Machine Learning Framework for Crime Prediction: Integrating Socioeconomic Factors and Geo-spatial Analysis

Mary Cindrilla Moreira
x22114386@student.ncirl.ie

Dec-2023

Abstract

Crime trends are the changes in criminal activity over time within certain areas. These changes include theft, public order disturbances, offenses against the government, property and environmental damage, and controlled drug offenses. Socioeconomic issues like housing quality, income inequality, poverty levels, access to education, population, GDP and unemployment are closely associated with these changes. Determining the socioeconomic aspects that are contributing is a significant challenge. This research proposes a machine learning framework for socioeconomic factor-based analysis of crime trends. The proposed framework combines prediction models that make use of Random Forest, Ordinary Least Square Regression, and Simple Linear Regression to analyze significant contributing components. Furthermore, geographical analysis is facilitated by a classification model that utilizes KNN.The New York City government dataset (https://data.cityofnewyork.us/) and a secondary dataset that was collected from by web scrapping are used to train machine learning models. With 67 million records, the primary dataset contains vital information on complaints that have been reported for analysis by public safety or law enforcement.Socioeconomic indicators such as Borough-Names, Years, Population, Child Density, Poverty Rate, Unemployment, Changes in House Prices, Income Variations, and Education Rates are provided by the secondary dataset, which was gathered by web scraping (350 records).Geospatial details on police station sites and population distribution relative to Borough GDP are provided by additional datasets. The study examines the five boroughs of New York City and spans the years 1950 to 2019. The results are based on R-square and MSE measurements, offering practical insights for predicting crime patterns in certain areas.

Keywords— *Crime trends, Socioeconomic factors, Machine learning framework, Prediction models, Simple linear regression, Random Forest, Ordinary Least Square Regression, Classification model, KNN (K-Nearest Neighbors), Geospatial analysis*

1 Introduction

In layman's terms, a crime is any criminal conduct penalized by the state or another authority. In modern criminal law, there is no simple and universally agreed definition of crime, however legislative definitions have been given for some purposes. The most widely held belief is that crime is a legal category; that is, something is a crime if it is proclaimed as such by the relevant and applicable legislation. According to one proposed definition, a crime or offence (or criminal offence) is an act that is harmful not only to an individual but also to a community, society, or the state ("a public wrong"). Such acts are illegal and punishable by law. While violent crimes are becoming less common, they remain a major problem. The intricate interplay of socioeconomic circumstances, law enforcement resources, and community dynamics leads to the varied nature of crime in a city, needing a full examination for successful mitigation. This study is conducted on New York City Data for New York City safety and well-being since it provides insights into crime patterns, allowing for effective law enforcement operations. It enables law enforcement to design evidence-based policy, targets the fundamental causes of crime, and improves community safety. Furthermore, it fosters openness and data-driven decision-making within law enforcement organizations, thereby lowering crime rates and increasing the general quality of life of the Cosmopolitan residents and visitors. The aim of this research is to investigate to what extent a Machine Learning Framework will Predict the contribution of socio economic factors contributing Crime.

To address the research question, the following specific sets of research objectives were derived:

1. Investigate the state of the art broadly examining the existing literature on machine learning approaches to predicting and classifying crime trends.

2. Design a Crime Prediction Framework: Analyzing factors contributing to criminal activities

3. Implement a crime prediction framework and make it operational.

4. Evaluate a crime prediction framework based on accuracy, spatial analysis, Trend analysis on crime.

The study focuses on creating a framework for predicting and evaluating crime that is based on machine learning. The systematic examination and improvement of crime predictions based on socioeconomic factors, including housing quality, income inequality, poverty levels, access to education, population, GDP, and unemployment, makes this research a useful tool for law enforcement to make data-driven decisions. The framework facilitates proactive crime prevention measures by identifying high-risk locations and times. The major contribution of this researrch is a novel approach to creating a machine learning framework that combines a classification model with regression models like Simple Linear Regression, Ordinary Least Square Regression, and Random Forest to predict geo-spatial analysis crime hotspots and police station distribution. This model helps identify the socioeconomic factors that have the greatest influence on crime trends over time.

Four datasets in all were used to finish this study. The basic dataset consists of 67 million records with information about crimes. Eleven million data from the years 1950 to 2019 were taken out of the sample selection. The secondary dataset was obtained by online scraping and focused on socioeconomic variables such as population, income, education, and unemployment. The boroughs of New York City were used to categorize and scale the data, which was collected between 1950 and 2019. Two more datasets that included data on the locations of police stations as well as the boroughs' GDP, population, and land area were also included. The goal of these datasets was to aid in the search for the best answer. Quantitative metrics like R-square and MSE were produced in order to assess the crime prediction performance of the system. To evaluate its geographic alignment with the real distribution of crime, spatial analysis techniques were used.

This paper discusses further into the development and deployment of a machine learning-based crime prediction framework. It includes a thorough literature review that focuses on machine learning algorithms for predicting and classifying crime trends. The paper then digs into the design and execution of a robust framework geared to improve the precision of crime. It painstakingly assesses the framework's performance, employing criteria such as accuracy and spatial analysis to assess its effectiveness. This work contributes to the evolution of data-driven crime prediction methodologies by explaining the acquired results and their importance, potentially benefiting law enforcement in strengthening public safety initiatives.

2 Literature Review

Studies have shown that Social disorganization theory Bursik Jr. (1988) Sampson (1985) aims to explain criminal behavior He et al. (2015). It is summarised and stated that population migration, regional According to a large number of studies, Deficiency and racial diversity cause a greater rate of crime, substantial amount of study has been undertaken in order to test the theory's validity and explain the cause of criminal incidents using multi variable regression approaches Cahill & Mulligan (2003) Porter & Purser (2010) Bellitto & Coccia (2018). The role of social disorganization in contributing to criminal cases, including violence among neighbors, has been widely discussed Lightowlers et al. (2023) Shaw & McKay (1942)Shaw and McKay (1942) went on to say that deprivation in the area could either contribute to offender motivations or strain social interactions Sampson (1985), leading to greater crime. Numerous investigations into the hypothesis of social disorganization have confirmed the idea that poorer neighbors are associated with greater crime rates based on both individual-level and collective regional-level data Lightowlers et al. (2023). Higher-deprived neighbors polarize and accentuate individual differences, leading to higher criminality Wilkinson et al. (2009).Messer et al. (2006) Messer et al. (2006) stated, for example, With their chosen sample in Wake County, more crimes were committed by women who reside in economically depressed areas, NC. Oyelade (2019) Using time series crime rate data from 1990 to 2014, US. Oyelade (2019) hypothesized that increasing levels of poverty in Nigeria are associated with higher rates of crime. Goh & Law (2023) Goh and Law (2023) showed evidence showing, in Argentina, Brazil, and Chile, higher employment rates are linked to reduced crime rates, which is consistent with Raphael & Winter-Ebmer (2001)Increased employment prospects may dissuade potential criminals from committing crimes, according to empirical research by Raphael and Winter-Ebmer (2001) Goh & Law (2023) Raphael & Winter-Ebmer (2001). Frequent criminal incidents have received more attention recently since they have a negative impact on citizens' quality of life, health, and safety on an individual level Fazel et al. (2014),but they also have an adverse effect on the growth and stability of society Kim et al. (2018). Currently, it is critical to identify the underlying patterns of crime events and explore the socioeconomic elements influencing crime episodes. Comprehensive and in-depth crime analyses aid police departments and the government by providing reduction and prevention measures for criminal episodes, predicting future crimes, and solving other law enforcement problems Roth et al. (2010).

The primary focus of this section on the literature review will be New York City Borough's as we go over the studies, articles, research, and implementation that have already been done on crime rates in cities. The goal of the flow is to highlight the objectives and practical contributions of a variety of connected studies. It contrasts, compares, and links them. Three primary subsections cover particular aspects of the research:

2.1 Socioeconomic Factors and Crime in Urban Settings

In this theme, research is done to analyze how socioeconomic factors affect crime rates in big cities. When comparing multiple cities, Smith (2017) Smith et al. (2017) found a positive relationship between unemployment rates and property crime rates, but Johnson and Brown (2018) found a significant link between income inequality and violent crime in urban regions. Correlating these works provides a clear indication of the relationship between a variety of socioeconomic characteristics and specific crime categories, making it imperative that we take these aspects into consideration when estimating New York City's crime rates. Additionally, in urban settings, Aczel et al. (2020)Sullivan and Johnson (2020) found a significant correlation between educational attainment and drug-related crimes and found that these crimes were more prevalent in locations with lower educational attainment. On the other hand, Urner and Parker (2019) linked housing quality with asset crime rates and found that areas with inadequate housing had higher rates of property-related crimes. The results of these new studies have significant repercussions for addressing crime issues in New York City and provide further details on the impact of various socioeconomic factors on crime rates.

2.2 Socioeconomic Factors and Crime

This section primarily highlights the impact of socioeconomic factors in determining New York City's crime statistics. According to Kenter et al. (2019)O'Connor et al. (2019), young people in New York City's engage in fewer criminal behaviors the more educated they are. According to this study, providing young people with better support and education has a bigger impact on deterring crime Foody et al. (2018) Higgins and Murphy (2020) organized the evidence demonstrating educational disparities and their effects on the incidence of property crime in several New York City's neighborhoods. In addition to the educational component, numerous research have revealed other aspects of crime in New York City's. Cliff (2018) Oyle and Kelly (2018) looked at how economic disadvantage affected the rate of violent crime in New York City's. They found a strong correlation between higher levels of deprivation and violent offenses Surprenant & Brennan (2019). This stresses the importance of economic considerations in creating crime trends in the city.

