Enhancing Public Safety: A Data-Driven Predictive Analysis

Table of Contents

Executive Summary

Background

The research presented in this report aims to enhance public safety through predictive analysis of crime data in Los Angeles. The motivation for this report is the concern of face recognition in predicting if a person will become a criminal or not. Thus this study pivots towards predicting victimhood by utilizing historical criminal data and assisting authorities in allocating resources efficiently and developing policies to enhance public safety. By doing so, we will be focusing on preventing crimes rather than accusing individuals based on a model that might have racial biases.

Methodology

This report uses a combination of qualitative and quantitive research techniques to identify the factors influencing an individual's risks of encountering a part 1 or part 2 crime within the state of Los Angeles in the United States. The dataset analyzed for this report was obtained from DATA.GOV (United States Government Data Website) and consisted of 499,999 records across 18 columns with variables such as victim age, gender, demographic, crime location, etc. Meticulous data preparation was performed by first using Excel to downsize the database due to software limitations, followed by identifying missing values on Python and importing them into SAS Viya for advanced cleansing and data refining. A logistic regression model was made to identify the significant variables that influence a part 1 or part 2 crime. Furthermore, predictive models such as Logistic Regression, Decision Trees, Support Vector Machines, Gradient Boosting, and Artificial Neural Networks were developed to identify the best-performing model. The dataset in the model was partitioned into 3 sets (60% training, 30% validation, and 10% testing) to ensure robustness and avoid overfitting. The random forest model emerged as the best-performing one as it scored the highest KS(Youden) value.

Key Findings

- Part 1 Crimes are the most common.
- Hispanics/Latin/
- Mexican's most victimized group
- The start of the year experienced the most crimes.
- Noon periods experienced the highest crimes.
- Active members are more susceptible to being a victim

Key Recommendations

- Increased resources surveillance
    - Noon Time regions
    - Start of the year (Holiday Period)

- Educating Individuals

  - Preventative measures
  - Surrounding Awareness
  - Reducing the risk of becoming victims

- Further Investments into Data Analytics
  - Advanced Predictive Models
  - Clean and up-to-date data
  - Continuous evaluation improvement

- Targeted Support for High-Risk Demographic

Introduction

Background

The motivation for this study originates from a research conducted at the Harrisburg University that claimed to develop a face recognition program that could predict if a person would become a criminal or not. However, for ethical reasons, the study was condemned by many researchers (Fussell, 2020). Thus, instead of predicting if a person will become a criminal or not and raising ethical concerns, this research will attempt to predict whether an individual will become a victim of part 1 or part 2 crime by leveraging historical crime data. In doing so, this report aims to predict future crimes, enhance public safety, and provide valuable insights to authorities to allocate their resources accordingly and make informed decisions.

Literature Review

The earliest machine learning model was developed in the 1950s by Arthur Samuel who developed a model that would calculate the winning chances for either side in checker (Bell 2022). Following this, Frank Rosenblatt developed the "Perceptron", a machine-learning model that identified the letters of the alphabet and is said to be the prototype of the modern artificial neural network (Rosenblatt 1957). Moving to the 21st century and the introduction of Big Data, new methods of data crunching were required which gave rise to the MapReduce and Hadoop technology by Google enabling us to process huge amounts of data (Ghazi & Gangodkar 2015). During this time, Nvidia emerged as a monopolist with their massive breakthrough in the GPU world (Pandey et al. 2022) and the price reduction of RAM fueled the development of more advanced and resource-heavy models (Dakalbab et al. 2022). Since then, Predictive machine learning models have gained massive success in fields such as face recognition (Basystiuk et al.), Voice recognition (Tahseen Ali et al.), automatic driving (Bachute & Subhedar 2021), etc.

The Federal Bureau of Investigation in the United States has divided crimes into two categories Part 1 crimes including Forcible rape, Robbery, Aggravated assault, etc, and Part-2 which include Fraud, Forgery and counterfeiting, and vandalism (FBI 2011). Furthermore, A criminal dataset is typically built on 2 aspects. The location aspect (where the crime occurred, the unemployment rate, population, etc) (Mittal et al. 2018). Whereas the other aspect talks about the crime itself (victim used, the weapon used, the day the crime occurred, etc) (Mary Shermila, Bellarmine & Santiago 2018). The early stages of research in criminology focused on the

psychological and sociological impacts they have on individuals and societies (Yildirimer, 2023). Other researchers focused on why crimes occur and identifying behavioral patterns (N. Mahmud et al. 2016), (Canter and Youngs, 2016). Since the introduction of advanced data analytics and machine learning, criminology has shifted towards predictive analysis with the most common one being the "Hotspot Analysis" (Butt et al. 2020) where historical crime data is uploaded onto a map allowing authorities to allocate resources to criminal hotspots. A study conducted by (Kim et al. 2018) modified the use of this heatmap approach to predict areas most likely to experience crimes in the city of Vancouver using boosted decision trees. The *"Domain Awareness System", "Predpol" and "Stingray (cellphone tracking)"* are similar technologies that are being used by the New York Police Department to transmit real-time data to identify hotspots and allocate resources accordingly (Meijer and Wessels, 2019), (Shah, Bhagat, and Shah, 2021). Some researchers have also modified machine learning models to cater to specific crimes. For instance, (Srivastava et al. 2008) in their research utilized the Hidden Markov Model to cater to credit card scams.

