Socioeconomic Attributes and Crime: Is Crime about the Economy?

Aojie Ju, Junan Mao, Cheng Wu, Dichuan Zheng

December, 2024

1 Introduction

This analysis investigates the influence of socioeconomic factors on crime patterns in San Francisco from 2018 to the present. By integrating spatially precise crime data with rich economic and demographic metrics from the ACS 5-year estimates (2017–2022), the study aims to identify significant predictors of crime and evaluate whether economic changes impact the frequency and type of criminal incidents.

1.1 Key Goals

1. Examine the role of socioeconomic dynamics—such as poverty, income inequality, and housing values—in shaping crime levels.
2. Assess the predictive capacity of economic, demographic, and transportation attributes for violent, property, and public order crimes.
3. Distinguish between within-tract and across-tract variations to pinpoint localized patterns.

2 Data Overview

2.1 Crime Data (2018–Present)

- **Source:** San Francisco's open data portal.
- **Records:** 913732 incidents categorized into violent crimes, property crimes, and public order offenses.
- **Processing:** Spatially joined to census tracts using geographic coordinates for precise regional analysis.

2.2 ACS 5-Year Estimates (2017–2022)

- **Variables:** Median household income, poverty rate, Gini index (income inequality), unemployment, home values, and demographic proportions (e.g., White, Black, Asian, and Hispanic populations).
- **Derived Metrics:** Rates of public transit usage, educational attainment, and year-over-year percentage changes in key socioeconomic indicators.

2.3 Spatial Integration

- Census tract boundaries were sourced from San Francisco shapefiles.
- Incident-level crime data was merged with ACS variables based on tract-year alignments to produce a comprehensive dataset.

3 Processing Steps

Crime data (2018–present) and socioeconomic data (ACS 2017–2022) were retrieved and cleaned, removing invalid records and imputing missing values. Key metrics, such as poverty and unemployment rates, educational attainment, and year-over-year changes, were calculated. Crime incidents were spatially linked to census tracts using geographic coordinates, and crimes were grouped into violent, property, and public order categories. The data was split into training (70%) and testing (30%) subsets, scaled for continuous variables, encoded for categorical ones, and aggregated at the tract-year level for regression models, ensuring completeness for longitudinal analysis.

4 Exploratory Data Analysis

4.1 Overview of Crime Data

- **Crime Incidents:** The dataset includes records from 2018 to the present, spanning various types of crimes categorized into:
  - **Property Crime:** Includes burglary, theft, and vandalism.
  - **Violent Crime:** Includes assault, robbery, and homicide.
  - **Public Order and Other Crime:** Includes drug offenses, disorderly conduct, and traffic violations.
- **Crime Dataset Column Descriptions:**To provide a clear understanding of the dataset used in this analysis, the following table outlines the key columns and their respective descriptions. Each column captures essential details related to crime incidents.

Table 1: Description of Incident Report Data Columns

Column Name	**Description**
Incident Date	The date the incident occurred.
Incident Year	The year the incident occurred.
Incident Day of Week	The day of week the incident occurred.
Incident ID	Generated identifier for incident reports.
Filed Online	Indicates whether the police report was filed through an online reporting system designed for non-emergency incidents.
Incident Category	A classification provided by the Crime Analysis Unit of the Police Department for organizing and reporting incidents.

Column Name	Description
Incident Subcategory	A detailed classification that further specifies the categories defined in the Incident Category.
Resolution	The resolution of the incident at the time of the report.
Police District	The Police District where the incident occurred.
Latitude	The latitude coordinate in WGS84.
Longitude	The longitude coordinate in WGS84.
Point	Geolocation in OGC WKT format.

4.2 Socioeconomic Indicators

- **Variables of Interest:** Median household income, poverty rate, unemployment rate, educational attainment (Bachelor's degree rate), and demographic proportions (e.g., racial composition).

4.3 Data Visualizations

Crime Distribution by Category

The analysis reveals that the most common type of crime in San Francisco is larceny theft. Within the larceny theft category, the subcategory with the highest frequency is "Larceny - From Vehicle," followed by other subcategories such as general theft and shoplifting.

