SAN DIEGO STATE UNIVERSITY

Big Data Analytics 594

ANALYSIS OF THE JUSTICE SYSTEM AT THE LAW ENFORCEMENT LEVEL

Final Project

Group III [Demo Group - 8]

Group Members

Maha Lakshmi

Anurima Saha

Fernanda Carrillo Escarcega

Manisha Radhakrishna

December 18, 2023

Abstract

According to a June 2023 CBS article that revealed that nearly half of U.S. murders go unresolved, this sparked our concern about public safety. Our project explores the efficacy of the justice system in handling crime incidents. While focusing on three major cities, we analyzed large datasets available from data.gov that helped us gain insights that we conveyed through meaningful visualizations. Our findings align with the news article showing a significant number of unsolved crime cases.

1. Problem Statement

In an era of increasing concern for public safety, our project tackles this problem through the lens of crime analytics. Our focus is on the cities of Los Angeles, Chicago, and San Francisco and the CrimeAlytics team aims to use the power of data to have a better understanding of criminal incidents. The goal is the creation of predictive models that could assess the probability of a crime case being resolved or unresolved based on specific features from the datasets. By immersing ourselves in the data of these cities, the project led us to discover patterns, and trends, as well as identifying factors that impact crime resolution rates. The exploration of crime resolution is key for shedding light on the competence of law enforcement efforts.

Our work adopts a comprehensive approach, combining advanced analytics with visualization techniques using tools like Tableau. The visualizations created offer valuable insights into criminal activities that could be a crucial resource for law enforcement agencies. The data-driven insights have the potential to create awareness and provide guidance on proactive decision-making processes. Ultimately, this project aims to improve community safety initiatives by providing a data-driven approach that could facilitate strategic interventions.

2. Literature review:

(i) Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques- URL

Cited by 40: W. Safat, S. Asghar and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," in *IEEE Access*, vol. 9, pp. 70080-70094, 2021, doi: 10.1109/ACCESS.2021.3078117.

This paper analyzed crime data of Chicago and Los Angeles, intending to predict future crime rates and areas with higher crime rates across the cities, and helping to improve strategies to eradicate crimes. They have used extensive machine learning models like support vector machine, K-nearest neighbor, multilayer perceptron, autoregressive integrated moving average, and some other predictive models to predict the future crime rates, out of which LSTM has performed well with reasonable values of root mean square error(RMSE) and mean absolute error(MSE). They have also performed exploratory analysis which resulted in identifying crime types and different trends of crime features.

(ii) Dynamic Network Model for Smart City Data-Loss Resilience Case Study: City-to-City Network for Crime Analytics- URL

Cited by 20: O. Kotevska, A. G. Kusne, D. V. Samarov, A. Lbath and A. Battou, "Dynamic Network Model for Smart City Data-Loss Resilience Case Study: City-to-City Network for Crime Analytics," in IEEE Access, vol. 5, pp. 20524-20535, 2017, doi: 10.1109/ACCESS.2017.2757841.

Some areas in cities are highly reliable on crime data, and during deployment, there is a chance of data loss. This paper has provided a network model that helps in crime data loss recovery. They have designed a network model with a data loss recovery system, showcasing crime rates across Montgomery County. They also discovered that weather is highly affecting the crime rates in the county. This has helped hospitals and police to know about the crimes in the cities quickly and act accordingly.

3. Data Collection

Data collection is a crucial step in our project. It is of utmost importance to collect reliable data for model building so that the Machine Learning models can identify the correct patterns in the data to produce trustworthy results.

For our project, we looked into arrest-related data from police departments across multiple cities in the United States of America. However, the data either had incomplete information or was not consistent with the purpose of our project. The three final datasets were collected from **data.gov** and three prime cities known for a record amount of crime- *Chicago, Los Angeles, and San Francisco*. The detailed data description available on the websites ensures the correctness of the data and ease of interpretability. Data considered for analysis for all three cities has a crime occurrence date from January 2018 - August 2023.

(i) Los Angeles

This dataset represents crime incidents reported within the City of Los Angeles. The information is transcribed from the original crime reports, which are initially documented on paper.

- Dataset dimension - 8+ million rows and 22 columns.
- Key Features - crime type, location description, time of the crime - occurrence and reported, area details, occurrence latitude and longitude, and victim demographics - age, sex, and descent.

(i) San Francisco

The San Francisco Crime dataset is an incident report dataset from the San Francisco Police Department (SFPD.) It consolidates information from the Crime Data Warehouse (CDW). It offers details on incident reports, which are submitted by SFPD officers or self-reported by members of the public through the SFPD online reporting system.

- Dataset dimension - 802 rows and 22 columns.
- Key Features - crime type, location description, time of the crime - occurrence and reported, area details, occurrence latitude and longitude, and victim demographics - age, sex, and descent.

(iii) Chicago

The Chicago dataset is sourced from data.gov, and provides a comprehensive overview of reported crime incidents (excluding murders with individual victim data) in the City of Chicago. The information is derived from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system and undergoes daily updates.

- Dataset dimension - 1.05 million rows and 28 columns.
- Key Features - crime type, location description, time of crime, ward details, arrest latitude and longitude

4. Data Processing

The datasets extracted required considerable cleaning and data treatment before they could be fed into the model. The primary steps involved in the data preparation have been summarized below:

- **Data Cleaning** - This step involved removing columns irrelevant to our analysis. Los Angeles data consisted of 22 columns out of which 11 columns were dropped. The final columns included victim demographics, crime occurrence and report dates, crime type, premise and geographic location, and weapon use. San Francisco and Chicago followed a similar data-cleaning process by sub-setting variables that provide similar attributes about crime occurrences and arrests. Furthermore, demographic variables like "Age" have inconsistent information(negative values and values>100) which were treated as missing.
- **Missing Value Treatment** - The datasets have missing values for multiple columns like descriptions of the type of crimes, premise, victim demographics, and weapon use. Each column has been treated differently. Victim demographics except age have been grouped into a separate category - missing. Age has been substituted with the median value.
- **Feature Engineering** - Date columns were changed from string formats to extract year, month, and day for analysis. The exact time of occurrence was also available which was grouped into - "Morning", "Afternoon", "Evening" and "Night", at an interval of 6 hours starting from 6 am. A variable "Type of Crime Group" was created, grouping crime type into 15 relevant categories. Similar groupings were done for "Premise", "Weapon Used" and "Victim Descent". Each categorical feature was then encoded using "One-hot encoding".

- **Target Variable Creation** - The target variable creation was different for each city. For Los Angeles, we have used the " Status" and "Status description" variables to create the data variable for our machine learning model. All crimes under the category "Investigation Continued" or "IC" have been considered as "Unresolved" and all other values have been considered as "Resolved". To justify the selection, we have taken data with a lag of 6 months from the date of occurrence to allow time for investigation and resolution. For Chicago, "arrest" had Boolean values that were directly used in the model. San Francisco data provides a variable "Resolution Category" where the "'Cite or Arrest Adult' and 'Exceptional Adult' categories were treated as resolved.

5. Analysis

5.1. Exploratory Data Analysis:

For the project analysis and implementation, we were able to understand the features and their correlations by constructing interactive tableau dashboards on different attributes of the dataset and the expected outcome variables. By analyzing and observing these treads we were able to understand the dataset better and also implement models to determine the expected outcome.

- **Analyzing the Total Crime Over the Years [2018-2023] (**Tableau Link**):** The below table visualizes the change in total crime across the cities over the period of time, we can deduce from the table that there was a slight decrease in the crime rate during covid but later increased after the lockdown. Each bubble in the map refers to a latitude and longitude of the neighborhood and the size of the bubble represents the total crime rate. Hence we can understand the relative changes over the years across the cities.

