code to assign a categorical label to each observation based on its numeric value. The target variable, auto theft, was binned to better analyze patterns in its distribution and simplify interpretation. We divided the continuous variable, representing the number of auto theft incidents, into three categories: low, medium, and high. The specific binning thresholds were determined based on the distribution of the following data: low: 0–50 incidents (inclusive), medium: 51–150 incidents, high: 151 or more incidents. These ranges were chosen by analyzing the data distribution using histogram visualizations and ensuring approximately equal representation across bins. Binning was implemented using conditional statements in Python, assigning categorical labels to each observation. This process allows us to explore relationships between socio-economic variables and crime rates at different severity levels more effectively. This preprocessing step allowed us to examine relationships between broader socio-economic categories and crime rates more

clearly, simplifying the interpretation of data trends.

To train and evaluate the predictive models, the dataset was divided into a training set (70%) and a testing set (30%). This ratio was chosen to provide sufficient data for training the models while keeping enough data for evaluation of their generalizability. Additionally, k-fold cross-validation was utilized to ensure the reliability of the models. Specifically, a 10-fold cross-validation approach was applied, as it offered a good balance between bias and variance by training on 90% of the data and testing on 10% in each iteration. This method reduces overfitting and ensures that all data points are used for both training and validation.

Model performance was evaluated using several metrics. The R^2 metric was employed to measure the proportion of variance in the target variable explained by the model. Mean Squared Error (MSE) was utilized to assess the average squared difference between observed and predicted values, while Root Mean Squared Error (RMSE) provided an interpretable measure of error magnitude in the same units as the target variable. These metrics were chosen because they all offer a comprehensive understanding of model performance in predicting auto theft rates.

For the random forest model, parameters such as the maximum depth and minimum samples required to split a node were fine tuned to improve accuracy and control overfitting. The gradient boosting model involved tuning parameters like the learning rate, which controls the contribution of each tree, and the number of estimators, which determines the number of boosting stages. Grid search was utilized to optimize these hyperparameters systematically by evaluating a range of predefined values for each parameter. This process ensured that the models were adjusted to achieve the highest predictive accuracy and generalization capability.

Our target variable was auto theft, represented by the variable autoTheft, which represented the number of auto thefts that occur in the city. For the target variable, we built a

predictive model to evaluate how well community social dynamics could predict crime rates. Social Disorganization Theory was used to assist in binning features from the dataset. To emphasize the impact of neighborhood characteristics on auto theft, socio-economic variables hypothesized to be key predictors were unemployment rates, poverty levels, and level of education. To explore the relationships between these variables and auto theft, we used correlation measures to identify which factors exhibited the strongest connections with our target variables. This allowed us to assess how effectively the social structures of a community predicted auto theft incidents.

We employed linear regression models to predict auto theft rates based on our pre-selected socio-economic indicators. Given the nature of our target variable and the continuous data involved, regression methods were well-suited to capture the relationships between community dynamics and crime. For the model creation process, we utilized a random forest model to try and predict auto theft rates, focusing on variables such as the total number of auto thefts and auto thefts per population. This model was designed to understand how socio-economic indicators, like unemployment, education, and poverty, are linked to auto theft rates. Once we trained the model, we evaluated their performance compared to a baseline model that only utilized the average auto theft rate as a predictor. The random forest model demonstrated stronger predictive capabilities, with higher accuracy values than other models, suggesting that socio-economic factors are indeed influential in forecasting auto theft. This study supports the social disorganization theory by demonstrating the strong correlation between crime rates and community factors such as unemployment and education levels. The models exhibited improved performance compared to the baseline.

Analysis

To enhance the model evaluation framework by integrating cross-validation, multiple supervised machine learning models, and cluster analysis. The objective is to assess the predictive performance of various models and identify the most suitable one for the features in the dataset. Incorporating cross-validation, specifically k-fold cross-validation, is crucial for evaluating model performance. It helps in assessing the model's ability to generalize to an independent dataset and prevents overfitting by ensuring that each data point has an opportunity to be in the training and validation set. This approach provides a comprehensive evaluation of model performance, leading to more robust and reliable predictive analytics. The models under consideration include Multiple Linear Regression for continuous target variables, Decision Trees, Random Forest, Naive Bayes, and Gradient Boosted Predictive Models. All the models discussed are a supervised learning algorithm that can be used for classification and regression. Each model comes with unique mechanisms and underlying assumptions:

Multiple Linear Regression relies on the assumption of a linear relationship between the target and predictor variables. It finds the best relationship using a straight line between an independent and dependent variable. This best fit line predicts the value of the dependent variable based on the independent variables and minimizes the differences between their predicted and actual values.