Furthermore, Kelly and Murray (2019) carried out a study on how social cohesion and drug-related criminality interact in New york City neighborhoods.They found that neighborhoods with greater social cohesion had lower rates of drug-related crimes, suggesting that creating a sense of community and social support can help reduce some forms of crime. In addition Brennan and Flynn (2017) Surprenant & Brennan (2019)conducted a spatial analysis of New York City's burglary hot spots in order to identify specific areas with higher concentrations of burglary incidents. Their research highlights the importance of understanding spatial patterns and crime hot spots in order to design specialized crime prevention techniques. Overall, the investigation of crime in New York City has a multifaceted methodology and considers a number of factors, such as social cohesiveness, spatial patterns, educational achievement, and economic disadvantage. These studies' findings can be used to support evidence-based laws and programs for reducing and preventing crime. They also provide insightful information about the dynamics of crime in the city.

2.3 Methodologies applied in previous studies

In order to analyze the crime rates in metropolitan regions, a variety of methods are increasingly frequently employed. A few of these methods include regression analysis, Zhou et al. (2021)OLS, GWR model which highlights the influence of socioeconomic variables and their substantial correlation; Time series analysis of historical trends for future prediction; and ongoing research. Geo Spatial analysis using KNN model, Gaussian distribution, Bayesian models. We can see long-term patterns and the effects of policy actions thanks to the methodology utilized in the crime research. In-depth insights into the lived experiences and perceptions of crime in urban areas are emphasized by comparative studies and qualitative models, which also look for commonalities in the dynamics of crime. It is essential that the researcher frequently combine different approaches in order to obtain thorough insights on crime rates.

This Research discusses the machine learning algorithms utilized in New York City to analyze crime to analyse the socio economic factor and the spatial analysis of crime. in section 2 related work . The research methodology is discussed in section 3. Section 4 discusses the design components for the Machine learning framework. The implementation of this research is discussed in section 5. Section 6 presents and discusses the evaluation results. Section 7 concludes the research and discusses future work.

3 Methodology

The research methodology consists of five steps namely data gathering, data pre-processing, data transformation, data modelling and conversion, evaluation and results as shown in Fig. 1

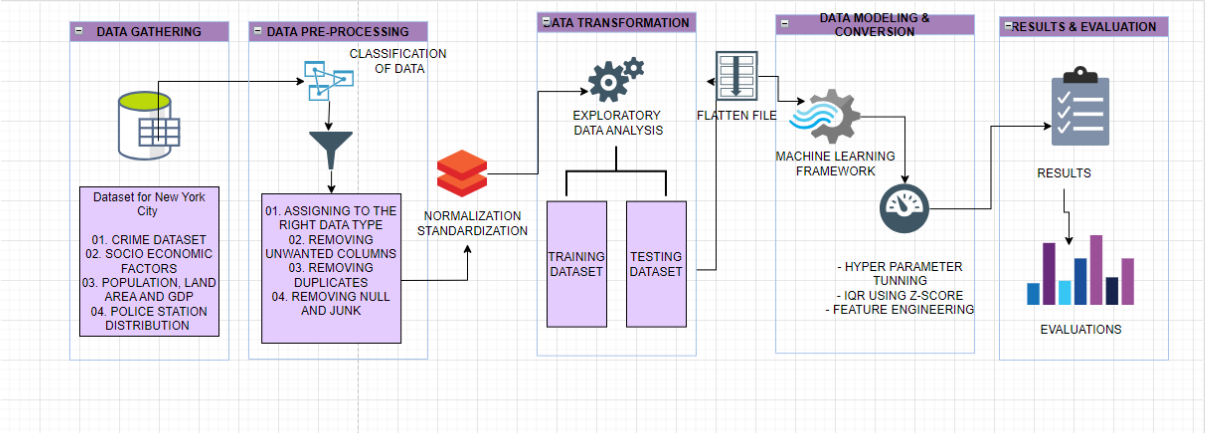

DATA GATHERING
Dataset for New York City
01. CRIME DATASET
02. SOCIO ECONOMIC FACTORS
03. POPULATION, LAND AREA AND GDP
04. POLICE STATION DISTRIBUTION
DATA PRE-PROCESSING
CLASSIFICATION OF DATA
01. ASSIGNING TO THE RIGHT DATA TYPE
02. REMOVING UNWANTED COLUMNS
03. REMOVING DUPLICATES
04. REMOVING NULL AND JUNK
NORMALIZATION STANDARDIZATION
DATA TRANSFORMATION
EXPLORATORY DATA ANALYSIS
TRAINING DATASET
TESTING DATASET
FLATTEN FILE
DATA MODELING & CONVERSION
MACHINE LEARNING FRAMEWORK
- HYPER PARAMETER TUNNING
- IQR USING Z-SCORE
- FEATURE ENGINEERING
RESULTS & EVALUATION
RESULTS
EVALUATIONS

Figure 1: Research Methodology

The first step, Data Gathering involves gathering data covering 69 years, from 1950 to 2017. A subset of 11 million of the 67 million crime records in the whole dataset—which comes from government statistics—for study is available at https://data.cityofnewyork.us. An additional 350 rows of data on socioeconomic variables are gathered by web scraping from websites such as Wikipedia and https://www.kaggle.com/code/assafco/nyc-crime-vs-education-geovisualization-tutorial/notebook. The following additional datasets are directly sourced: https://www.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page; they cover GDP, land area, population, and police station locations. It's interesting to notice that webscraping several online sources was used to generate the secondary dataset. A thorough foundation for the study is ensured by this aggregation of multiple datasets, which allows for a multimodal examination of crime patterns and their relationship to socioeconomic aspects.

The second step, Data Preprocessing, Several critical procedures were conducted during the preprocessing and cleaning of the dataset related to crime in New York City in order to prepare the data for analysis. The dataset was carefully cleaned, with duplicate entries found and eliminated to guarantee data accuracy, avoid redundancy, and preserve the general integrity of studies by getting rid of potential biases and inaccuracies brought on by repeated inputs. In order to deal with missing values, a wide range of imputation techniques were used, including mean imputation for numerical features, ”Not Available” for categorical data, ”Zero” for particular numeric fields, and the addition of labels for columns pertaining to boroughs. Furthermore, the date and time columns were set to default values. These various imputation methods sought to maintain statistical properties and improve interpretability while establishing dataset consistency and eligibility for further studies.

Furthermore, by avoiding mistakes in numerical operations and enabling the correct processing of categorical and temporal data, the assignment of suitable data types to each column was crucial in guaranteeing accurate analysis. This action greatly improved the dataset's general dependability and interpretability. To ensure uniform scaling across features, standardization and normalizing procedures were applied in simultaneously. By encouraging convergence and reducing sensitivity to input feature scales, this not only mitigated the unwarranted influence of some factors but also improved the performance and stability of machine learning models. These scaling strategies strengthened the analytical framework even more and produced a dataset that was more reliable and strong for later investigations.

The third step, Data Transformation, The exploratory data analysis (EDA) of the New York crime dataset was a complex investigation of the subtle relationship between socioeconomic characteristics and crime patterns across the city's boroughs. The investigation began by examining temporal features, offering fascinating insights into how crime rates evolved over time, suggesting probable seasonality and long-term tendencies. Following that, geospatial analysis was used to map the distribution of crimes, effectively locating crime hot spots and regions with lower crime occurrence. A careful investigation of crime types shed light on the prevalence of various transgressions, identifying the most and least prevalent crimes and tracing their shifts throughout time. Demographic analysis added to the understanding by looking at the age, gender, and race of both victims and suspects, with the goal of identifying probable links to certain criminal behaviors. The impact of GDP, unemployment, and poverty levels on crime rates was investigated, revealing socioeconomic influences. Police response times and crime clearance rates were also examined, providing vital insights on law enforcement efficiency and the criminal justice system's effectiveness. Statistical summaries, data visualization, and correlation analyses were critical tools throughout this EDA journey, providing a solid platform for additional research and modeling.

Moreover, a machine learning model is trained on most of the dataset in this step by dividing it into 80% training and 20% testing. This ensures that the model generalizes effectively to new, unseen data. By using this technique, over fitting can be identified and models that work well in real-world situations can be chosen.

The code creates a 'ComplaintYear' column, converts dates to datetime format, generates dummy variables for criminal categories like 'FELONY,' 'MISDEMEANOR,' and 'VIOLATION,' and sets the Year and Borough names as the index for analysis. After adding the "ComplaintYear" column to a DataFrame, the data is aggregated for crime categories and location coordinates, and the data is grouped by year and borough. A CSV file is saved with the generated DataFrame. where 350 records out of 11 million records from 1950 to 2019 are flattened based on year and Borough names.

The fourth step, Data Modeling and Conversion, includes both model training and conversion. The prediction model makes use of Random Forest, Simple Linear Regression, and Ordinary Least Squares (OLS). OLS and SLR provide more straightforward and comprehensible insights into linear relationships, but Random Forest is selected because to its capacity to represent complex, non-linear relationships and the interplay between different socioeconomic elements that impact criminal activity. In order to investigate hotspots among the Boroughs, K-Nearest Neighbors (KNN) is applied for the classification model.