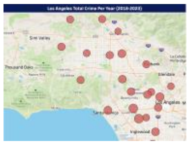
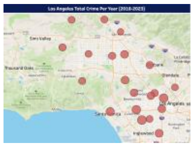
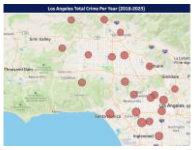
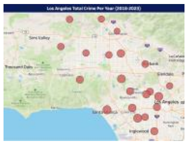
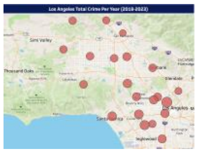
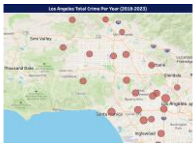
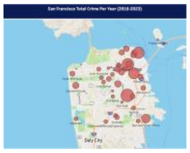
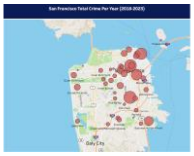
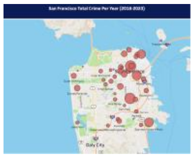
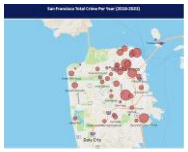
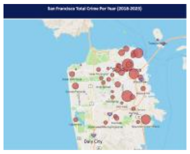
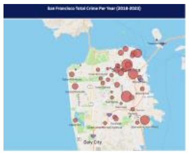
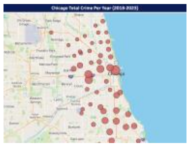
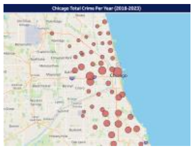
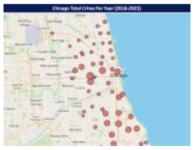
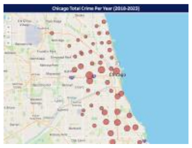
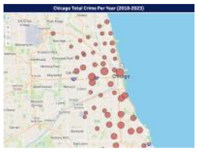
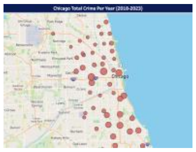
	2018	2019	2020	2021	2022	2023
Los Angeles	 	 	 	 	 	
San Francisco	 	 	 	 	 	
Chicago	 	 	 	 	 	

Fig 1: Time Series plot for LA, SF and Chicago

- **Distribution of Data based on the Outcome Variable [Resolved & Unresolved] (**Tableau Link**):** This graph shows how both resolved and unresolved cases vary over months from 2018 to 2023. The gap between resolved and unresolved is increasing over time as the number of resolved cases shows a downward trend and there is a surge in unresolved crime. There is a decrease in the number of unsolved crimes from 2020 April to 2021 April, which is Covid time.

- Los Angeles: There is a sharp increase in unresolved crimes from February 2021, a steep decrease in resolved crimes overall, with the highest in July 2018.
- San Francisco: There is an all-time low of unresolved crimes in April 2020. and an increase from April 2020 to October 2021. But can observe almost no difference in resolved cases in this period.
- Chicago: There is a cyclical trend from 2021 to 2023, with the least amount of crimes occurring in February and most crimes occurring in October.

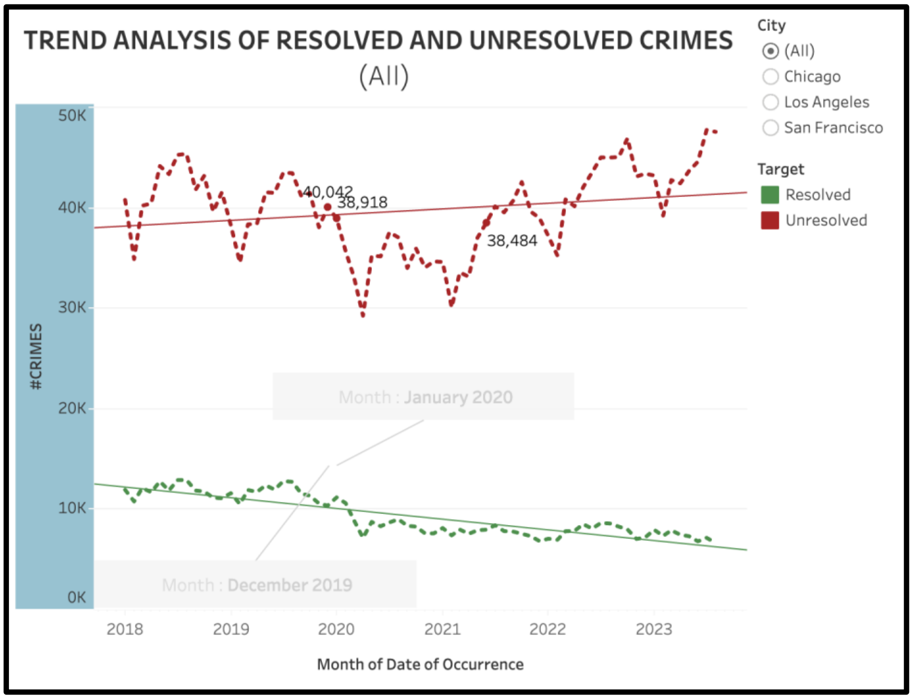


Fig 2: Trend

Analysis between Resolved and Unresolved

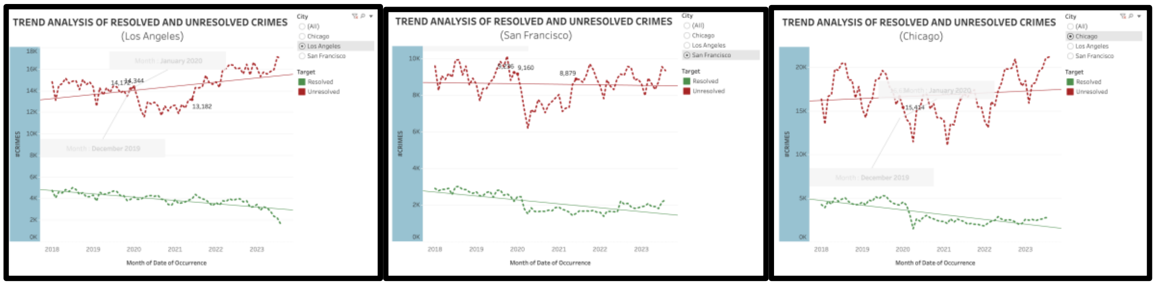


Fig 3: Trend Analysis across LA,SF and Chicago

- **Exploring Crime Types and Location Premises(Tableau Link):**

  The crimes mostly occurred in domestic, roads, streets and business areas. In Chicago, apart from domestic and roads, business areas and public places are most crime occurring areas. In Los Angeles, we can observe that a major number of crimes are occuring. Apart from, domestic and street areas, business, parking and garage areas also take a significant part.

  The above pie chart shows that most frequent crimes are theft, grand theft, other crimes and assaults. Sex offense, serious and weapon assault being the least occurring crimes. In Chicago, grand theft, assault, others and vandalism are most frequent crimes. In Los Angeles, theft and grand theft are the top 2 crimes and assault stays in 3rd place.

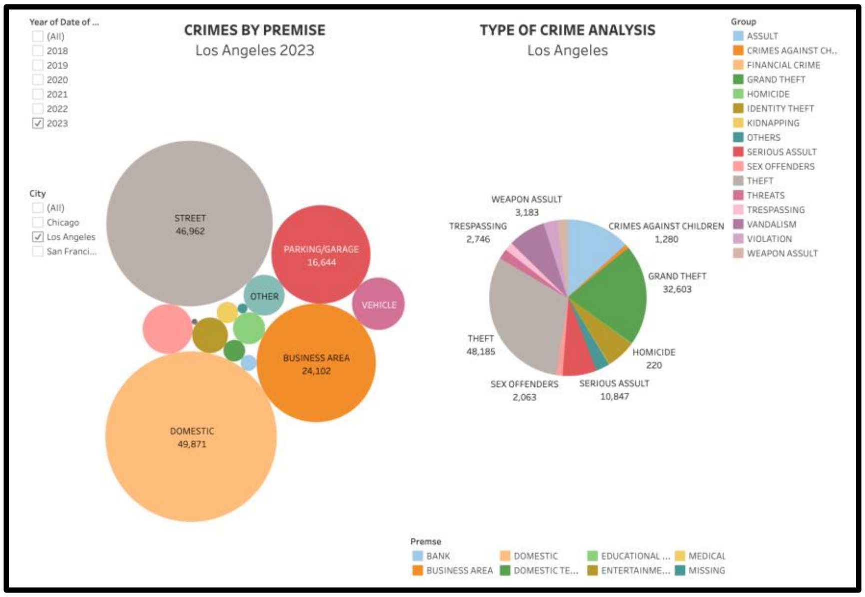


Fig 4: Crime Type and Location Premise Analysis

- **Comparing Crime Resolution Status based on Time of Day(Tableau Link):**

From the first graph it is clear that most of the crimes are taking place in the afternoon and evening and less crimes are occurring in each city. There is a significant amount of crime occurring in Chicago at night. Compared to Chicago and Los Angeles, San Francisco people are experiencing the least number of crimes.

