Vehicle Theft in Charlotte, North Carolina

A Research Study on Predictive Policing

Mahmoud Alhousseiny, Tacoma Jones, Mason Keziah, Gray Lumsden

Department of Data Science, University of North Carolina at Charlotte

DTSC 3602

Dr. Nadia Najjar & Dr. Shannon Reid

December 8th, 2023

Problem Statement and Background

Problem Statement

Crime involving vehicle-related theft is a pervasive and growing issue in urban areas across the United States. According to the Federal Bureau of Investigation, in 2019 alone, an estimated 721,885 incidents of motor vehicle thefts were reported nationwide, with an estimated rate of 219.9 vehicle thefts per 100,000 residents (Uniform Crime Reporting Program, 2019). Regrettably, Charlotte, North Carolina, is not exempt from this trend. Recent data suggests a total of 738 incidents of vehicle theft in 2022, increasing by nearly 50% the following year, with 1,468 incidents in 2023 (Charlotte-Mecklenburg Police Department, 2023). Other previous CMPD data, accumulated between 2011 and 2016, further highlights the overwhelming presence of motor vehicle theft in the Charlotte region, with 11,305 incidents spread across the six years, around half of which occurred at a private residence.

With the utilization of predictive modeling methodologies that are adjacent to those commonly used within the realm of predictive policing, this study seeks to delve deeper into the issue of vehicle theft-related crime in Charlotte, NC, focusing on crime that occurs within a private residence. Utilizing relevant local and national data, this study considers a range of predictors to forecast whether a theft transpired within a private or non-private residence. This research hopes to contribute meaningfully to the knowledge that informs crime prevention and contributes to the ongoing efforts of developing predictive policing tools. We may foster safer and more secure urban environments. Understanding the propensity of vehicle-related thefts in different areas of Charlotte could aid in resource allocation for its police departments.

Impacts & Stakeholders

The consequences of vehicle theft affect individuals, communities, and the broader society. For individuals, the loss or damage to their vehicle can result in substantial financial and emotional distress. Communities where vehicle crimes are rampant may experience depreciated property values, increased insurance premiums, and a general atmosphere of insecurity and mistrust (Ihlanfeldt, 2010).

As a metropolitan region, Charlotte unsurprisingly scores low on walkability, with critiques ranking Charlotte 26 out of 100 (Palmisano, 2023). This ranking underscores the city's heavy reliance on vehicular transportation. Consequently, elevated vehicle theft rates pose a significant threat to its residents' overall quality of life. Addressing the problem is paramount to fostering safer, more enjoyable communities and mitigating the economic and social repercussions of vehicle crimes.

Background

Integrating modern data-driven methodologies, particularly those encompassed within predictive policing, has dramatically reshaped the criminal justice landscape. This evolution offers the potential to avert specific criminal acts and presents a paradigm shift towards preemptive action. When deployed, predictive policing strategies can substantially mitigate incidents of vehicle theft, thereby optimizing resource distribution among local law enforcement and fostering more secure communities. However, this complex challenge affects the communities' quality of life and mobility. This portion of the current study examines the role of predictive policing (PP) in addressing vehicle-related crimes, focusing on the multifaceted nature of vehicle theft, incorporating victim data, seasonality, vehicle specifics, and socioeconomic factors.

Vehicle theft transcends mere property loss; it directly attacks the victim's well-being and daily functioning, especially in cities like Charlotte, NC, where the dependency on cars is high due to limited public transportation options. Lee and Sener (2015) emphasize the profound impact of vehicle theft on long-term transportation needs, which is particularly relevant in Charlotte, given its low walkability score (Palmisano, 2023). The Charlotte-Mecklenburg Police Department (2023) reports a dramatic increase in vehicle thefts, with a 50% rise in one year, underscoring a growing problem within the city. This escalation reflects a local security issue and places Charlotte in the lower safety percentile nationally, with 83% of cities being safer than Charlotte. (CrimeGrade.org, 2023).

Vehicle-related crimes fluctuate with seasons, with places like Vancouver experiencing the lowest rates during harsh winters (Andresen & Malleson, 2013). This pattern indicates that weather and seasonality could also influence crime trends in Charlotte despite the milder winters. Moreover, studies by Sugiharti et al. (2023) reveal that wealth inequality and expenditure patterns are significant motivators of criminal behavior, with poverty being positively related to crime. These findings suggest that incorporating census-derived income and poverty metrics could enhance the predictive accuracy of PP concerning vehicle theft.

Predictive policing has been shown to reduce crime rates, including vehicle-related thefts, to varying degrees (Mugari & Obohia, 2021). However, implementing predictive policing must be handled with care to address ethical concerns, such as privacy and the potential for systemic bias. La Vigne & Lowry (2011) demonstrate that strategic camera surveillance in public parking lots can be an effective deterrent, while Weisel et al. (2006) point out that specific vehicle models are more likely to be stolen, suggesting that car characteristics should be factored into predictive models. However, Benson et al. (2021) caution that without careful consideration, predictive policing could perpetuate class and racial disparities. Similarly, the use of biometrics, such as face, fingerprint, and iris, has been proposed to act as authentication for prevention (Zhou, Du, Thomas, & Delp, 2012); however, face issues of differing biological characteristics from baseline authentication, due to an accident or aging. Surveillance systems have been shown to reduce crime, but only in specific contexts, and whether they merely displace crime requires more study (Kille, 2014).

