Support Strategies for Victims of Crimes in Los Angeles: A Data-Driven Approach

By: Franky Araujo | Email: araujofranky@gmail.com

Objective

By decoding crime patterns, we can strategically channel resources to enhance safety and support in key areas. The objective here is to construct predictive tools that will anticipate victim demographics to enhance law enforcement practices, facilitate victim support initiatives, and enable focused efforts by service providers in critical areas

Overview

This report outlines a project that utilizes LAPD crime records to predict the age and gender of individuals affected by criminal activities in Los Angeles. To tackle this multifaceted task, two distinct models were developed: a Gradient Boosting classifier model for gender classification and a Gradient Boosting regressor model for age prediction. The methodology, parameters, and nuanced approaches applied to each model will be discussed in detail. Additionally, the report will present performance evaluation metrics to illustrate the rationale behind choosing Random Forest and Gradient Boosting models. Finally, the report will outline potential improvements for the models in the concluding section.

Quick Note: This report covers my data analysis journey with some techy details and code snippets. If you're short on time or just want the highlights, head to page 8 for the results and real-world impact.

Data Insights

The data was made available by the Los Angeles Police Department via this website. Data source: Los Angeles Open Data The data was a combination of Traffic Collision Data from 2010 to Present And Crime Data from 2020 to Present. Although the dataset was a good size at 293.9+MB, there were unique challenges present that needed to be navigated such as:

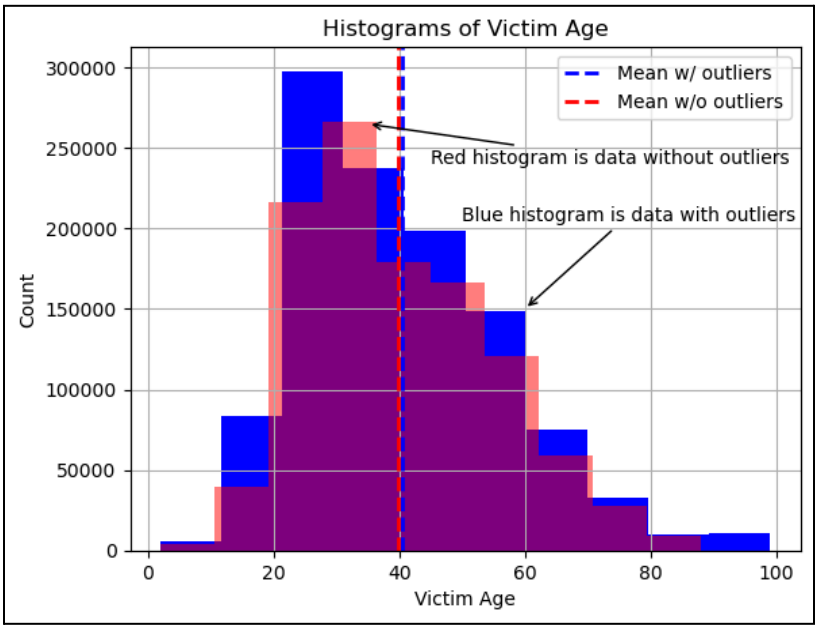
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1375881 entries, 0 to 1375880
Data columns (total 28 columns):
 #  Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0  Unnamed: 0          1375881 non-null  int64
 1  DR Number           1375881 non-null  int64
 2  Date Reported       1375881 non-null  object
 3  Date Occurred       1375881 non-null  object
 4  Time Occurred       1375881 non-null  int64
 5  Area ID             1375881 non-null  int64
 6  Area Name           1375881 non-null  object
 7  Reporting District  1375881 non-null  int64
 8  Crime Code          1375881 non-null  int64
 9  Crime Code Description  1375881 non-null  object
 10 MO Codes            1181546 non-null  object
 11 Victim Age          1290351 non-null  float64
 12 Victim Sex          1263940 non-null  object
 13 Victim Descent      1262997 non-null  object
 14 Premise Code        1374913 non-null  float64
 15 Premise Description 1374460 non-null  object
 16 Weapon Used Cd      271192 non-null   float64
 17 Weapon Desc         271192 non-null   object
 18 Status              779803 non-null   object
 19 Status Desc         779803 non-null   object
 20 Crm Cd 1            779793 non-null   float64
 21 Crm Cd 2            57524 non-null    float64
 22 Crm Cd 3            1919 non-null     float64
 23 Crm Cd 4            57 non-null       float64
 24 Address             1375881 non-null  object
 25 Cross Street        692955 non-null   object
 26 LAT                 1375881 non-null  float64
 27 LON                 1375881 non-null  float64