Feature engineering is used, together with a logarithmic adjustment of the population column, to improve the accuracy of OLS and Simple Linear Regression models. In addition to normalizing data, addressing skewed distributions, and lessening sensitivity to extreme values, this transformation makes the analysis more reliable and understandable.

Moreover, it is imperative to manage outliers. Extreme data points are identified and eliminated using statistical and visualization techniques like the Z-score and Interquartile Range (IQR). Through the reduction of anomaly effects on statistical measures and the improvement of modeling process quality overall, this approach guarantees more robust and trustworthy studies.

The fifth step, Evaluation and Results, Both Mean Squared Error (MSE) and R-square (R^2) are used to evaluate the performance of any machine learning prediction and classification model. Model fit is indicated by R^2, which offers insights into the percentage of variance explained; prediction accuracy is measured by MSE, which quantifies the squared difference between the predicted and actual values. After conducting experiments, the best machine learning prediction model is determined by comparing and visualizing the three models using Python. The classification approach additionally assesses crime hotspots with an emphasis on the highest level of criminality (FELONY). Driven by key feature importance results from the model summary, the analysis takes into account the proportionate match with population and socioeconomic characteristics. This makes sense since it guarantees a thorough assessment of the performance of the prediction and classification models.

4 Design Specification

In order to evaluate the importance of socioeconomic factors impacting crime rates in New York City, the hybrid machine learning framework architecture integrates prediction and categorization models. While KNN is used for classification to determine which boroughs have the highest concentration of police station hot spots and felonies, traditional approaches such as random forest, simple linear regression, and Ordinary Least Square regression model are applied for prediction. Socioeconomic elements such as GDP, poverty, housing price, unemployment, education, and income are all part of the architecture. Crime hot spots, model accuracy, crime trends over time, and feature importance analysis are all included in the model findings. Part 4.1 of the components includes a Hybrid Prediction model, while Section 4.2 delves deeper into criminal pattern analysis.

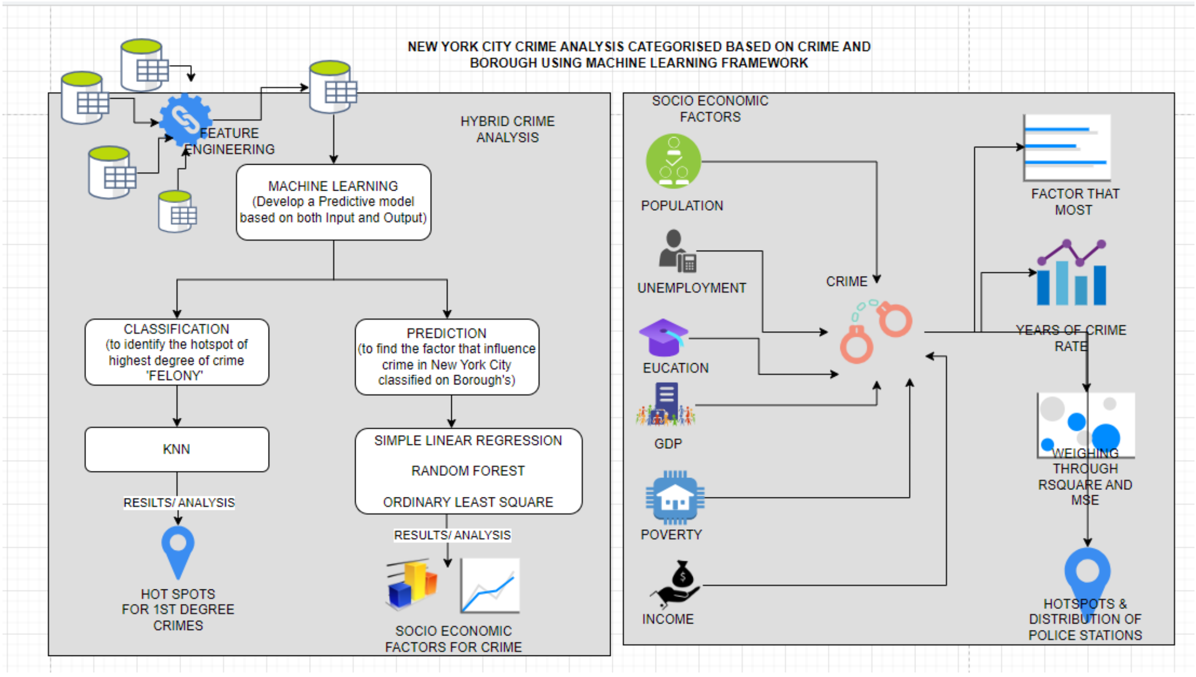

NEW YORK CITY CRIME ANALYSIS CATEGORISED BASED ON CRIME AND BOROUGH USING MACHINE LEARNING FRAMEWORK
FEATURE ENGINEERING
HYBRID CRIME ANALYSIS
MACHINE LEARNING (Develop a Predictive model based on both Input and Output)
CLASSIFICATION (to identify the hotspot of highest degree of crime 'FELONY'
PREDICTION (to find the factor that influence crime in New York City classified on Borough's)
KNN
SIMPLE LINEAR REGRESSION
RANDOM FOREST
ORDINARY LEAST SQUARE
RESILTS/ ANALYSIS
RESULTS/ ANALYSIS
HOT SPOTS FOR 1ST DEGREE CRIMES
SOCIO ECONOMIC FACTORS FOR CRIME
SOCIO ECONOMIC FACTORS
POPULATION
UNEMPLOYMENT
EUCATION
GDP
POVERTY
INCOME
CRIME
FACTOR THAT MOST
YEARS OF CRIME RATE
WEIGHING THROUGH RSQUARE AND MSE
HOTSPOTS & DISTRIBUTION OF POLICE STATIONS

Figure 2: Design Specification For Crime Prediction In New York City

4.1 Hybrid Crime Analysis

To thoroughly evaluate New York City's crime trends, the hybrid crime analysis framework combines prediction and classification models. The advantages of Simple Linear Regression, Random Forest, Ordinary Least Square Regression, and K-Nearest Neighbors (KNN) for classification are combined in this hybrid technique. Included in the design are the following elements:

4.1.1 Prediction Models

Random Forest:It is applied because it may identify intricate, nonlinear connections between different socioeconomic conditions and patterns of crime. ***Ordinary Least Square Regression (OLS)*** used to offer clear-cut explanations of the linear correlations between particular socioeconomic factors and crime rates. ***Simple Linear Regression (SLR)*** utilized for the analysis of specific socioeconomic characteristics and how they affect crime linearly.

4.1.2 Classification Model

K-Nearest Neighbors (KNN) developed with an emphasis on the most serious crimes, such as felonies, in order to pinpoint crime hotspots within boroughs.

4.1.3 Feature Engineering

Logarithmic Adjustment used to improve the OLS and SLR models' accuracy in the population column. ***Outlier Management*** Robust model performance is ensured by identifying and managing outliers using statistical techniques such as Z-score and Interquartile range.

4.1.4 Model Evaluation

Mean Squared Error (MSE) Quantifies the squared difference between expected and actual data to determine the accuracy of the forecast. ***R-square (R^2)***shows the percentage of variance that can be explained, rating the overall fit of the model.

4.2 Socioeconomic Factors Considered and Evaluation

The socioeconomic elements taken into account in the crime analysis are thoroughly examined in the design specification. When assessing their influence on trends in crime, the following elements are critical:

4.2.1 Socioeconomic Factors

GDP An economic indicator's relationship to crime rates was examined. ***Poverty Rate*** investigated to determine the connection to illegal activities. ***Housing Price Changes*** examined for possible impact on crimes involving property. ***Unemployment*** examined to see how it affects crime, particularly in light of the state of the economy. ***Education Rates*** investigated in relation to crime, taking into account how society's conduct is influenced by education. ***Income Variations*** conducted research to determine how income levels affect criminal behavior. ***Population Density*** taken into account because of its possible relationship to crime rates.

4.2.2 Evaluation Metrics

Model Accuracy guarantees that classifications and predictions are accurate. ***Feature Importance Analysis*** determines the proportional significance of every socioeconomic component in crime pattern prediction. ***Spatial Analysis*** used to map areas of high crime and evaluate regional trends.

The goal of the design specification is to develop a thorough framework that will not only properly forecast crime trends but also shed light on the socioeconomic variables impacting criminal activity in New York City.

5 Implementation

Using Python 3.12 and Visual Studio Code as the integrated development environment, the procedure of this thorough implementation begins with data cleansing and exploration. The dataset—which is most likely connected to crime statistics—is put into a Pandas DataFrame, and any missing values are filled in with the proper methods. The next step is to perform exploratory data analysis (EDA) in order to find patterns or trends and obtain an understanding of the structure of the dataset. To comprehend the connections between various variables, correlation analysis, visualizations, and descriptive statistics are used.

After the EDA stage, interaction terms, squared population values, and log transformations are created for the dataset through feature engineering. Consistent scales across features are ensured by applying standardization approaches as Min-Max scaling, Standard scaling, and Z-score normalization. Improving the performance of machine learning models requires this preprocessing step.

Following that, the analysis moves on to the model development phase, when KNN classification models, random forest regression, and linear regression are used. Metrics like R-squared and mean squared error are used to assess and train the models. Scatter plots showing expected versus actual values for each target variable are used to illustrate each model's performance.