The daily impact of vehicle theft is significant, and predictive policing has the potential to mitigate this issue in car-dependent cities like Charlotte. By analyzing variables such as seasonality, vehicle characteristics, victim data, and socioeconomic information, law enforcement can make informed predictions to prevent car theft. However, it is crucial to approach predictive policing with an awareness of the social implications, ensuring that the technology serves to protect without contributing to existing disparities. Further research is needed to refine these methods and articulate a clear, concise, and equitable strategy for predictive policing implementation.

The Data

Data Acquisition

The initial datasets were shared with research groups by the academic instructors of the course, who obtained the data exclusively from the Charlotte Mecklenburg Police Department. This study uses incident reports from 2011 to 2016, detailing attributes relative to vehicles, victims, seasonality, and geospatial characteristics. After initializing with this dataset, the data is processed through ArcGIS to accompany its geographical characteristics with other newly acquired characteristics, including census tract attributes and police station proximity. Integrating census tract attributes allowed for the introduction of new socio-economic factors provided by the 2011-2016 American Community 5-Year Estimate, sourced from the official website of the United States Census Bureau (U.S. Census Bureau, 2015).

Data Cleaning

In the initial data cleaning and integration phase, datasets from individual years (2011-2016) were merged into a comprehensive dataset spanning all six years. A new column, "year," was introduced to this unified dataset, indicating the year each incident occurred.

During this initial integration, the 2014 incidents file had anomalies in its features, which were subsequently removed. The 2011 incidents file lacked the columns 'Incident_From_Time' and 'Incident_to_Time'. To maintain consistency, these columns were added, albeit with null values. It was later ascertained that these columns were not pivotal for our model, but the correction was preserved as a standard practice.

After addressing these discrepancies, the dataset was filtered based on a feature called 'NIBRS_Hi_Class'. This categorical variable indicates the type of crime specific to each incident. Initially, the intent was to exclude any entry where the 'NIBRS_Hi_Class' value was unrelated to vehicle theft-related crime.

The dataset was subsequently imported into ArcGIS. Here, features related to geospatial characteristics, based on census tracts, were aggregated using the 'X-Coordinate' and 'Y-Coordinate' columns of the dataset. This integration facilitated the visualization of vehicle theft-related incident reports in the Charlotte area and the incidents' proximity to police stations. Moreover, it streamlined the process of incorporating additional census tract data-related characteristics sourced from other databases.

For this integration, the transition of the dataset from the VSC coding environment to ArcGIS encountered data corruption issues. While the file was initially standardized to a .csv format, it is presumed that the sizable nature of the dataset led to specific data being misrepresented. Among the most significantly impacted was the 'Complaint_No' column, a crucial primary key for the study's dataset. This posed a substantial challenge. By this stage in the data integration and cleaning process, vehicle and victim characteristics were planned to be merged with the current dataset, both of which depended on the 'Complaint_No' primary key for accurate integration. Alarmingly, this corruption condensed over 10,000 unique complaint numbers into 13 across all rows. To address the issue of data loss, the files were transitioned to a .xlsx format, which proved more resilient against such data corruption.

With the inclusion of census tract data for each entry, additional information was aggregated to the dataset grouped by census tract. This new data encompasses each tract's poverty rate, unemployment rate, and median household income. Once these new features were implemented into the dataset, the study moved on to correlation analysis.

Methods

Variable Exploration

Target Variable

The study's target variable is "Place2," which aims to predict vehicle theft incidents within specific census tracts. "Place2" is a dichotomous variable distinguishing between 'Private Residence' and 'Non-Private Residence'. Intriguingly, upon analyzing the training dataset, vehicle theft incidents are evenly distributed between these two classifications; or in other words, about 60% of the reported thefts occurred within private residences. In contrast, the remaining half occurred in non-private residences.

Predictor Variables

In this research, a diverse set of predictors is incorporated to model the vehicle theft-related incidents within the specified census tracts of Charlotte. The predictors are segmented into categorical and numerical variables:

Categorical Predictors

- 'Vehicle_Make' - The vehicle manufacturer.
- 'Vehicle_Model' - The vehicle model, based on manufacturer.
- 'Vehicle_Color' - The color of the vehicle.
- 'Vehicle_Body' - The category/type of vehicle.

- 'Victim_Race' - The race of the victim.
- 'Victim_Gender' - The gender of the victim.
- 'Victim_Ethnicity' - The ethnicity of the victim.
- ‘In_Police_Buffer’ - Whether the crime occurred within the vicinity of a police station.
- ‘Seasonality’ - Season that the crime occurred.