Decision Trees operate by recursively searching and splitting the data based on feature values or conditions to form a tree-like structure, aiming to minimize impurity. Gini impurity determines the probability of misclassifying a random observation in a dataset. Lower impurity results in better splits in the decision tree. Decision trees are one of a few models that do not

require normalization or scaling and handle missing values well. However, a single decision tree is prone to overfitting the test data and sensitive to changes in the data.

Random Forest combines multiple decision trees to improve predictive accuracy and controls overfitting by creating non-correlated individual models. These trees vote on the best classification or regression model which allows it to reduce overfitting and handle imbalance datasets. However, generating these trees can demand significant memory and computational resources, making it slower for large datasets.

Naive Bayes applies Bayes' theorem with the assumption of feature independence. It uses the probability of each feature in relation to the target class to make predictions. By calculating the conditional probability of each feature given each class, the model classifies based on the highest probability. The assumption of independence among predictors simplifies computations and is useful for large datasets. It is important to understand that independence of features does not always hold.

Gradient Boosting combines many weak single decision trees to create a stronger tree. The trees are connected, and each tree tries to minimize the error of the previous tree. So, the final model includes the results from the prior step and a better result is achieved. The best results come at the cost of taking more time to train the model and the model can overfit.

K-fold cross-validation was completed for k = 5 and k = 10. Cross-validation is important because it helps improve the reliability of the model’s performance. Recall, cross-validation works by dividing the data into k equal-sized folds. The model is trained on k-1 folds and tested on the remaining folds. The process is repeated k times where each time a different fold is used for testing. By averaging across all k iterations, it creates performance metrics which are more reliable. Besides improving reliability, cross-validation prevents overfitting and maximizes data

utilization. For all five models, the best performance was observed when k = 10. With 10-fold cross-validation, each fold is using 90% of the data for training and 10% for testing, providing a better training process compared to 80% training set used in 5-fold cross-validation. Additionally, a larger number of folds reduces the variance in the model. The findings for this process are summarized in Table 2.

Table 2

Baseline Model: k-Fold Cross Validation using k =10 and random state = 21

	Mean R-squared	Mean MSE	Mean RMSE
Multiple Linear Regression	0.9131	389608.9792	583.2387
Decision Tree	0.4819	6040805.7781	1797.3123
Random Forest	0.8282	3111806.6436	1217.0095
Naive Bayes	0.0796	10404117.0793	2550.8160
Gradient Boosting	0.7784	2959268.1188	1227.3824

The best model is evaluated based on the mean R^2, mean MSE and mean RMSE. The mean R^2 is the average of the R^2 values across all folds, providing an overall measure of how the model explains the variability in the data. The model with the highest R^2 indicates a better fit. The mean MSE is the average MSE across all folds and measures of prediction error. A lower mean MSE indicates better performance. The mean RMSE is the square root of the MSE and provides the average magnitude of the error.

Overall, the best model performance is achieved with multiple linear regression when k-fold = 10 with a mean R^2 of 91.32% across all folds. This means over 91% of the variance in autoTheft, the target variable, is explained by the selected features in the model. Besides the highest mean $R^{2,}$ random forest also has the lowest mean MSE and RMSE. However, despite

these metrics, linear regression may not be the best model. After plotting residuals, examining their distributions, and analyzing the Q-Q plot, it appears that linear regression does not adequately fit the dataset. These diagnostic plots revealed patterns and deviations that suggest the assumptions underlying linear regression are not fully met. Therefore, alternative models, such as random forest, which better handle the complexity and potential non-linearity in the data, should be considered.

Random forest performed second best with a mean R^2 of 82.82%. It performed better by building and combining the results from multiple decision trees which reduces the chances of overfitting. By averaging the results, random forest performs better on new, unseen data.