The implementation, of particular note, compares two distinct regression models, namely random forest and linear regression, in order to illustrate the advantages and disadvantages of each. Additionally, a KNN classifier is used to predict crimes by utilizing geographic information such as latitude and longitude. The model's accuracy in predicting crime sites can be visually evaluated by examining the predictions presented on an interactive Folium map.

A thorough and reliable study is required, which drives the selection of models and approaches. When it comes to predicting performance and capturing non-linear patterns, random forest regression outperforms linear regression in terms of comprehending the links between characteristics and target variables. By using a KNN classifier for geographical prediction, the study gains a spatial dimension and becomes more applicable to real-world scenarios.

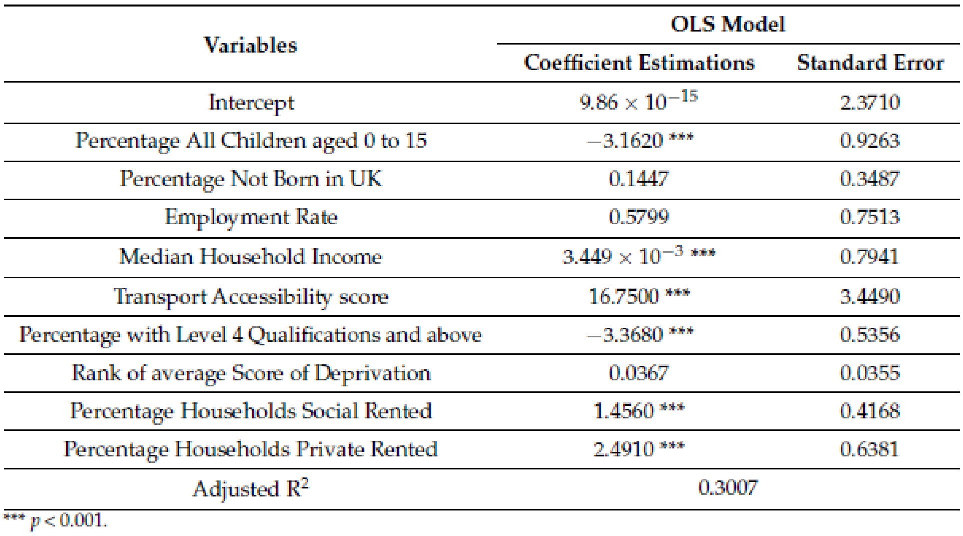
Using Python 3.12 and Visual Studio Code throughout the implementation guarantees that the newest tools and language features are compatible. Modular coding organization prioritizes readability and maintainability. To extract valuable insights and prediction capabilities from the crime dataset and contribute to a well-rounded and influential thesis, a thorough approach to data cleaning, exploratory data analysis, and model construction is employed.

6 Results and Discussion

The aim of this experiment is to identify socioeconomic factors that influence criminal activity by utilizing linear regression and random forest regression to create prediction models for crime outcomes. Furthermore, a K-nearest neighbors (KNN) classifier is used to predict regional crimes using spatial characteristics. Using measures such as mean squared error and R-squared for a comparison study, model performance is systematically assessed, offering insights into the advantages and disadvantages of each modeling strategy in comprehending and forecasting crime trends.

6.1 Experiment 1: Replication of State of Art

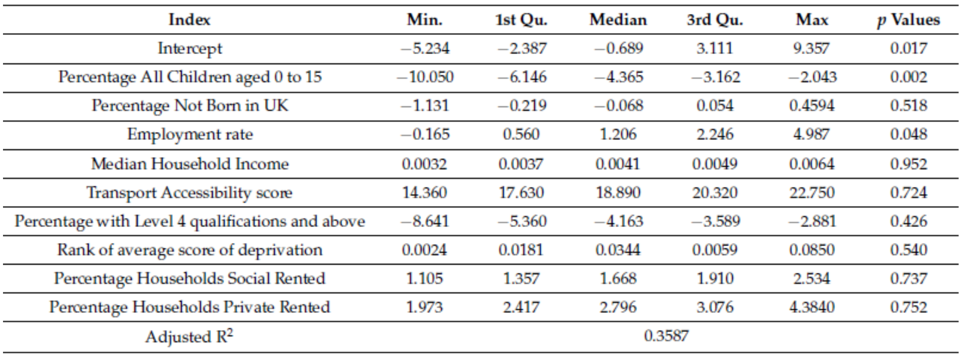
This study aims to reproduce the state-of-the-art analysis of crime rates in London Zhou et al. (2021) by using Geographically Weighted Regression (GWR) and Ordinary Least Squares (OLS) models. These models have produced insightful findings on the spatial patterns and socioeconomic element affects. The positive influence of transport accessibility scores and the complexity of factors like education levels, deprivation scores, and housing arrangements were among the important predictors of crime rates that the OLS model found. The spatial heterogeneity inherent in crime patterns may be oversimplified by the OLS model's (ref. Figure. 3) "one size fits all" approach. On the other hand, the spatially non-stationarity-aware GWR model fared better than OLS by displaying clear spatial differences in the impacts of important components. The GWR model showed, among other things, that the relationship between the percentage of children in London between the ages of 0 and 15 and employment rates affected crime rates differently in each district. This more complex understanding emphasizes how crucial spatial factors are to crime analysis, as seen by the GWR model's higher adjusted R-square.(ref. Figure. 4) These results can be used by law enforcement and policymakers to customize tactics and interventions for particular areas, taking into account the spatial dynamics seen in crime trends.



Variables	OLS Model	
	Coefficient Estimations	Standard Error
Intercept	9.86×10^{-15}	2.3710
Percentage All Children aged 0 to 15	-3.1620 ***	0.9263
Percentage Not Born in UK	0.1447	0.3487
Employment Rate	0.5799	0.7513
Median Household Income	3.449×10^{-3} ***	0.7941
Transport Accessibility score	16.7500 ***	3.4490
Percentage with Level 4 Qualifications and above	-3.3680 ***	0.5356
Rank of average Score of Deprivation	0.0367	0.0355
Percentage Households Social Rented	1.4560 ***	0.4168
Percentage Households Private Rented	2.4910 ***	0.6381
Adjusted R^2	0.3007	

*** $p < 0.001$.

Figure 3: Variation of coefficient estimations and adjusted R-square of OLS model



Index	Min.	1st Qu.	Median	3rd Qu.	Max	*p* Values
Intercept	−5.234	−2.387	−0.689	3.111	9.357	0.017
Percentage All Children aged 0 to 15	−10.050	−6.146	−4.365	−3.162	−2.043	0.002
Percentage Not Born in UK	−1.131	−0.219	−0.068	0.054	0.4594	0.518
Employment rate	−0.165	0.560	1.206	2.246	4.987	0.048
Median Household Income	0.0032	0.0037	0.0041	0.0049	0.0064	0.952
Transport Accessibility score	14.360	17.630	18.890	20.320	22.750	0.724
Percentage with Level 4 qualifications and above	−8.641	−5.360	−4.163	−3.589	−2.881	0.426
Rank of average score of deprivation	0.0024	0.0181	0.0344	0.0059	0.0850	0.540
Percentage Households Social Rented	1.105	1.357	1.668	1.910	2.534	0.737
Percentage Households Private Rented	1.973	2.417	2.796	3.076	4.3840	0.752
Adjusted R^2			0.3587			

Figure 4: Variation of coefficient estimations and adjusted R-square of GWR model.

6.2 Experiment 2: Exploratory Data Analysis of each Dataset

To glean important insights, a thorough analysis was done on every dataset. To comprehend time patterns, crime kinds, and spatial distributions, the crime dataset from https://data.cityofnewyork.us/ was carefully examined. Potential correlations were identified with the help of demographic context provided by population statistics. For spatial analysis to link crime occurrences to particular geographic areas, land area information was essential. In order to evaluate the impact of the economy on crime patterns, GDP data was also examined.

Comprehensive study was performed on the socio-economic datasets, which spanned 69 years, that were gathered through web scraping. We looked at factors that might have an effect on crime rates, including economic disparity, unemployment rates, and educational attainment. A detailed grasp of the intricate interactions between socioeconomic determinants and criminal outcomes is made possible by the thorough research conducted across datasets that serves as the basis for strong predictive modeling.

6.2.1 Dataset 01: PoliceStationsOfNewYorkCity.csv

The project started by carefully cleaning the "PoliceStationsOfNewYorkCity.csv" dataset, using Pandas to handle duplicates and missing information, and Seaborn to visually present the results. Due to the comprehensive statistics provided for "Precinct" and "Borough," the EDA phase highlighted the uniqueness of the values. A count plot representing the distribution of police stations throughout the boroughs was created using categorical analysis, and cross-tabulations were used to examine the connections between "Borough" and "Precinct." Using geometry data from a GeoDataFrame, geocoding police station locations allowed for inspection within precinct polygons through geospatial analysis driven by geopy and geopandas. GeoJSON data from NYC precincts improved the spatial context. The comprehensive methodology of the project guaranteed openness, bolstering significant findings and suggestions for the police station dataset.

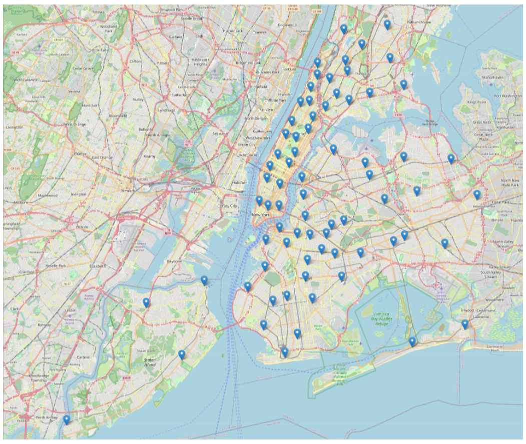


Figure 5: Map for the Distribution of Police Stations in New York City.