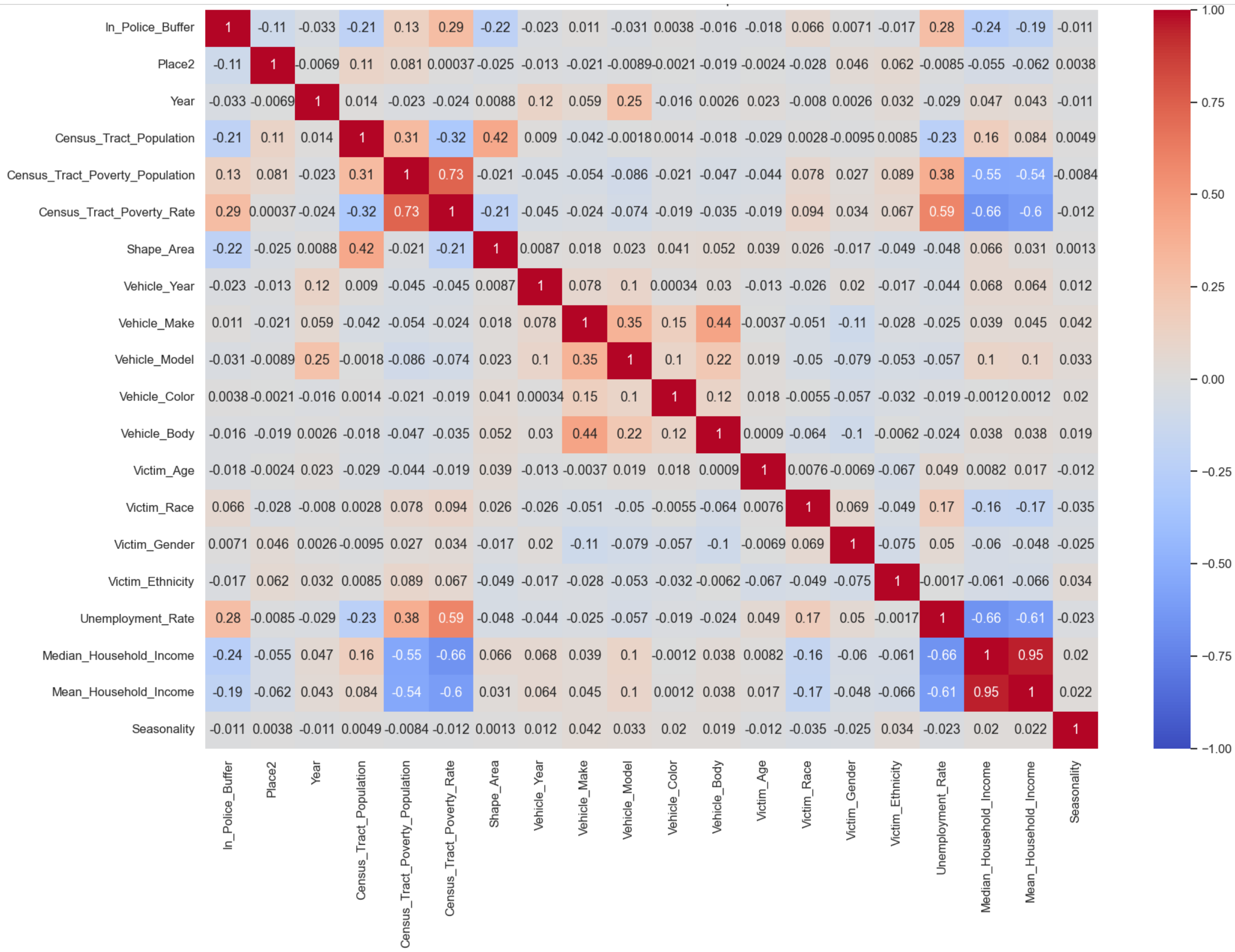
These categorical predictors are all attributes that are involved with the individuals, as well as their vehicles, that may be victimized by a vehicle related theft, whether it be in a private or non-private residence and are thus theorized to help predict the target variables value.

Numerical Predictors:

- 'Vehicle_Year' - The year vehicle was made
- 'Census_Tract_Poverty_Rate' - The rate of poverty based on census tract
- 'Victim_Age' - The age of the victim
- 'Unemployment_Rate' - The rate of unemployment based on census tract
- 'Median_Household_Income' - Median household income based on census tract

For this study, no binning was conducted on the numerical predictors. However, future considerations may involve the potential binning of variables such as 'Census_Tract_Poverty_Rate' and 'Median_Household_Income' to provide more granular insights or to handle non-linearity.

Correlation Analysis



Census_Tract_Poverty_Population vs. Census_Tract_Poverty_Rate:

Correlation Value: 0.73

Analysis: The robust correlation between the Census_Tract_Poverty_Population and the Census_Tract_Poverty_Rate suggests that these variables convey largely overlapping information. Including both in a predictive model might introduce multicollinearity, potentially affecting the stability and interpretability of the model. Given the potential redundancy, retaining only the 'Census_Tract_Poverty_Rate' may be prudent, eliminating the population count from subsequent analyses.