The models were further fine-tuned by applying grid search to Lasso Regression (L1 penalty), Ridge Regression (L2 penalty), Decision tree, random forest, gradient boosting, and Naive bayes. Hyperparameter tuning is an important step in optimizing machine learning models (ML). By using predefined hyperparameter values, grid search makes sure that the model is tuned to achieve the highest accuracy. The findings for this process are summarized in Table 3.

Lasso regression (L1 penalty) also adds a penalty which is equal to the absolute value of the coefficients to the cost function. Unlike ridge regression, lasso regression can shrink coefficients to zero which effectively removes those features from the model. It also helps with reducing overfitting.

Ridge regression (L2 penalty) is a linear regression model that adds a penalty to the coefficients of the cost function. This penalty term helps to shrink the coefficients towards zero which helps to reduce overfitting. In ridge regression, the coefficient never becomes exactly zero so all features remain in the model.

After hyperparameter tuning, the best models were Lasso and Ridge regression with R^2 of nearly 79%. Despite their ability to reduce and remove less relevant features, these models did not outperform multiple linear regression.

Table 3

Hyperparameter Model: Grid Search using cv = 10 and random state =21

	Mean R-squared	Mean MSE	Mean RMSE	Best Parameter
Lasso Regression	0.7947	100869.6811	317.5999	alpha = 1
Ridge Regression	0.7947	100873.5410	317.6060	alpha = 100
Decision Tree	0.7885	103918.8886	322.3645	max_depth: 8, min_samples_leaf: 1, min_samples_split: 2
Random Forest	0.6914	151606.4145	389.3667	max_depth=3, max_features='sqrt', n_estimators=15,
Naive Bayes	0.7824	106932.8281	327.0059	var_smoothing = 0.0026826958
Gradient Boosting	0.7885	103918.8886	322.3645	n_estimators=200

Note. with Cross validation

Clustering involves grouping items with common characteristics into a group. K-means is an unsupervised algorithm used for clustering. It works by dividing a group of observations into a predetermined number of clusters. This number of clusters (k) is determined by the user, who can either randomly select a number or use a method such as the "elbow method" to calculate the optimal number of clusters. Once k is decided, k data points are randomly selected to be the centroids (centers of the clusters). Each observation is then assigned to the nearest centroid based on distance. After all observations are assigned, the centroids are recalculated by averaging the points assigned to each cluster. This process of determining distances and reassigning observations is repeated until the centroids no longer move, no data points change clusters, or the maximum number of iterations is reached. This iterative process ensures that the clusters are as compact and well-separated as possible. Overall k-means is relatively simple to implement, and scales well to large data sets. The drawbacks include k must be chosen manually and it has

difficulty scaling data with a large number of features. It is important to use PCA to reduce the number of features before running the algorithm. After using k-means clustering, the "elbow method" and Silhouette analysis was applied to further confirm the best cluster for the data set. According to the elbow method, the optimal cluster is 5, as shown Figure 2.