The GeoDataFrame, also known as gdf_nyc_precincts, is an essential dataset that serves as a basis for mapping and analyzing the precincts' geographic distribution throughout New York City. This spatial dataset gives scholars and analysts the fundamental data they need to carry out spatial queries, investigate spatial correlations (Ref. Figure 5), and gain insightful knowledge about the geographic dynamics of law enforcement throughout the city. Examining the figure directly reveals that Staten Island has the least concentration of police stations, with Brooklyn and Manhattan having the largest dispersion. This visual depiction provides a quick and clear understanding of the diverse law enforcement presence densities in the various New York City boroughs.

6.2.2 Dataset 02: PopulationAreaWithGDP.csv

A thorough examination of the "PopulationAreaWithGDP.csv" dataset was conducted throughout the thesis' implementation phase. The investigation started with graphics showing the population distribution by borough from the 2020 Census, providing a handy summary in the form of a sorted bar plot. A pie chart was used to depict how the land area was distributed among all the boroughs, with the goal of determining which one has the most land area. After the data were sorted in descending population order, a thorough comparison of land area and population was performed. By strategically utilizing Seaborn, a dual-axis bar plot was produced, offering a picture of each borough's land size and population.

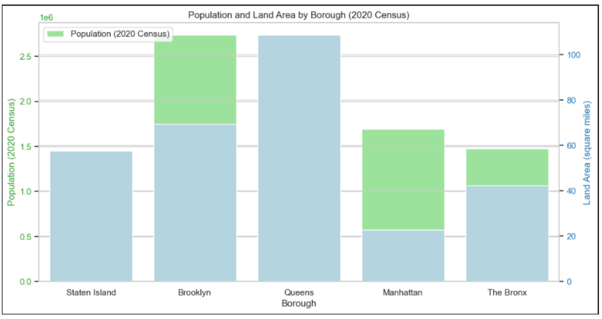


Figure 6: Dual Axis Bar-Plot Distribution of Population and Land Area.

The 2020 Census population and land area of New York City boroughs are displayed in a dual-axis bar plot (Ref. Figure 6) based on the "PopulationAreaWithGDP.csv" dataset. Queens holds the top spot with 36.2% of the total land area, followed by Manhattan (7.6%), Staten Island (19%), Brooklyn (23%), and The Bronx (14%). Notably, the population distribution does not exactly match the size of the population; Brooklyn has the largest population, followed by Queens, Manhattan, The Bronx, and Staten Island. The illustration skillfully draws attention to the discrepancy between land extent and population density, offering insightful information on the geographic and demographic makeup of every borough in New York City.

6.2.3 Dataset 03:CrimeNYC.csv

The "CrimeNYC.csv" does a detailed analysis of a portion of the large NYPD complaint dataset, which consists of 11 million records obtained between 1950 and 2019. The program carefully preprocesses the data, taking care of missing values, outliers, and temporal trends, using pandas and visualization tools. It visualizes linkages, crime hierarchies, and suspect-victim dynamics using Plotly Express. Regardless of the dataset's size, our multimodal analysis guarantees a thorough comprehension of crime patterns, categories, and linkages.

With Plotly, you can generate an interactive bar chart and line plot (Ref. Figure 7) that shows the annual crime rates for every borough in New York City. The data is filtered and processed to highlight trends from 2006 to 2018 after being taken from the 'CMPLNT_FR_DT' (complaint date) column. To improve visual clarity, each borough is represented by a different hue. The interactive display that results lets visitors examine changes in crime rates across the given years. Remarkably, Queens, The Bronx, Manhattan, and Brooklyn had the lowest number of complaints in 2017, while Staten Island had the most. There were fewer complaints altogether in the following year, 2018.

Sankey diagram (Ref. Figure 8) that makes use of Plotly, which shows the connections and patterns between the various categories in the crime data. Categories like victim age group (VIC_AGE_GROUP), victim race (VIC_RACE), and crime type (LAW_CAT_CD) are used to arrange the data. With each link denoting the flow of events between the designated categories, the Sankey diagram graphically depicts

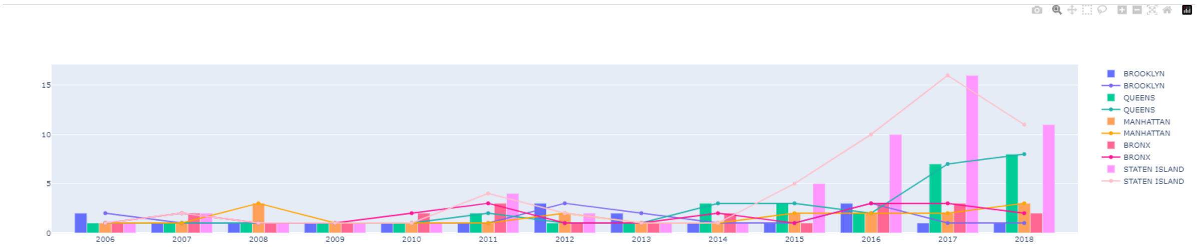


Figure 7: bar chart and line plot that shows the annual crime rates.

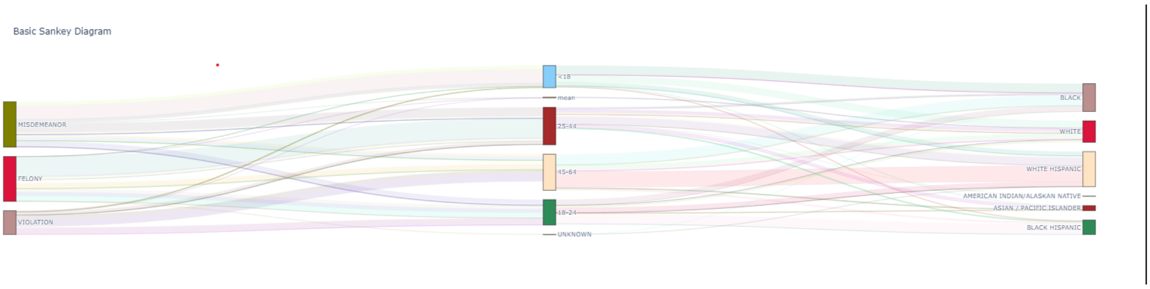


Figure 8: Sankey diagram connections and patterns.

the relationships between these categories. The graphic that results shows the relationships between various features of criminal episodes and offers insights into trends and interactions within the crime dataset. The color-coding of links makes the links easier to distinguish.

6.2.4 Dataset 04:CompleteDs.csv

"CompleteDs.csv," the dataset, contains socioeconomic data for the boroughs of New York City from 1950 to 2019. Data on the population, age groups, population density, unemployment rate, poverty rate, changes in housing prices, income distribution, and level of education are all included. With the use of multiple visualization approaches, the offered Python code analyzes and represents these socio-economic elements, providing insights into the dynamics and patterns that exist within the city's boroughs during the given timeframe. Sankey graph (Ref. Figure 9)is used to investigate and understand socioeconomic

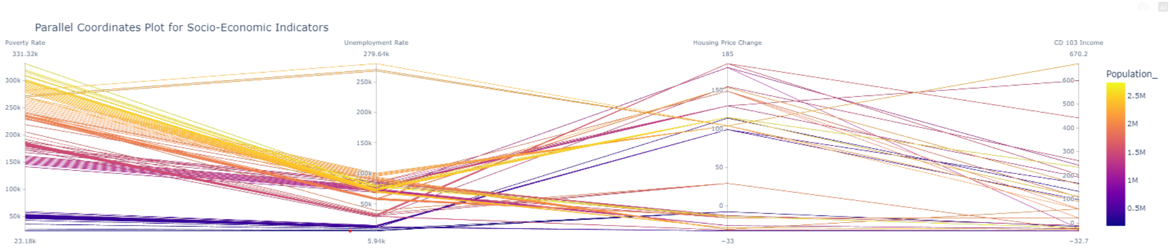


Figure 9: Sankey diagram.

parameters in the boroughs of New York City. The multidimensional linkages between four important indicators—the unemployment rate, the percentage of CD 103 income, the change in housing price, and the poverty rate—are highlighted by the Parallel Coordinates Plot, which was created using Plotly Express. Every borough is depicted by a unique line, with colors denoting population size variances. This map offers a thorough overview, making comparisons and pattern recognition easier. At the same time, the Seaborn Pair Plot uses scatter plots, histograms, and kernel density estimations to provide a thorough analysis of paired correlations between particular markers. The boroughs are distinguished by the color-coded markers, which allow for a more in-depth investigation of the socioeconomic dynamics both inside and across boroughs. When viewed as a whole, these visualizations help to provide a comprehensive overview of the dataset by highlighting complex relationships and patterns in the socioeconomic trends that vary among the boroughs of New York City.