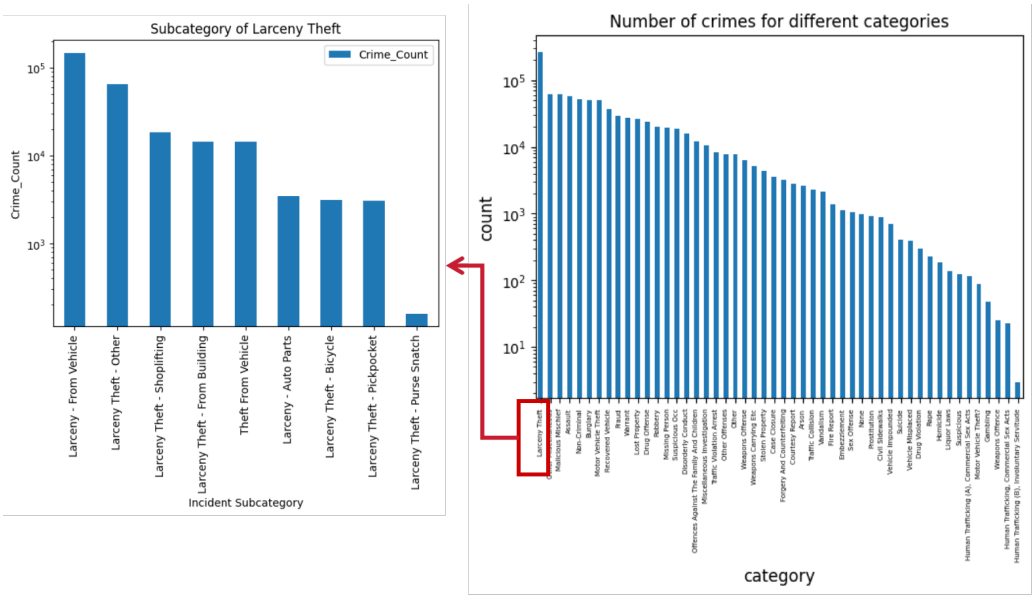

Figure 1: Crime Distribution by Category

Crime Distribution by District

The distribution of crime by police districts indicates that the Central District has the highest number of reported crimes. Within the Central District, the top two types of crime are larceny theft and malicious mischief, which are depicted in the pie chart.

Monthly Crime Trends

The temporal trend of total crime counts and larceny theft incidents reveals significant fluctuations over time. Notably, during the first half of 2020, there was a sharp decline in both total crimes and larceny theft. After mid-2020, crime counts gradually increased, displaying seasonal variations but remaining below pre-2020 levels. Both total crimes and larceny theft

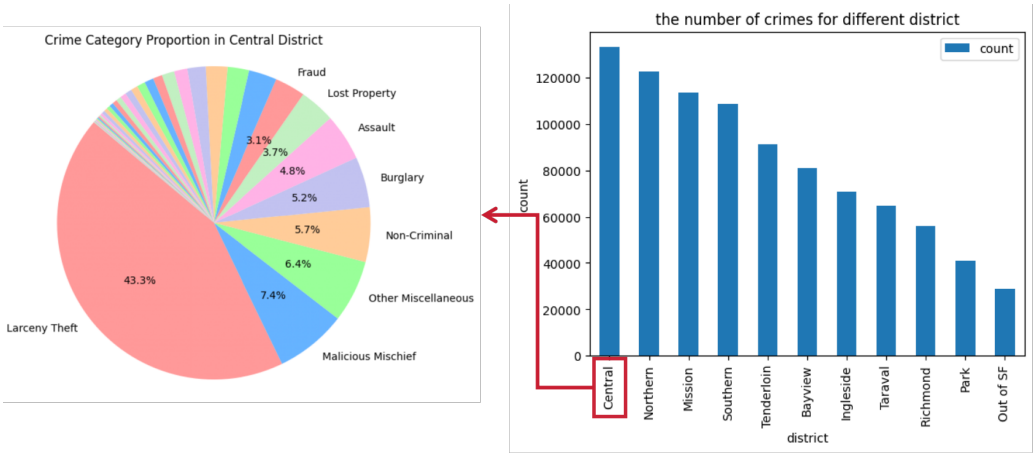

Figure 2: Crime Distribution by District

exhibit similar patterns, with peaks and troughs occurring at consistent intervals, highlighting a strong correlation between these two metrics.

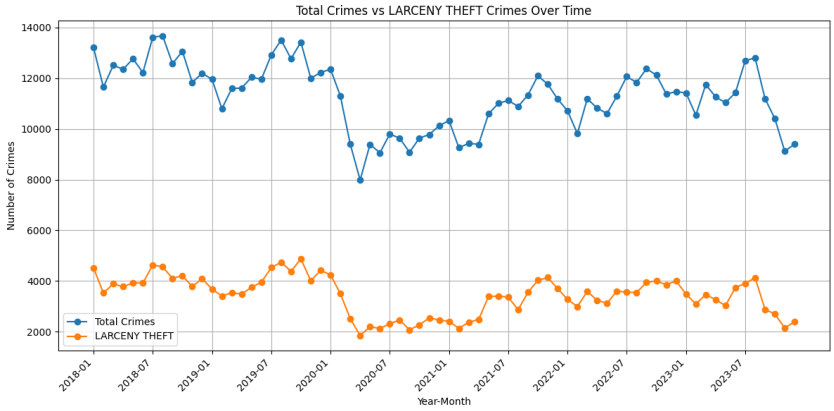

Figure 3: Monthly Crime Trends

Key Insights from EDA

Larceny theft is the most common crime in San Francisco, with "Larceny - From Vehicle" being the leading subcategory. The Central District reports the highest number of crimes, primarily larceny theft and malicious mischief. Crime counts dropped sharply in early 2020 but gradually increased afterward. Socioeconomic factors, such as income, poverty, and education, play a role in shaping crime patterns.

5 Machine Learning Analysis

5.1 Introduction of the Process

Our machine learning analysis follows a structured workflow, including the following steps:

1. **Data Preprocessing:** Cleaning and preparing the raw data, including handling missing values, outliers, and feature scaling, to ensure the data is suitable for modeling.

2. **Single Model Development and Comparison:** Training individual models, such as linear regression and random forests, to establish baseline performance. After modeling, we use accuracy as the evaluation metric to compare models.