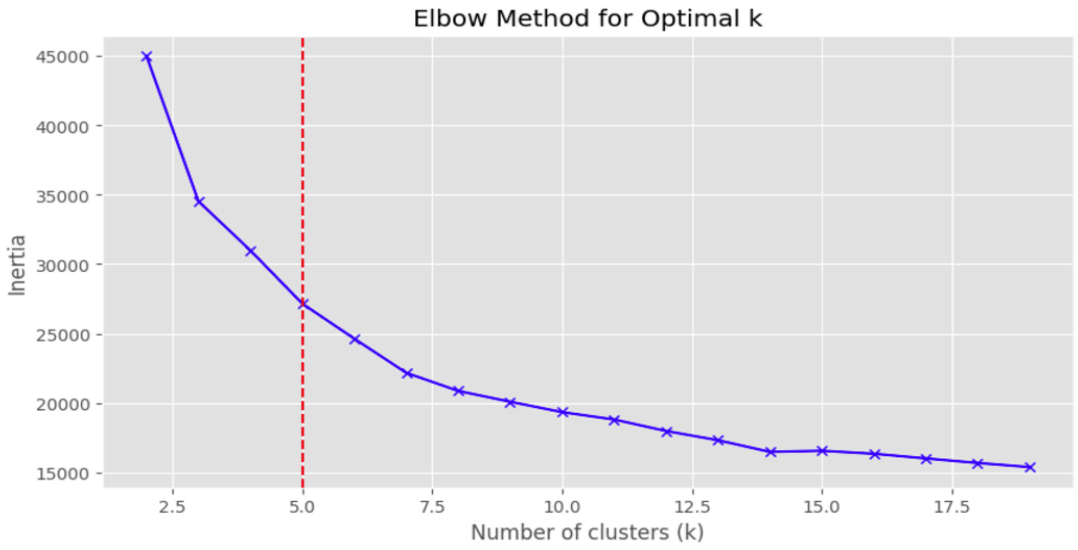

Figure 2

To confirm the selection from the elbow method, silhouette analysis was applied to the model. Silhouette measures similarities and differences within the clusters compared to other clusters where a higher value indicates a better-defined cluster; however, it is important to observe the quality or shape of the clusters, as a higher number does not always indicate the best model. The results from the analysis are displayed in table 4. The highest score was achieved with n = 2 clusters, yielding a score of 0.5638, as shown in Figure 3. Despite this high score, the silhouette plot for two clusters reveals that the clusters are not uniform. One cluster is considerably larger than the other indicating that the data within that cluster could be further divided into small clusters.

Table 4

Silhouette Analysis for k-Means Clustering

Cluster	Score
n = 2	0.5638
n = 3	0.5171
n = 4	0.5331
n = 5	0.5353
n = 6	0.5062
n = 7	0.4484
n = 8	0.5186
n = 9	0.4639
n = 10	0.4743

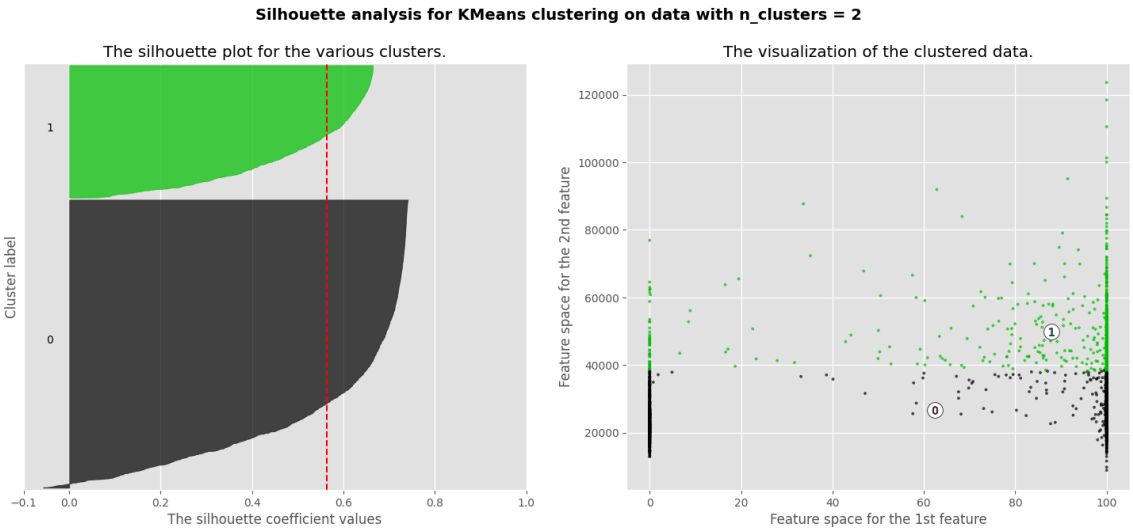

Figure 3

Ideally, all clusters should have similar average silhouette scores, indicating uniform separation. In addition, values should be positive with a narrow spread implying tight clusters. In the visualization for n = 5, which was selected as optimal using the elbow method, the clusters are

not uniform, despite the silhouette score being 0.5353 as shown in Figure 5. This could suggest several possibilities: the data does not naturally form well-defined clusters, the chosen clustering algorithm and parameters are not capturing the true structure of the data, or data has overlapping or complex patterns that are not well-suited to simple clustering techniques. Cluster analysis is not the best tool for analysing the data.