We can observe a similar trend of crime occurrence for premises at each time of day. Crimes at the roads, parking, and garages most frequently occur in the evening time. We can also observe more crimes on the streets occurring in the evening. The areas like banks, medical centers, and domestic temporarily are the areas where crimes are occurring in fewer numbers.

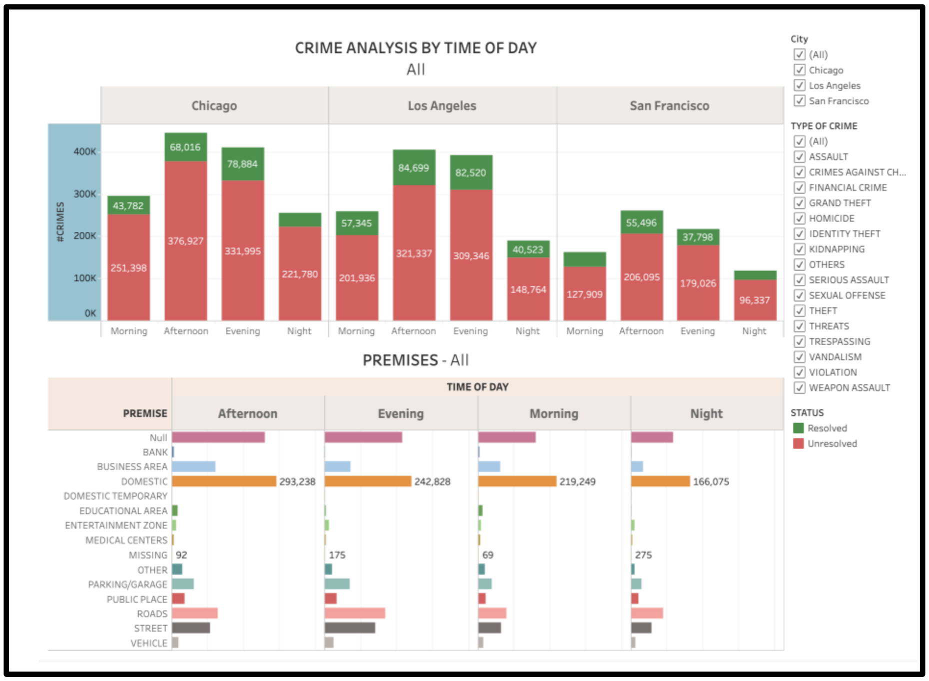


Fig 5: Crime Analysis by Time of Day

- **Distribution of Crime Type across the City(Tableau Link):**

This map shows the top crime category by area in each of the cities, exploring San Francisco, Los Angeles, and Chicago by clicking on the tabs.

It can be observed that there are a couple of top crime types per city, San Francisco has miscellaneous and property crimes, and most crimes happen in the northeast part of the city. Los Angeles shows grand theft and theft as the most common categories with theft being the most dominant one, the crime rate is constant across Los Angeles city. Lastly, Chicago was the only city that had multiple crime categories as a top crime in each area including assault, grand theft, identity theft, and others. Grand theft is more prominent in the northern part of the city, and assault is present in most parts of the cities.