dtypes: float64(9), int64(6), object(13)
memory usage: 293.9+ MB
```

- The data, transcribed from paper reports, introduces discrepancies or errors during transcription, potentially affecting data accuracy and reliability.
- Mixed data types present a challenge, as difficulty in standardizing and analyzing diverse data types may lead to misinterpretations or inconsistencies in results.
- Text elements with multiple labels pose challenges, resulting in difficulty extracting and interpreting meaningful information accurately, impacting the precision of analyses.
- The dataset is limited to reported crimes which may lead to an incomplete representation of overall criminal activity, potentially skewing analysis results.
- The presence of pre-encoded values (i.e., crime codes and status codes) requires consideration to ensure accurate interpretation, as misinterpreting encoded information may lead to flawed conclusions.
- Finally, a consideration for the Covid pandemic and its potential impact on any trends had to be noted.

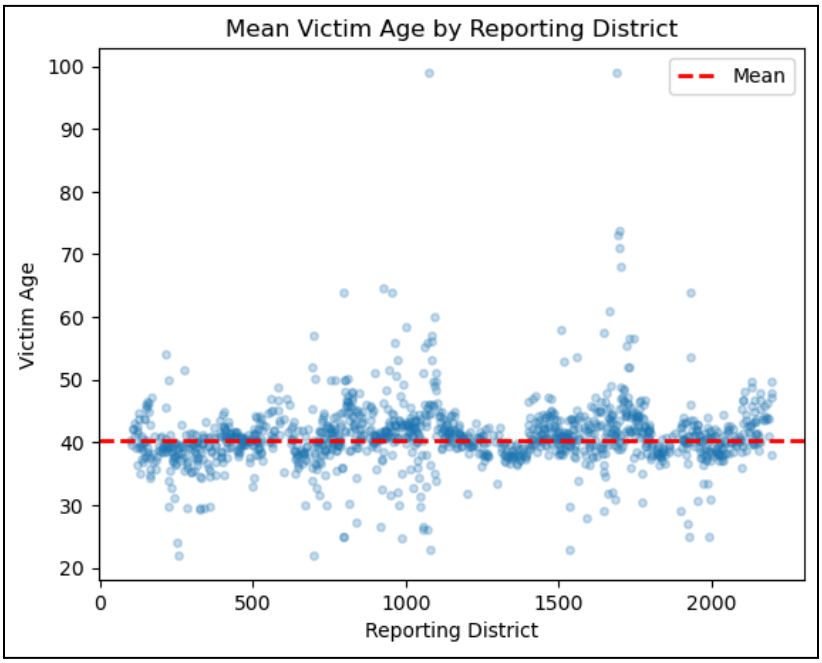
Addressing these challenges through comprehensive data preprocessing, exploration, and interpretation strategies was essential to maintain the integrity of the dataset and enhance the reliability of the analysis results

Initial Findings

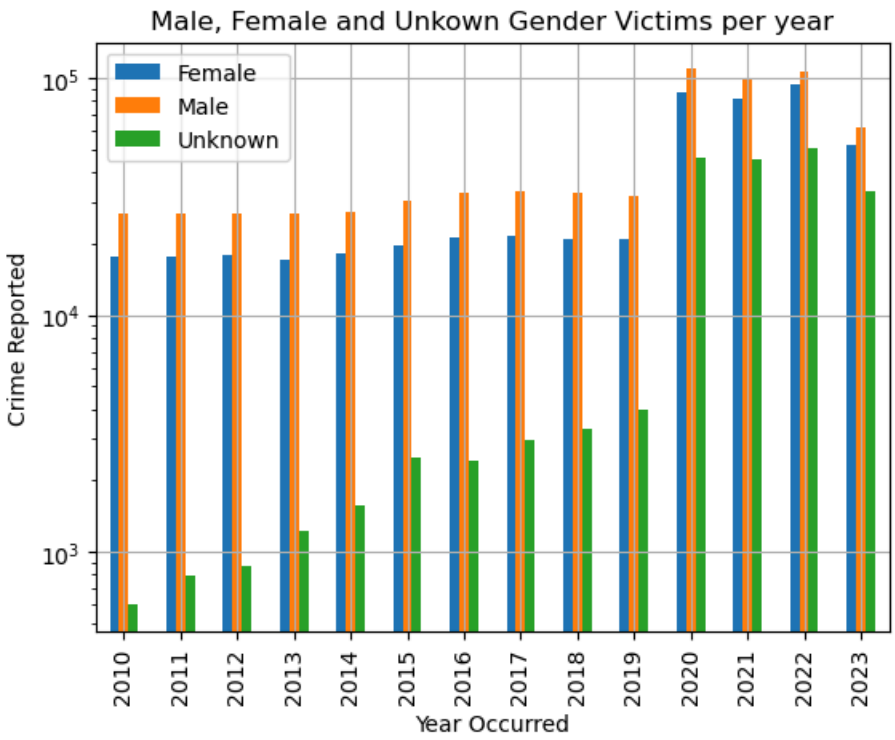
Fortunately, the raw data still presented insights during the exploratory data analysis step. For instance, the average age of victims in the dataset is around 40, as depicted in the overlay of histograms below.



This consistency was also seen across reporting districts in Los Angeles, as highlighted by the density of points along the horizontal red dashed line in the scatter plot below.



Another notable insight was that most of the reported crimes were involving a male victim (see graph below).



Note: The y-axis was scaled to enhance the visual comparison between bars and provide a more accurate representation of the patterns.

Initial insights from the raw data, including an average victim age of around 40 and a consistent male victim trend, prompted the shift to preprocessing. This marks the transition to the preprocessing steps.

The Preprocessing Approach

In this section, we will discuss various preprocessing steps, model selection, and training approaches, including:

- Data Cleaning
- Train/Test Split Approach
- Encoding & Scaling
- Using a Baseline Model
- Model Selection & Performance Metrics

Data Cleaning

To begin the preprocessing step, features with missing values were evaluated and the following approach was taken:

1. Removal of Features: Features like Crm Cd 4, Crm Cd 3, and Crm Cd 2, with over 90% missing values, were removed due to limited contribution.