6.2.5 Dataset 05:merged_dataFinal.csv

"merged_dataFinal.csv," the combined dataset, integrates socioeconomic indicators (from "CompleteDs.csv") and crime data (from "CrimeNYC.csv") for the boroughs of New York City. It is built by squaring 11 million records, using a 1950–2019 time frame. The dataset contains geographic coordinates, crime counts for FELONY, MISDEMEANOR, and VIOLATION, ComplaintYear, BORO_NM (borough names), and a number of socioeconomic variables, including population, age groups, poverty rate, unemployment rate, and education rate, among others. By combining the crime statistics for every borough in 1950, the code offers a thorough examination of how crime trends connect to socioeconomic variables during the designated period of time.

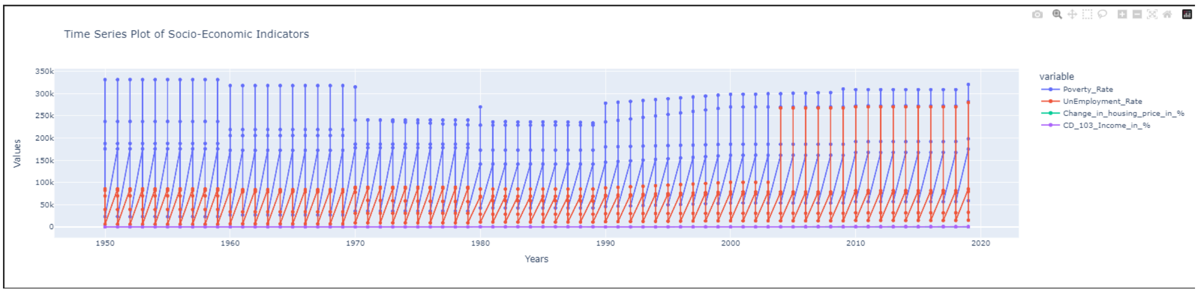


Figure 10: time series plot.

An interactive time series plot(Ref. Figure 10) that graphically depicts the variations in the chosen socioeconomic variables throughout the given time period is the end result. The rates of poverty, unemployment, changes in housing prices, and CD 103 income % over time will all be trended and fluctuated.

6.3 Experiment 3: Machine Learning Model Simple Linear Regression

The assumption of linearity between socio-economic indicators like GDP or population density and crime rates, as well as the model's interpretability and computing efficiency, led to the choice to use a basic linear regression model for crime prediction. By using advanced feature engineering and preprocessing approaches, the offered code improves predictive accuracy. This entails applying logarithmic transformations, interaction terms, Z-scores for outlier removal, and Min-Max and Standard scaling for feature standardization. Assessment measures and scatter plots demonstrate significant improvements, suggesting that further work is required to improve the model further or explore other models, especially given the 0.2 split constraints that have been noted.

6.3.1 Initial Model Performance

A Linear Regression model is used in this analysis to predict crime rates (Felony, MISDEMEANOR, and VIOLATION) based on socioeconomic characteristics. The goal variables are crime rates, and key aspects include demography and economic factors. With an R-squared value of 0.38 and a Mean Squared Error of roughly 9.56 million, the model, which was trained on a split dataset, performs moderately. With scatter plots, prediction accuracy is visually evaluated. Features such as 'Working_age_16_64,' 'Children_aged_0_15,' and 'CD_103_Income_in_%' have significant contributions, according to feature importance analysis. Positive correlations between these traits and crime rates are suggested by the model, highlighting possible impacts. But because of its limited explanatory capacity, more research using more sophisticated models should be taken into account in order to improve accuracy.

6.3.2 Feature Engineering

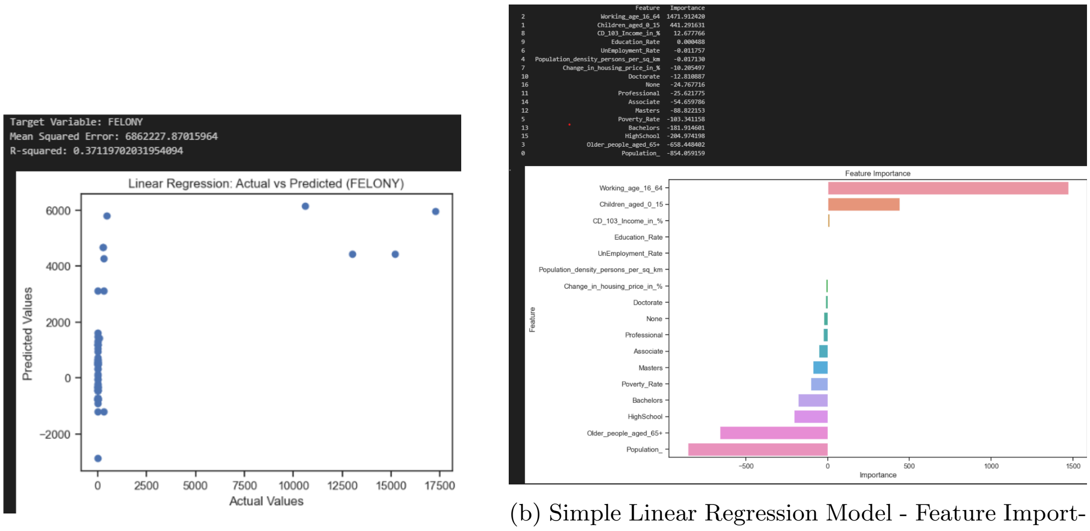
Novel variables are incorporated in the feature engineering process, such as ”SquaredPopulation,” which is acquired by squaring the population, and ”InteractionTerm,” which is computed by multiplying population density and housing price change. Additionally, 'Population_' undergoes a log transformation to produce 'LogPopulation.' With these improvements, we hope to better capture the complex interactions that exist between socioeconomic variables and criminal outcomes.

Z-scores are calculated for each characteristic, and outliers outside of a 3-standard deviation range are eliminated, in an effort to enhance the quality of the data. This measure guarantees that severe data points do not unnecessarily impact the model.

The chosen features are then normalized using two scaling approaches: conventional scaling and min-max scaling. While normal scaling yields features with a mean of 0 and a standard deviation of 1, min-max scaling transforms data to a predetermined range (often [0, 1]). By improving the stability and interpretability of the model, these scaling techniques help to produce socioeconomic factor-based crime predictions that are more accurate.

6.3.3 Model Performance Post Feature Engineering

Based on FELONY crime rates, the prediction model shows a modest level of accuracy with an R-squared (R^2) (ref. Figure. 11) value of approximately 0.37 and a Mean Squared Error (MSE) of roughly 6.86 million. These measures show how well the model can account for differences in FELONY crime rates according to particular socioeconomic characteristics. Both the positive and negative effects of specific variables on predictions are revealed by the feature importance analysis. Features such as 'Working_age_16_64' and 'Children_aged_0_15' have positive coefficients, indicating a positive link with FELONY rates. On the other hand, low coefficients indicate a negative correlation for variables such as 'Unemployment_Rate' and 'Poverty_Rate'. 'CD_103_Income_in_%' and 'Education_Rate,' two noteworthy contributors, highlight the impact of both income and education on the results of criminal activity.



```
Target Variable: FELONY
Mean Squared Error: 6862227.87015964
R-squared: 0.37119702031954094
```

(a) Simple Linear Regression Model.

```
                                Feature   Importance
2                     Working_age_16_64  1471.912420
1                     Children_aged_0_15   441.291631
8                      CD_103_Income_in_%    12.677766
9                        Education_Rate     0.000488
6                      UnEmployment_Rate    -0.011757
4   Population_density_persons_per_sq_km    -0.017130
7            Change_in_housing_price_in_%   -10.205497
10                             Doctorate   -12.810887
16                                  None   -24.767716
11                          Professional   -25.621775
14                             Associate   -54.659786
12                               Masters   -88.822153
5                           Poverty_Rate  -103.341158
13                             Bachelors  -181.914601
15                            HighSchool  -204.974198
3                   Older_people_aged_65+  -658.448402
0                            Population_  -854.059159
```

(b) Simple Linear Regression Model - Feature Importance.

Figure 11: Simple Linear Regression Model.

The findings offer insightful information on the socioeconomic variables affecting FELONY crime rates, and they can direct future research and development towards a more comprehensive understanding.

6.4 Experiment 4: Machine Learning Model Random Forest

Random forest regression models are preferred for crime prediction due to their ability to capture complex correlations between many socio-economic components and to account for non-linear patterns. It is superior than simple linear regression when handling complicated datasets with a variety of factors. Predictive accuracy and robustness to outliers are improved by the ensemble approach of the model. Notably, I performed hyperparameter tuning and increased the number of trees to reduce overfitting, guaranteeing a reliable and accurate crime prediction model.

6.4.1 Initial Model Performance

At 1.35 million and R-squared at 0.91, the Random Forest Regressor forecasts crime rates with remarkable accuracy using a sample of 100 trees. Still, exercise caution—especially when dealing with a large number of trees—despite Random Forest's built-in resistance to overfitting. Noise in training data may be captured by overfitting. Achieving a balance between generalizability and complexity requires careful tuning of hyperparameters and performance monitoring of the model. The importance table's highlighted important socioeconomic characteristics are what fuel the model's capacity for prediction.

