Predicting Crime Factors in Philadelphia Using Gradient Boosting and XGBoost

Ramsha Perwez
College of Computing & Informatics
Drexel University
Philadelphia, PA 19104, USA
rp955@drexel.edu

Abstract

Leveraging published crime data and various socioeconomic and demographic statistics can be used in combination with machine learning techniques to predict the underlying factors contributing to crime in Philadelphia, Pennsylvania. Using data from the United States Census Bureau and the Philadelphia Police Department, this paper establishes the relationship between criminal activity trends and socioeconomic factors. Gradient Boost Machine and Extreme Boost Machine models are used to identify consistently performing variables. By measuring the Root Mean Squared Error and feature importances, insights are gained into the key factors associated with each crime type. This paper will compare machine learning models Gradient Boost Machine and Extreme Boost Machine to observe if any variables perform well consistently between the methods. Based on recent years' statistics, the study aims to provide insights into potential improvements that can be made to improve the quality of life in Philadelphia.

1 Introduction

Exploring the crime rates in the city of Philadelphia can have profound beneficial insights into improving the livelihoods of the impacted populations. The Federal Bureau of Investigation examined multiple characteristics and found factors such as median income, availability of public transportation, and population density to be prime prerequisites of various types of crime occurring from jurisdiction to jurisdiction [1]. In this paper, non-aggravated theft crimes including burglary of both residential and commercial property, auto-related, individual, and retail thefts will be evaluated against socioeconomic factors. The relationship between these crime rates and five socioeconomic factors will be examined: family unemployment, ineligibility of SNAP benefits despite incomes below the poverty level, poverty class qualification, lack of a bachelor's degree among those over 25, and uninsured health coverage. The results, as measured by Root Mean Squared Error (RMSE) and feature importances, provide insights into which factors are most strongly associated with each type of crime.

2 Data Preparation

The data for this study was sourced from two reputable government agencies: the Philadelphia Police Department and the United States Census Bureau. Both datasets are publicly accessible through their respective official websites. While the US Census Bureau offers API extensions, the specific granularity of the required data made it more efficient to manually select the relevant information from each available

chart by year instead.

The Philadelphia Police Department (PPD) provided crime data, which encompassed reported incidents of burglary, auto-related thefts, theft from person, and retail theft [2]. This data is available in two formats: a 52-week cycle breakdown and a comprehensive annual compilation. It is important to recognize that the data is not provided at a granular level, meaning that it is not broken down by individual neighborhoods. Instead, the data is presented at an overall city level, which helps to maintain consistency and comparability across the datasets. To further reduce the risk of over-fitting in the machine learning models, similar types of criminal activities were grouped together. For example, commercial burglary and residential burglary were combined into a single category, ensuring that the models could capture broader patterns and trends in the data without the results being overly influenced by related subcategories.

Socioeconomic data was obtained from the United States Census Bureau's American Community Survey (ACS). The ACS provides detailed information on a wide range of demographic and economic characteristics at various geographic levels. For this study, ACS data at the overall city level was utilized, as the PPD did not have data available that corresponded to individual census tract level. The specific socioeconomic variables collected from the ACS included:

1. Family unemployment rate [3]
2. Percentage of population below the poverty level who did not receive SNAP benefits [4]
3. Overall poverty rate [5]
4. Percentage of population over 25 years old without a bachelor's degree [6]
5. Percentage of uninsured population [7]

To ensure consistency in the machine learning models, the selected data reflected the overall city statistics and was aggregated annually for the years 2013 through 2022. This approach guarantees that the models operate with uniform parameters across all available datasets, enabling a more accurate and reliable analysis of the city-wide trends over the specified period.

All data was standardized by subtracting the mean and dividing by the standard deviation for each feature. This standardization process helps to improve the performance and interpretability of the machine learning models.

3 Experimentation

3.1 The Mechanics

The code is written in Python 3 and implemented using Jupyter Notebook, an interactive development environment that allows for the creation and execution of code cells alongside markdown documentation. For the statistical calculations, the code uses the following Python libraries: pandas, NumPy, scikit-learn. The following Machine Learning libraries were used for modeling purposes: GradientBoostingRegressor and xgboost. Data was manually loaded into a pandas data frame to make it accessible.