Median_Household_Income vs. Mean_Household_Income:

Correlation Value: 0.95

Analysis: As anticipated, a substantial correlation exists between the median and mean household incomes. By its inherent nature, the median is resilient to outliers, rendering it potentially more reliable in heterogeneous datasets. To mitigate multicollinearity concerns, the 'Mean_Household_Income' variable might be excluded in favor of the median.

Unemployment_Rate's Relationship with Median_Household_Income and Mean_Household_Income:

Correlation Values: -0.66 and -0.62, respectively

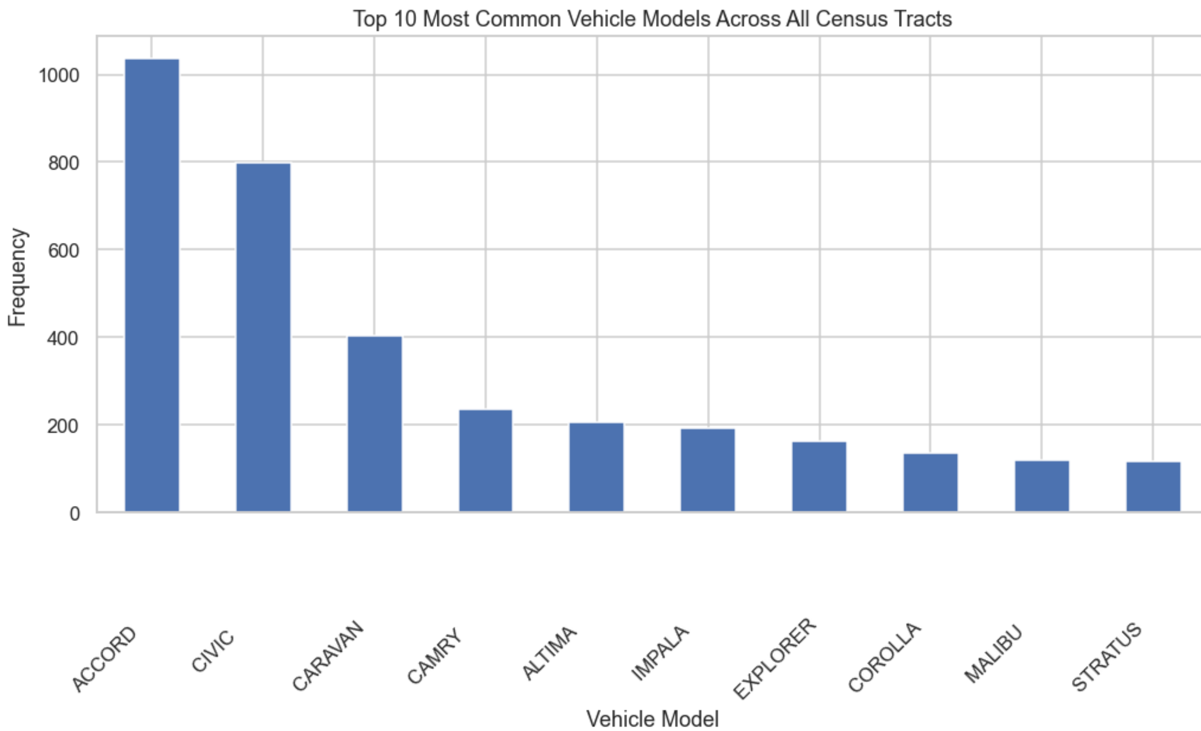
Analysis: The observed negative correlations suggest that regions with elevated unemployment rates typically manifest reduced income levels. This is a logical association, underscoring the economic interdependence of these metrics. Given the relatively moderate strength of these correlations, they offer complementary insights. Retain these variables in the model to encapsulate the nuanced socio-economic landscape of the regions under scrutiny.

Shape_Area's Relationship with Census_Tract_Population:

Analysis: The correlation between tract area and census tract population is anticipated, highlighting that population density is somewhat tethered to the geographical constraints of an area. However, from a predictive standpoint, the 'Census_Tract_Population' likely encapsulates the salient information also conveyed by 'Shape_Area.' Favor the inclusion of 'Census_Tract_Population' over 'Shape_Area,' attributing greater analytical weight to population dynamics.