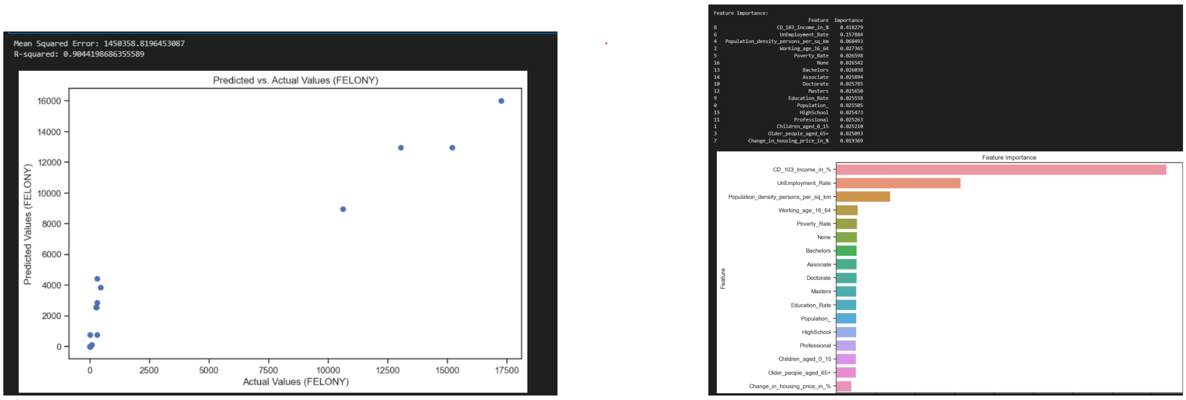
6.4.2 Feature Engineering

The decision to adjust hyperparameters and add more trees to the Random Forest is motivated by the need to prevent overfitting and achieve optimal generalization, even with a noteworthy 91% accuracy rate. The robustness of the model is increased during this process by adjusting important hyperparameters like the maximum depth of trees. 10,000 more trees have been added as part of an exploration to find the ideal balance. Visualizations of expected vs. actual values and feature importance shed light on the effectiveness of the model and the key elements involved in the prediction process.

6.4.3 Model Performance Post Feature Engineering

Feature importances are calculated and presented visually to reveal the predictive power of the model. Unexpectedly, 'CD_103_Income_in_%' is the feature that has the biggest impact, followed by 'UnEmployment_Rate' and 'Population_density_persons_per_sq_km.' These results are in line with expectations because employment and income levels are usually related to the dynamics of crime.

In the next investigation, we will increase the Random Forest model's tree count to 10,000. With an MSE of over 1.45 million and an R-squared value of roughly 0.90 (ref. Figure. 12), the modified model impressively maintains high performance. In predicting crime rates, wealth, unemployment, and population density continue to be the most important factors, as seen by the unchanged hierarchy of feature importances.



```
Mean Squared Error: 1450358.8196453087
R-squared: 0.9044198686355589
```

```
Feature Importance:
                                Feature  Importance
8                     CD_103_Income_in_%    0.418279
6                      UnEmployment_Rate    0.157884
4   Population_density_persons_per_sq_km    0.068493
2                      Working_age_16_64    0.027365
5                           Poverty_Rate    0.026598
16                                  None    0.026542
13                             Bachelors    0.026038
14                             Associate    0.025894
10                             Doctorate    0.025785
12                               Masters    0.025650
9                         Education_Rate    0.025558
0                            Population_    0.025505
15                            HighSchool    0.025473
11                          Professional    0.025263
1                       Children_aged_0_15    0.025210
3                   Older_people_aged_65+    0.025093
7              Change_in_housing_price_in_%    0.019369
```

(a) Random Forest Model.

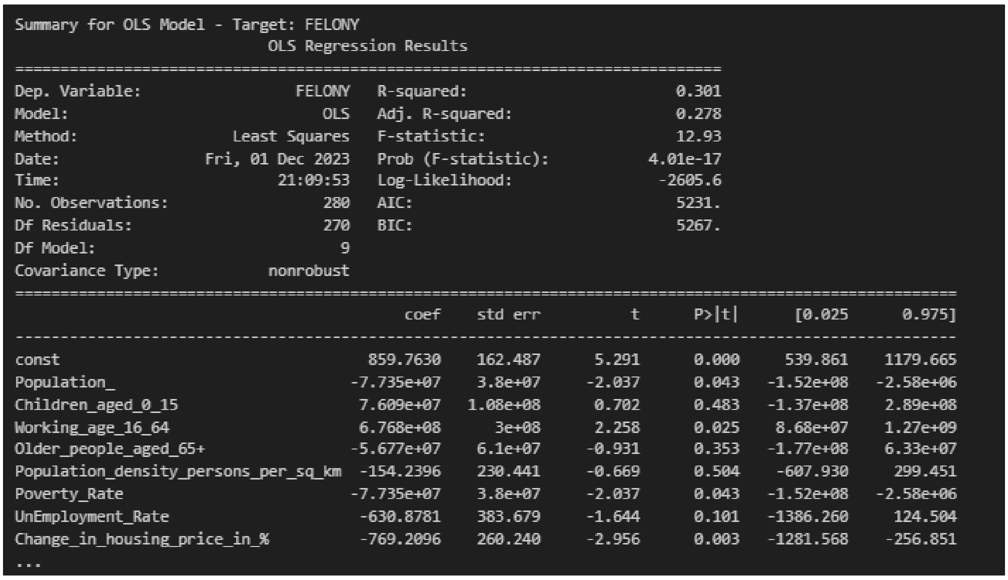
(b) Random Forest Model - Feature Importance.

Figure 12: Random Forest.

The correctness of the model is clearly demonstrated by visual examination using scatter plots that compare actual and anticipated values for each target variable. Put together, the code provides a comprehensive examination of the effectiveness of the Random Forest Regressor, highlighting the significance of features and demonstrating its robustness to changes in the tree count—a priceless tool for crime prediction based on socioeconomic variables.

6.5 Experiment 5: Machine Learning Model Ordinary Least Square Model

The statistical dependability, interpretability, and simplicity of Ordinary Least Squares (OLS) make them a popular choice for crime prediction. By minimizing the sum of squared differences between observed and anticipated values, OLS offers a clear understanding of how socioeconomic characteristics and crime rates are related. The coefficients obtained using OLS provide simple interpretations, which facilitate the comprehension of the influence of specific predictors. Enhancing the evaluation of predictor importance, OLS also offers crucial statistical inference instruments including confidence intervals and hypothesis testing. Although OLS is a useful baseline model, it is important to recognize its assumptions and, for a more accurate analysis, take into account other approaches in case these assumptions are not satisfied.



```
Summary for OLS Model - Target: FELONY
                            OLS Regression Results
==============================================================================
Dep. Variable:                 FELONY   R-squared:                       0.301
Model:                            OLS   Adj. R-squared:                  0.278
Method:                 Least Squares   F-statistic:                     12.93
Date:                Fri, 01 Dec 2023   Prob (F-statistic):           4.01e-17
Time:                        21:09:53   Log-Likelihood:                -2605.6
No. Observations:                 280   AIC:                             5231.
Df Residuals:                     270   BIC:                             5267.
Df Model:                           9
Covariance Type:            nonrobust
=========================================================================================================
                                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                 859.7630    162.487      5.291      0.000     539.861    1179.665
Population_                         -7.735e+07    3.8e+07     -2.037      0.043   -1.52e+08   -2.58e+06
Children_aged_0_15                   7.609e+07   1.08e+08      0.702      0.483   -1.37e+08    2.89e+08
Working_age_16_64                    6.768e+08      3e+08      2.258      0.025    8.68e+07    1.27e+09
Older_people_aged_65+               -5.677e+07    6.1e+07     -0.931      0.353   -1.77e+08    6.33e+07
Population_density_persons_per_sq_km -154.2396    230.441     -0.669      0.504    -607.930     299.451
Poverty_Rate                        -7.735e+07    3.8e+07     -2.037      0.043   -1.52e+08   -2.58e+06
UnEmployment_Rate                    -630.8781    383.679     -1.644      0.101   -1386.260     124.504
Change_in_housing_price_in_%         -769.2096    260.240     -2.956      0.003   -1281.568    -256.851
...
```

Figure 13: OLS Model.

The offered Python code predicts crime rates (FELONY, MISDEMEANOR, VIOLATION) based on socioeconomic characteristics by using Ordinary Least Squares (OLS) regression (ref. Figure. 13). Population density and poverty rate are important factors that contribute to the model's 30.1% variance explanation for FELONY. Improvement is shown by the MSE. Similar explanatory power is shown by the VIOLATION and MISDEMEANOR models, which are 29.6% and 31.9%, respectively. The socioeconomic status and demography of the population are important variables. The MSEs for VIOLATION are $10\hat{6}$, whereas those for FELONY and MISDEMEANOR are $10\hat{7}$. Accuracy and generalizability could be improved through iterative refinement, such as the addition of interaction terms or polynomial characteristics.

6.6 Experiment 6: Machine Learning Model KNN for Hot-Spot Analysis

Although K-Nearest Neighbors (KNN) has an inherent capacity to handle spatial data and identify patterns, it is the method of choice for hotspot analysis and crime category classification. Because it doesn't impose strong assumptions, its non-parametric character makes it suited for a variety of difficult crime datasets. While KNN's simplicity and lack of a separate training phase make it successful for dynamic crime datasets, its ability to adapt to local patterns makes it useful for hotspot identification. The ease of use of KNN in identifying geographical nuances and classifying crimes according to geographic characteristics is improved by its simple implementation.