3.2 Machine Learning Models

Gradient Boost Machines (GBMs) and Extreme Gradient Boosting (XGBoost) are ensemble machine learning algorithms that combine weak learners, typically decision trees. GBMs train trees sequentially, with each tree correcting the errors of the preceding ones, enabling the capture of complex relationships and higher accuracy compared to single decision trees. GBMs handle non-linear relationships and diverse data types well but can be prone to overfitting and require more computational resources. XG-

Boost is an iteration of GBM which enhances performance through regularization, advanced tree pruning, and various implementation optimizations. Using both GBM and XGBoost in this study provides a more comprehensive analysis of the data, as each algorithm offers unique strengths, allowing for a better understanding of the relationships between socioeconomic factors and crime rates in Philadelphia [8].

3.3 K-Fold Cross Validation

This study utilizes only 10 years of compiled data. Therefore, K-Fold Cross-Validation is employed as a technique to evaluate the model's generalization ability and mitigate the risk of overfitting. In this implementation, the code utilizes 10-fold cross-validation. Data is split into 10 subsets or folds, with each fold used once as the testing set and the remaining 9 as the training set. This ensures thorough training and testing across all data points, providing robust performance assessment for the model. The average RMSE and feature importances are then calculated across all folds.

4 Results

4.1 Using RMSE For Analysis

RMSE (Root Mean Squared Error) measures how well a model predicts outcomes. A small RMSE indicates accurate predictions, where the model closely matches actual data. For crime factor prediction, a small RMSE suggests the model effectively links socioeconomic factors with crime rates. Meanwhile, a large RMSE signifies less accurate predictions, implying the model struggles to account for all influencing factors or handles complex relationships poorly.

4.2 GBM Results

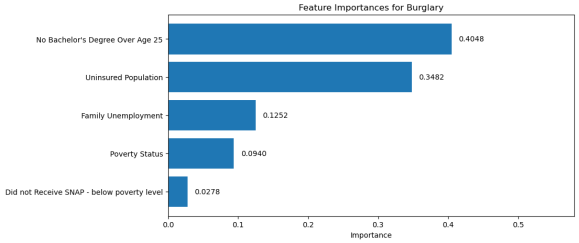
The model predicts auto-related thefts most accurately (RMSE: 0.3943) and struggles the most with theft from person (RMSE: 0.8331). Educational attainment (No Bachelor's Degree Over Age 25) is the strongest predictor of theft across all categories, with feature importance values ranging from 0.4048 (Burglary) to 0.5103 (Auto Related Thefts).

4.2.1 Feature Importance

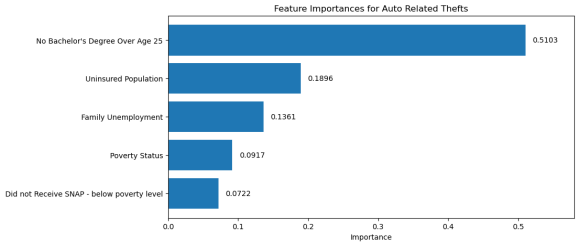
While a lack of higher education (0.4048-0.5103) and family unemployment (0.1252-0.1516) consistently predict theft, the importance of other factors varies. The uninsured population significantly influences burglary (0.3482) and retail theft (0.2070), while its impact on theft from person (0.1157) is less pronounced. Auto-related thefts are most sensitive to family unemployment (0.1361), suggesting a heightened association with immediate financial needs compared to other types of theft (0.1252-0.1405). Poverty status consistently contributes to all types of theft, but its strongest association is with theft from person (0.1457), highlighting a potential link to broader economic hardship.

4.3 XGBoost Results

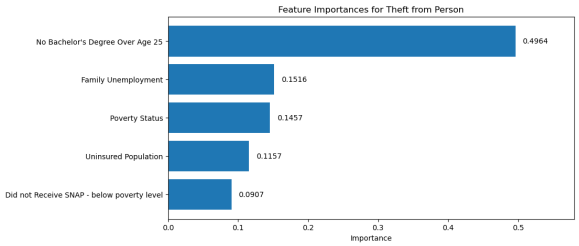
The model predicts burglary most accurately (RMSE: 0.5088) and struggles most with theft from person (RMSE: 0.9616). In the XGBoost model, the key predictors of theft vary significantly across categories. For burglary (RMSE: 0.5088), family unemployment is the most influential factor. Auto-related thefts (RMSE: 0.6933) are primarily driven by poverty status. Theft from person (RMSE: 0.9616) is most strongly associated with a lack of higher education ("No Bachelor's Degree Over Age 25"). For retail theft (RMSE: 0.8398), family unemployment emerges as the primary factor. Notably, "Did not Receive SNAP - below



