Automobile Theft Crime Report Data Analysis

Princely Jonas Lopes, Thinh Huynh, Zahira Ghazali, Saloni Sandeep Ingle

CMPE 255, San Jose State University

March 24, 2024

***Abstract*—Predictive policing, an emerging approach in law enforcement, utilizes data analysis and machine learning algorithms to forecast criminal activity and allocate resources accordingly.Most developed nations have implemented predictive policing, albeit with mixed reactions over its effectiveness. Whilst at its inception, predictive policing involved simple heuristics and algorithms, it has increased in sophistication in the ever-changing technological environment. This project provides an overview of predictive policing, examining its evolution, methodologies, and applications. It explores the technological advancements that have facilitated its implementation, including the utilization of big data, artificial intelligence, and predictive analytics.**

Keywords—Machine learning algorithms; crime analysis; predictive policing;Big data;predictive analytics

I. Introduction

Crime prevention strategies have evolved over the years. Historically, the emphasis was largely on reactionary measures in crime management, with policing primarily centered around responding to incidents after they occurred. As a result, police evaluation predominantly revolved around assessing their responses to and resolutions of reported crimes. The predictive models utilized in this project can help reduce response time and reduce crime rates. Lower response time can be critical in emergency situations in order to minimize harm. However, the scope of crime, the type of crime, location, time of occurrence and other details remain unknown until after the incident actually happens. By identifying potential crime locations, law enforcement agencies can use their resources effectively by being available at the right place at the right time. Predictive policing is meant to eliminate the period of time between the occurrence of a crime and law enforcement being on site to deal with the crime. This project specifically focuses on creating data analytics to implement predictive policing for automobile related crime.

II. Related works

A. Series Finder

A previous work in predictive policing was a machine learning model called “Series Finder” created by Wang and Rudin at MIT in 2013 using Cambridge Police Department’s crime data from Cambridge, Massachusetts. Series Finder was trained to detect housebreak patterns, and it “learned” how to do this using historical data from the Cambridge Police Department’s crime analysis unit.The algorithm tries to construct a modus operandi (M.O.) of the offender. The M.O. is a set of habits that the offender follows and is a type of behavior used to characterize a pattern. The M.O. for the burglaries included factors like means of entry (front door, back door, window), day of the week, characteristics of the property (apartment, single family house), and geographic proximity to other break-ins. As Series Finder grew the pattern from the database, the M.O. for the pattern became better defined.

B. PredPol

The Los Angeles Police Department (LAPD) implemented a predictive policing system called PredPol in 2011. The system used machine learning algorithms to analyze data from past crimes to identify potential hotspots and predict where crimes are likely to occur in the future.The epidemic-type aftershock sequence model is known as PredPol, and recently became Geolitica.The ETAS model is a branching process: first generation events occur according to a Poisson process with constant rate μ, then events (from all generations) each give birth to N direct offspring events, where N is a Poisson random variable with parameter θ. As events occur, the rate of crime increases locally in space, leading to a contagious sequence of “aftershock” crimes that eventually dies out on its own, or is interrupted by police intervention.However, the program came to an end amid reports on how it led to the over-policing of black and brown communities.

C. Strategic Subject List (SSL)

The Chicago Police Department implemented a predictive policing system called Strategic Subject List (SSL). The system used machine learning algorithms to analyze data from past crimes and identify individuals who are most likely to commit violent crimes in the future.SSL uses a list and

everyone on the list gets a risk score, reflecting their predicted likelihood of being involved in a shooting and are ranked on a score between 1 and 500, and the scores are recalculated every day.However, the system has faced criticism for issues of data bias and lack of transparency.

III. Data analysis/ Observations

A. Preprocessing and Datasets Chosen

Firstly, it should be noted that there are two datasets that will be discussed in this paper: an Oakland and Los Angeles dataset. Both datasets specifically focus on automobile related crimes. In the case of the Oakland dataset, other types of crimes that weren't necessarily auto related were included and were then filtered out to ensure that the dataset that was being dealt with only concerned auto related crimes. The Los Angeles dataset was also similarly preprocessed. These datasets were then used to make the deductions discussed later below.