3. **Optimization and Ensemble Modeling:** Leveraging Optuna for hyperparameter tuning and constructing ensemble models to improve prediction accuracy.

4. **Model Weight Analysis:** Analyzing the weights of individual models in the ensemble to understand their contributions to the final prediction.

5.2 Data Preprocessing

Our machine learning analysis follows a structured workflow, including the following steps:

1. **Data Preprocessing:** The raw data was processed to be suitable for modeling. The preprocessing steps included:

    - Removing missing values (`dropna`) to ensure a clean dataset for analysis.
    - Converting the `Filed Online` into a binary variable called `Emergency`, with values `True` or `False`.
    - Transforming the day of the week into a numeric variable named `Day Numeric`, with values ranging from 1 to 7 referring to Monday to Sunday.
    - Encoding the `Season` variable into numeric values from 1 to 4, representing Spring, Summer, Fall, and Winter respectively.
    - Encoding categorical variables such as `Incident Category` and `Police District` into numeric representations using label encoding.
    - Categorizing the time of the incident into a new variable named `Time Category`, with values from 1 to 3, representing morning, afternoon, and evening respectively.
    - Resolved, marked as 0,1 referring to unresolved and resolved case. It's the target variable.

5.3 Single Model Development

In this stage, we trained multiple machine learning models to establish baseline performance and gain insights into their predictive capabilities. The models used included:

- **Random Forest:** A robust ensemble-based model that is well-suited for handling both numerical and categorical data.
- **Logistic Regression:** A simple yet effective linear model for binary classification tasks.
- **Gradient Boosting Trees:** An ensemble method that builds models sequentially to minimize prediction errors.
- **Naive Bayes:** A probabilistic model based on Bayes' theorem, particularly effective for text or categorical data.
- **LightGBM:** A fast and efficient gradient boosting framework optimized for large datasets.
- **XGBoost:** A popular gradient boosting library known for its performance and scalability.

Each model was trained using the preprocessed dataset, and their accuracy was evaluated on a test set. The comparison of accuracy across these models helped identify their performance.

When selecting models, we considered a balance between simplicity and complexity to achieve both computational efficiency and performance. Logistic Regression and Naive Bayes

were chosen as simple models due to their computational efficiency and ability to provide interpretable results. In contrast, Random Forest, Gradient Boosting Trees, LightGBM, and XGBoost were included as ensemble models with better predictive performance, but with a higher risk of overfitting.

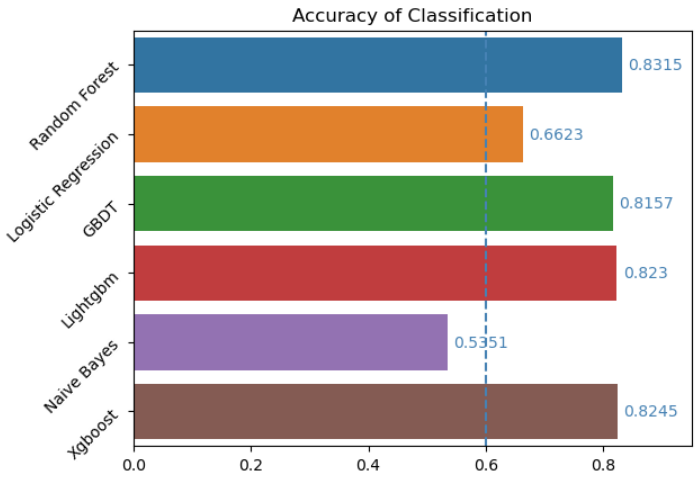

Figure 4: Comparison of Modeling Performance by Accuracy

The bar graph illustrates the classification accuracy of six machine learning models. Overall, ensemble models such as Random Forest, LightGBM, XGBoost, and GBT achieved significantly higher accuracy, all exceeding 80 percent, showing their effectiveness in capturing complex patterns within the data. In contrast, simpler models like Logistic Regression and Naive Bayes exhibited lower accuracy, indicating their limitations in handling complex data.

According to the result of feature importance of the data, ensemble models consistently ranked the encoded Incident Category as the most important feature, while Logistic Regression identified Emergency as the top predictor. Notably, Emergency ranked second across all ensemble models, suggesting Emergency is also an important predictor.

5.4 Optimization and Ensemble Modeling

In this section, we employed a two-step approach process: building a meta-model and optimizing hyperparameters for each sub-model.

First, we constructed a meta-model, also known as a stacked ensemble, by using the prediction probabilities from each base model(except logistic regression) as input features. These probabilities were fed into a **Logistic Regression** model, which served as the meta-model to combine and weigh the predictions from individual models. This approach allows the strengths of multiple models to complement each other, thereby improving overall prediction accuracy.

To further refine the performance of the base models, we used **Optuma**, a powerful hyperparameter optimization framework. Optuna performs an efficient search over the hyperparameter space by employing advanced techniques like Bayesian optimization and pruning. This enabled us to identify the optimal hyperparameters for each sub-model, ensuring that they performed at their best.

At the end of the optimization process, we obtained not only the best hyperparameters for each sub-model but also the relative weights assigned to each model in the meta-model. These weights reflect the contribution of each sub-model to the final ensemble, providing insights into their predictive power and reliability.