Methodology

Research Questions

What factors (environmental, Demographic, Location, Presence of a weapon, etc.) most effectively predict an individual's susceptibility towards part-1 or part-2 crime, and how machine learning models can be used to enhance the accuracy of these predictions?

Data Collection/ Preparation

Data Background

The dataset for this research was obtained from the United States Government Website (DATA.GOV). It consists of details on the crimes, arrests, and victims from the City of Los Angeles. The dataset has a total of 18 columns and 499999 rows.

Data Preparation

Upon acquisition, the dataset underwent a rigorous cleansing process to prepare it for further analysis. Initially, the dataset was imported into Excel to manage the dataset's dimensions. Due to the import file size limitations of the analytical software, redundant columns were removed whilst keeping the columns that were crucial for the study.
Furthermore, the dataset was imported into Python to identify missing values. The following missing values were identified:

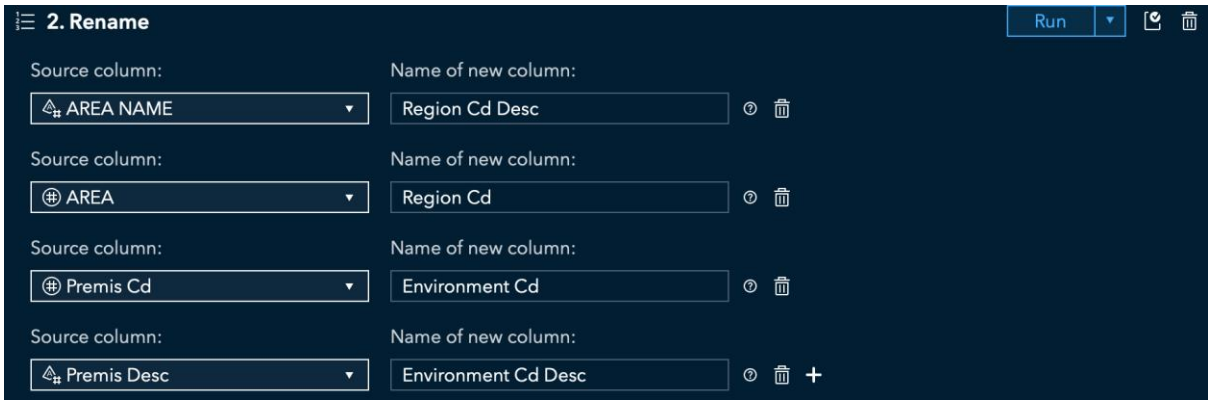
```
df.isnull().sum()

DR_NO                0
DATE OCC             0
TIME OCC             0
AREA                 0
AREA NAME            0
Rpt Dist No          0
Part 1-2             0
Crm Cd               0
Crm Cd Desc          0
Vict Age             0
Vict Sex         66836
Vict Descent     66840
Premis Cd            5
Premis Desc        393
Weapon Used Cd  332053
Weapon Desc     332053
Crm Cd 1             7
LOCATION             0
dtype: int64
```