B. Basic Assumptions and Generalizations

For all occurrences of automobile related crime in both the Los Angeles and Oakland datasets, there were several areas that were specifically targeted for data analysis. Namely these were as follows: the day of occurrence, the hour of occurrence, the month of occurrence, and the general location of occurrence. Furthermore, the general assumptions for both datasets were the same and were as follows: crime would most occur on the days closest to/on the weekend, the times of day in which crime would occur would likely be the extremely late or early hours of the day, and it would likely be the months towards the end of the year (i.e. December, November, October) which had the highest rates of crime as it is in those months that most holidays in the United States- and thus people using their vehicle as a means of transit whether it is for commerce or for meeting others- occur. However, in terms of assumptions regarding the location or the most frequently occurring areas in which crime would occur - due to lack of knowledge on the general areas present within Oakland and Los Angeles and thus which areas would be more prone to crime or a higher crime rate - no prior generalizations or estimations were made before actually conducting the data analysis. The only general assumption that was issued is that it would likely be smaller roads that would have higher frequency of automobile related crime as supposed to larger ones. After the data analysis results were found, research on the region with the highest auto crime frequency in both the Oakland and Los Angeles dataset was conducted to determine what qualities may have made the area more likely to be an area with an increased crime rate.

C. Oakland Dataset Observations and Data Analysis

For Oakland, the highest frequency of crimes occurred on Fridays as opposed to any other day of the week. Moreover, late night hours, particularly midnight, witnessed the peak occurrence of automobile related crimes. There is also a general increase in the frequency of crime in all areas up until the peak occurrences at midnight as can be seen depicted in the heat map. Additionally, the late months – specifically January and December – had the highest frequency of crime when compared to all other months in the calendar year which did match the original assumption that it . Also, the 1st, 15th, and 28th of the month in any month exhibited the highest frequency of automobile related crime. The year 2023 was found to be the year with the highest frequency of auto related crimes and it also had noticeably higher levels of frequency when compared to the previous two years 2022 and 2021. Lastly, in terms of area or "regions", District 7 or Downtown Oakland was found to be the area with the highest frequency of auto related crime and furthermore it was found that crime tended to occur on an avenue or street as supposed to other street types.

D. Los Angeles Dataset Observations and Data Analysis

Similar to the Oakland dataset, the highest frequency of auto related crime in Los Angeles occurred on Friday and Saturday. In terms of month, both October and December had the highest frequency of auto related crime. However, it must be noted that all the months did seem to have similar rates when compared to one another – all the months were roughly within several hundred cases of each other – so it can be argued that no month tended to have more cases in comparison to another. In terms of the rate of auto theft per hour, the highest frequency occurred during 22:00 or 10:00 PM. The area with the highest frequency of auto crime was West LA. In terms of the crime frequency by area and hour, it did appear that – across all areas from Los Angeles – the highest frequency appeared to occur in the evenings. Similarly to the Oakland dataset as well, there does appear to be a general trend of the frequency of crime increasing in the later hours, though it is not as strong of a correlation when compared to the previous dataset. However, 77th street did have the highest frequency even in the later hours. For crime frequency by area and day of the week, there weren't any specific days that appeared to have more crime than the others. Once again, 77th street had a higher frequency of crime even when compared to all other areas. For crime frequency by area and by month, in some areas, there does appear to be an increase towards the end of the year, however, again, the margin is very small so it is difficult to entirely state that the later months have a higher

frequency of auto crime. 77th street does continue to also have higher rates than other areas in this chart as well. In terms of crime frequency by day of the week and hour, it does appear that the later hours in the day do have higher frequencies of crime when compared to previous hours.