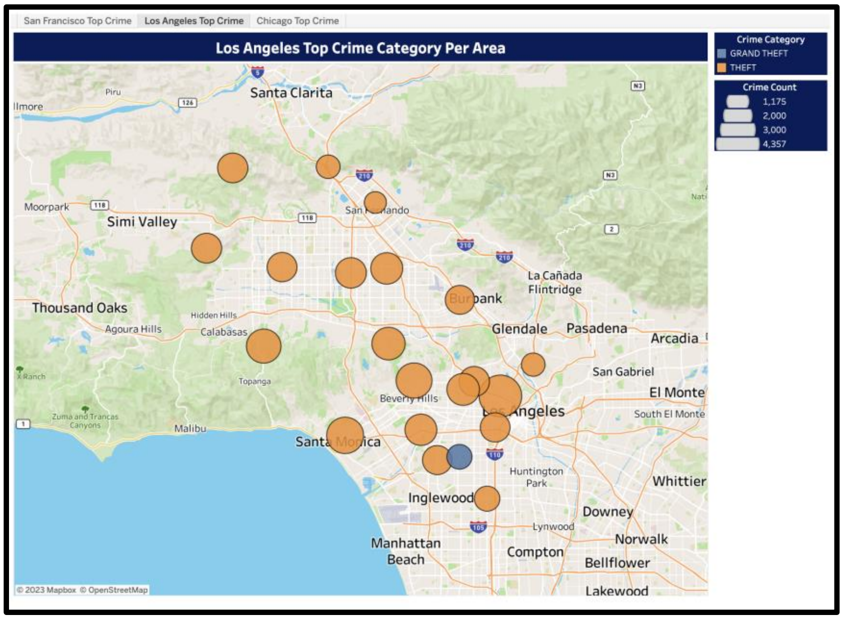


Fig 6: Crime Category Distribution in Los Angeles

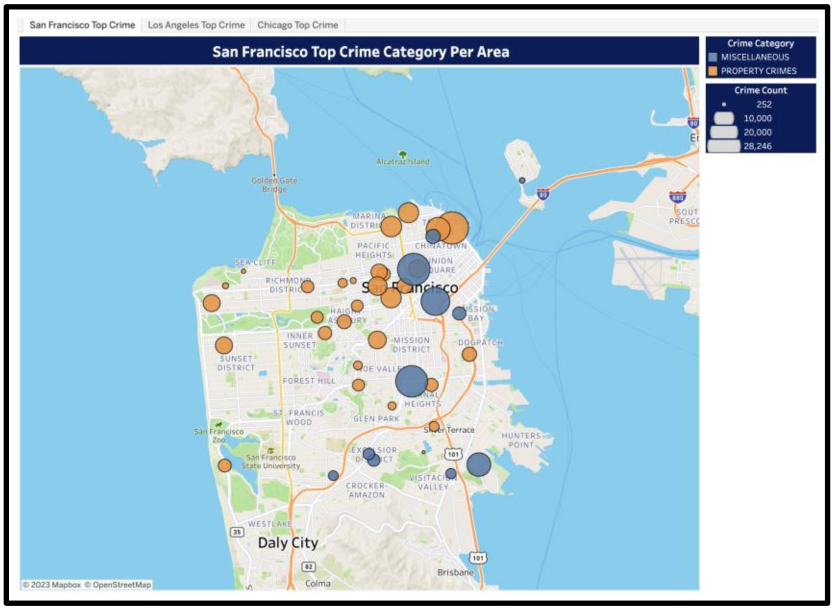


Fig 7: Crime Category Distribution in San Francisco

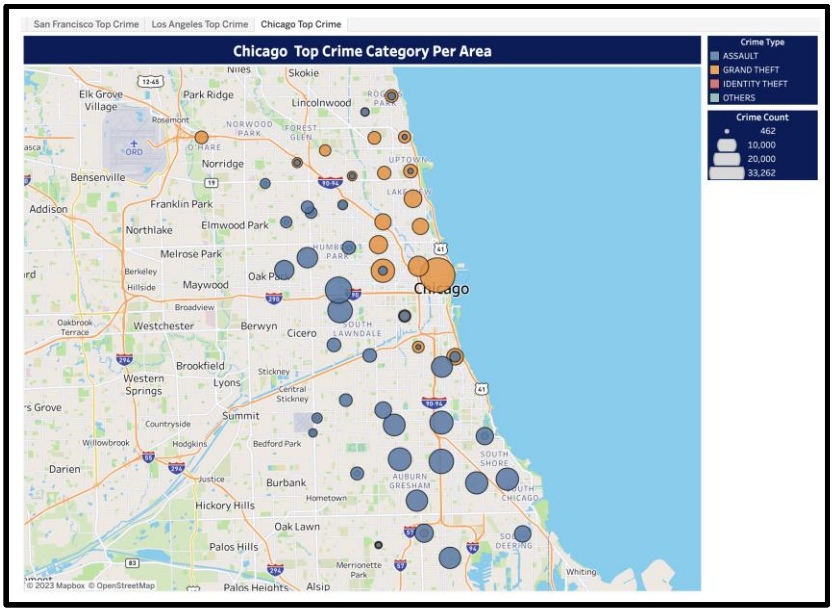


Fig 8: Crime Category Distribution in Chicago

- **Exploring Victim Demographics based on Weapons Type (Tableau Link):**

  The above visualization describes how different gender and race groups are affected by crimes in Los Angeles. The conclusion is that Hispanic males and females are most affected by crimes, followed by black and white race people at almost all premises and weapon types used. Most of the Hispanic people's crimes are unresolved, followed by black and white crime cases that seem equally unresolved. Compared to Hispanics, Blacks, and Whites, other races do take a significant amount in experiencing different crimes but as their population is generally less, their crime experience numbers are less. When we consider the unresolved crimes on the streets other genders and people of unknown races are most affected compared to a particular gender male and female with particular race. Overall, the weapons are mostly used in thefts, grand theft, and assault. When we consider weapons in crimes, Sharp objects are mostly used in serious assaults and thefts. Firearms are used in serious assaults, weapon assaults, thefts, and homicide crimes.

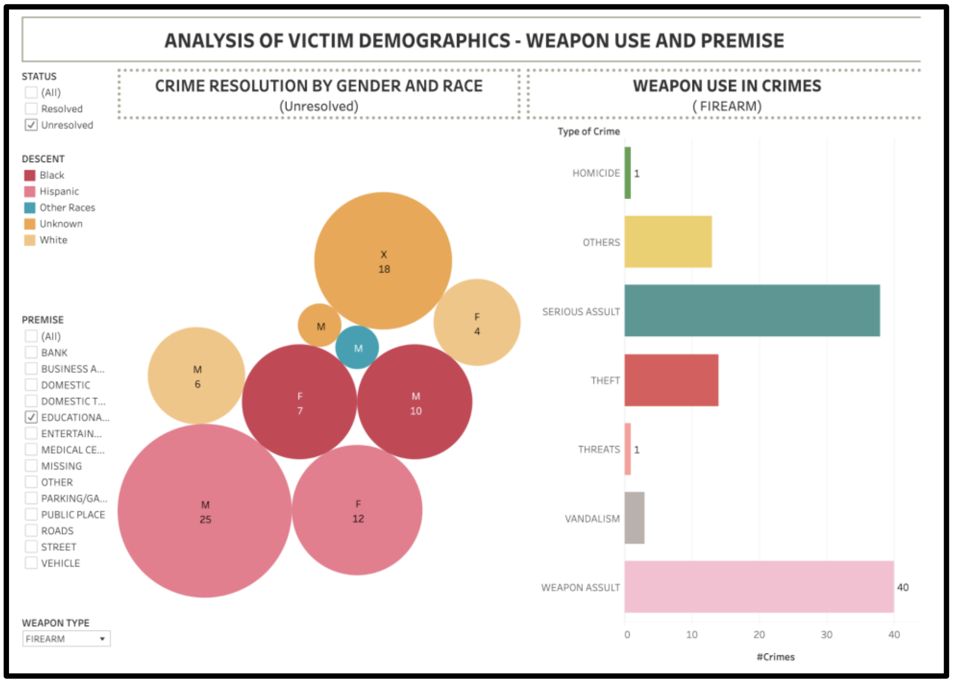
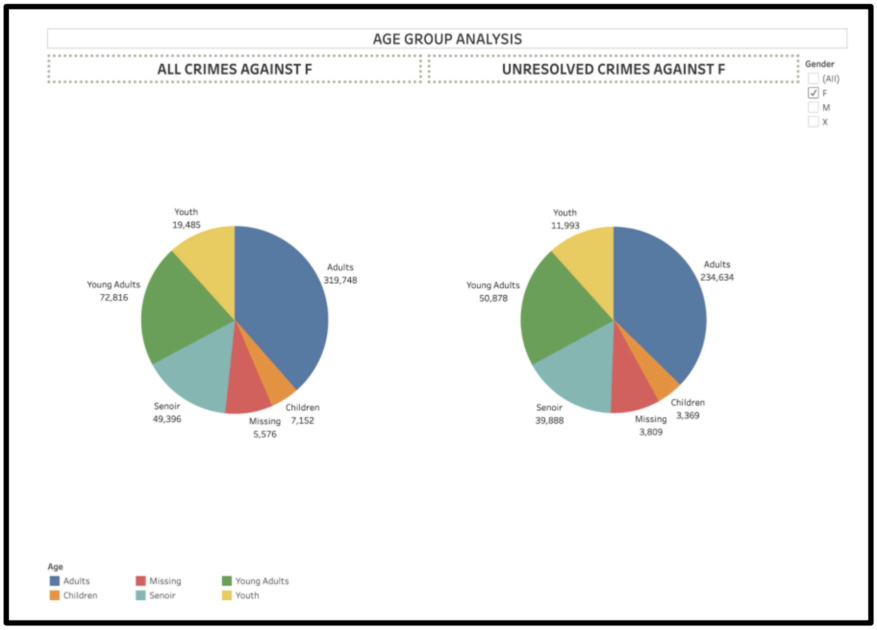


Fig 9: Victim Demographics based on Weapon Type for LA

- **Exploring Total Crime based on Age and Gender(Tableau Link):**

The first pie chart provides a complete snapshot of crimes against individuals in Los Angeles from 2018 to 2023, it includes filters to visualize data categorized by age and gender. The chart visually breaks down the distribution of reported crimes, highlighting patterns and trends across different demographics. The second graph illustrates the proportion of unresolved crimes. This crucial aspect provides insights into the challenges and gaps in the criminal justice system, emphasizing areas that may require attention and improvement.

The data reveals a notable trend, a higher incidence of crime victims among adults, closely followed by young adults. Intriguingly, this pattern aligns consistently with the prevalence of unresolved crimes, indicating a correlation between the demographic groups experiencing a higher number of criminal incidents and those facing challenges in case resolution.



Gender
□ (All)
☑ F
□ M
□ X

Fig 10: Age Group Analysis over Gender in LA

6. Predictive Model:

The project's goal is to predict whether a crime will be resolved or unresolved based on various input features such as the location of the crime, the day it occurred, the type of crime, etc. This problem statement has been defined as a **classification mode**l with **binary targets** - *Resolved(0)* and *Unresolved(1)*. We have experimented with multiple classification models that are mentioned below:

- **Logistic regression** - parametric modeling of probability of a discrete outcome given input features.

