UMBC

Introduction to Data Science

AUTHORS

Sai Krishna Vishnumolakala

Venkat Pantham

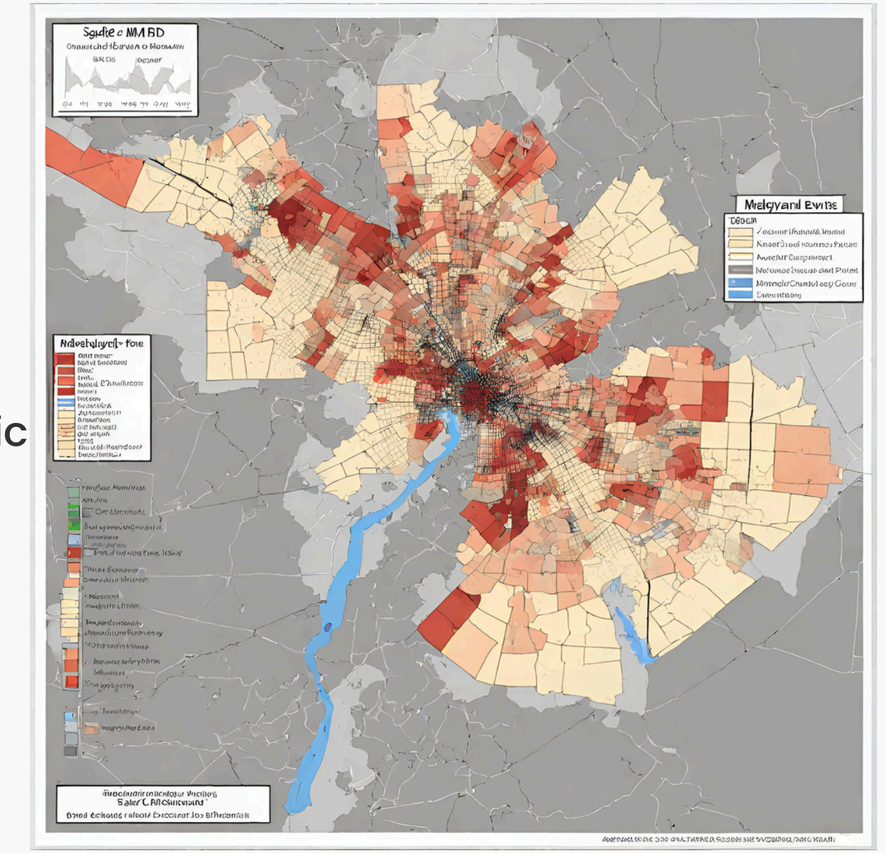
# SAFE-MD: Statistical Analysis and Forecasting of Crime Events in Maryland

# Table of Contents

# Statistical Analysis and Forecasting of Crime Events in Maryland(SAFE-MD)

## Research Motivation

- Crime Impact Mitigation: To develop strategies that reduce crime rates effectively by understanding the complex interplay between socio-economic factors and crime dynamics in Maryland.
- Data-Driven Policy Making: To empower policymakers with predictive insights that enable proactive rather than reactive approaches to public safety, ensuring resources are allocated efficiently.
- Community Safety Enhancement: To provide communities with actionable data that informs interventions, enhances safety, and improves overall quality of life.



- ## What is SAFE-MD?

SAFE-MD is a predictive analytics tool developed as part of this research to forecast crime rates across Maryland counties. Utilizing advanced statistical models and machine learning algorithms, SAFE-MD processes historical crime data along with socio-economic indicators to predict future crime trends.

**How SAFE-MD Helps:**

- **Proactive Planning:** SAFE-MD allows law enforcement and public safety organizations to anticipate crime hotspots and allocate resources more strategically.
- **Community Engagement:** By making crime predictions accessible through an interactive Streamlit application, SAFE-MD fosters a better understanding among community members about the factors influencing their safety.
- **Policy Development:** The insights provided by SAFE-MD support policymakers in crafting laws and regulations that address the root causes of crime, potentially leading to more sustainable crime reduction.

SAFE-MD stands as a cornerstone in using technology and data science to enhance public safety through informed decision-making and strategic planning.