2. Redundancy Check: Weapon Desc and Weapon Used Cd, both with 80% missing data, were deemed redundant. 'Weapon Desc' containing key information was retained.
3. Imputation for MO Codes: MO Codes, with 40% missing values, were imputed using empty strings, preserving present values for vectorization.
4. Careful Removal of Features: Features with <1% missing values (Premise Description, Premise Code) had corresponding rows removed.
5. Target Variable Imputation: Victim Age had missing values imputed with the mean for comprehensive representation.

Train/Test Split Approach

With a non-null dataset, the next step involved laying the groundwork for modeling. A Train/Test Split (70/30) was implemented, ensuring models will be trained on a representative 70% sample and assessed on an independent 30% set.

The target variable, 'victim sex,' being categorical, required label encoding before the split. As mentioned earlier, two models were developed for different target variables, requiring that the train_test_split be applied twice. Stratification was used to maintain balanced class distribution in both training and testing sets for model training and evaluation.

Encoding & Scaling

The encoding of categorical variables posed a challenge during the preprocessing phase. To address the diverse text values, a combination of one-hot, label, and word2vec encoding methods was employed.

For variables with a low number of unique values, such as 'Area Name' and 'Victim Descent,' one-hot encoding was applied. On the other hand, text-based features with higher uniqueness, including 'Crime Code Descriptions,' 'MO Codes,' 'Premise Descriptions,' and 'Addresses,' underwent word2vec vectorization.

To maintain transparency and facilitate future reference, feature names were preserved during encoding by adding prefixes. This allowed us to store an unscaled version of the encoded data for potential later use.

Since a combination of encoding methods was used for categorical variables, StandardScaler() was then applied to the numerical features . to contribute to the overall robustness and effectiveness of the training process.

Post-scaling, the preprocessed data, both scaled and unscaled, were combined and converted into dataframes for Victim Age and Victim Sex. These DataFrames were stored in CSV files, ready for further analysis.

Baseline Model & Performance Metrics

A performance benchmark was needed for subsequent model evaluation so a baseline model was established using a dummy regressor for predicting Victim Age and a dummy classifier for predicting Victim Sex. Metrics like MAE (10.42), MSE (206.23), RMSE (14.36), accuracy (39.45%), precision, recall, F1 score, and multilabel confusion matrix were computed. These metrics served as benchmarks for subsequent model evaluations.