The final weight results are surprising: Naive Bayes received the highest weight despite it being the weakest among these models. In the ensemble model, the simplest model, Random

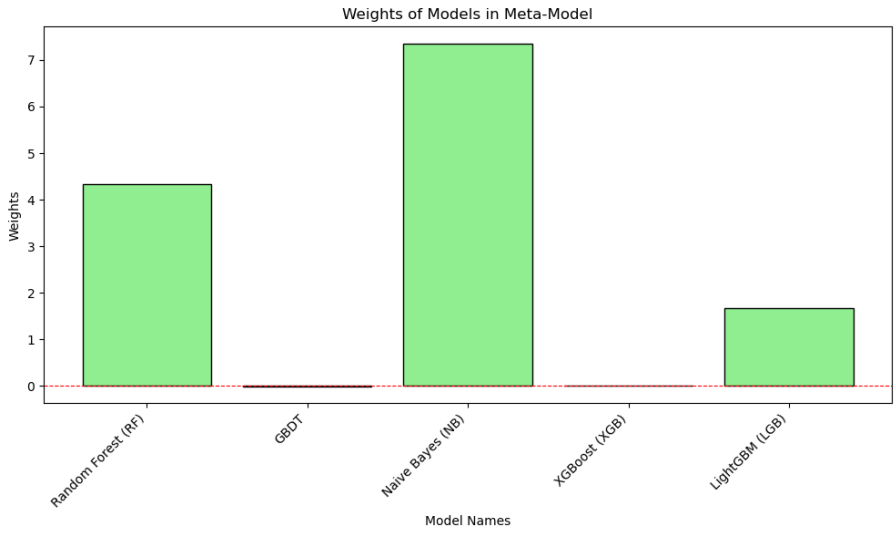

Figure 5: Weight of each Sub-model

Forest, achieved the highest weight. Based on our analysis, the meta-model's complexity may cause a high risk of overfitting. To mitigate this risk, higher weights were assigned to relatively simpler models.

According to the result of feature importance, Encoder Incident Category is considered as the most important feature by most models, indicating its significant contribution to whether the case by resolved.

6 Time Series Analysis

6.1 Introduction of Process

This section focuses on the time series analysis of the dataset, which involved the following key steps:

1. **Data Preprocessing:** We performed comprehensive data preprocessing, including feature engineering to create relevant variables, splitting the dataset into training (pre-2024) and testing (2024) subsets, and conducting a time series analysis of the target variable to understand its patterns.

2. **Modeling:** We employed three modeling approaches to capture the temporal dynamics:

   - **ARIMAX:** A time series model incorporating explanatory variables to improve prediction accuracy.
   - **SARIMAX:** An extension of ARIMAX that accounts for seasonality in the data.
   - **Linear Regression:** A baseline model to evaluate feature importance and temporal trends.

3. **Feature Subset Selection:** To enhance model performance, we applied a backward stepwise selection method for each model to identify the optimal feature subset for prediction.

4. **Prediction and Evaluation:** The best-performing model was used to make predictions on the testing dataset. The predictions were compared with the real data to evaluate model accuracy.

6.2 Data Preprocessing

1. **Daily Aggregation:** Using a 'groupby' operation, we aggregated the data by each day, extracting the following metrics for every date:

- **Resolve Rate:** The percentage of cases resolved on that day.
- **Emergency Rate:** The percentage of emergency cases among all cases for that day.
- **Time Category Rates:** The proportion of cases in each of the three time categories (morning, afternoon, evening).
- **Day of the Week:** Represented numerically as values from 1 (Monday) to 7 (Sunday).
- **Season:** Represented numerically as values from 1 to 4 (Spring, Summer, Fall, Winter).
- **Most Happen Case:** The category of cases with the highest occurrence on that day.
- **Most Happen District:** The district with the highest number of cases on that day.

2. **Target Variable:** We selected the **Resolve Rate** as the target variable for our time series analysis.

3. **Dataset Splitting:** The data was split into training and testing datasets based on the year. Data from 2018 to 2023 was used as the **training dataset**, while data from 2024 was reserved as the **testing dataset**. This ensures that the model is trained on historical data and tested on the most recent, unseen data, maintaining the integrity of the time series structure.

4. **Time Series Pre-analysis** Before Modeling, we checked the ACF and PACF plot of the model to decide the parameters of ARIMAX and SARIMAX models, also, ADF test was used to test whether the target variable was stationary.

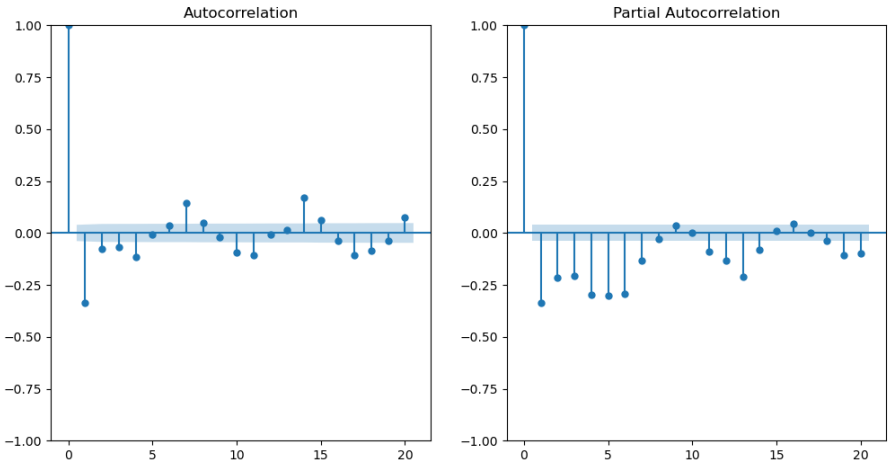

Figure 6: ACF and PACF Plot of Target Variable

Both the ACF and PACF plots cut off at lag 1, indicating AR(1) and MA(1) is suitable for the data. The ADF test shows the target variable is stationary therefore we do not have to do our transformation to make it stationary.