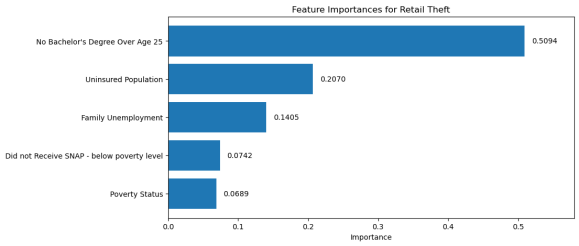
(a) Burglary



(b) Auto Related Thefts



(c) Theft from Person



(d) Retail Theft

Figure 1: Bar charts showing feature importance for different types of theft.

poverty level" consistently holds minimal importance in predicting any type of theft in the XGBoost model.

4.4 Differences in Results

The striking differences in the results obtained from XGBoost and GBM can be attributed to several key factors, including the limited training size, the linearity of the available data, and the distinct methodologies employed by each model. With only 10 years of data available, the models are tasked with learning complex relationships between socioeconomic indicators and crime rates from a relatively small sample size. XGBoost employs advanced regularization and optimization techniques which in practice leads to more accurate predictions and insights. GBM may be effective with this dataset. The data presents a relatively linear relationships between crime rates and socioeconomic factors, a scenario in which simpler models such as GBM excel. GBM's straightforward approach may avoid overfitting the limited data, offering a more generalized model that captures the essential trends without getting bogged down in noise or outliers.

4.5 Erractic Jumps in the Predicted Means

The analysis of crime trends using GBM and XGBoost revealed challenges in predicting mean values due to limited data, potential model overfitting, and feature importance variance across cross-validation folds. These factors, coupled with inherent fluctuations in crime rates and potential outlier events, contributed to erratic predictions in certain years. Despite these challenges, the model identified some influential features, such as family unemployment and poverty status, in explaining variations in crime types. Further investigation is needed to refine the model, potentially through gathering more data and exploring alternative feature combinations. The purpose of this study was to analyze the features affecting crime however, the faultiness of the predictions does not limit our

understanding.

4.6 Conclusion

This study in Philadelphia, utilizing GBM and XGBoost models to predict crime rates based on socioeconomic factors, parallels the challenges and insights highlighted by Li et al. (2022) in their examination of poverty prediction in Kyrgyzstan.

Both studies acknowledge the limitations of relying on a few key variables for prediction. Li et al. (2022) discovered that, even with the advanced XGBoost algorithm, the most influential predictors of poverty varied considerably, suggesting a complex interplay of factors that cannot be easily reduced to a simple formula. Similarly, our analysis found that while factors such as unemployment and education level held predictive power for burglaries, the significance of these factors differed for other crime types. The models themselves differed in their results. Although GBM suggests that education attainment can help attain a safer city, the XGBoost model suggests that all socioeconomic factors can help Philadelphia.

The findings show we must use many approaches, not just one policy change, to solve problems like unemployment, poverty, and unequal education. If we face these challenges directly and use machine learning, we can improve plans to reduce theft and build a stable, prosperous future for everyone.

References

[1] U.S. Department of Justice—Federal Bureau of Investigation. Variables affecting crime (no date). Available at: https://ucr.fbi.gov/hate-crime/2011/resources/variablesaffectingcrime_final.pdf (Accessed: 14 June 2024).

[2] Philadelphia Police Department. Crime Maps & Stats (no date). Available at: https://www.phillypolice.com/crimestats (Accessed: 14 June 2024).

[3] U.S. Census Bureau (2022). Employment Characteristics of Families, American Community Survey, ACS 1-Year Estimates Subject Tables, Table S2302. Available at: https://data.census.gov/table/ACSST1Y2022.S2302?q=PhiladelphiaCounty,PennsylvaniaS2302 (Accessed: 14 June 2024).

[4] U.S. Census Bureau (2022). Receipt of Food Stamps/SNAP in the Past 12 Months by Poverty Status in the Past 12 Months for Households, American Community Survey, ACS 1-Year Estimates Detailed Tables, Table B22003. Available at: https://data.census.gov/table/ACSDT1Y2022.B22003?q=PhiladelphiaCounty,PennsylvaniaB22003 (Accessed: 14 June 2024).