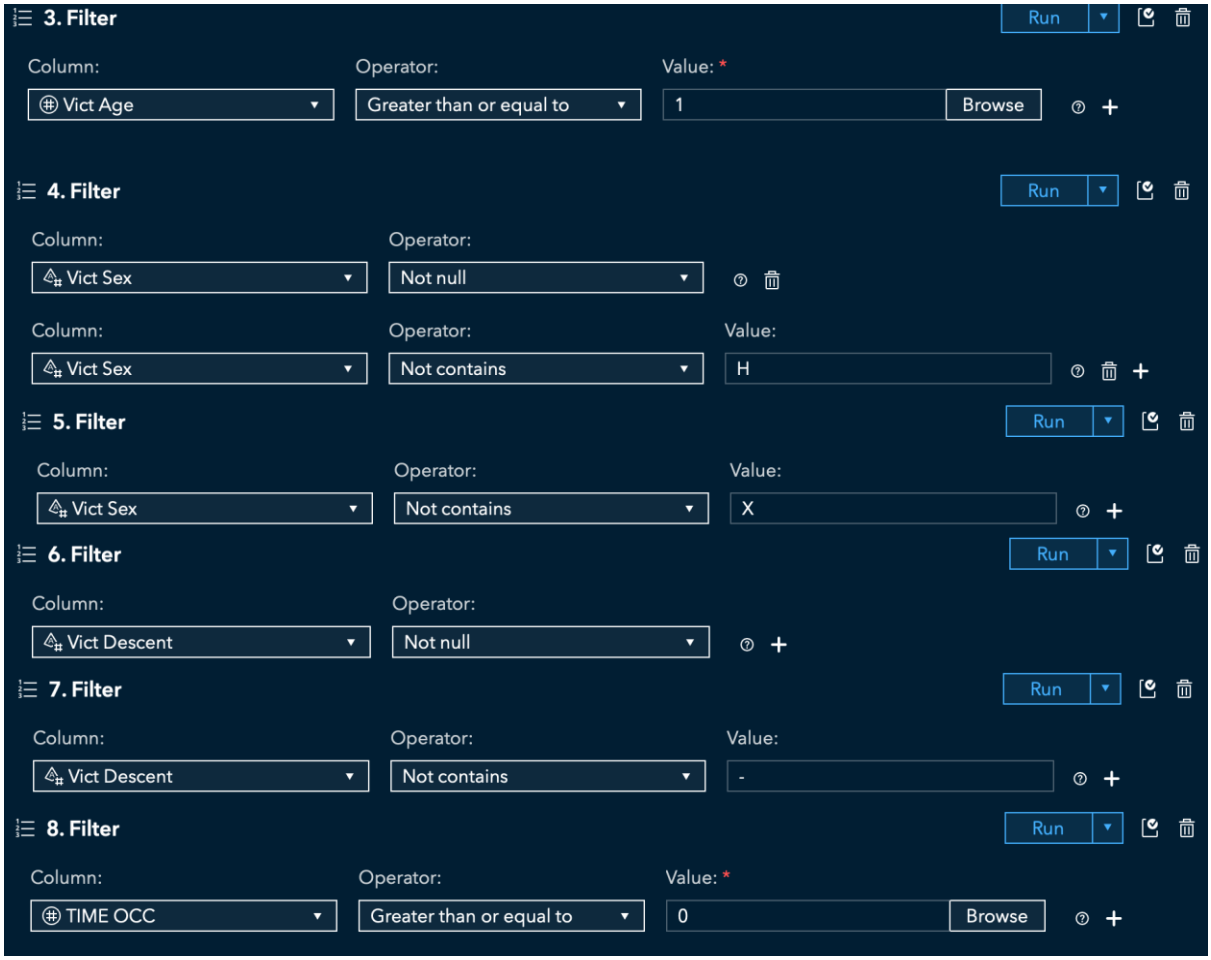
Following the missing value identification, The dataset was imported into SAS Viya for further cleansing. This cleansing process included:

- Removal of LOCATION and CRM cd 1

- Renaming of AREA NAME, AREA, Premis Cd, Premis Desc (For ease of understanding)



- Filtering out the missing values of the following tables (Weapon desc missing values were not cleaned as it means "No weapons were used in the crime")



Lastly, the key variables chosen for this study include:

- **Part 1-2** – Indicated the type of crime. (Dependent Variable)
- **DATE OCC** – *The date the crime occurred.*
- **Environment Cd Desc** – *The type of structure, vehicle, or location where the crime occurred.*
- **Region Cd Desc** – *The geographical area within the LAPS's jurisdiction.*
- **TIME OCC** – *The time when the crime occurred.*
- **Vict Age** – *The victim's age*
- **Vict Descent** – *The victim's Ethnicity (Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian)*
- **Vict Sex** – *Victims Gender*
- **Weapon Desc** – *The description of the weapon used.*

Data Analysis

Exploratory Analysis
The refined dataset was then used to generate charts and graphs to understand the data structure, and distribution and gain some insights into the dataset.

Model Development
A deeper and more comprehensive analysis was done using SAS Viya's analytical capabilities to generate sophisticated machine-learning models that would enable predictive analysis. The models that were generated for this report include:

- Logistic Regression (To explore the relationships between the dependent and independent variables)
- Decision Tree
- Support Vector Machine
- Gradient Boosting
- Artificial Neural Network

These models were then run through a pipeline for model comparison to identify the best-performing model which was then deployed as the final model.

Limitations

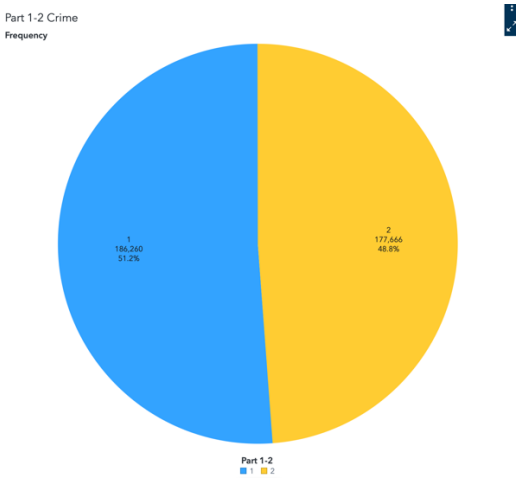
Data Limitations
One of the biggest data limitations for crime analysis studies is the "Bias in Reporting" limitation. This means that some crimes are underreported as victims are reluctant to come forward whilst other crimes are overemphasized. This can lead to the machine learning model to have inaccurate predictions.