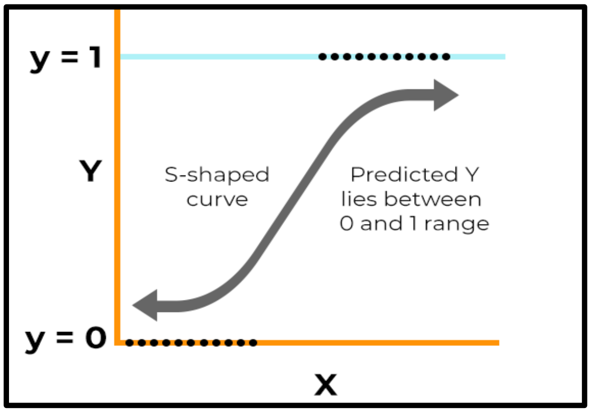


Fig 11: Basic Architecture of Logistic Regression Model

- **Decision Tree** - non-parametric supervised learning model that predicts the value of a target variable by learning simple decision rules

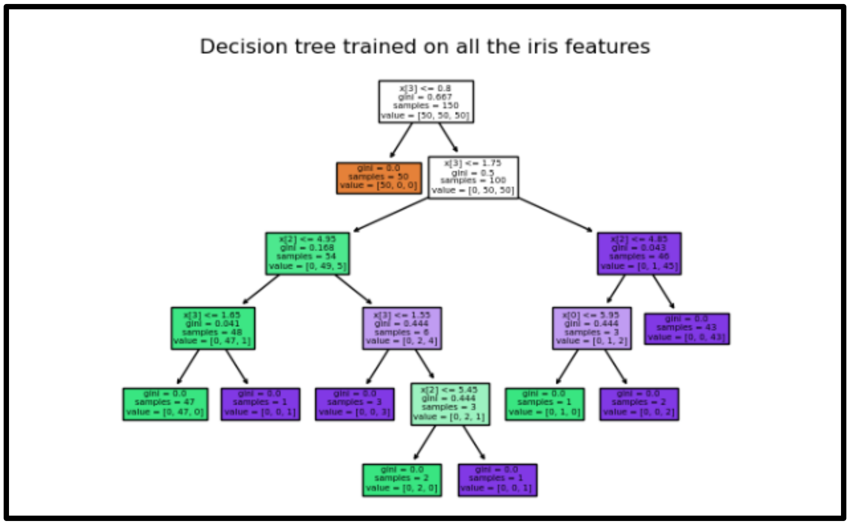

Decision tree trained on all the iris features
x[3] <= 0.8 gini = 0.667 samples = 150 value = [50, 50, 50]
gini = 0.0 samples = 50 value = [50, 0, 0]
x[3] <= 1.75 gini = 0.5 samples = 100 value = [0, 50, 50]
x[2] <= 4.95 gini = 0.168 samples = 54 value = [0, 49, 5]
x[2] <= 4.85 gini = 0.043 samples = 46 value = [0, 1, 45]
x[3] <= 1.65 gini = 0.041 samples = 48 value = [0, 47, 1]
x[3] <= 1.55 gini = 0.444 samples = 6 value = [0, 2, 4]
x[0] <= 5.95 gini = 0.444 samples = 3 value = [0, 1, 2]
gini = 0.0 samples = 43 value = [0, 0, 43]
gini = 0.0 samples = 47 value = [0, 47, 0]
gini = 0.0 samples = 1 value = [0, 0, 1]
gini = 0.0 samples = 3 value = [0, 0, 3]
x[2] <= 5.45 gini = 0.444 samples = 3 value = [0, 2, 1]
gini = 0.0 samples = 1 value = [0, 1, 0]
gini = 0.0 samples = 2 value = [0, 0, 2]
gini = 0.0 samples = 2 value = [0, 2, 0]
gini = 0.0 samples = 1 value = [0, 0, 1]

Fig 12: Basic architecture of Decision Tree

- **Random Forest** - A commonly used Machine Learning model that combines the output of multiple decision trees and aggregates into a single result.

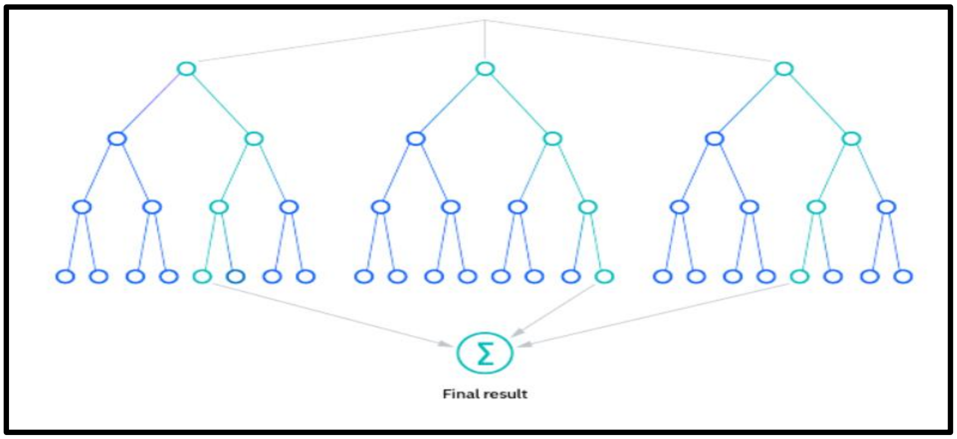


Fig 13: Basic architecture of Random Forest Model

- **Extreme Gradient Boosting** - a machine learning algorithm that uses gradient-boosted decision trees to enhance speed and performance.

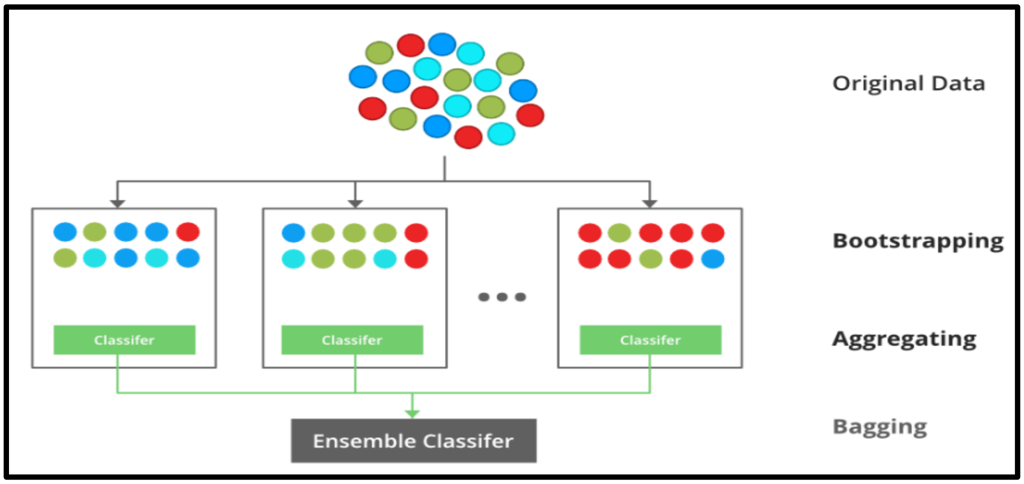


Fig 14: Basic architecture of XGBoost Model

7. Result

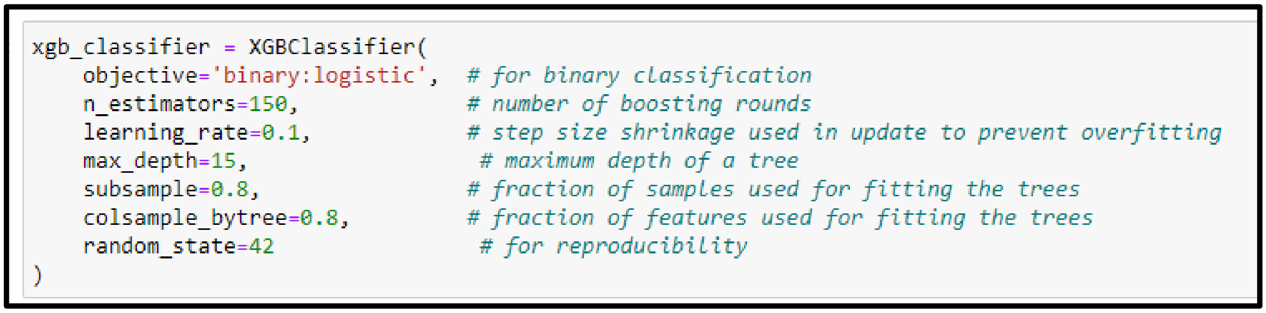
Models for three different cities have been trained and tested separately to obtain city-specific optimized results. The results of training and testing different models along with model details have been summarized below.

(i) Los Angeles

We began by training the data for Los Angeles on a *Decision Tree Classifier* with hyperparameters - max_depth =20 and min_samples_split=100 which gave an accuracy score of 97.74% and a score of 97.48%. However, on the test data, the scores were 71.92% and 70.21% respectively which is a clear indication of overfitting.