6.3 Modeling and Feature Selection

In this section, we employed ARIMAX, SARIMAX, and Linear Regression models to analyze and predict the resolve rate. The modeling process is described as follows: In this section,

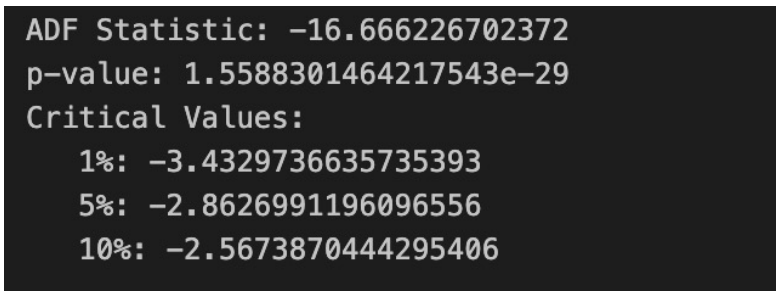

Figure 7: ADF Test Result

we employed ARIMAX, SARIMAX, and Linear Regression models to predict the resolve rate. Given the complexity of our data, we first chose ARIMAX, which incorporates both temporal dependencies and explanatory variables for more accurate predictions. Considering the periodic nature of criminal activity, we further introduced the SARIMAX model with a weekly cycle (period = 7) to account for seasonality. Lastly, we included a Linear Regression model as a baseline for comparison. To optimize each model, we applied backward selection using the Bayesian Information Criterion (BIC) as the evaluation metric, iteratively removing the variable with the largest p-value until the BIC no longer decreased. This process allowed us to identify the best-performing model and the most significant features for prediction.

Model	Trian MSE	Test MSE
ARIMAX	0.00144	0.00228
SARIMAX	0.00148	0.00415
Linear Regression	0.00220	0.00395

Figure 8: Model Comparison

The result of model comparison showcases that the ARIMAX model obtained the best test MSE, and linear regression had the highest MSE. ALL there models had greater test MSE than Train MSE, indicating they may have problems of overfitting.

Model	Train MSE	Test MSE	Variables Eliminated
ARIMAX	0.00142	0.00147	Time Cate 1,Time Cate 2, Season
SARIMAX	0.00144	0.00207	Time Cate 1,Time Cate 2, Season
Linear Regression	0.0022	0.00395	Most Happen Case,Day Numeric

Figure 9: Result of Backwards Step

Backward selection significantly improved the performance of ARIMAX and SARIMAX, reducing the Test MSE and mitigating overfitting risks. However, Linear Regression showed no improvement, likely due to its insensitivity to redundant features. Apart from Linear Regression, the overfitting risk was effectively reduced across the other models.

6.4 Prediction and Evaluation

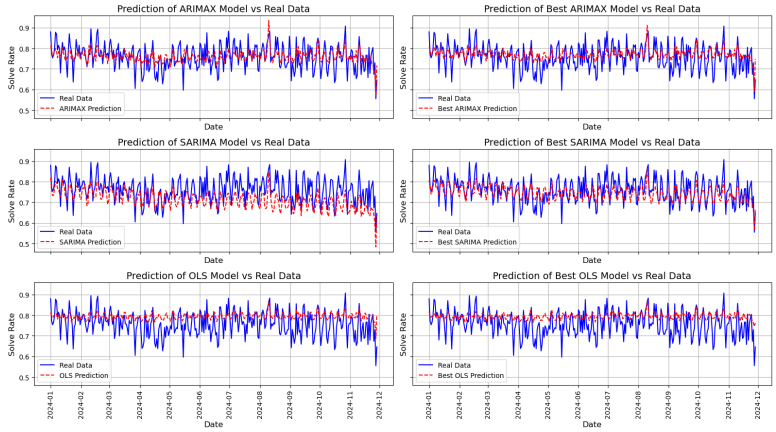

Figure 10: Predicted Value vs Real Data

The time series plot shows that comparison of predicted values and real values.

- **Best ARIMAX Model:** After optimization, the ARIMAX model performed the best. It effectively tracked the fluctuations and trends in the real data, demonstrating strong predictive ability.
- **Best SARIMA Model:** The SARIMA model showed improvements in capturing seasonal patterns. However, it still struggled to fit short-term fluctuations, performing slightly worse than ARIMAX.
- **OLS Model:** Even after optimization, the OLS model remained overly smooth and failed to capture the complex dynamics of the time series data.

Overall, the **ARIMAX model** is the most suitable for this dataset, as it provides the most accurate predictions and aligns closely with the real data trends.