Ethical Concerns and Privacy
Machine learning models are susceptible to biases present in the training set that can lead to unfair and wrong predictions for certain demographics. Additionally, the use of individual personal data such as demographic, age, gender, and region raises privacy concerns. However, the selected dataset does not contain the names of the victims which keeps their privacy intact.

Results

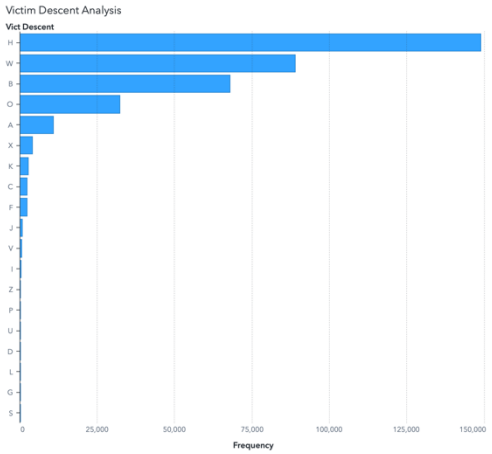
Exploratory Analysis

What Type of Crime has the highest occurrence?



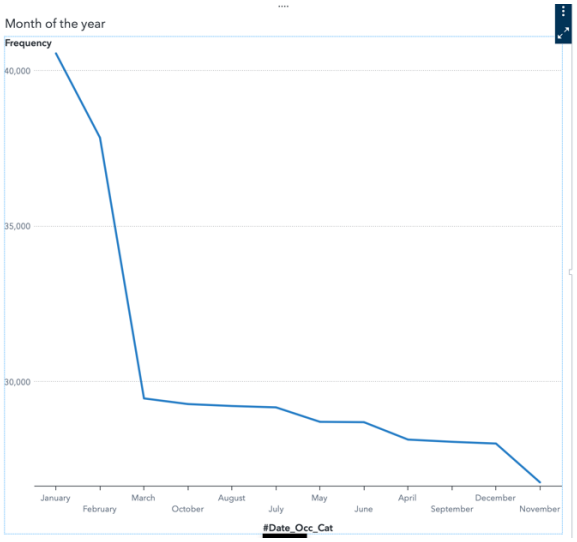
The results show us that the crimes categorized under Part 1 were the most frequent occurring 186,260 times, whereas Part 2 crimes had a total frequency of 177,666. This shows us that more severe crimes (Forcible rape, Robbery, Aggravated assault, etc) are statistically more occurring. But the distribution is almost 50-50.

What Demographic has the highest Victim rate?



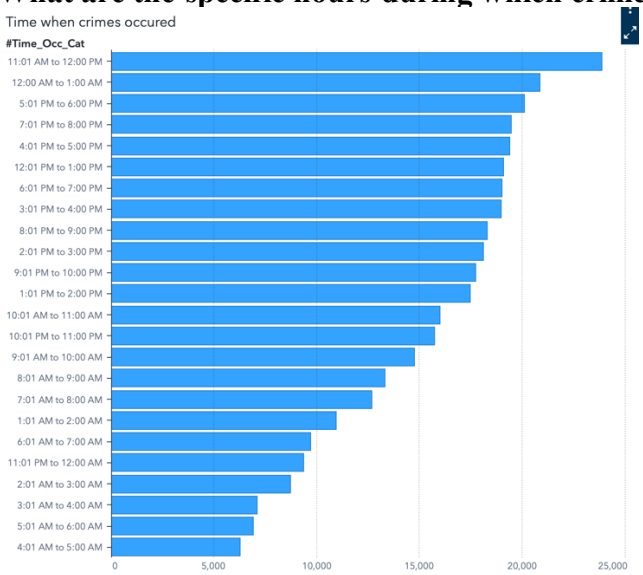
The data analysis revealed that “Hispanics/Latin/Mexicans” (149,011) experienced the highest number of crimes followed by “White” (89,100), “Black” (68,017), and “Other Asians” (32,287).

What Month of the Year had the most Crime Rates?