Then we moved on to training a Random Forest Classifier with similar hyperparameters - max_depth=30, min_samples_split=10, and n_estimators=150 which gave an accuracy and f1-score of about 88.51%. On the test set, the accuracy was 78.98% and 80.17% as f1-score.

In order, to obtain a more optimized performance we have also trained our model on Extreme Gradient Boosting. The hyperparameters are shown in the picture below.

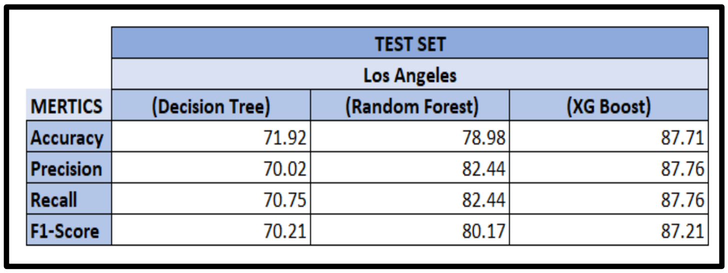


```
xgb_classifier = XGBClassifier(
    objective='binary:logistic',  # for binary classification
    n_estimators=150,             # number of boosting rounds
    learning_rate=0.1,            # step size shrinkage used in update to prevent overfitting
    max_depth=15,                  # maximum depth of a tree
    subsample=0.8,                # fraction of samples used for fitting the trees
    colsample_bytree=0.8,         # fraction of features used for fitting the trees
    random_state=42                # for reproducibility
)
```

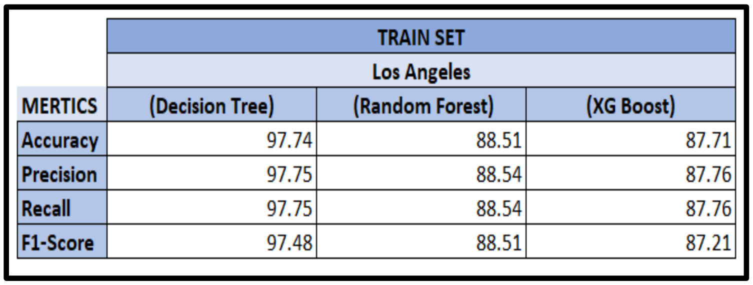
Fig 15: The hyperparameters of the XGBoost model used as the final model Los Angeles data.

The model has an accuracy of 87.72% on the training set and 78.77% on the test set. This was selected as the final model as the precision of the model was ~83% as compared to the random forest model where precision was ~82%. Please see the details of model training and testing in the table below.

Fig 16: Shows the evaluation metrics of Los Angeles Dataset



	TEST SET		
	Los Angeles		
MERTICS	(Decision Tree)	(Random Forest)	(XG Boost)
Accuracy	71.92	78.98	87.71
Precision	70.02	82.44	87.76
Recall	70.75	82.44	87.76
F1-Score	70.21	80.17	87.21



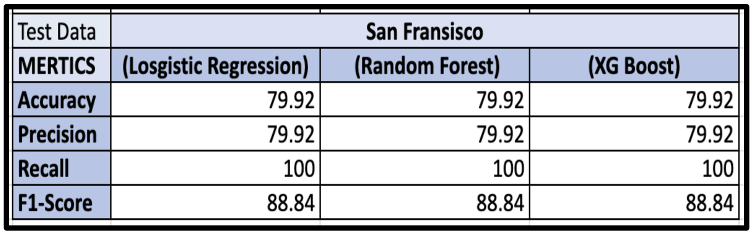
	TRAIN SET		
	Los Angeles		
MERTICS	(Decision Tree)	(Random Forest)	(XG Boost)
Accuracy	97.74	88.51	87.71
Precision	97.75	88.54	87.76
Recall	97.75	88.54	87.76
F1-Score	97.48	88.51	87.21

(i) San Francisco

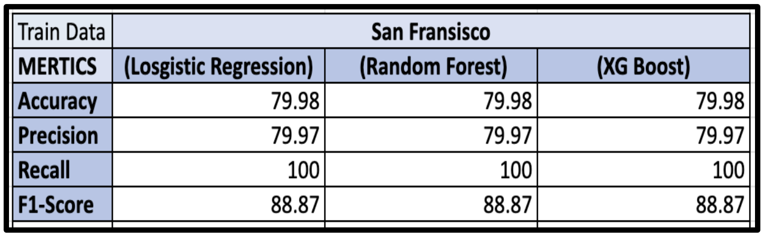
The results provided for Random Forest, Logistic Regression, and XGBoost models for the San Francisco dataset show consistent performance metrics across all three algorithms. This could be because the features provided might not capture all patterns.

With the information provided we might need to consider hyper tuning the parameters to identify specific triads, which would be considered as the future scope of this application. As all three models provide the same output, we are moving forward with Logistic Regression to reduce computational resources.

Fig 17: Shows the evaluation metrics of San Francisco Dataset



Test Data	San Fransisco		
MERTICS	(Losgistic Regression)	(Random Forest)	(XG Boost)
Accuracy	79.92	79.92	79.92
Precision	79.92	79.92	79.92
Recall	100	100	100
F1-Score	88.84	88.84	88.84



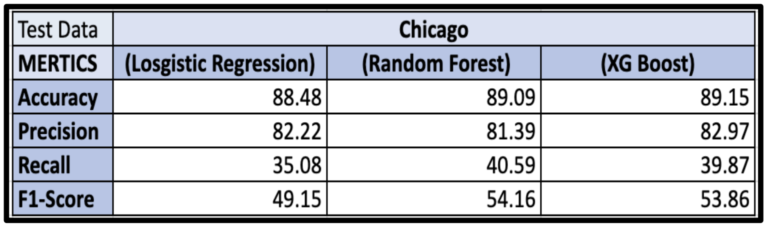
Train Data	San Fransisco		
MERTICS	(Losgistic Regression)	(Random Forest)	(XG Boost)
Accuracy	79.98	79.98	79.98
Precision	79.97	79.97	79.97
Recall	100	100	100
F1-Score	88.87	88.87	88.87

(ii) Chicago

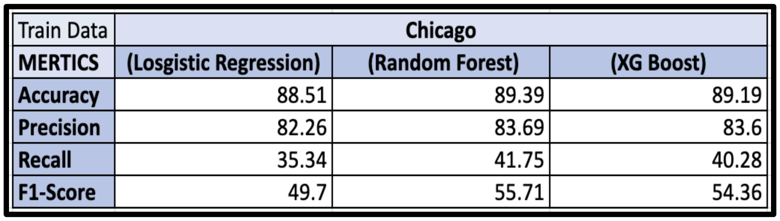
For the Chicago dataset Random Forest provides high accuracy on booth training and test dataset(89.09% on test, 89.39% on train) along with balanced precision and recall (F1 score 0.5417 on test, 0.5571 on train). Logistic regression provides less accuracy compared to Random Forest and XGBoost performs similarly but has comparatively less F1 score.

Random Forest performance on both test and train data is good, thus proving the model is not overfitting or underfitting, it is also known for its robustness to noisy data which is advantageous for real-world samples.

Fig 18: Shows the evaluation metrics of Chicago Dataset



Test Data	Chicago		
MERTICS	(Losgistic Regression)	(Random Forest)	(XG Boost)
Accuracy	88.48	89.09	89.15
Precision	82.22	81.39	82.97
Recall	35.08	40.59	39.87
F1-Score	49.15	54.16	53.86



Train Data	Chicago		
MERTICS	(Losgistic Regression)	(Random Forest)	(XG Boost)
Accuracy	88.51	89.39	89.19
Precision	82.26	83.69	83.6
Recall	35.34	41.75	40.28
F1-Score	49.7	55.71	54.36

8. Application developments

In a world where personal safety is paramount, our Crime-Alytics Web Application emerges as a groundbreaking tool designed to provide individuals with insights into the likelihood of resolving criminal cases. By combining user inputs such as city, crime location, crime type, day of the week, and other pertinent features, our application employs advanced predictive analytics to assess the probability of successfully resolving a given case.