II Data Origin

## Datasets

Maryland Crime Data by county - FBI Crime Lab

Maryland Unemployment Data by County - US Labour Dept.

Maryland Schooling Enrollment Data by county - Data.gov

The data for analyzing crime rates and socio-economic factors across Maryland counties were meticulously gathered from three key sources: the FBI Crime Lab provides detailed county-level crime statistics via the Uniform Crime Reporting (UCR) Program; the U.S. Department of Labor offers unemployment rates by county, presenting crucial socio-economic indicators; and Data.gov supplies schooling enrollment data across various educational levels for each county. These datasets collectively enable a comprehensive analysis of the interplay between crime dynamics and socio-economic variables, facilitating the development of predictive models and informed policy interventions.

- Data Sources: Our comprehensive dataset integrates crime statistics from the FBI's Uniform Crime Reporting (UCR) and socio-economic data from the U.S. Census.
- Scope: Includes detailed records from multiple counties in Maryland, encompassing a range of years to facilitate both cross-sectional and time-series analysis.
- Key Variables:
  - Crime Data: Categories include murder, rape, robbery, aggravated assault, burglary, larceny, and motor vehicle theft.
  - Socio-Economic Data: Includes population, unemployment rates, and educational enrollment across different grade levels (Pre-K, K-5, 6-8, 9-12).

# III Data Transformation, Description and EDA

Purpose: Examine the distribution and trends of crime across various counties to identify patterns.

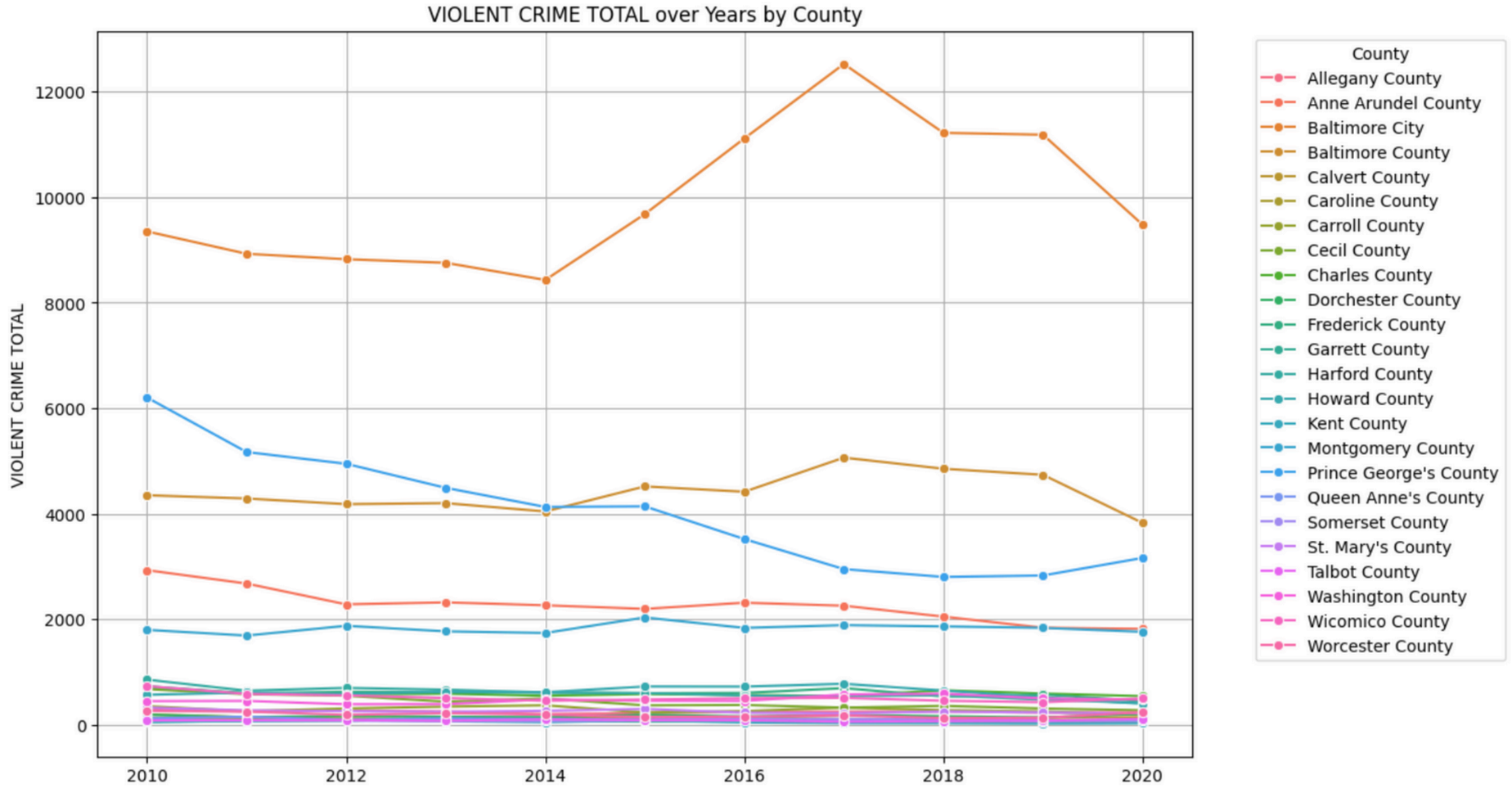
Approach:

- Calculated descriptive statistics (mean, median, mode) to understand central tendencies in crime rates.
- Analyzed year-over-year crime trends to spot increases or decreases across categories.

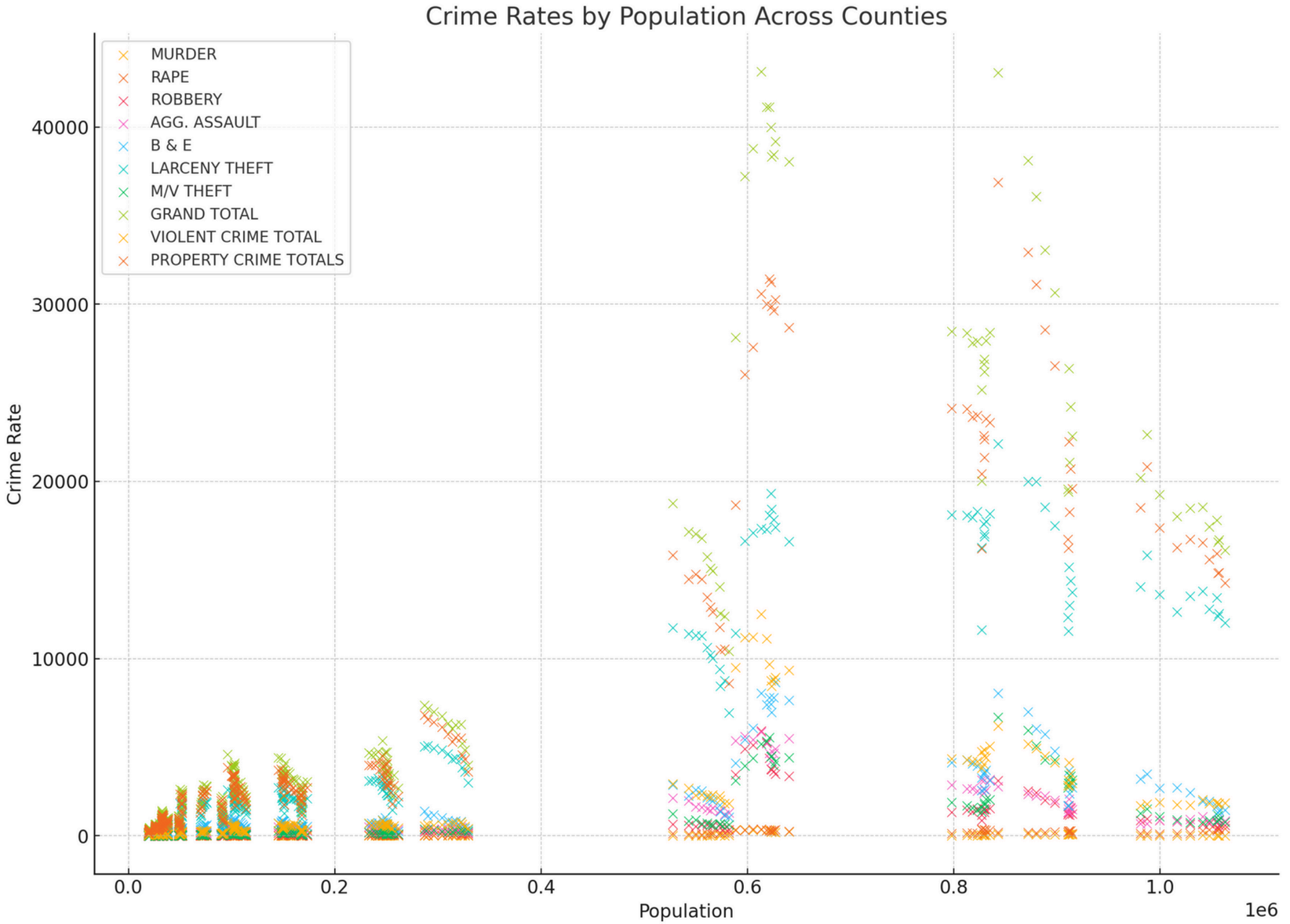
Findings:

- Notable differences in violent crime rates suggest significant local influences.
- Correlation observed between economic downturns and rises in property crime rates in several counties.

# III Data Transformation, Description and EDA

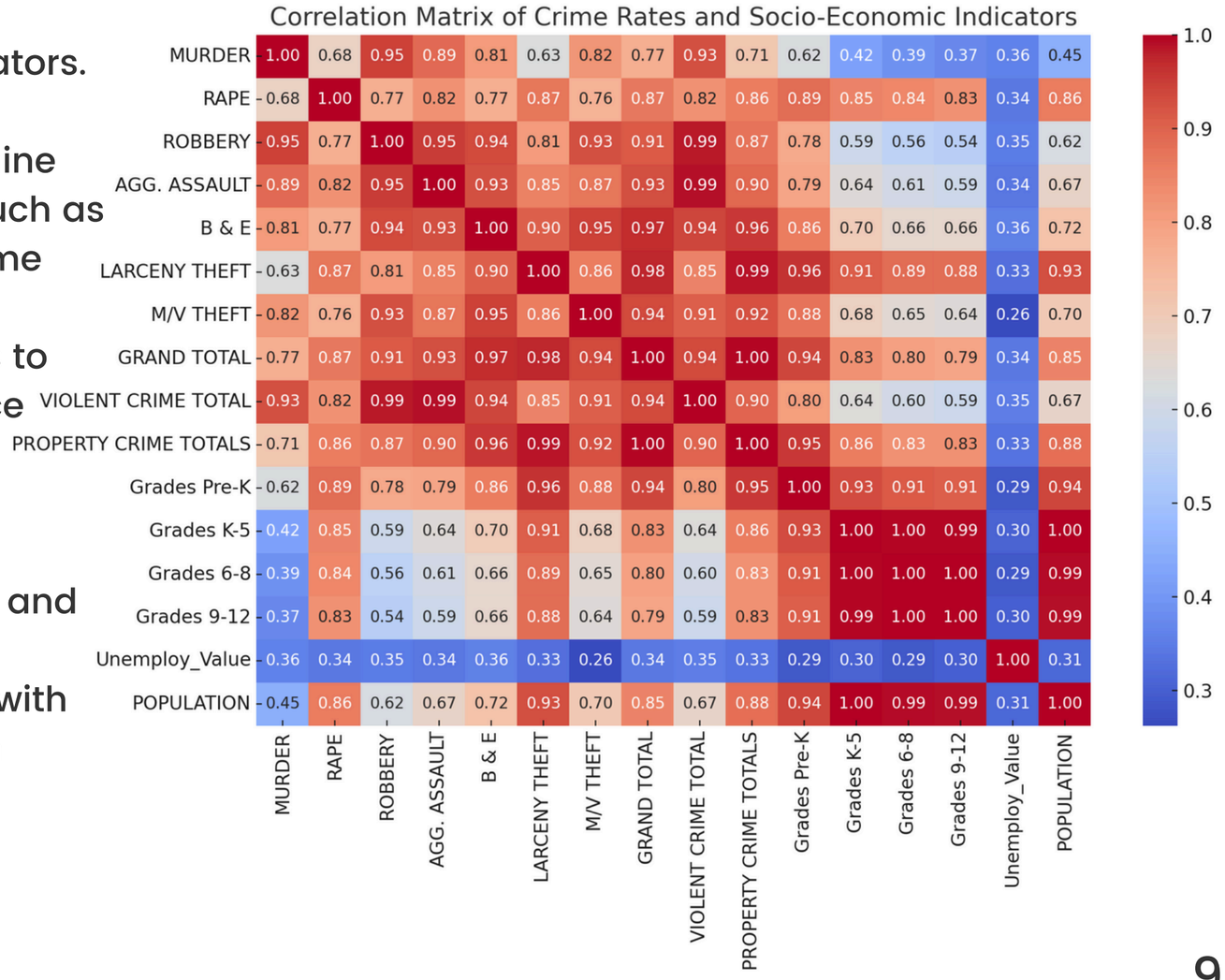


# III Data Transformation, Description and EDA



# III Data Transformation, Description and EDA

- Objective: Investigate how socio-economic factors correlate with crime rates to identify predictive indicators.
- Methods:
  - Performed Pearson correlation analysis to determine the strength of relationships between variables such as unemployment rates, population density, and crime rates.
  - Conducted comparative analysis across counties to see how socio-economic disparities may influence crime statistics.
- Insights:
  - Strong positive correlation found between unemployment rates and increased rates of theft and burglary.
  - Higher population densities are often associated with more frequent violent crimes, underscoring urban crime dynamics.



# IV Model

**RandomForestRegressor**

The RandomForestRegressor is a type of ensemble machine learning model that operates by building a multitude of decision trees at training time and outputting the average prediction of the individual trees. This model is particularly effective for regression tasks because:

**Robustness to Overfitting:** Unlike individual decision trees, which can easily overfit to the data, the averaging approach of RandomForest helps in generalizing better to unseen data.

**Handling of Non-linear Data:** It can manage non-linear relationships between features effectively, thanks to its ensemble nature.

**Feature Importance:** RandomForest provides insights into which features are most important in predicting the outcome.

**GradientBoostingRegressor**

The GradientBoostingRegressor works by building trees one at a time, where each new tree helps to correct errors made by previously trained trees. This model is often preferred for regression because:

**Predictive Power:** It has a strong predictive performance, especially where the relationship between variables is complex and non-linear.

**Flexibility:** It can optimize on different loss functions and provides several hyperparameter tuning options that can make a substantial difference in model performance.

**Handling of Heteroscedasticity:** Unlike standard regression models, gradient boosting can handle heteroscedasticity (non-constant variance in the response variable).

# IV Model

**Comparison with Neural Networks**

While neural networks are powerful for a wide array of tasks, especially those involving high-dimensional data like image and speech recognition, they may not always outperform tree-based models in structured/classical regression tasks. This can be due to:

- **Data Size and Complexity:** Neural networks generally require large amounts of data to perform well and avoid overfitting. In scenarios where data is not exceedingly large or complex (typical in structured datasets), tree-based models often perform better.
- **Interpretability:** Tree-based models offer more straightforward interpretability compared to the often 'black-box' nature of neural networks, which can be crucial in settings where understanding the decision-making process is important.

| | MSE | MAE | Model |
|---|---|---|---|
| **0** | 7.007950e+06 | 887.101868 | Model 1 NN |
| **1** | 1.583621e+07 | 1353.278687 | Model 2 NN |
| **2** | 3.516494e+05 | 188.333887 | Random Forest |
| **3** | 2.426961e+05 | 164.076400 | Gradient Boosting |

# IV Model

**What is an Ensemble?**

An ensemble method uses multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. In machine learning, ensembles help in reducing variance (bagging), bias (boosting), or improving predictions (stacking).

**What is Stacking?**

Stacking involves taking multiple different models and using their outputs as inputs for a final predictor. This model can be thought of as a "meta-model" that refines the predictions made by individual models. The reasons for its effectiveness include:

**Diversity of Models:** Stacking leverages the strengths of various models, making it robust to a variety of data irregularities.