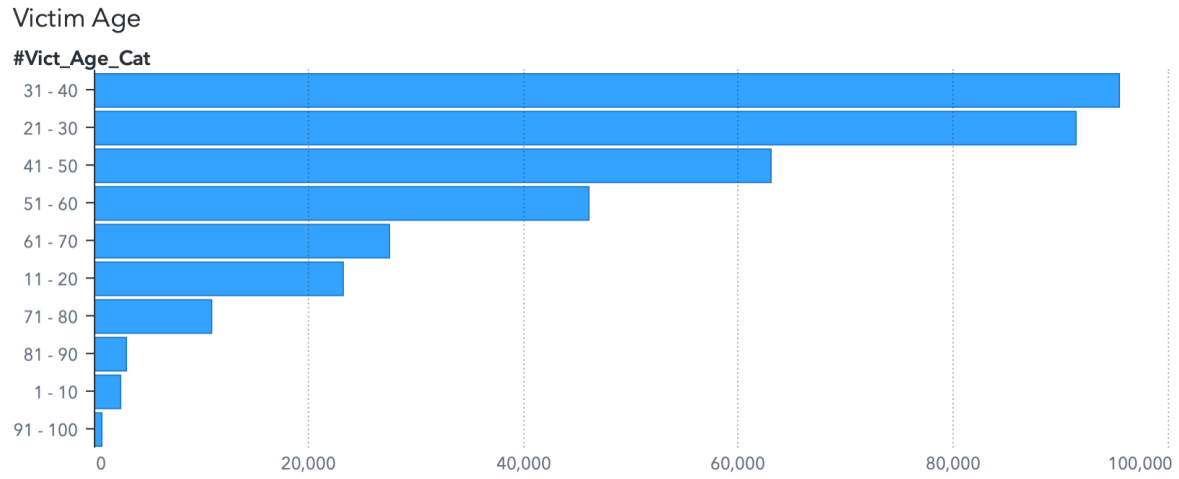
Upon analyzing the Data, it was found that the start of the year (January, February, and March) experienced the most crime rates.

What are the specific hours during which crime rates are highest?



A surprising discovery was made upon analyzing the time when the crimes occurred. One would expect night times to have the highest density of crimes. However, it was found that the highest density of crimes was experienced during the Noon period (11:01 Am to 12:00 Pm) (23,872) which was then followed by the expected Night-time (12:00 Am to 1:00 Am) (20,861).

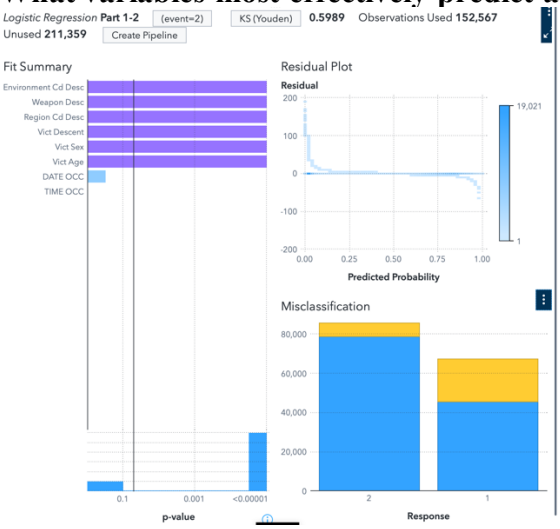
What age group experiences the most crimes?



Individuals between the age of 31 – 40 (95,430) experienced the most crimes followed by 21 – 30 (91,440) and 41- 50 (63,011). This shows us that the young and middle-aged populations (active members of socity) are more susceptible to becoming victims of crime.

Model Development

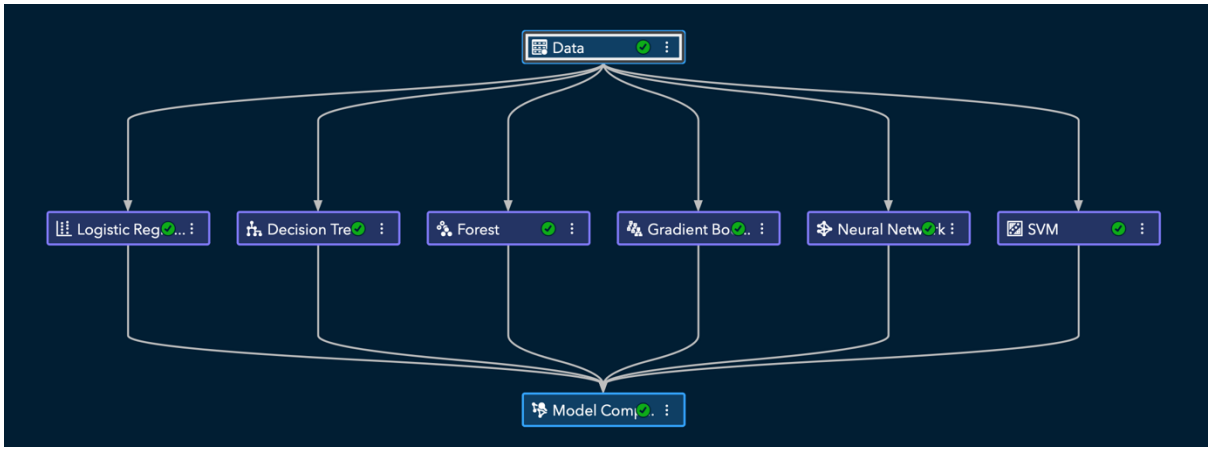
What variables most effectively predict a part 1-2 crime?