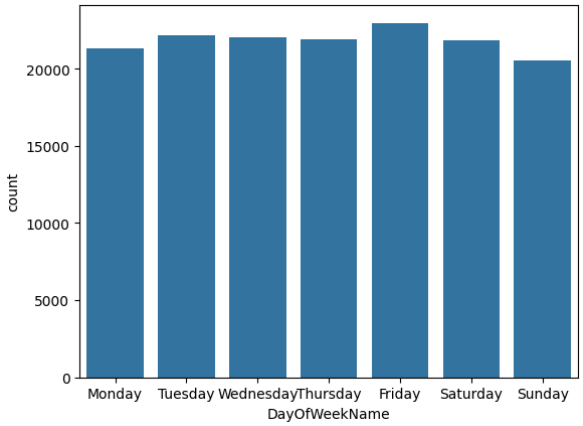

Fig. 1. Frequency of Crimes per Day of the Week - Oakland

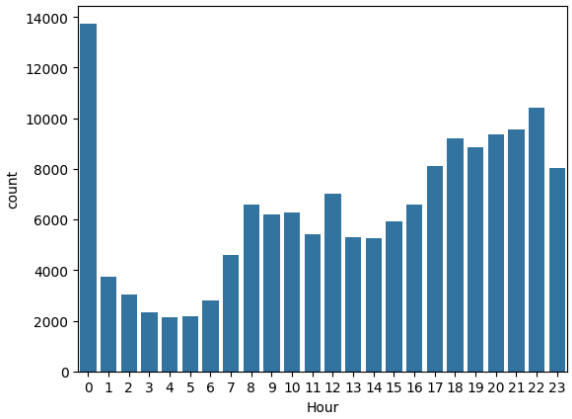

Fig. 2. Frequency of Crimes per 24 Hours - Oakland

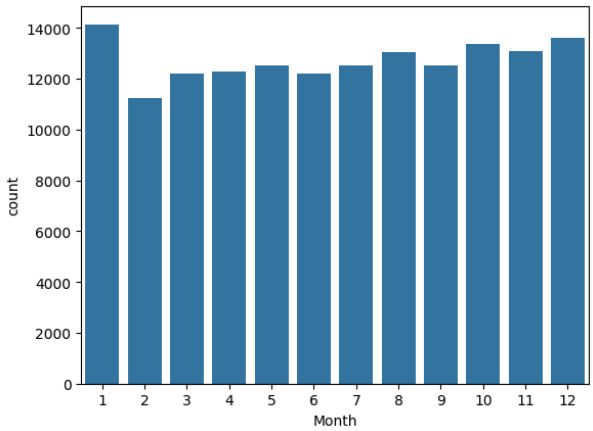

Fig. 3. Frequency of Crime per Month - Oakland

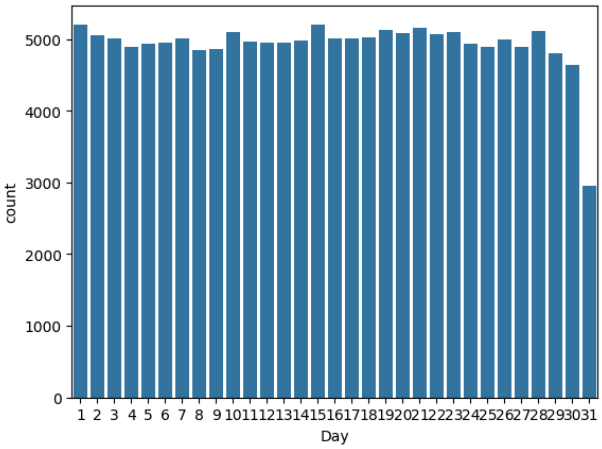

Fig. 4. Frequency of Crime per Day - Oakland

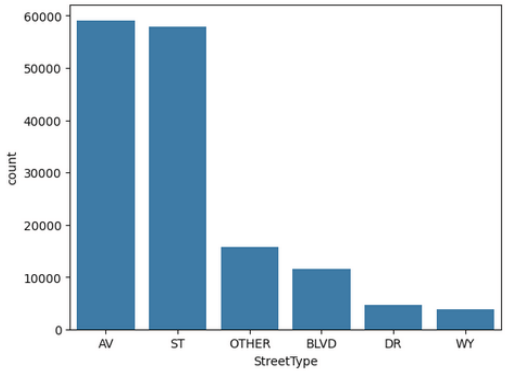

Fig. 5. Frequency of Crime by Street Type - Oakland

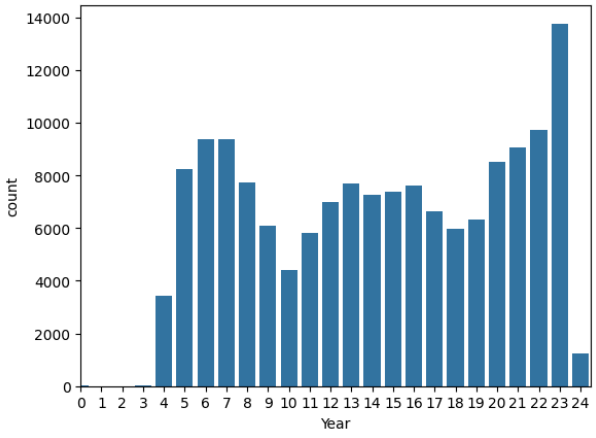

Fig. 6. Frequency Of Crime per Year

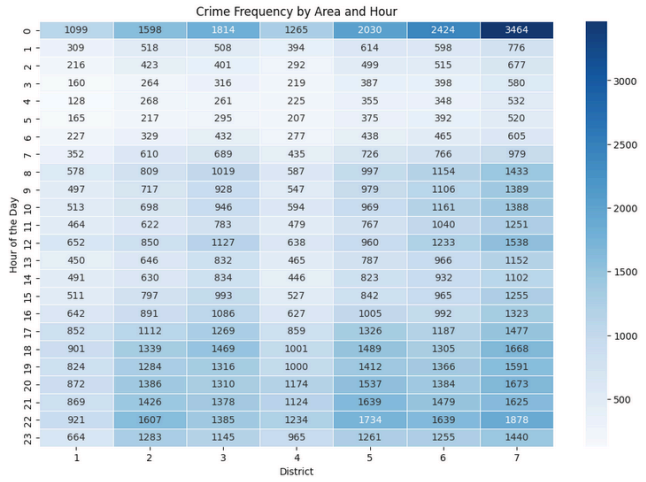

Fig. 7. Crime Frequency by Area and Hour - Oakland

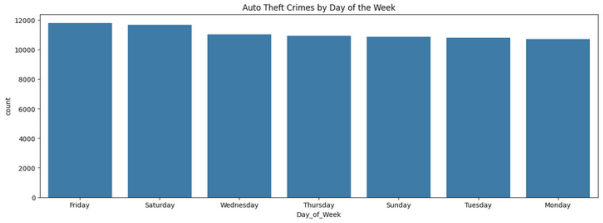

Fig. 8. Auto Theft Crime by Days of the Week - Los Angeles

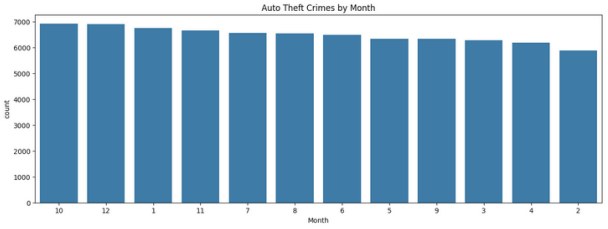

Fig. 9. Auto Theft Crime by Month - Los Angeles

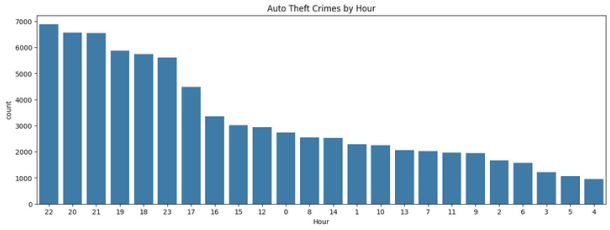

Fig. 10. Auto Theft Crime by Hour - Los Angeles

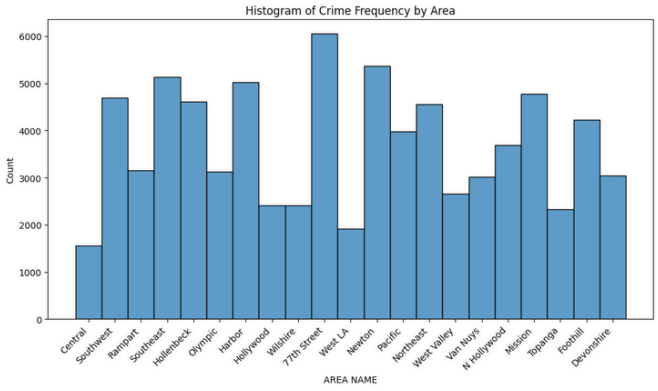

Fig. 11. Histogram of Crime Frequency by Area - Los Angeles

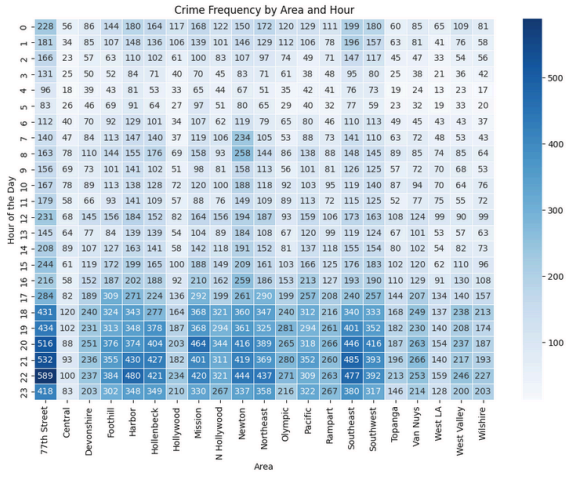

Fig. 12. Crime Frequency by Area and Hour - Los Angeles

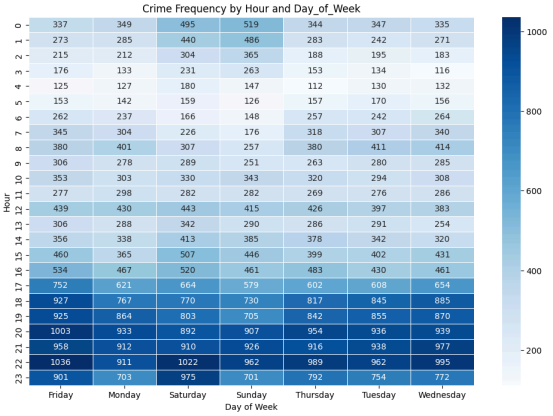

Fig. 13. Crime Frequency by Hour and Day of the Week - Los Angeles