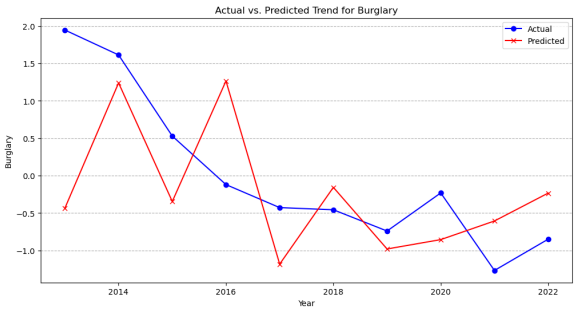
[5] U.S. Census Bureau (2022). Poverty Status in the Past 12 Months, American Community Survey, ACS 1-Year Estimates Subject Tables, Table S1701. Available at: https://data.census.gov/table/ACSST1Y2022.S1701?q=PhiladelphiaCounty,PennsylvaniaS1701 (Accessed: 14 June 2024).

[6] U.S. Census Bureau (2022). Educational Attainment, American Community Survey, ACS 1-Year Estimates Subject Tables, Table S1501. Available at: https://data.census.gov/table/ACSST1Y2022.S1501?q=PhiladelphiaCounty,PennsylvaniaS1501 (Accessed: 14 June 2024).

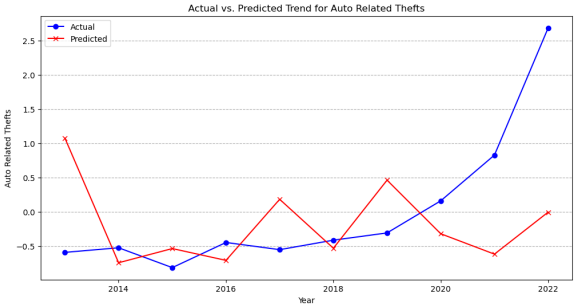
[7] U.S. Census Bureau (2022). Selected Characteristics of Health Insurance Coverage in the United States, American Community Survey, ACS 1-Year Estimates Subject Tables, Table

S2701. Available at: https://data.census.gov/table/ACSST1Y2022.S2701?q=PhiladelphiaCounty,PennsylvaniaS2701 (Accessed: 14 June 2024).

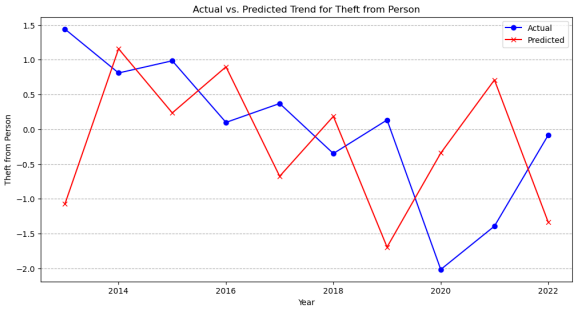
[8] NVIDIA Data Science Glossary. What is XGBoost? (no date). Available at: https://www.nvidia.com/en-us/glossary/xgboost/ (Accessed: 14 June 2024).