```
Dummy Regressor - Mean Absolute Error: 10.417887179842625
Dummy Regressor - Mean Squared Error: 206.23319913904604
Dummy Regressor - Root Mean Squared Error: 14.360821673534074
Dummy Regressor - R-squared: -2.3358559531061474e-10
```

```
Dummy Classifier - Accuracy: 0.39445447167274356
Dummy Classifier - Precision: 0.39445447167274356
Dummy Classifier - Recall: 0.39445447167274356
Dummy Classifier - F1 Score: 0.39445447167274356
Dummy Classifier - Multilabel Confusion Matrix:
[[[164439  96094]
  [ 96395  55837]]

 [[107744 103227]
  [103269  98525]]

 [[303399  50627]
  [ 50284   8455]]]
```

Model Selection & Performance Evaluation

The preprocessing steps set the foundation for fitting, training, and evaluating regression and classification models. To enhance efficiency, the models were trained on sampled data as it reduced the training time to approximately 3 minutes. It's important to acknowledge this step, as the performance metrics presented here may differ from those obtained before hyperparameter tuning below within *The Models* section.

	Model	Accuracy
0	RandomForest	0.644039
1	GradientBoosting	0.669210
2	LogisticRegression	0.488214

	Model	MSE
0	RandomForest	199.234778
1	GradientBoosting	192.366243
2	LinearRegression	197.955067

The GradientBoosting model performed the best for both the regression and classifier models with an MSE of 192.36 and an accuracy of 66.92%.

The Models

Model Description(s)

The Gradient Boosting model was selected so two distinct models were developed: a Gradient Boosting Classifier model for gender classification and a Gradient Boosting Regression model for age prediction. The approaches and parameters employed in the models will be detailed in this section.

Imagine building a predictive team with friends. Starting with an initial guess, each new member (a weak predictor) learns from mistakes, enhancing overall predictions. This iterative process continues until the team becomes highly skilled, resulting in a powerful ensemble model that combines individual strengths for improved results.

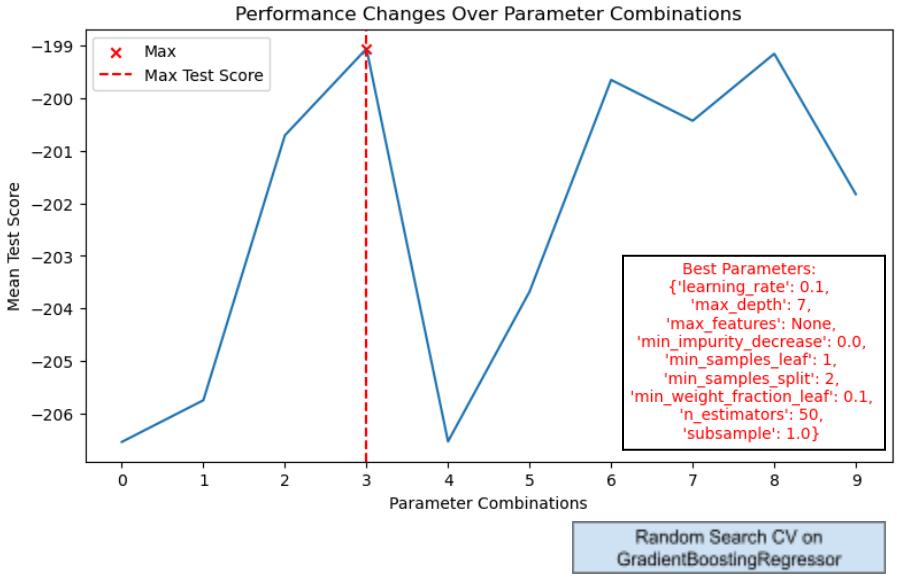
Hyper Parameter Tuning

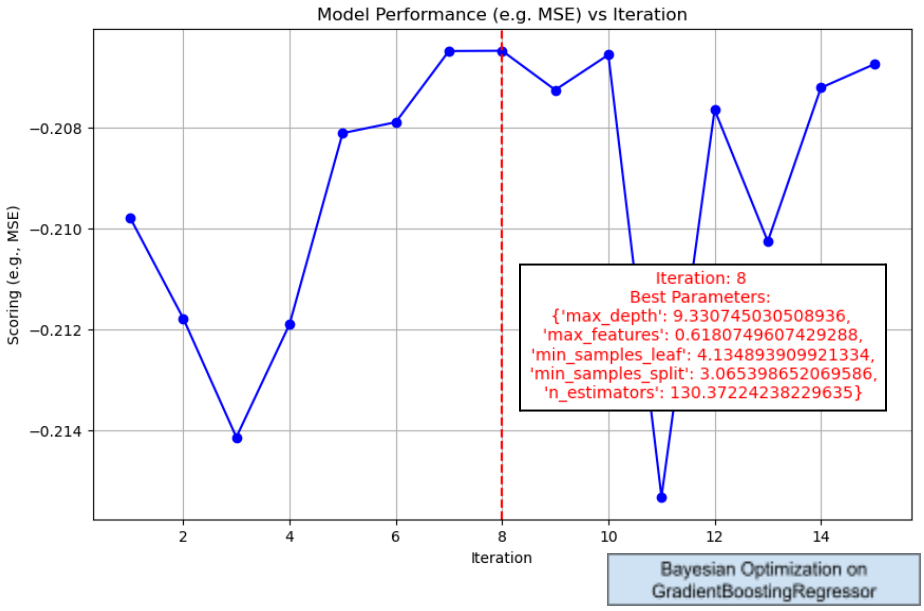