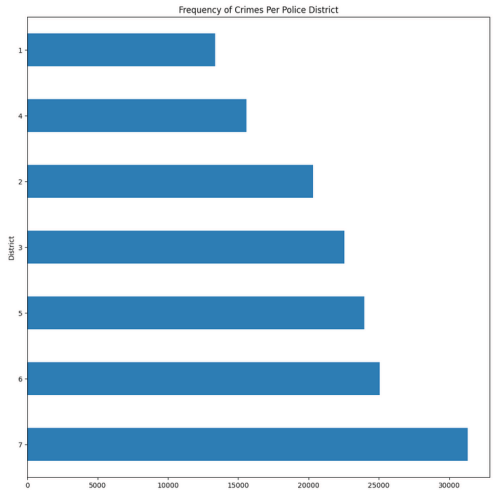

Fig. 14. Frequency of Crimes per Police District - Oakland

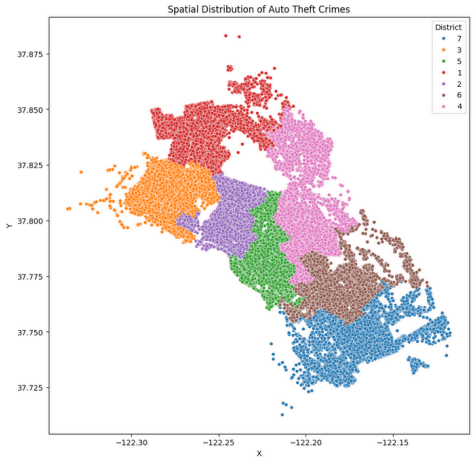

Fig. 15. Spatial Distribution of Auto Theft Crimes - Oakland

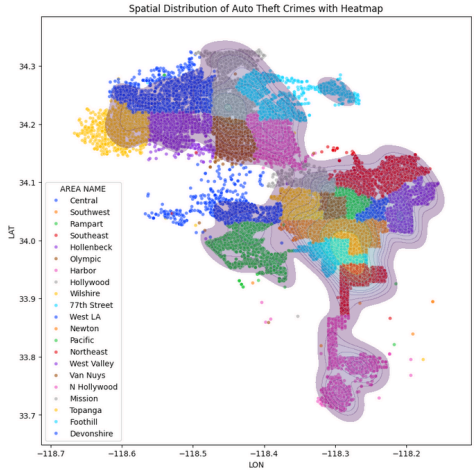

Fig. 16. Spatial Distribution of Auto Theft Crimes - Los Angeles

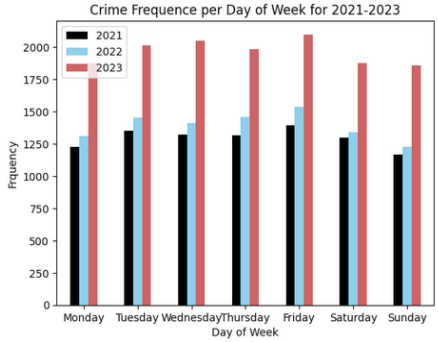

Fig. 17. Crime Frequency per Day of the Week (2021-2023) - Oakland

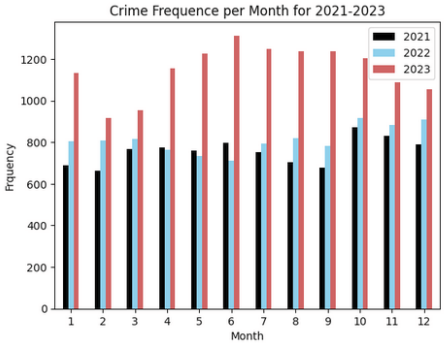

Fig. 18. Crime Frequency per Month (2021=2023) - Oakland

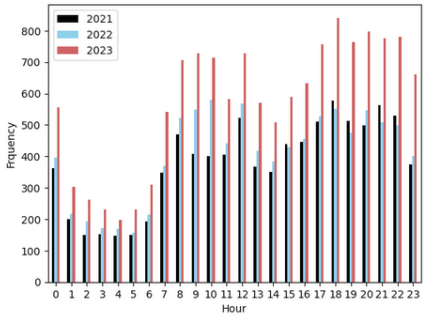

Fig. 19. Crime Frequency per Hour (2021-2023) - Oakland

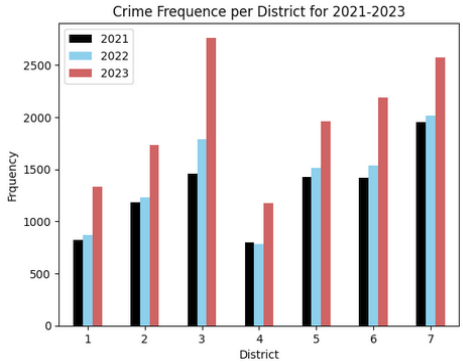

Fig. 20. Crime Frequency per District (2021-2023) - Oakland

IV. Discussion

A. Assumptions vs. Results

Overall, the general assumptions stated prior to any data analysis were upheld; for both datasets, the highest frequency of crime tended to occur in the later hours of the day with the highest frequencies also occurring during the weekend (i.e. Friday, Saturday, Sunday). In terms of area, smaller streets did seem to report higher crime frequency than larger ones for both datasets.

B. Notable Deviations from Assumptions