An interactive map is created using the Folium library and a K-Nearest Neighbors (KNN) (ref. Figure. 14) classifier to show the locations of real and anticipated crimes in New York City. The borough names and a variety of spatial characteristics, such as latitude, longitude, population density, poverty rate, and unemployment rate, are chosen and ready for training. The dataset is used to train the KNN classifier, and the resultant map, which is centered on New York City and has an initial zoom level of 11, shows the locations of actual crimes in green icons inside one MarkerCluster and expected sites in red icons within another. This graphic illustrates that the Bronx and Queens have the highest rates of felony crime, respectively.

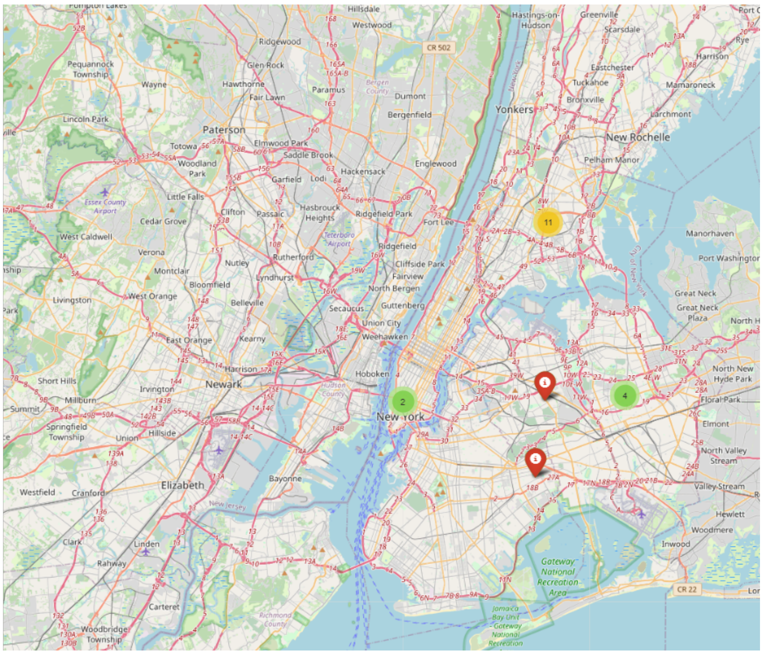


Figure 14: Hotspots Using KNN.

7 Evaluation

By carefully designing features and performing preprocessing operations, the offered code improves the performance of the linear regression model. Together with the addition of new features like "Squared Population," this also involves the addition of interaction phrases. The code also applies important preprocessing procedures for the data, including log transformation, Z-score-based outlier elimination, and scaling with the Min-Max and Standard Scaling approaches.

The linear regression model is then thoroughly assessed for crime prediction across several categories (FELONY, MISDEMEANOR, VIOLATION) in the wake of these improvements. The R-squared values indicate that even with the improvements, there may still be difficulties for the linear model to completely capture the complexities of the underlying relationships.

A thorough examination that contrasts the two models' performance indicators and visualizations is advised when examining Random Forest's potential as an alternative. Whether Random Forest performs better than the revised linear regression model will be determined by using a comparison method, given its intrinsic capacity to handle complex interactions. Ongoing improvement still depends on constant model tuning, refining, and investigation of different strategies.

8 Conclusion and Future Work

The aim of the research is to provide useful data regarding crime trends in order to enhance the safety and well-being of New York City. This makes it feasible to develop evidence-based policy and implement targeted tactics to deal with the root causes of crime, which improves the effectiveness of law enforcement activities. The research proposed a machine learning architecture that combines a classification model to analyze crime hotspots and a prediction model to identify the socioeconomic factors that contribute to crime. The project aims to promote transparency and data-driven decision-making within law enforcement institutions, with the goal of reducing crime rates and enhancing the quality of life for both residents and visitors to the multicultural metropolis.

This research reveals a significant socioeconomic component that has contributed to the growth in crime rates during the 69-year data set. Spatial analysis focuses on boroughs that have been identified as having committed felonies, or big crimes, by utilizing three different models to identify factors with a high degree of accuracy. The Random Forest model outperforms the OLS and Simple Linear models with an incredible 90% accuracy in the findings. The accuracy of the other models is significantly lower. It appears that income is the primary factor increasing crime rates, with population density and unemployment coming in second and third, respectively.

This research can potentially enhance the prediction tools that foresee negative patterns in police behavior. This work can be improved by However, one of the study's drawbacks must be acknowledged because it only included a sample of the 67 million records. The use of cutting-edge technologies is essential for the development of the crime prediction framework in next studies. It may be possible to optimize data processing, storage, and accessibility by incorporating cloud-based technologies. This enables easy incorporation of the most recent data for more accurate projections and real-time analysis. Deep learning and ensemble techniques are two of the newest machine learning algorithms that can be used to increase the prediction power of the model. Examining state-of-the-art data sources such as Internet of Things (IoT) devices and social media feeds could enhance the framework's accuracy by adding real-time socioeconomic information to the dataset. Additionally, the geographical analysis component may be improved by incorporating geospatial technologies, such as GIS mapping or spatial analytics, which would enable more accurate identification of high-risk areas and help targeted crime prevention programs.The adoption of these advancements will ensure that the crime prediction framework is future-proof, ensuring its effectiveness and usefulness in the dynamic domains of law enforcement and crime prevention.

References

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharsky, S., Benjamin, D. et al. (2020), 'A consensus-based transparency checklist', *Nature human behaviour* **4**(1), 4–6.

Bellitto, M. & Coccia, M. (2018), 'Interrelationships between violent crime, demographic and socioeconomic factors: A preliminary analysis between central-northern european countries and mediterranean countries', *Journal of Economic and Social Thought* **5**, 230–246.

Bursik Jr., R. J. (1988), 'Social disorganization and theories of crime and delinquency: Problems and prospects', *Criminology* **26**(4), 519–552.

Cahill, M. E. & Mulligan, G. F. (2003), 'The determinants of crime in tucson, arizona', *Urban Geography* **24**, 582–610.

Cliff, B. (2018), *Irish Crime Fiction*, Springer.

Fazel, S., Zetterqvist, J., Larsson, H., Långström, N. & Lichtenstein, P. (2014), 'Antipsychotics, mood stabilisers, and risk of violent crime', *The Lancet* **384**, 1206–1214.

Foody, M., Murphy, H., Downes, P. & O'Higgins Norman, J. (2018), 'Anti-bullying procedures for schools in ireland: principals' responses and perceptions', *Pastoral Care in Education* **36**(2), 126–140.

Goh, L. T. & Law, S. H. (2023), 'The crime rate of five latin american countries: Does income inequality matter?', *International Review of Economics & Finance* **86**, 745–763.

He, L., Páez, A., Liu, D. & Jiang, S. (2015), 'Temporal stability of model parameters in crime rate analysis: An empirical examination', *Applied Geography* **58**, 141–152.

Kenter, J. O., Raymond, C. M., Van Riper, C. J., Azzopardi, E., Brear, M. R., Calcagni, F. et al. (2019), 'Loving the mess: navigating diversity and conflict in social values for sustainability', *Sustainability Science* **14**(6), 1439–1461.

Kim, S., Joshi, P., Kalsi, P. S. & Taheri, P. (2018), Crime analysis through machine learning, *in* '2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)', IEEE, pp. 1–7.

Lightowlers, C., Pina-Sánchez, J. & McLaughlin, F. (2023), 'The role of deprivation and alcohol availability in shaping trends in violent crime', *European Journal of Criminology* **20**, 738–757.

Messer, L. C., Kaufman, J. S., Dole, N., Savitz, D. A. & Laraia, B. A. (2006), 'Neighborhood crime, deprivation, and preterm birth', *Annals of Epidemiology* **16**, 455–462.

Oyelade, A. O. (2019), 'Determinants of crime in nigeria from economic and socioeconomic perspectives: A macro-level analysis', *International Journal of Health Economics and Policy* **4**, 20–28.

Porter, J. R. & Purser, C. W. (2010), 'Social disorganization, marriage, and reported crime: A spatial econometrics examination of family formation and criminal offending', *Journal of Criminal Justice* **38**, 942–950.

Raphael, S. & Winter-Ebmer, R. (2001), 'Identifying the effect of unemployment on crime', *The Journal of Law and Economics* **44**, 259–283.

Roth, R. E., Ross, K. S., Finch, B. G., Luo, W. & MacEachren, A. M. (2010), A user-centered approach for designing and developing spatiotemporal crime analysis tools, *in* 'GIScience', Vol. 15, Zurich, Switzerland.

Sampson, R. J. (1985), 'Race and criminal violence: A demographically disaggregated analysis of urban homicide', *Crime Delinquency* **31**(1), 47–82.

Shaw, C. R. & McKay, H. D. (1942), *Juvenile Delinquency and Urban Areas: A Study of Rates of Delinquents in Relation to Differential Characteristics of Local Communities in American Cities*, University of Chicago Press, Chicago.

Smith, L. N. (2017), Cyclical learning rates for training neural networks, *in* '2017 IEEE winter conference on applications of computer vision (WACV)', IEEE, pp. 464–472.

Surprenant, C. & Brennan, J. (2019), *Injustice for all: How financial incentives corrupted and can fix the US criminal justice system*, Routledge.

Wilkinson, R., Pickett, K. & Scott Cato, M. (2009), *The Spirit Level: Why More Equal Societies Almost Always Do Better*, Allen Lane, London.

Zhou, Y., Wang, F. & Zhou, S. (2021), 'The spatial patterns of the crime rate in london and its socio-economic influence factors', *Social Sciences* **12**(6), 1340.