A logistic regression analysis was performed to identify the variables that significantly influenced the likelihood of an individual experiencing a part 1-2 crime. The analysis revealed several critical predictors such as Environment Cd Desc, weapon Desc, Region Cd Desc, Vict Descent, Vict Sex, and Vict Age resulted as statistically significant variables with a p-value of <0.00001 whereas DATE OCC and TIME OCC are comparatively not significant.

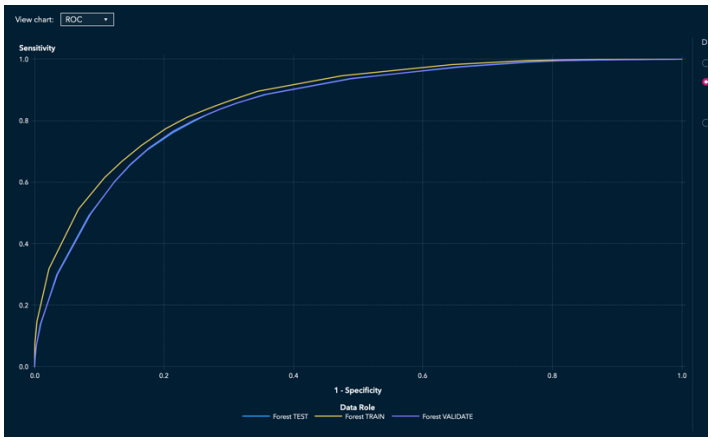
Model Comparison

Upon analyzing the significant variables, a pipeline was created with multiple machine-learning models (Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boosting, and Artificial Neural Network) built to identify the best-performing model. The dataset was divided into the Training (60%), Validation (30%) and testing (10%) sets.



Result

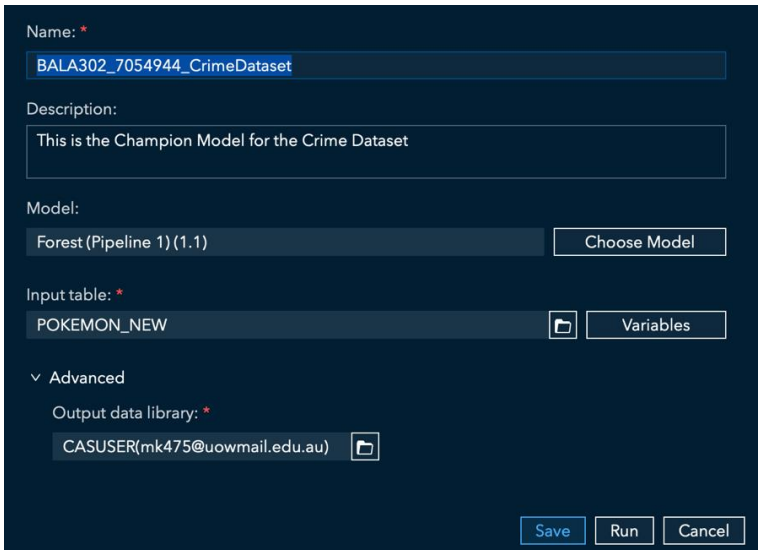
Algorithm Name	False Positive Rate	Name	Misclassification Rate	KS (Youden)	Area Under ROC	Champion
Forest	0.244926446902	Forest	0.22449372132	0.553061271531	0.850196230818	TRUE
Neural Network	0.359765918608	Neural Network	0.439809853543	0.116510323864	0.558255161932	FALSE
SVM	0.112047675293	SVM	0.336987882285	0.315443211838	0.707365298657	FALSE
Logistic Regression	0.114785783314	Logistic Regression	0.33624598137	0.316800866093	0.716932463242	FALSE
Decision Tree	0.3200901965	Decision Tree	0.260764432721	0.481339420206	0.800966499549	FALSE
Gradient Boosting	0.233490819285	Gradient Boosting	0.247987250295	0.503324625078	0.823296925893	FALSE



The results show us that the Forest model is the best-performing one with a KS (Youden) score of 0.5531 and a misclassification rate of 0.2245. The ROC curve also complements the KS score.

Deployment

Lastly, the Forest Mode model was deployed for public use.