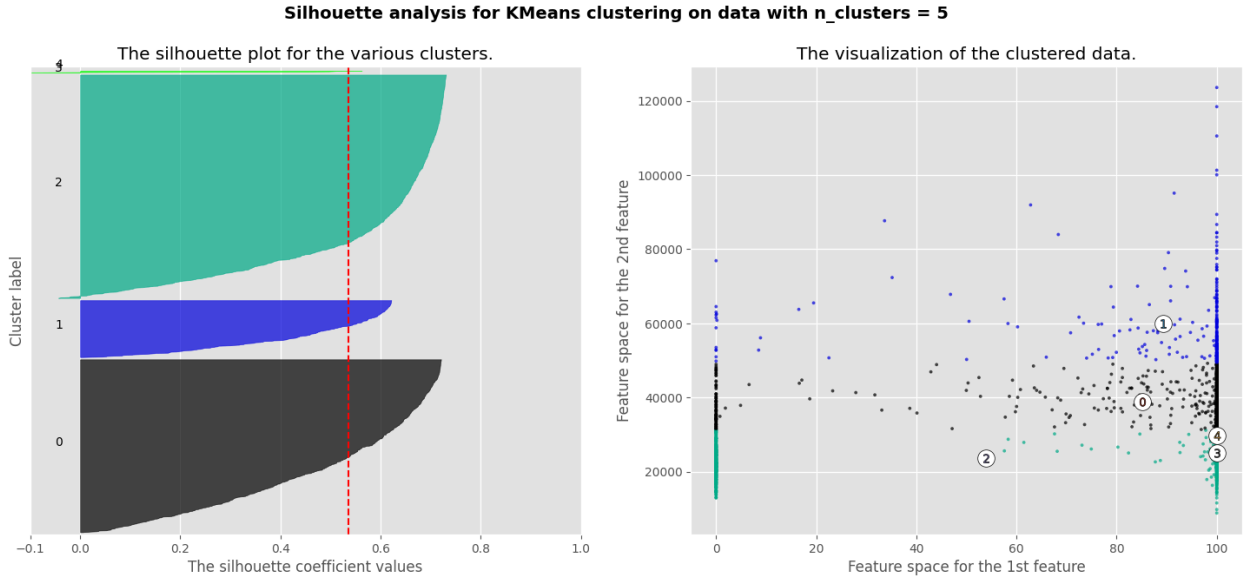

Figure 4

Discussion

This research aimed to address the factors influencing auto theft through the lens of social disorganization theory by examining socio-economic predictors and applying machine learning models. Our random forest model exhibited the strongest predictive performance across its metrics, the model highlighted key variables—housing vacancy, unemployment rates, and education levels—as significant contributors to the incidence of auto theft. Our methodological adjustments, including feature selection and binning, aimed to simplify data representation and

improve interpretability but may have contributed to the limitations of our models. To mitigate auto theft effectively, we propose two actionable solutions. The first being targeted community revitalization programs, based on the findings that housing vacancy strongly correlates with increased auto theft rates, community revitalization programs focusing on reducing vacant properties could help address this issue. Evidence suggests that policies incentivizing property rehabilitation, affordable housing development, and community engagement can strengthen neighborhood stability and reduce crime. For instance, neighborhood watch programs combined with environmental design strategies, such as improved street lighting and surveillance systems, have shown a 30% reduction in property crimes, including auto theft (Welsh & Farrington, 2008). Our second proposed solution is technological interventions and awareness campaigns, enhancing vehicle security through the adoption of anti-theft technologies, such as immobilizers and GPS tracking systems, is another effective solution. Studies report that vehicles equipped with such technologies exhibit a recovery rate of over 90% when stolen compared to a general recovery rate of about 50%. Additionally, public awareness campaigns focused on educating residents about locking vehicles, removing valuables from sight, and reporting suspicious activity can further deter opportunistic theft (Prasetyo & Hidayat, 2021).

Despite these proposed solutions, the limitations of this research should be acknowledged. The dataset used, while comprehensive, is dated and lacks representation of modern vehicle technologies and socio-economic dynamics. This may have affected the models' ability to capture contemporary patterns in auto theft. Furthermore, excluding incomplete rows may have reduced the generalizability of our findings.