For the Oakland dataset, it was found that 2023 had an overall higher crime frequency than all previous years. It must be noted that the Oakland dataset is fairly comprehensive and spans the previous two decades up until the start of the twenty-first century. Considering the only crime type that is being examined is auto related crime, this means auto related crime in general increased drastically when compared to previous years in Oakland. While one can state that this spike is likely still due to that of the pandemic, it should be noted that the previous two years – both 2021 and 2022 – are also post-pandemic yet did not have nearly as high an auto crime frequency as 2023. Also, for the Oakland dataset, there did appear to be a dip in frequency during the month of February. This is likely due to the fact that February has a lower number of days when compared to every other month, however it is still notable as this was the case for every year present in the Oakland dataset. Lastly, the highest frequency of auto crime took place at midnight. While this was in line with the general assumptions, the sheer volume of cases that made midnight the primary crime occurrence is somewhat surprising. Also notable is that there is a sharp dip in auto crime frequency past this point. While it was assumed that the later hours would have higher frequency, it was also assumed that the early hours of the morning would also have higher frequency which is not corroborated by the finding for this dataset.

For the Los Angeles dataset, there didn't appear to be any specific month that could be firmly stated to have a higher crime frequency than another. Again, the margin between each month was only a few hundred cases; all the months were within several hundred of each other and were all roughly in the same region as depicted on the bar chart. As such, while it can be stated that the original general assumption of later months having a higher frequency is upheld by the fact that October and December having the overall highest frequencies, it is not truly possible to state that this would always be the case. The other anomaly was specifically for 77th street which consistently tended to have a higher frequency than all other areas present in the Los Angeles dataset across all months and for every day of the week. While it is reasonable that certain areas may have more of a tendency for automobile crime, the extremely high frequency of automobile crime in 77th street to any other region is somewhat surprising.