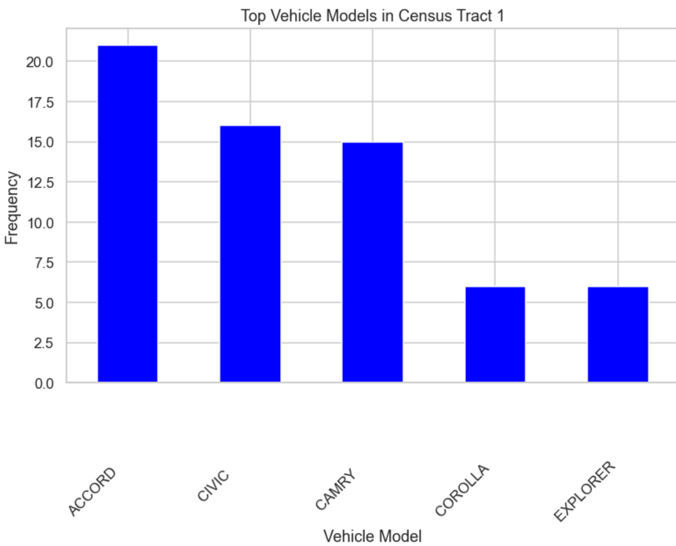
Key Observations

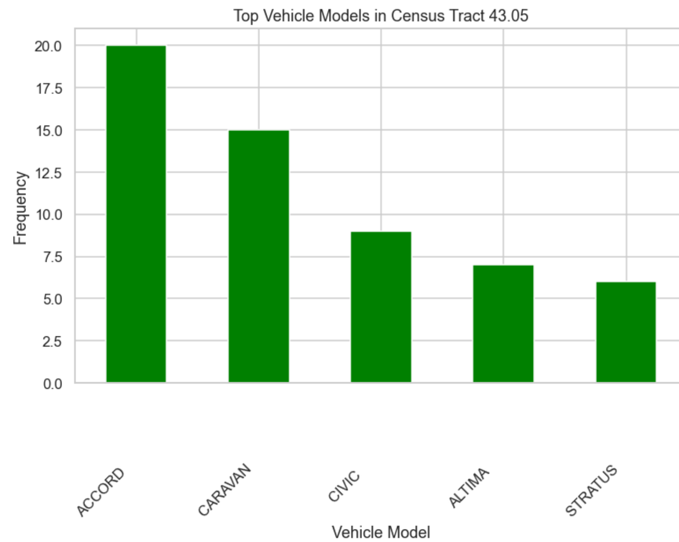
Commonality by Vehicle Type



Naturally, the most common vehicles from the Census data are the affordable ones. Considering the relationship between vehicle crime and socioeconomic factors, it can be rationalized that the vehicles in the above visualization are at the greatest risk of theft, as their general affordability relates to their proximity to the conditions that create vehicle crime.

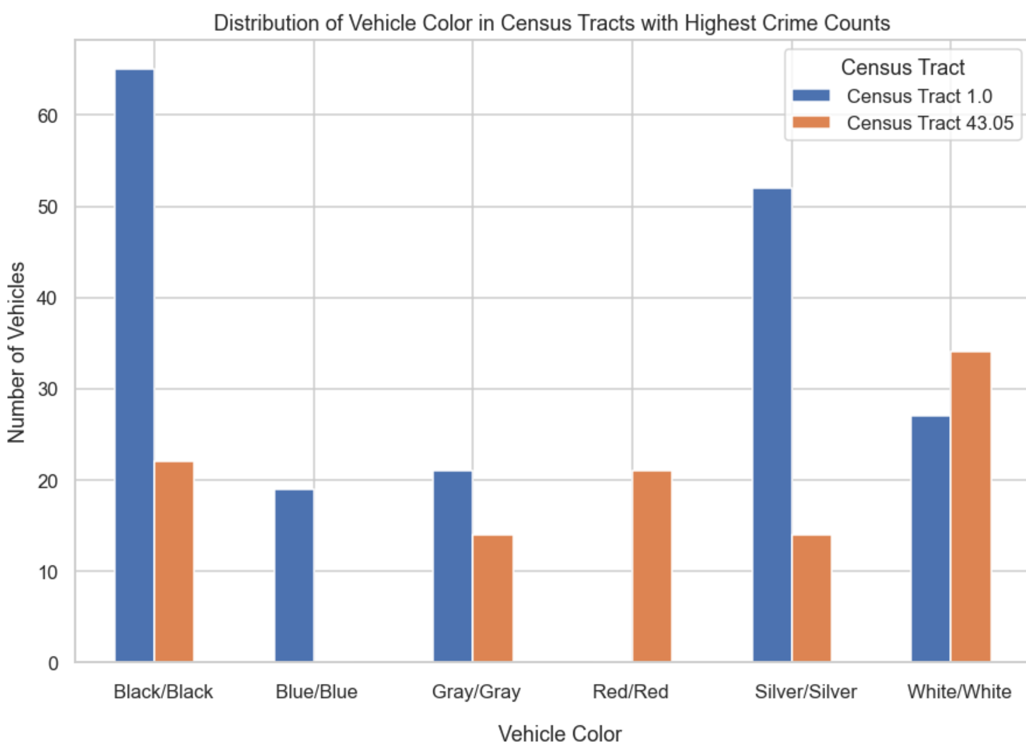
Commonality by Vehicle Types by Tract





This visualization is similar to the previous chart but offers a direct comparison between two different census tracts and therefore different areas of the city. The census tracts chosen for this visualization contain the highest concentrations of vehicle theft. While the Honda Accord and Honda Civic are still highly prevalent, knowledge about vehicle affordability can be utilized in looking at different tracts to determine the vehicle makeup of each area, as well as some of their subsequent socioeconomic conditions.