Hyperparameters are model settings that are set before training and impact the model performance. Using the example above, Imagine building a predictive team of friends. Starting with an initial guess, each new member (a weak predictor) learns from mistakes, enhancing overall predictions. This iterative process continues until the team becomes highly skilled, resulting in a powerful ensemble model that combines individual strengths for improved results.

Recall, the classifier model scored an accuracy of 66.92% the regression model had a Mean Squared Error (MSE) of 192.36. The purpose here is to improve those metrics to increase the predictive power of the models by selecting the best model parameters.

RandomSearchCV and BayesianOptimization were employed in this pursuit of better results.. RandomSearchCV randomly samples a set of hyperparameter combinations from a predefined list and evaluates them to find the combination that results in the best model performance.

To demonstrate a glimpse into this, here is a visualization showing how performance is measured as RandomSearchCV attempts different combinations of parameters.





Bayesian Optimization uses a probabilistic surrogate model to predict the next set of hyperparameters within a range of continuous numbers likely to improve performance. Think of having more guidance as to what range the best combination of hyperparameters will most likely be within.

Note: Notice the parameter values as floats since this optimization approach looks at values in between integers whereas RandomSearchCV will look at combinations of the int values you give it.

Model Performance

To maintain simplicity, I opted for a single metric for each model during the evaluation process. The regression model's performance was assessed using Mean Squared Error (MSE), while the classifier model's performance was measured using Accuracy.

Performance Results:

For the regression model, the Mean Squared Error (MSE) initially stood at 191.40 and demonstrated improvement after hyperparameter tuning, reaching a lower value of 184.98 with BayesianOptimization. Lower MSE values indicate enhanced accuracy in predicting numerical outcomes.