Future research should incorporate real-time data, such as recent vehicle theft trends and socio-economic updates, and explore advanced machine learning techniques to improve model

accuracy. Moreover, integrating insights from community stakeholders could ensure interventions are equitable and contextually relevant, addressing systemic inequalities while fostering community trust and collaboration. By emphasizing both root causes and technological deterrents, these findings provide a foundation for multi-faceted interventions aimed at reducing auto theft, reinforcing the critical role of community stability in crime prevention.

Conclusion

In conclusion, this study set out to examine the relationship between social disorganization and auto theft using socio-economic and law enforcement data. By employing predictive models and focusing on key variables such as unemployment rates, poverty levels, housing vacancy, and educational attainment, we identified significant socio-economic predictors of auto theft in various communities. The findings reinforce the importance of social structures in crime prevention, showing that weakened community cohesion and economic instability contribute to higher rates of auto theft.

The integration of supervised and unsupervised machine learning models enhanced our understanding of these dynamics. While the supervised models quantified the relationships between specific variables and auto theft rates, the unsupervised models provided deeper insights into community-level patterns, uncovering unique clusters and unexpected anomalies. These methods together emphasized the complexity of the factors driving auto theft and bring out the need for context-specific interventions.

Furthermore, this research underscores the ethical responsibility to address systemic inequalities and engage with stakeholders when designing crime prevention strategies. Policies that strengthen communities, reduce housing vacancy, and create economic opportunities can address the base causes of auto theft, rather than simply its symptoms. Working with local

governments, law enforcement, urban planners, and community members will be essential to bringing on equitable and effective solutions.

References

Bennett, T., Holloway, K., & Farrington, D. P. (2006). Does neighborhood watch reduce crime? A systematic review and meta-analysis. *Journal of Experimental Criminology*, 2(4), 437–458. https://doi.org/10.1007/s11292-006-9018-5

Bowers, K. J., Johnson, S. D., & Hirschfield, A. (2004). Closing off opportunities for crime: An evaluation of alley-gating. *European Journal on Criminal Policy and Research, 10*(4), 285-308. https://link.springer.com/article/10.1007/s10610-005-5502-0

Chelsea M. Spencer, et al. "The Role of Income Inequality on Factors Associated with Male Physical Intimate Partner Violence Perpetration: A Meta-Analysis." *Aggression and Violent Behavior*, Pergamon, 22 Aug. 2019, www.sciencedirect.com/science/article/abs/pii/S1359178918302726.

FBI. (2016, August 16). *Motor vehicle theft*. FBI. https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/offenses-known-to-law-enforcement/motor-vehicle-theftTheft

Lee, B., Lee, J., & Hoover, L. (2016). Neighborhood characteristics and auto theft: An empirical research from the social disorganization perspective. *Secur J, 29*, 400–408. https://doi.org/10.1057/sj.2013.35

National Insurance Crime Bureau. (2020). *Vehicle theft prevention tips*. NICB. https://www.nicb.org/prevent-fraud-theft/vehicle-theft-prevention

Prasetyo, E., & Hidayat, R. (2021). GPS-Based Vehicle Tracking and Theft Detection Systems using Google Maps and SMS. *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 7-12.

https://ieeexplore.ieee.org/document/9501928?arnumber=9501928&utm_source=

Sampson, R. J., & Groves, W. B. (1989). Community structure and crime: Testing social-disorganization theory. *American Journal of Sociology, 94*(4), 774-802. https://doi.org/10.1086/229068

Straughan, D. (2024, September 12). *Car theft statistics 2024*. MarketWatch. https://www.marketwatch.com/guides/insurance-services/car-theft-statistics/

Suresh, G., & Tewksbury, R. (2013). Locations of motor vehicle theft and recovery. American Journal of Criminal Justice: *AJCJ, 38*(2), 200-215. https://doi.org/10.1007/s12103-012-9161-7

Welsh, B. C., & Farrington, D. P. (2008). *Effects of improved street lighting on crime*. Office of Justice Programs. https://www.ojp.gov/ncjrs/virtual-library/abstracts/effects-improved-street-lighting-crime

Wickert, C. (2023, November 28). *Social Disorganization Theory (Shaw & McKay)*. SozTheo. https://soztheo.de/theories-of-crime/social-disorganization/soziale-desorganisation-shaw-mckay/?lang=en