**Error Reduction:** Each model in the stack makes different assumptions and learns different aspects of the data, leading to error reduction when the meta-model corrects these individual errors.

**Improved Accuracy:** By effectively combining different models, stacking often results in higher accuracy than any single model and often outperforms simpler ensembling techniques like simple averaging.

# IV Model

## Ensemble Results

| | Crime Type | MSE | MAE |
|---|---|---|---|
| 0 | MURDER | 44.807687 | 3.680496 |
| 1 | RAPE | 499.283241 | 14.063895 |
| 2 | ROBBERY | 4104.124017 | 32.692012 |
| 3 | AGG. ASSAULT | 6367.767854 | 56.920012 |
| 4 | VIOLENT CRIME TOTAL | 17008.610450 | 78.804852 |

## Stacking Results

| | Crime Type | MSE | MAE |
|---|---|---|---|
| 0 | MURDER | 3.586364e+01 | 4.017983e+00 |
| 1 | RAPE | 5.726569e+02 | 1.377585e+01 |
| 2 | ROBBERY | 1.951384e+03 | 3.131732e+01 |
| 3 | AGG. ASSAULT | 5.816035e+03 | 5.043573e+01 |
| 4 | VIOLENT CRIME TOTAL | 1.407145e-20 | 1.004707e-11 |

**Why Stacking Worked Better**

In the context of your RandomForest and GradientBoosting regressors, stacking can yield better performance as it combines the robustness of random forests with the powerful correcting ability of gradient boosting. The final model, therefore, can mitigate weaknesses specific to each base model, leading to more accurate and reliable predictions.

# IV Streamlit Model

Purpose: The Streamlit application is developed to enable policymakers and analysts to dynamically interact with the crime prediction model. It facilitates real-time adjustments of socio-economic variables and immediate visualization of how these changes could influence crime rates in specific counties.

Functionality:
Interactive Controls: Users can manipulate inputs such as population density, unemployment rates, and educational enrollments through intuitive sliders and dropdown menus.
Immediate Feedback: Upon adjusting the input variables, the model processes the new data, and the predicted crime rates are instantly updated and displayed.
Visualization Tools: The application integrates powerful visualization capabilities, presenting the predicted crime rates through bar charts, which makes comparative analysis straightforward and visually engaging.

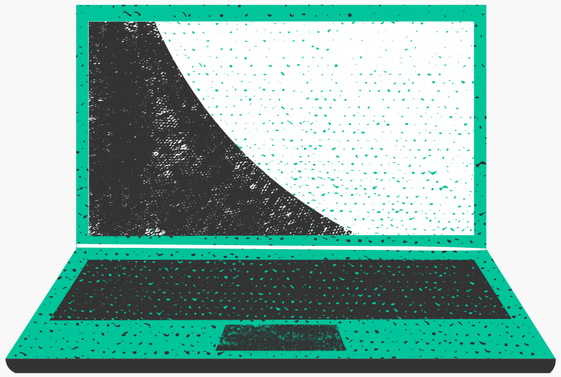
User Experience:
Designed with a focus on usability, the interface is clean and user-friendly, ensuring that even individuals without technical expertise can easily navigate and utilize the tool.
Provides a practical demonstration of the potential impact of socio-economic changes on crime rates, aiding in strategic planning and resource allocation.
Visuals: Include screenshots of the Streamlit application showing the interactive sliders and the resulting visualizations of crime predictions. Optionally, a brief video or animated gif demonstrating the interaction process could enhance engagement.

Impact:
Strategic Planning: By allowing for scenario testing, the application helps policymakers understand potential outcomes of changes in socio-economic conditions, supporting more informed decision-making.
Resource Allocation: Insights derived from the model can guide more effective distribution of law enforcement and community resources to areas where they are most needed.

# V Limitations & Future Work

## Limitations of the Study

- **Data Limitations**
- **Complexity of Social Phenomena**
- **Overfitting Risk**



## Future Work

- **Incorporating Additional Data Sources**
- **Model Expansion and Refinement**
- **Community Engagement and Policy Maker Interface**

AUTHORS
Sai Krishna Vishnumolakala
Venkat Pantham

# Q/A and Feedback

# Thank You