7 Socioeconomic Factors of Crime

How does socioeconomic development or recession influence crime occurrences, both on individual and aggregate levels? In criminology, researchers have long examined this relationship, integrating insights from political economy, statistics, and social theory. Before the 1970s, anomie theory dominated the field. Rooted in the Marxist tradition, anomie theory posits that crime arises when rapid social changes disrupt cultural norms, particularly during sharp economic downturns or upturns (Merton 1938, Messner and Rosenfeld 2012).

Since the mid-1970s, criminology shifted from macro-structural explanations to focusing on individual, situational, and community factors. This pragmatic turn was driven by demands for actionable insights that could inform micro-level interventions without requiring sweeping socioeconomic reforms to address the "root causes" of crime.

Despite this shift, economic causes remain central to the study of crime. Empirical research has consistently found modest-to-strong positive relationships between inequality (Box 1987), unemployment (Currie 1998; Fielding et al. 2000; Hooghe et al. 2011; Albertson and Fox 2012), and crime, particularly in deindustrialized contexts. These studies suggest that financial insecurity motivates property crimes. This study applies these criminological theories to San Francisco, examining how socioeconomic predictors influence crime occurrences.

7.1 Data Sources and Methodology

For individual crime incidents, we used San Francisco Police Department Incident Reports data from 2018 to 2023. This dataset contains 913,732 observations, including time, location, and crime categories. Socioeconomic predictors were sourced from the American Community Survey (ACS) 5-year estimates (2017-2022). Crime reports were geocoded to census tracts and combined with tract-level ACS data based on year and tract.

The outcome variable we examine is incident categories. There are 49 incident categories within the original data, and we aggregated into three types: property crime, violent crime, and public order and other crimes. Predictors include changes in median household income, Gini index, poverty rate, unemployment rate, and median home value. Additional variables include transportation metrics (proportions using public transit, bicycles, or walking), demographic proportions (White, Black, Asian, Hispanic), and educational attainment (Bachelor's degree or higher). A COVID-19 dummy variable indicates incidents occurring during the pandemic (2020-2022).

We conducted both individual- and tract-level analyses. For individual-level analysis, we used logistic regressions with tract-level fixed effects. To decide between mixed-effects and fixed-effects models, we performed Intraclass Correlation Coefficient (ICC) tests. Small ICC values (< 0.1) for all categories indicated that fixed-effect modeling was appropriate. Given small ICC statistics (< 0.1) for all three categories, we knew that variations in each category were mainly within-tracts, and fixed-effects modeling is more appropriate.

7.2 Model Specification

The fixed-effects logistic regression model is specified as:

$$logit(P(y_i = 1|X)) = \alpha + \sum_{k=1}^{k} \beta_j X_{ik} + \mu_{j[i]} + \epsilon_i \tag{1}$$

where y_i is the crime category for incident i, X_{ik} is the kth predictor, $\mu_{j[i]}$ is the fixed effect for tract j, and ϵ_i is the error term.

Train-test splits were conducted by stratifying data by census tract, with 70% for training and 30% for testing. Model performance was evaluated on the test set using various metrics.

The logistic regression results bear several notable implications. Most economic predictors have neither strong nor consistent impacts on these crime categories. Only income growth increases public order crimes and more unemployment increases violent crimes but decreases property crimes. In terms of demographics, educational attainment significantly reduces violent and public order crimes but increases property crimes. COVID-19 decreases public order crimes while increasing property crimes. In terms of racial categories, White, Black, and Asian proportions are negatively associated with public order crimes but are positively associated with property crimes.

Table 2 displays the model performance on out-of-sample data using several metrics. The model performs better in predicting property crimes given non-zero precision, recall, and F-1 score. However, all models present weak predictive powers, as the AUC-ROC scores are all only slightly above 0.5, suggesting not much overperformance than random guessing. Therefore, we find out that the individual-level logistic model does not perform ideally in terms of using economic predictors to predict individual crime incidents.

Our next approach aggregates individual crime incidents into tract-level counts. Given this conversion to a tract-year panel structure, we employed Poisson models and, to address potential overdispersion, negative binomial models for our predictions. The Poisson model is specified as:

Table 2: Logistic Regression Results (Excluding Fixed Effects)

Variable	**Public Order**	**Violent Crime**	**Property Crime**
Intercept	55.36*** (15.35)	20.90 (21.71)	-60.13*** (14.51)
Median Household Income Change	0.0095* (0.0051)	-0.0017 (0.0072)	-0.0078 (0.0049)
Gini Index Change	-0.0031 (0.0053)	0.0077 (0.0078)	-0.0010 (0.0050)
Poverty Rate Change	-0.0053 (0.0056)	-0.0117 (0.0089)	0.0087 (0.0052)
Unemployment Rate Change	0.0150 (0.0104)	0.0282* (0.0153)	-0.025* (0.0097)
Median Home Value Change	-0.0067 (0.0052)	0.0054 (0.0073)	0.0045 (0.0048)
Bachelor Rate	-0.059** (0.0185)	-0.061* (0.0266)	0.076*** (0.0172)
Public Transit Rate	0.0336** (0.0137)	-0.0178 (0.0191)	-0.0192 (0.0129)
Bicycle Rate	-0.0313 (0.0242)	-0.0254 (0.0363)	0.0436 (0.0227)
Walking Rate	-0.0096 (0.0122)	0.0201 (0.0169)	-0.0047 (0.0116)
White Proportion	-0.074* (0.0346)	-0.0570 (0.0475)	0.093** (0.0327)
Black Proportion	-0.122*** (0.0254)	0.0342 (0.0346)	0.090*** (0.0240)
Asian Proportion	-0.098** (0.0348)	-0.0395 (0.0482)	0.110*** (0.0327)
Hispanic Proportion	-0.0333 (0.0222)	0.0301 (0.0301)	0.0118 (0.0210)
Covid Years	-0.136*** (0.0186)	0.0345 (0.0266)	0.110*** (0.0175)
Total Population	0.000025 (0.000023)	-0.00007* (0.00003)	0.000006 (0.000022)
Year	-0.028*** (0.0076)	-0.0118 (0.0108)	0.030*** (0.0072)