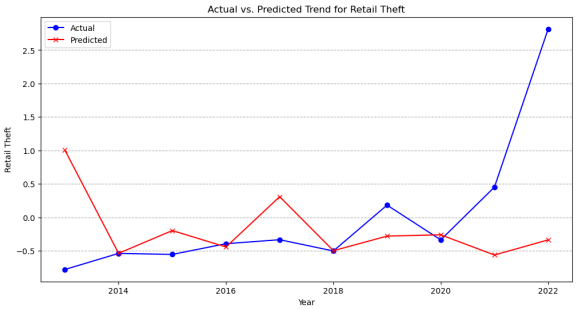
(a) Burglary



(b) Auto Related Thefts

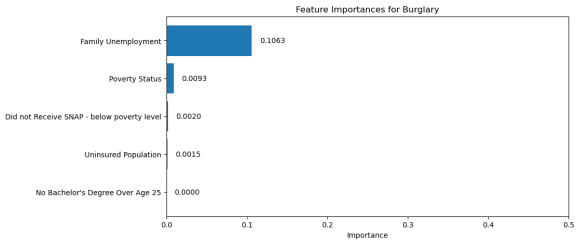


(c) Theft from Person

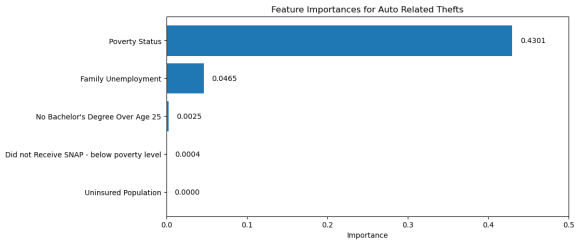


(d) Retail Theft

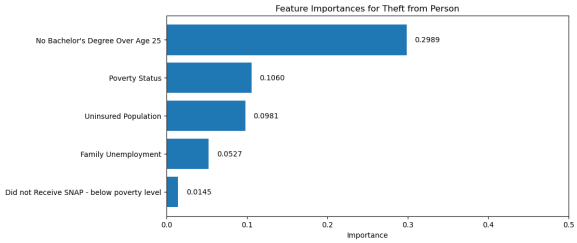
Figure 2: Line charts showing Actual vs. Predicted Means for GBM.



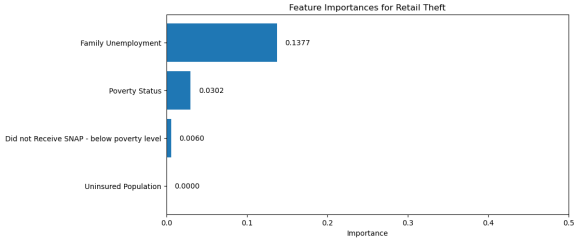
(a) Burglary



(b) Auto Related Thefts

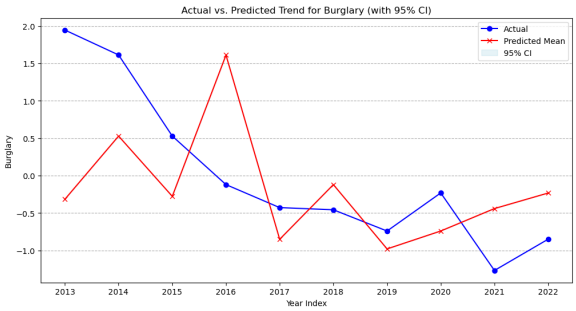


(c) Theft from Person

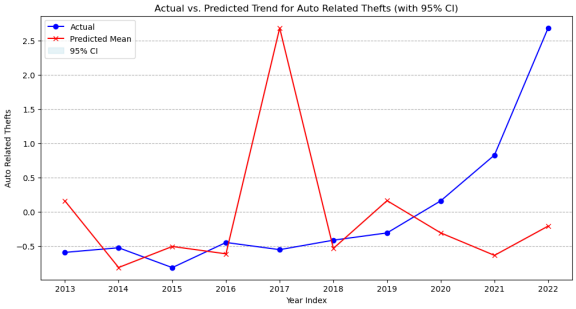


(d) Retail Theft

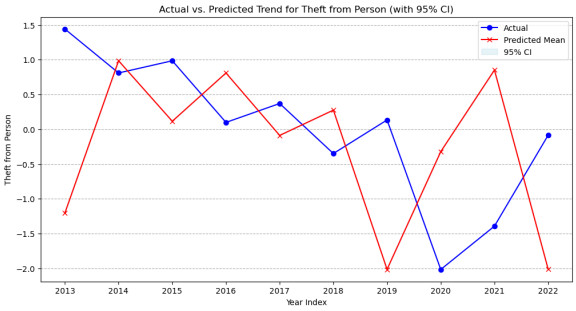
Figure 3: Bar charts showing feature importance for different types of theft for XGBoost.



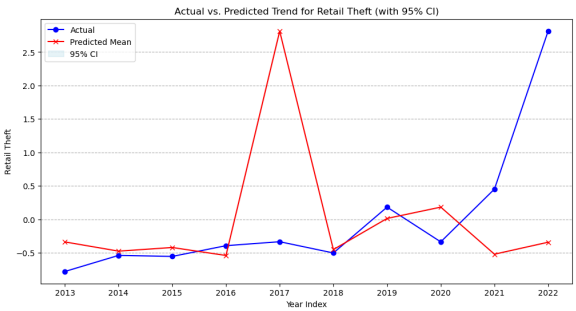
(a) Burglary



(b) Auto Related Thefts



(c) Theft from Person



(d) Retail Theft

Figure 4: Line charts showing Actual vs. Predicted Means for XGBoost.