V. Models Training and Performance Metric

Our group created a feature named “safe” which took into account the frequency of crime based on the hour, day, month and the location of the crime to generate a target for prediction. We then train the 8 models listed in the image below and get the accuracy score for our model prediction

	Model	Accuracy
0	Random Forest	0.816351
1	Logistic Regression	0.781452
2	Support Vector Machine	0.781452
3	Decision Tree	0.712386
4	GuassianNB	0.779256
5	MLPClassifier	0.781452
6	SGDClassifier	0.781452
7	AdaBoostClassifier	0.781330

Fig 20. Models Prediction Accuracy

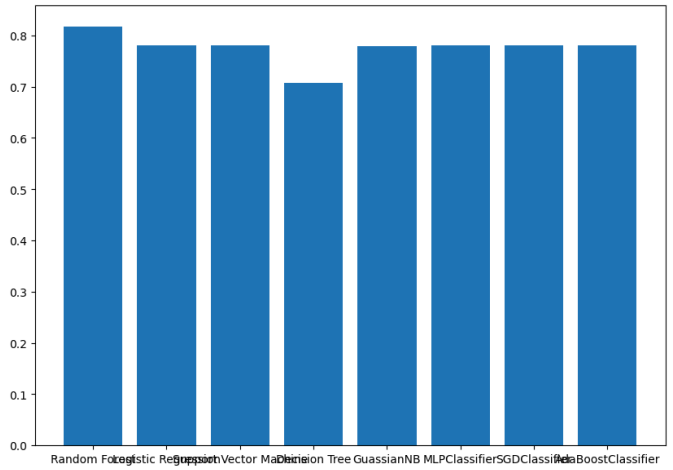

Fig 21. Models Prediction Accuracy Bar Plot

The trained models showed a very high accuracy for the predictions (Fig 20). As a group for future work, we could train the models using different hyperparameters since these models are running on default settings.

VI. Community Contribution

This data could again be utilized in predictive policing. The current issue with the response to crime is that law

enforcement and police must react to crime. Predictive policing removes that barrier by allowing for some level of prediction as to where or when there is likely to be a crime occurring which lessens the time needed for police or law enforcement to react to the crime and make a successful arrest. Thus objectively good predictive policing would help incredibly in lessening the ramifications of all crime -- and most certainly automobile crime. For instance, 77th Street had a higher crime rate overall when compared to all other areas of Los Angeles at most hours of the day. As such, it would make sense for that to be an area that police and law enforcement pay special attention to that region for auto related crime. However, this assignment of priority - which areas should be targeted first or what times of day should be prioritized over others -- need to be objective. Should there be a level of bias present in the datasets which hold records of all arrests, meaning some of the arrests present in the dataset are due to racial prejudice or other factors, then the validity of the model and thus any predictions used for predictive policing will also be nil. This then begs the question of how should one know whether the dataset is free of this bias? Unfortunately, there is no clear answer to this question. The obvious answer is to state that the dataset must then only consist of valid crime reports. However, this is not something that one can enforce easily and, indeed, in the case of the two mentioned datasets used to make the previous analyses, it is not a quality that is ensured.

VII. Conclusions

The purpose of this paper was to identify trends in automobile crime for potential use in predictive policing. As observed, the general assumptions regarding location, hourly, daily, and monthly trends were largely consistent across both datasets. The potential expansion into predictive policing holds promise, provided that the underlying data remains objective. Such a tool could prove invaluable for law enforcement, aiding in crime prevention and ultimately benefiting potential victims of crime.

VIII. References

[1] "Advisory concerning the Chicago Police Department's predictive risk models," Oversight.gov, https://www.oversight.gov/report/state-local/IGCHICAGO/Advisory-Concerning-Chicago-Police-Department%E2%80%99s-Predictive-Risk-Models (accessed Mar. 24, 2024).

[2] L. Meliani, "Machine Learning at PredPol: Risks, Biases, and Opportunities for Predictive Policing," *Technology and Operations Management*. https://d3.harvard.edu/platform-rctom/submission/machine-learning-at-predpol-risks-biases-and-opportunities-for-predictive-policing/

[3] T. Wang, C. Rudin, D. Wagner, and R. Sevieri, "Detecting Patterns of Crime with Series Finder." Accessed: Mar. 24, 2024. [Online]. Available: https://users.cs.duke.edu/~cynthia/docs/WangRuWaSeAAAI13.pdf