Distribution of Vehicle Color



From this visualization, it can be seen that vehicles with neutral colors such as black, silver, or white stand a much higher chance of being stolen when compared to others such as red or blue. This is likely due simply to the prevalence of these colors among vehicle supply and demand, but could also be due to a lack of visibility or 'shock-value' in the perspective of perpetrators. Since black, white, and silver cars are more common they will obviously be stolen more often, but those colors are also less likely to stand out amongst other vehicles.

Modeling

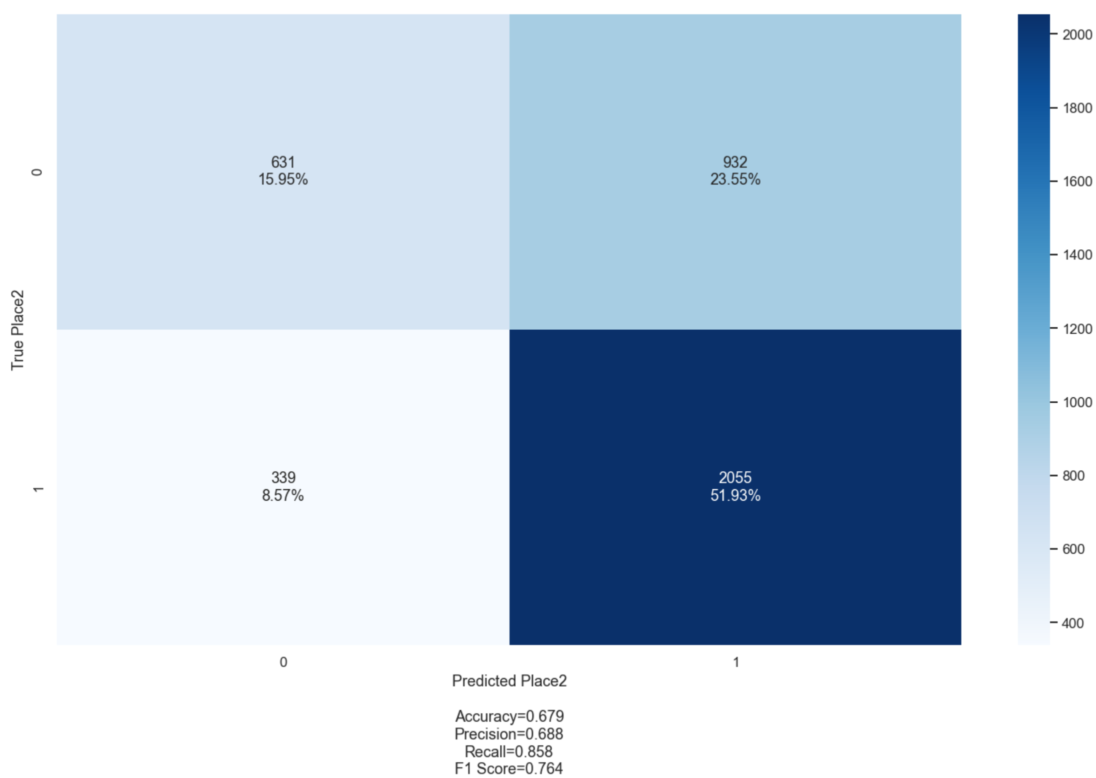
For the study, xgboost, random forest, cart, sgd, and c5 classifiers were used to predict the target variable. From the results, the xgboost model better represented the data than the other models' accuracy, standard deviation accuracy of folds, and average accuracy of folds. The closest-performing model was the random forest model. The xgboost model was used over the random forest model because the xgboost model had more performance, speed, and flexibility. Overfitting was a particular problem for the random forest model as it seemed to overfit the training data, thus achieving an accuracy of 99%. The xgboost model avoided this issue entirely. The xgboost model also handles missing values and has more tuning capabilities. As a result, the focus shifted to relying purely on the xgboost model for data interpretation. xgboost stands for eXtreme Gradient Boosting and is a widespread and efficient open-source implementation of the gradient-boosted trees algorithm.

The model was paired with K-Fold cross-validation from the sklearn library. The module is initialized with ten splits and a shuffle parameter set to true, ensuring the data is randomly shuffled before splitting the folds. The cross-validation helps produce a reliable estimate of a model's performance. A confusion matrix and features of importance visualization were used to understand the models' performance. The confusion matrix and features of importance help with understanding where the model is lacking and help direct future iterations of the model.

Results

The baseline model accuracy is 60.5%, based on the value counts of our target variable's data distribution. Afterward, the first iteration of the xgboost model before hyperparameter tuning achieved an accuracy of 66% on the test set, 0.01 standard deviation of folds, and 0.68% average accuracy of folds. Once the results were retrieved, the next step was to tune the model. For the final model, the accuracy on the test set marginally increased to 67.9%, the standard deviation of the accuracy of folds was 0.015, and the average accuracy of folds was 66.4%. The model marginally improved in 2 out of 3 of the initial metrics that were used.