```
MSE for GradientBoost_reg is: 191.3965280777735
MSE for GradientBoost_reg with RandomSearchCV is: 187.86463081127204
MSE for GradientBoost_reg with BayesianOptimization is: 184.98259362367278
```

For the classification model, the accuracy fluctuated during hyperparameter tuning, starting at 0.6709, decreasing to 0.6660 with RandomSearchCV, and then increasing to 0.6763 with BayesianOptimization. While accuracy provides an overall measure of correct predictions, further analysis using metrics like Confusion Matrix, Precision & Recall, and AUC/ROC could offer more nuanced insights into the model's performance across different classes and thresholds.

```
Accuracy for GradientBoost_clf is: 0.6709968141678679
Accuracy for GradientBoost_clf with RandomSearchCV is: 0.6660254624301963
Accuracy for GradientBoost_clf with BayesianOptimization is: 0.67625888822938
```

Model Findings

Exploring the outcomes of the Gradient Boosting models, this section discusses key findings resulting from developing these models.

1. The models successfully decoded crime patterns in Los Angeles, providing actionable insights for strategic resource allocation. By anticipating victim demographics, law enforcement practices can be enhanced, and victim support initiatives can be more directed to critical areas.
2. The Gradient Boosting model demonstrated promising results in gender classification. Leveraging LAPD crime records, the model predicted gender, contributing to the project's objective of providing an understanding of victim demographics.
3. The Gradient Boosting model, tailored for age prediction, demonstrated promising results. By considering various factors, the model predicted victim age, facilitating more understanding of the age distribution among individuals affected by criminal activities.
4. Navigating through challenges such as discrepancies introduced during transcription, diverse data types, and text elements with multiple labels, the utilization of data preprocessing strategies is outlined in this report.
5. Recognizing the dataset's constraints, especially its focus on reported crimes, we explore potential biases and gain insights into the intricacies of interpreting encoded information. Considering the influence of the Covid-19 pandemic on crime trends adds depth to our understanding of potential contributing factors.

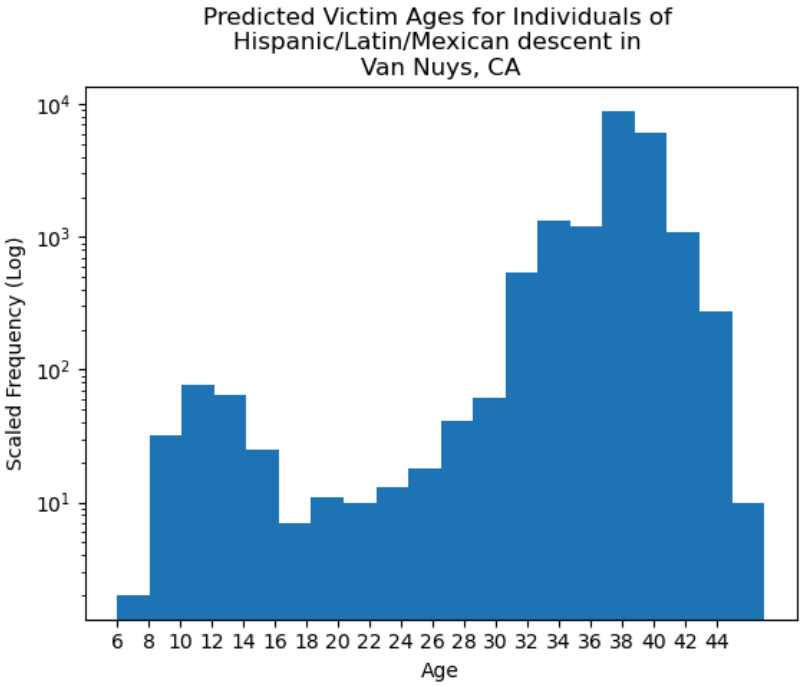
By translating data-driven predictions into strategic actions, the Gradient Boosting models have the potential to significantly impact safety and support strategies in Los Angeles. The findings offer a pathway for law enforcement, service providers, and community initiatives to collaboratively address the nuanced needs of crime-affected individuals.

Potential Impact

The predictive models developed for identifying victims' age and sex in Los Angeles hold tremendous potential for influencing resource allocation, community outreach, and strategic partnerships. The impact of these models is exemplified through two scenarios, showcasing how they provide valuable insights to optimize support services and guide decision-making for nonprofit organizations.

Scenario 1 | Crime Victims of Latin Descent in Van Nuys, CA

In this scenario, a nonprofit organization strategically utilized the age distribution predictive analysis to inform staffing and support service planning for the upcoming year. The model's ability to predict a significant concentration in the 30-40 age group served as a crucial guide for tailored program planning. This strategic insight allowed the organization to assess its capabilities, ensuring the provision of a diverse range of services to meet the specific needs of this demographic.

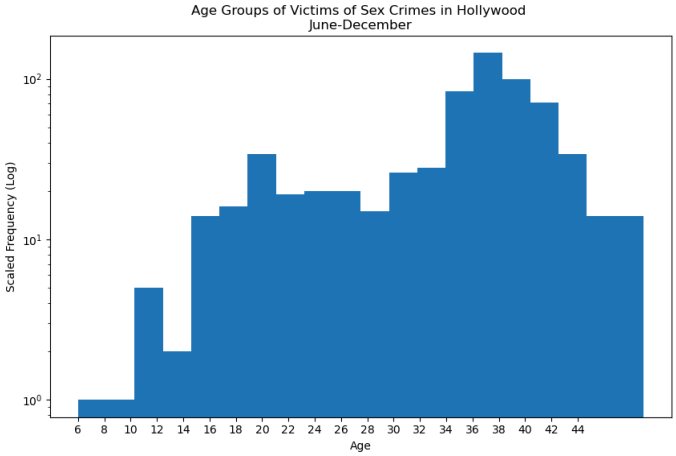


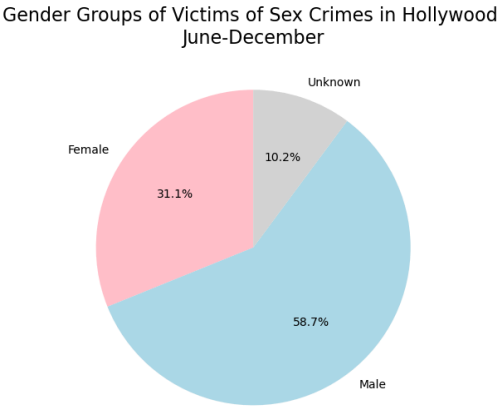
Furthermore, the identification of clusters in the 8-16 and 18-26 age groups prompted the nonprofit to explore these demographics further, ensuring a comprehensive understanding of community needs. The distinct focus on individuals under 20 underscored the importance of providing tailored services for the younger community members.