Built using the StreamLit Python package and hosted on its Community Cloud for better accessibility.

Link : https://crimealytics-bda594.streamlit.app/?city=la

Key Features

- City Selection: Users can choose the city where the crime occurred from the sidebar's drop-down menu. The application dynamically adjusts its features based on the selected city. The webpage design varies for each city, considering the unique feature parameters present in the dataset for each location.

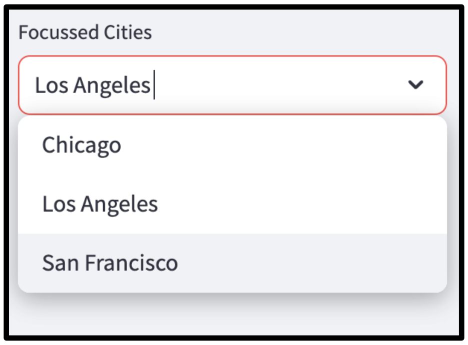
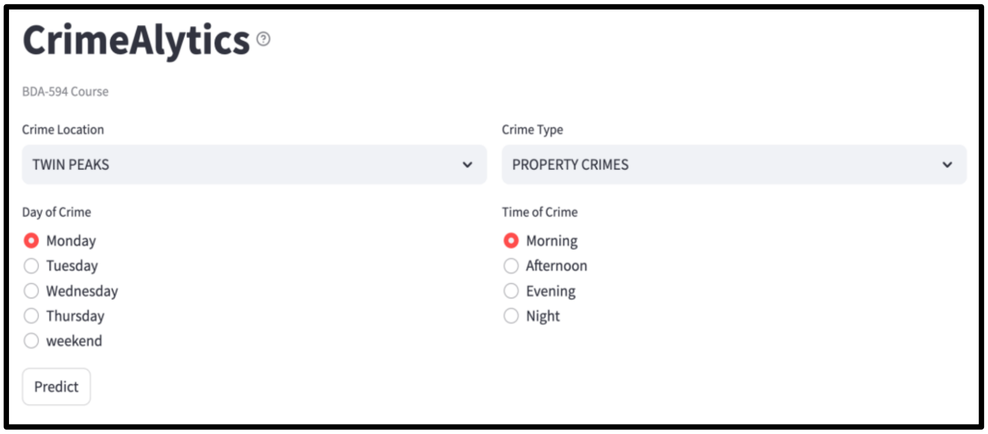


Fig 19: City Filter

- Crime Details Input: Victims can input specific details about the crime, including the location, premise, type, weapon used, and the time it occurred. Additionally, they can provide information about their race, sex, age, and the dates of the incident and reporting.



CrimeAlytics

BDA-594 Course

Crime Location: TWIN PEAKS

Crime Type: PROPERTY CRIMES

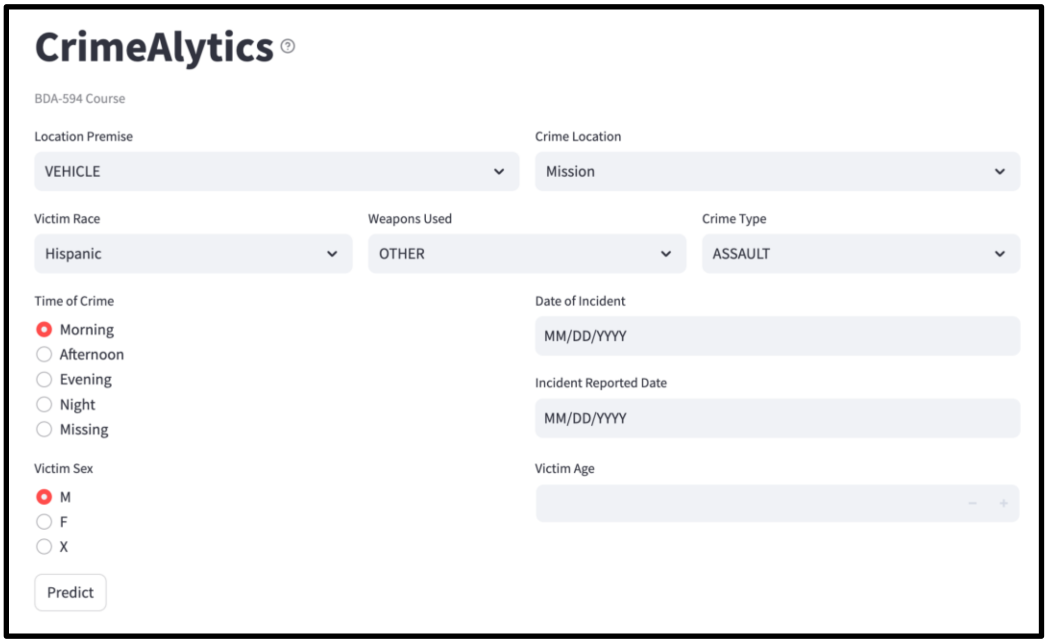
Day of Crime

● Monday
○ Tuesday
○ Wednesday
○ Thursday
○ weekend

Time of Crime

● Morning
○ Afternoon
○ Evening
○ Night

Fig 20: San Francisco Web Page



CrimeAlytics

BDA-594 Course

Location Premise
VEHICLE

Crime Location
Mission

Victim Race
Hispanic

Weapons Used
OTHER

Crime Type
ASSAULT

Time of Crime
● Morning
○ Afternoon
○ Evening
○ Night
○ Missing

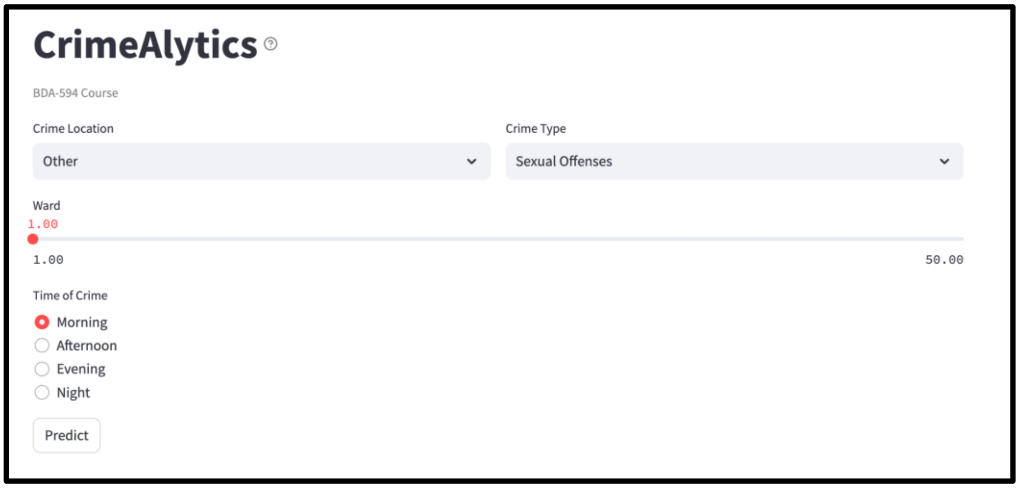
Date of Incident
MM/DD/YYYY

Incident Reported Date
MM/DD/YYYY

Victim Sex
● M
○ F
○ X

Victim Age

Fig 21: Los Angeles Web Page



CrimeAlytics

BDA-594 Course

Crime Location
Other

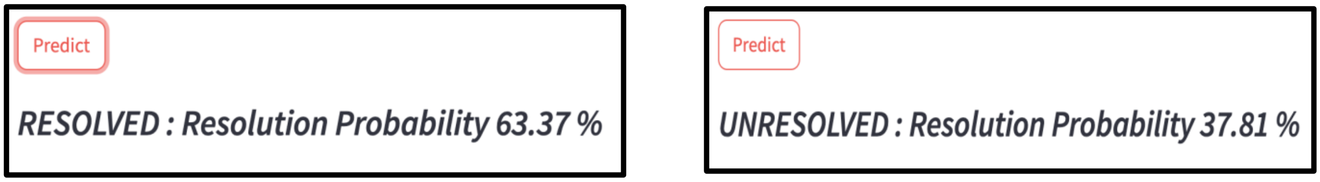
Crime Type
Sexual Offenses

Ward
1.00
1.00 50.00

Time of Crime
● Morning
○ Afternoon
○ Evening
○ Night

Fig 22: Chicago Web Page

- Prediction Analysis: After selecting all relevant features, users can click the predict button to initiate the analysis. The output provides insights into whether the crime is likely to be resolved or unresolved, along with the corresponding probability.