Output Table

Into: Part 1-2	Predicte...	Predicte...	Weapon Desc	Region Cd ...	Vict Sex	Vict Age	TIME OCC	Vict Descent	DATE OCC
1	0.6968317529	0.3031682471		West Valley	M	50	1300	H	2022 11 12T12 00 00 AM
1	0.695794786	0.304205214		Topanga	M	74	1745	H	2022 06 09T12 00 00 AM
1	0.7340372671	0.2659627329	KNIFE WITH BLADE OVER 6 INCHES IN LENGTH	Wilshire	M	59	1635	O	2022 10 22T12 00 00 AM
1	0.7771341873	0.2228658127		Olympic	M	24	1840	W	2022 10 17T12 00 00 AM
1	0.7086417751	0.2913582249		Northeast	M	62	1100	W	2022 04 30T12 00 00 AM
1	0.6960131903	0.3039868097		77th St	M	36	1000	H	2022 11 23T12 00 00 AM
1	0.6704791256	0.3295208744		Southeast	F	34	430	H	2022 04 04T12 00 00 AM
1	0.7138286779	0.2861713221		West la	F	73	800	W	2022 10 21T12 00 00 AM
1	0.6980531503	0.3019468497	HAND GUN	Southeast	F	19	240	B	2022 01 23T12 00 00 AM
2	0.1856187754	0.8143812246	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	Southeast	M	14	1500	B	2022 08 30T12 00 00 AM

Discussion

The comprehensive analysis conducted on various variables of the crime dataset has highlighted significant points. This enables us to gain a deeper insight into when crimes occur the most and what variables significantly affect the type of crime. For instance, it was found that Part-1 crimes occurred the most. Indicating a higher frequency of more serious crimes such as Forcible rape, Robbery, Aggravated assault, etc. Secondly, The demographic analysis tells us that Hispanics/ Latins/Mexicans are the most victimized groups which can be explained due to their socioeconomic and regional factors.

The timeline analysis revealed that the start of the year (January, February, and March) experienced the highest number of crimes possibly due to seasonal economic changes and post-holiday periods contributing to an increase in the crime rate. A shocking discovery was made when the highest crime rate density was recorded around the noon period contradicting the normal assumption that most crimes occur at night times. This trend suggests that criminals might be exploiting busy mid-day periods due to lower guard levels and high potential victim density.

The age group analysis indicated that individuals between the ages of 21 and 40 are the most susceptible to crimes. This analysis highlights the impact of an individual's social life on the likelihood of them becoming a victim of a crime as individuals between these age groups tend to have an active lifestyle with a lot of social exposure and a tendency to engage in nighttime activities.

The Logistic regression model revealed significant variables that impact the likelihood of a part 1 or 2 crime and the pipeline analysis revealed the Forest model to be the best performing and most fit among several other models. The ROC curve revealed that the model is stable and is likely to have a similar performance on unseen data suggesting the lack of overfitting. However, the results presented limitations as the dataset contained missing values that needed to be removed in turn reducing the training size. Additionally, the misclassification rate although relatively low (0.2245) suggests room for improvement and refinement.

In conclusion, this study highlights the potential of a data-driven approach to enhance public safety by predicting whether an individual is subjected to a part-1 or part-2 crime. This form of study is more ethical as compared to predicting whether an individual will become a criminal or not.

Recommendations

Based on the results from the comprehensive analysis of the crime dataset of Los Angeles, the following recommendations are made to enhance public safety:

Increased Midday patrol
Given the surprising discovery of high-density crime rates during the noon period, the LAPD should invest more resources and officers during midday, especially in high-density areas where potential victims are the highest.

Targeted Support for High-Risk Demographic
The study showed that Hispanics/ Latins/Mexicans experienced the highest number of crimes. Thus, law enforcement agencies can look into investing more resources into communities with high densities of these demographics and focusing on preventive measures, educating these individuals on self-defense, and providing assistance

Seasonal Crime Preventive Measures
Since the start of the year experiences the most crimes, law enforcement agencies should organize crime preventive measures that are tailored to these times. Examples of such initiatives are awareness campaigns and increased patrol.

Education on Public Safety
Public safety and education campaigns can be held to educate young individuals on public safety, evaluate their surroundings, and reduce the risk of them becoming victims.

Utilizing Predictive Models
The performance of the Forest model shows us the significant impact machine learning models can have in crime prediction. Thus the Los Angeles law enforcement agencies can develop more predictive models by utilizing additional and more accurate data ethically and responsibly. Additionally, proper and continuous evaluation of the model must be conducted and clean new data should be fed into the model for more accurate predictions. Lastly, they should address data limitations and ensure that no biases are present in the training set.