Significance: The age distribution analysis, despite acknowledging potential model limitations, emerged as a valuable tool for various stakeholders, including NGOs, law enforcement, victim service providers, and government funding bodies. By leveraging these insights, organizations can enhance their ongoing efforts to address community needs effectively. This strategic approach positions the nonprofit to adapt to the dynamic needs of its diverse clientele, contributing to the overall welfare of the community served.

Scenario 2 | Sex Crime Victims in Hollywood, CA

In this example, a nonprofit organization tailored its programs to meet the distinct needs of diverse client groups, recognizing the variations between minors and adults. The model played a pivotal role in decision-making regarding funding allocations for programs, considering factors such as the timing of services, service areas, and support for victims of specific crime types.





Examining Sex Crimes in Hollywood, CA, between June and December revealed a distribution of 31.1% female, 58.7% male, and 13% unknown victims. The age distribution analysis displayed peaks and dips, such as a slight dip between ages 26-29 and a subsequent spike between 30-38. The accompanying pie chart illustrates this gender distribution. The clustering within age groups may provide additional insights into the demographics of the area and how to best render services.

Significance: Aligning services with these insights, particularly for specific age groups, emerged as a strategic move to support ongoing efforts in addressing community needs. The age distribution analysis served as a foundation for targeted program planning, while the predictive model highlighted demographics and timing. This, in turn, contributed to the nonprofit's strategic service delivery efforts, ensuring impactful and tailored support for diverse communities.

These scenarios showcase the tangible impact of predictive models in guiding strategic decisions and enhancing the overall effectiveness of support services for victims in Los Angeles.

What now?

As I currently work in the victim services space, I thought I'd share what further actions I intend to take or recommend to the agency as I consider applications (and variants) of these models.

- Regularly update and enhance the predictive models by incorporating fresh data and accounting for emerging patterns. Continuous refinement ensures that the models stay adaptive to the evolving dynamics of victim demographics.
- Facilitate collaboration and data sharing among victim service providers, law enforcement, and relevant organizations. Establishing a collaborative platform can foster a collective understanding of the nuances in victim demographics, leading to more comprehensive and impactful support strategies.
- Engage directly with the communities being served. Solicit feedback, concerns, and insights from community members to ensure that support services align with their evolving needs. Community engagement is crucial for building trust and tailoring interventions for maximum effectiveness.
- Maintain a strong commitment to ethical considerations and privacy standards in handling sensitive victim data. Regularly review and update data protection protocols to ensure compliance with evolving legal and ethical standards.
- Launch public awareness campaigns to educate the community about available victim services and resources. Increased awareness fosters a proactive approach to seeking help and contributes to the overall well-being of potential victims.

I encourage continuous collaboration among nonprofits, law enforcement, and relevant agencies to refine and adapt these models. This ongoing partnership ensures the sustained effectiveness of our support systems, fostering a more resilient and responsive approach to addressing the evolving needs of victims in the community.

Project Links

GitHub Repository

- Link to Your GitHub Repository

Jupyter Notebooks

1. Data Wrangling Initial Exploration & Cleaning
2. Exploratory Data Analysis
3. Pre-processing and Training Data Development
4. Modeling

Data

- Link to Raw Data - Crime Data from 2020 to Present | Los Angeles
- Link to Raw Data - Traffic Collision Data from 2010 to Present | Los Angeles
- Link to Project Data

Helpful Resource Links

- RandomSearchCV Explained
- Bayesian Optimization Explained