Note: Standard errors are in parentheses. *p<0.10, **p<0.05, ***p<0.01.

Table 3: Performance Metrics for Crime Categories

Crime Category	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
Public Order	0.69	0.00	0.00	0.00	0.57
Violent Crime	0.88	0.00	0.00	0.00	0.61
Property Crime	0.60	0.61	0.85	0.71	0.60

$$log(\lambda_i) = \alpha + \sum_{k=1}^{p} \beta_k x_{ik} \tag{2}$$

where λ_i represents the expected count for each category of tract i, and x_{ik} denotes the k_{th} predictor for tract i. The negative binomial model assumes a negative binomial distribution of counts and adjusts the variance to account for overdispersion.

Table 3 displays the model coefficients for each crime category. Both models produce similar estimates, indicating robustness in predicting crime counts. At the tract level, Gini index and unemployment rate negatively impact public order and violent crimes in the Poisson model, and both variables show significantly negative impacts on property crimes across both models. This finding contrasts with our hypothesis, as it suggests that narrower income gaps and lower unemployment rates are associated with increased crime counts.

In terms of transportation, public transit rate and bicycle rate consistently show positive associations with crime counts across all models. This aligns with our hypothesis that areas with higher utilization of public transit and bicycles tend to have higher crime rates. Regarding racial demographics, higher proportions of White and Asian populations are consistently associated with lower crime counts across all categories.

Table 4 provides the performance metrics for both Poisson and negative binomial models.

Table 4: Regression Results for Crime Categories (Poisson and Negative Binomial Models)

	Public Order		**Violent Crime**		**Property Crime**	
Variable	**Poisson**	**NegBin**	**Poisson**	**NegBin**	**Poisson**	**NegBin**
Intercept (const)	0.0022***	0.0022***	0.0017***	0.0017***	-0.0601***	-0.0512***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0145)	(0.0152)
Median Household Income Change	0.0031	-0.0017	0.0063	-0.0055	0.0062*	0.0049
	(0.004)	(0.037)	(0.006)	(0.038)	(0.003)	(0.0042)
Gini Index Change	-0.0368***	-0.0218	-0.0298***	-0.0057	-0.0146***	-0.0182***
	(0.004)	(0.036)	(0.006)	(0.037)	(0.003)	(0.0038)
Poverty Rate Change	-0.0023	-0.0036	0.0019	-0.0002	0.0431***	0.0358***
	(0.004)	(0.039)	(0.007)	(0.039)	(0.003)	(0.0039)
Unemployment Rate Change	-0.0111***	-0.0168	-0.0154***	-0.0254	-0.0174***	-0.0147***
	(0.003)	(0.025)	(0.005)	(0.025)	(0.002)	(0.0027)
Median Home Value Change	0.0019	-0.0243	0.0288***	0.00004	0.0114***	0.0097**
	(0.004)	(0.041)	(0.006)	(0.041)	(0.003)	(0.0032)
Bachelor Rate	-0.0714***	-0.0488	-0.0600***	-0.0627	0.1577***	0.1483***
	(0.007)	(0.066)	(0.011)	(0.067)	(0.005)	(0.0072)
Public Transit Rate	0.1444***	0.1484***	0.2065***	0.2101***	0.0537***	0.0491***
	(0.005)	(0.045)	(0.007)	(0.046)	(0.003)	(0.0035)
Bicycle Rate	0.5193***	0.5199***	0.5297***	0.5336***	0.3832***	0.3714***
	(0.004)	(0.049)	(0.007)	(0.049)	(0.003)	(0.0035)
Walking Rate	0.1109***	0.0738	0.1129***	0.0972**	0.0919***	0.0892***
	(0.004)	(0.047)	(0.007)	(0.048)	(0.003)	(0.0031)
White Proportion	-0.5404***	-0.4784***	-0.5818***	-0.5844***	-0.5491***	-0.5384***
	(0.016)	(0.167)	(0.025)	(0.169)	(0.012)	(0.0158)
Black Proportion	-0.0517***	-0.0037	0.0410***	0.0827	-0.0652***	-0.0671***
	(0.008)	(0.088)	(0.012)	(0.088)	(0.006)	(0.0072)
Asian Proportion	-0.4010***	-0.3754**	-0.3178***	-0.3590**	-0.4169***	-0.4223***
	(0.014)	(0.149)	(0.022)	(0.150)	(0.010)	(0.0135)
Hispanic Proportion	-0.0225**	0.0074	0.1206***	0.1368	-0.1270***	-0.1198***
	(0.008)	(0.084)	(0.013)	(0.085)	(0.006)	(0.0083)
COVID Years	-0.1861***	-0.1298	-0.0044	0.0680	-0.0780***	-0.0692***
	(0.008)	(0.084)	(0.013)	(0.085)	(0.006)	(0.0071)
Total Population	8.26e-05***	0.0001***	4.43e-05***	7.07e-05**	2.99e-06	3.47e-06
	(3.07e-06)	(3.05e-05)	(4.97e-06)	(3.09e-05)	(2.27e-06)	(2.52e-06)
Year	0.0022***	0.0022***	0.0018***	0.0017***	0.0027***	0.0025***
	(6.78e-06)	(6.77e-05)	(1.1e-05)	(6.86e-05)	(4.95e-06)	(6.55e-06)