References

- Bachute, MR & Subhedar, JM 2021, 'Autonomous Driving Architectures: Insights of Machine Learning and Deep Learning Algorithms', *Machine Learning with Applications*, vol. 6, p. 100164.
- Basystiuk, O, Melnykova, N & Rybchak, Z n.d., *Machine Learning Methods and Tools for Facial Recognition Based on Multimodal Approach*.
- Bell, J 2022, 'What Is Machine Learning?', *Machine Learning and the City*, pp. 207–216.
- Butt, UM, Letchmunan, S, Hassan, FH, Ali, M, Baqir, A & Sherazi, HHR 2020, 'Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review', *IEEE Access*, vol. 8, pp. 166553–166574.
- Canter, D & Youngs, D 2016, 'Crime and society', *Contemporary Social Science*, vol. 11, no. 4, pp. 283–288.
- Dakalbab, F, Abu Talib, M, Abu Waraga, O, Bou Nassif, A, Abbas, S & Nasir, Q 2022, 'Artificial intelligence & crime prediction: A systematic literature review', *Social Sciences & Humanities Open*, vol. 6, no. 1, p. 100342.
- FBI 2011, *Offense Definitions*, FBI.
- Fussell, S 2020, *An Algorithm That 'Predicts' Criminality Based on a Face Sparks a Furor*, Wired.
- Ghazi, MR & Gangodkar, D 2015, 'Hadoop, MapReduce and HDFS: A Developers Perspective', *Procedia Computer Science*, vol. 48, pp. 45–50.
- Kim, S, Joshi, P, Kalsi, PS & Taheri, P 2018, 'Crime Analysis Through Machine Learning', *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*.
- Mary Shermila, A, Bellarmine, AB & Santiago, N 2018, *Crime Data Analysis and Prediction of Perpetrator Identity Using Machine Learning Approach*, IEEE Xplore, pp. 107–114.

- Meijer, A & Wessels, M 2019, 'Predictive policing: Review of benefits and drawbacks', *International Journal of Public Administration*, vol. 42, no. 12, pp. 1031–1039.
- Mittal, M, Goyal, LM, Sethi, JK & Hemanth, DJ 2018, 'Monitoring the Impact of Economic Crisis on Crime in India Using Machine Learning', *Computational Economics*, vol. 53, no. 4, pp. 1467–1485.
- N. Mahmud, K. I. Zinnah, Y. A. Rahman and N. Ahmed 2016, 'Crimecast: A crime prediction and strategy direction service', *International Conference on Computer and Information Technology (ICCIT),* pp. 414–418.
- Pandey, M, Fernandez, M, Gentile, F, Isayev, O, Tropsha, A, Stern, AC & Cherkasov, A 2022, 'The transformational role of GPU computing and deep learning in drug discovery', *Nature Machine Intelligence*, vol. 4, no. 3, pp. 211–221.
- Rosenblatt, F 1957, 'The Perceptron, A Perceiving and Recognizing Automaton, Project Para Report', *Project Para Report, Cornell Aeronautical Laboratory (CAL).*, pp. 85–460.
- Shah, N, Bhagat, N & Shah, M 2021, 'Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention', *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1.
- Srivastava, A, Kundu, A, Sural, S & Majumdar, AK 2008, 'Credit Card Fraud Detection Using Hidden Markov Model', *IEEE Transactions on Dependable and Secure Computing*, vol. 5, no. 1, pp. 37–48.
- Tahseen Ali, A, Abdullah, HS & Fadhil, MN 2021, 'WITHDRAWN: Voice recognition system using machine learning techniques', *Materials Today: Proceedings*.
- Yildirimer, K 2023, 'A Review of Sociological, Individual and Psychological Crime Theories. '.

Appendices

Data Information

```
[8] df.shape

(499999, 18)

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 499999 entries, 0 to 499998
Data columns (total 18 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   DR_NO           499999 non-null  int64
 1   DATE OCC        499999 non-null  object
 2   TIME OCC        499999 non-null  int64
 3   AREA            499999 non-null  int64
 4   AREA NAME       499999 non-null  object
 5   Rpt Dist No     499999 non-null  int64
 6   Part 1-2        499999 non-null  int64
 7   Crm Cd          499999 non-null  int64
 8   Crm Cd Desc     499999 non-null  object
 9   Vict Age        499999 non-null  int64
 10  Vict Sex        433163 non-null  object
 11  Vict Descent    433159 non-null  object
 12  Premis Cd       499994 non-null  float64
 13  Premis Desc     499606 non-null  object
 14  Weapon Used Cd  167946 non-null  float64
 15  Weapon Desc     167946 non-null  object
 16  Crm Cd 1        499992 non-null  float64
 17  LOCATION        499999 non-null  object
dtypes: float64(3), int64(7), object(8)
memory usage: 68.7+ MB
```

```
df.isnull().sum()

DR_NO                0
DATE OCC             0
TIME OCC             0
AREA                 0
AREA NAME            0
Rpt Dist No          0
Part 1-2             0
Crm Cd               0
Crm Cd Desc          0
Vict Age             0
Vict Sex         66836
Vict Descent     66840
Premis Cd            5
Premis Desc        393
Weapon Used Cd  332053
Weapon Desc     332053
Crm Cd 1             7
LOCATION             0
dtype: int64
```

Partition Data