Confusion Matrix



Confusion Matrix Components

- True Positive (TP): 2055 instances where the model predicted that a vehicle related theft occurs at a private residence (1) and it actually does.
- True Negative (TN): 631 instances where the model predicted that a vehicle related theft does not occur at a private residence (0) and it actually doesn't.
- False Positive (FP): 932 instances where the model predicted that a vehicle related theft occurs at a private residence (1), but it doesn't.
- False Negative (FN): 339 instances where the model predicted that a vehicle related theft doesn't occur at a private residence (0), but it does.

Model Metrics

- Accuracy: 67.9% of the total predictions made by the model are correct.
- Precision: Out of all instances predicted as "vehicle related theft occurs at a private residence", 68.8% are correct.
- Recall (Sensitivity): Out of all actual instances of "vehicle related theft occurring at a private residence", the model correctly identified 85.8%.
- F1 Score: A harmonic mean of Precision and Recall which is 76.4%. It helps balance Precision and Recall in situations where one might be more relevant than the other.

Model Conclusions

The model has a slightly better ability to identify instances where vehicle-related theft occurs at a private residence (Recall = 85.8%) than to be correct when it claims a vehicle-related theft occurs at a private residence (Precision = 68.8%). False positives (FP) are slightly higher than false negatives (FN). This means the model is more likely to incorrectly predict that a vehicle-related theft occurs at a private residence than to miss an actual instance of this event. The model is moderately accurate but only slightly higher than the baseline in terms of accuracy.

Tools

As previously discussed, many tools were used, such as the correlation matrix, target, input, and xgboost model paired with K-Fold cross-validation. The model was then followed up with hyperparameter tuning and data interpretation, such as a confusion matrix and features of importance visualization. Yet, regarding data wrangling, label-encoding was used to transform features with many inputs into simple, easy-to-use numbers. ArcGIS Pro spatially mapped the census data and police buffer zones to the initial CMPD dataset via ‘x’ and ‘y’ coordinates. The dataset also grew many times over as the CMPD dataset combined all the different years of CMPD data to form a single dataset for ease of use. The initial csv format was changed to xlsx to avoid any data loss that occurred with the csv format. In this instance, examples of data loss included complaint numbers, all having the same number, columns missing, etc. Geopandas was used to help visualize our target variables’ distribution throughout Charlotte. The map was then used to focus on specific census tracts with many private/non-private residence car-related crimes.

Regarding some tools that were used but later dropped due to ineffectiveness, many regression models were initially used to predict our target of ‘Place2’ before ‘Place2’ was feature-engineered to be binary for private vs non-private. Initially, the study aimed to predict the most likely areas where vehicle-related crimes would occur. A standard scalar was initialized before the hyperparameter tuning the random forest regressor. A grid search with cross-validation for random forest with 3-fold cross-validation was also employed. Regarding regression models, xgboost regressor, random forest regressor, support vector regressor, k-nearest neighbors regressor, decision tree regressor, and a linear regression model were used. A dictionary was set to store results before executing the models. To interpret the results, mean squared error, root mean squared error, mean absolute error, and r-squared were implemented alongside features of importance graph. Model performance was poor as it failed to outpace the baseline performance of the model. As such, the regression models were omitted and archived. Perhaps if the study were to be iterated upon further, the regression model could change to have some use by way of more feature engineering, more data wrangling, and more data exploration.

Ethical implications

In the realm of predictive policing, particularly in the context of vehicle-related thefts in Charlotte, North Carolina, it is crucial to navigate the ethical implications meticulously and examine how specific variables can influence the project's ethical and practical outcomes. This section of the report reflects on how these ethical considerations were addressed in the study and the breakdown of critical variables and their ethical implications. While predictive models offer valuable insights, they cannot capture the full complexity of human behavior and societal dynamics; thus, human oversight is necessary for interpreting and applying any analysis to the predictive model.

Addressing Bias and Fairness

The initial data set potentially carries the bias inherent in police operations and reporting. To counteract this, efforts were made to ensure the data represented the entire city, including diverse neighborhoods and socio-economic backgrounds. The study aimed to include a broader spectrum of socio-economic conditions by integrating census tract data. Using machine learning algorithms, like xgboost, raises concerns about perpetuating systemic biases. This was ensured by critically examining the features used in the models. Features such as 'Victim_Race', 'Victim_Gender,' and 'Victim_Ethnicity' were included with an understanding of their potential to introduce bias.

The study involved handling sensitive personal data, including victim information. Strict data handling protocols were adhered to to address privacy concerns, ensuring anonymity and data security. Ethical guidelines were adhered to under university and legal standards. Ensuring the interpretability of our models was also a critical ethical concern. Transparency was a key consideration in the methodologies and findings, enabling oversight and accountability.