Notes: Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5: Performance Metrics for Poisson and Negative Binomial Models by Crime Category

Crime Category	**Poisson Model**				**Negative Binomial Model**			
	RMSE	**MAE**	**Deviance**	**OOS R^2**	**RMSE**	**MAE**	**Deviance**	**OOS R^2**
Public Order Crime	78.3825	61.8770	-179982.8351	0.1224	78.3883	61.9696	-180043.6100	0.1221
Violent Crime	32.3128	24.1359	-43935.1546	0.1487	35.3065	24.8180	-43267.8766	0.1616
Property Crime	147.0322	120.7313	-418001.4674	0.1168	154.9175	123.2958	-415457.8819	0.1222

Comparing the two models, the negative binomial model demonstrates better performance for violent and property crimes, while the Poisson model slightly outperforms in predicting

public order crimes. Both models achieve their best predictive accuracy with violent crimes, as indicated by the lowest RMSE, MAE, and deviance, along with the highest out-of-sample R^2 values.

These findings imply that while both models provide comparable predictive power, the choice between them may depend on the specific crime category under investigation. The negative binomial model's ability to handle overdispersion makes it preferable for crime types with highly variable counts, such as violent and property crimes.

7.3 Conclusion and Policy Implications

Our study explores the socioeconomic factors influencing crime occurrences in San Francisco, utilizing both individual- and tract-level analyses through logistic, Poisson, and negative binomial models. The results indicate limited predictive power for individual-level models, with weak performance metrics suggesting that socioeconomic predictors alone may not fully explain variations in crime at the individual level. On the other hand, tract-level models, particularly the negative binomial model, provide more robust insights into crime patterns across neighborhoods.

Notably, socioeconomic factors like income inequality and unemployment produced mixed effects on crime categories, sometimes contradicting conventional criminological theories. For instance, narrower income gaps and lower unemployment rates were associated with increased property crimes, which might indicate our models' limitations when generalizing in San Francisco's urban context. Additionally, transportation variables such as higher public transit and bicycle usage consistently correlated with increased crime rates, suggesting the role of urban mobility in shaping neighborhood crime patterns.

These findings suggest important policy implications. First, targeted interventions addressing transportation and mobility patterns—such as enhanced public transit safety measures and better urban design to reduce crime-prone areas—could mitigate crime in high-risk neighborhoods. Second, given the complicated relationship between economic factors and crime, policymakers should exercise caution in assuming linear relationships between economic development and crime reduction. Programs fostering community cohesion and localized interventions may complement broader economic policies to achieve more meaningful crime reductions.

Future research following our paths should address potential endogeneity issues in socioeconomic predictors, such as reverse causality between crime and economic variables like unemployment or housing values. Incorporating instrumental variable approaches might help disentangle causal relationships. Additionally, extending the analysis to include interaction effects between socioeconomic, demographic, and urban design factors may provide deeper insights. Tests for spatial autocorrelation and dynamic models incorporating temporal lags in crime predictors could refine predictive accuracy. Finally, integrating richer data sources, such as police deployment records, behavioral surveys, or real-time mobility data, would enable a more comprehensive understanding of crime patterns and allow policymakers to implement evidence-based, multidimensional strategies for crime prevention.

References

[1] Albertson, K. and Fox, C. 2012. *Crime and Economics*, London: Routledge.

[2] Box, S. 1987. *Recession, Crime and Punishment*, London: Macmillan.

[3] Currie, E. 1997. "Market, Crime and Community: Toward a Mid-range Theory of Post-industrial Violence," *Theoretical Criminology*, 1(1): 147-72.

[4] Field, N., Clarke, A., and Witt, R. (eds) 2000. *The Economic Dimensions of Crime*, London: Palgrave.

[5] Hooghe, M., Vanhoutte, B., Hardyns, W., and Bircan, T. 2011. "Unemployment, Inequality, Poverty and Crime: Spatial Distribution Patterns of Criminal Acts in Belgium 2001-6," *British Journal of Criminology*, 51(1): 1-20.

[6] Merton, R. 1938. "Social Structure and Anomie." *American Sociological Review* 3(5): 672-82.

[7] Messner, S. and Rosenfeld, R. 2012. *Crime and American Dream*, 5th edn, Belmont, Cal.: Wadsworth.