RESOLVED : Resolution Probability 63.37 %

UNRESOLVED : Resolution Probability 37.81 %

Fig 23: Prediction Outcomes

This user-friendly and intuitive application leverages machine learning models to provide victims with valuable insights into the potential resolution of their reported crimes. The transparent and interactive interface enhances accessibility and empowers users to make informed decisions regarding their cases.

Benefits of the App

The Crime-Alytics application offers a range of benefits, providing users with valuable insights and assistance in navigating the complexities of crime resolution. Here are some Key Advantages

- Empowering Victims by Offering a Digital Support System: The app empowers crime victims by giving them a tool to understand the potential outcomes of their cases. It provides a sense of control and transparency in a situation that can often be overwhelming and uncertain, offering assistance and clarity during what can be a challenging and emotionally charged experience..
- Informed Decision-Making: Users can make more informed decisions about their next steps based on the predicted resolution outcomes. This assists individuals in planning and preparing for the potential scenarios that may unfold in their cases.

- City-Specific Analysis: The app's city-specific design ensures that predictions are tailored to the unique characteristics of each location. This localized approach enhances the accuracy of predictions by considering the distinct features of different cities. Thus enabling lawmakers and other government officials to enable resources that increase the crime resolution rate.

In summary, the Crime-Alytics application is a valuable resource that leverages technology to provide meaningful support and insights to crime victims, contributing to a more informed and empowered community.

9. Conclusion

This project provided valuable information into the dynamics of criminal activities in the major U.S. cities of Los Angeles, San Francisco, and Chicago. The analysis of large datasets allowed us to uncover patterns, trends, and factors impacting crime resolution rates. It helped us to understand some of the challenges faced by law enforcement in solving criminal cases.

Our findings showed a significant amount of unresolved cases, echoing worries shared in recent news reports. The models developed as part of our project provide a proactive approach to determining case resolution, leveraging a possible tool for law enforcement agencies to properly use resources and improve crime resolution rates.

Additionally, the visualizations created made large datasets more understandable so members of our communities can engage with and comprehend the distinct crime patterns in their respective cities. As we continue to develop this project, the results of our research highlight the relevance of leveraging data analytics for improving public safety.

10. Challenges and Limitations

Few challenges we faced during the development of this application are given below

(i) Collecting Dataset: Identifying datasets across all cities in the US with our project requirements was a tedious task as in many cities they are not publishing complete data for few crimes. We had to gather all metadata details and then finalize on these 3 datasets.

(ii) Processing and Loading Datasets: Each city dataset has a huge set of rows for data at least ranging from 2018 to 2023. Thus the dataset was easily about hundreds of MBs to a few GBs. Thus loading them into the machine and mounting them for processing was a time consuming task.

(iii) Integrating Dataset: For the San Francisco dataset, the crime location feature was derived after joining with a different dataset. Similarly for Los Angeles, we had to join datasets for deriving data from 2018 to 2023. These required additional data validation processes to check the compatibility of the datasets.

(iv) Feature Extraction: As the datasets are derived from data sources of each city, the attributes present in each were inconsistent, which made it extremely difficult for us to identify similar features for building a comparative analysis among all cities. Hence, we were only able to correlate few features and retain few exclusive features for every city.

(v) StreamLit Application: While deploying and hosting the StreamLit application on its community cloud, the main issue was the memory space allocated for storing and processing (limited to 1GB). Hence, we had to optimize our package import and the files deployed on the server to avoid running out of memory.

11. Future Scope

The further development of this project would include additional cities across the country and delivering a more comprehensive understanding of crime and resolution patterns on a national level. The methodologies for this would follow our current process if data is available. This would include data collection, data processing, exploratory data analysis, and the creation of predictive models and visualizations. Potential challenges include the availability of data so collaboration with law enforcement agencies is crucial to ensure the gathering of accurate data, and the adaptation of models for each city if needed. Further steps involve seeking answers to detailed questions based on the initial analysis, like crime patterns over time, and peak times for specific types of crimes. Likewise, the team would conduct a deeper examination of crime hotspots or cold spots in each city to understand the factors that have an impact on crime rates. Lastly, we will analyze how different demographic variables such as age, gender, and race can be correlated with crime patterns which will likely involve working with crime datasets as well as demographic datasets. Taking this approach would allow our project to contribute further to crime analytics by offering insights that are detailed on a national level.

12. Tools Used

- Python Packages for complete Code Base
  - Seaborn
  - Sklearn
  - Pandas
  - Os module
  - Scipy
  - Pickle
  - Xgboost
  - Matplotlib
  - StreamLit
- StreamLit Community Cloud for website Deployment and Hosting
- Tableau for Data Analysis and Visualizations.
- Google Sites for Website Development and Hosting.

Fig 24: Tools Used for building the Application

13. Reference

- [LA, chicago Datasets] https://www.data.gov/

- [SF Datasets] https://www.sanfranciscopolice.org/
- Mandalapu, V. (2023). [Crime Prediction Using Machine Learning and Deep Learning](https://arxiv.org/pdf/2303.16310). arXiv.
- Shah, N. (2021). [Crime forecasting: a machine learning and computer vision](https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z). Vciba SpringerOpen.
- Jenga, K. (2023). [Machine learning in crime prediction](https://link.springer.com/article/10.1007/s12652-023-04530-y). Journal of Ambient Intelligence and Humanized Computing.
- Saeed, R. M. (2023). [A study on predicting crime rates through machine learning](https://www.degruyter.com/document/doi/10.1515/jisys-2022-0223/html?lang=en). De Gruyter.
- [Crime Prediction Using Machine Learning and Deep Learning](https://www.researchgate.net/publication/369623700_Crime_Prediction_Using_Machine_Learning_and_Deep_Learning_A_Systematic_Review_and_Future_Directions). ResearchGate.
- Dakalbab, F. (2022). [Artificial intelligence & crime prediction: A systematic](https://www.sciencedirect.com/science/article/pii/S2590291122000961). ScienceDirect.
- [Binary Classification](https://www.learndatasci.com/glossary/binary-classification/). LearnDataSci.
- [Target Variable](https://h2o.ai/wiki/target-variable/). H2O.ai.

- [Logistic Regression](https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/#:~:text=Logistic%20regression%20is%20a%20supervised%20machine%20learning%20algorithm%20that%20accomplishes,1%2C%20or%20true%2Ffalse.). Spiceworks.
- [Decision Tree Algorithm](https://scikit-learn.org/stable/modules/tree.html). Scikit-learn.
- [Introduction to Decision Tree Algorithm](https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/). Analytics Vidhya.
- [Random Forest Algorithm](https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/). Section.
- [Random Forest Algorithm Tutorial](https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm). Simplilearn.
- [XGBoost](https://www.nvidia.com/en-us/glossary/data-science/xgboost/). NVIDIA.
- [Extreme Gradient Boosting in Python](https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/). Machine Learning Mastery.
- [XGBoost - Extreme Gradient Boosting](https://www.geeksforgeeks.org/ml-xgboost-extreme-gradient-boosting/). GeeksforGeeks.
- [Streamlit](https://streamlit.io/). Streamlit.
- [Streamlit Tutorial](https://builtin.com/machine-learning/streamlit-tutorial). BuiltIn.
- [Predictive Crime Software](https://www.predpol.com/predictive-crime-software/). PredPol.

14. Team Members

- Anurima Saha [asaha8669@sdsu.edu]
- Maha Lakshmi [mpillalamarri9369@sdsu.edu]
- Fernanda Carrillo Escarcega [mcarrillo6082@sdsu.edu]
- Manisha Radhakrishna [mradhakrishna9449@sdsu.edu]

Thank You

—