Addressing Specific Variables

Victim Race, Gender, and Ethnicity

The inclusion of demographic variables like race, gender, and ethnicity in predictive models can be ethically contentious due to the potential for reinforcing systemic biases. Biases were considered in our model's predictions related to these variables. Steps were taken to ensure these features did not result in discriminatory outcomes, such as unfairly targeting specific demographic groups. Including these variables was justified by their potential relevance to understanding patterns in vehicle-related thefts. However, their use was balanced with ethical considerations, ensuring they did not become the primary determinants of predictive outcomes.

Socio-Economic Factors

These variables (Poverty Rate, Unemployment Rate, Median Household Income) are crucial for understanding the socio-economic context of crime but can be sensitive and must be handled properly. These factors provided context and a deeper understanding of the areas prone to vehicle thefts rather than to stereotype or stigmatize specific neighborhoods. To avoid this, considerations were made to ensure that these socio-economic indicators did not skew the model towards overly simplistic correlations between poverty and crime.

Geospatial Data

The spatial aspect of crime is critical for predictive policing, but it also raises concerns about potential over-surveillance in certain areas. In order to recommend an equitable distribution of police resources, as is intended with this predictive analysis, Insights gathered from geospatial components must be ensured that they were gathered in a fair manner and lack bias so that over-policing and under-policing are avoided. In carefully considering the impact this would have on Charlotte's community, the study fosters trust and cooperation between law enforcement and residents.

Vehicle Information

While these details are relevant to identifying patterns in vehicle theft, they must be used judiciously to avoid unintended consequences, such as profiling specific types of vehicles or owners. The analysis focused on objective patterns, such as which vehicle types are more prone to theft, without implying anything about victim characteristics. Insights regarding vehicle information were intended for preventive measures and public awareness rather than for targeting specific vehicle owners.

By carefully considering the ethical implications of these variables and the overall methodology, the project aimed to contribute to the development of more just, effective, and community-aligned predictive policing practices.

Lessons Learned

Key Takeaways

In Charlotte, NC, predictive policing can strategically address the surge in vehicle thefts, particularly at private residences, thereby enhancing the efficiency of resource distribution and the effectiveness of preventative measures. A strong correlation exists between socioeconomic indicators such as poverty and unemployment rates and the incidence of vehicle-related theft; targeted interventions in these areas may lead to a decline in crime rate. Incorporating ethical considerations into predictive policing is crucial to prevent the perpetuation of existing social and economic inequalities within the community.

For the Future

Throughout this research, many lessons were learned on how the project could be improved through further iteration. The research could be refined further through the variables by increasing granularity and reducing non-linearity through binning, as previously mentioned with the ‘Census_Tract_Poverty_Rate’ and ‘Median_Household_Income’ variables. Considering research of this nature, increased and more involved data collection also has the potential to tune the project for its direct purpose since most of the crime data for research was already provided. Lastly, further iterations would see improvement in a more precise, more defined problem statement and additional assessment of the chosen target variable in both application and purpose.

References

Charlotte Mecklenburg Police Department, Crime statistics report. Charlotte. (n.d.). https://www.charlottenc.gov/cmpd/News-Resources/Newsroom/Crime-Statistics-Report

CrimeGrade, The safest and most dangerous places in Charlotte, NC ... - crimegrade.org. CrimeGrade. (n.d.). https://crimegrade.org/safest-places-in-charlotte-nc/

G., L. V. N., & Lowry, S. S. (2011). Evaluation of camera use to prevent crime in commuter parking facilities: A randomized controlled trial. Urban Institute, Justice Policy Center

Ihlanfeldt, K., & Mayock, T. (2010). Panel data estimates of the effects of different types of crime on housing prices. Regional Science and Urban Economics, 40, 161–172. https://doi.org/10.1016/j.regsciurbeco.2010.02.005

Kille, L. W., & Maximino, M. (2014, February 11). The effect of CCTV on public safety: Research roundup. The Journalist's Resource. Retrieved from https://journalistsresource.org/politics-and-government/surveillance-cameras-and-crime/

Palmisano, K. L. (n.d.). Is Charlotte North Carolina a walkable or car ... - charlotteobserver.com The Charlotte Observer. https://www.charlotteobserver.com/living/living-here-guide/article276899918.html

Uniform Crime Reporting Program. (2019, September 13). Motor vehicle theft. FBI. https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/motor-vehicle-theft

U.S. Census Bureau. (2015). SELECTED ECONOMIC CHARACTERISTICS. American Community Survey, ACS 5-Year Estimates Data Profiles, Table DP03. Retrieved November 3, 2023, from https://data.census.gov/table/ACSDP5Y2015.DP03?t=Employment&g=050XX00US37119$1400000&y=2015

Zhou, Z., Du, E. Y., Thomas, N. L., & Delp, E. J. (2012). A new human identification method: Sclera recognition. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 42(3), 571-583. https://doi.org/10.1109/TSMCA.2011.2170416