Project Poirot: Machine Learning based Predictive Crime Analysis

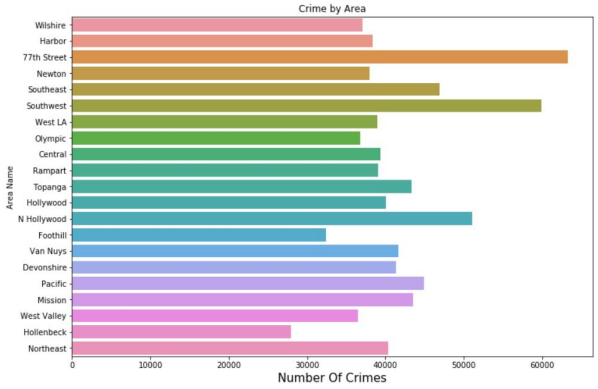
Mansi Ganatra[1], Naga Ritwik Indugu [2] and Uma Kanumuri [3]

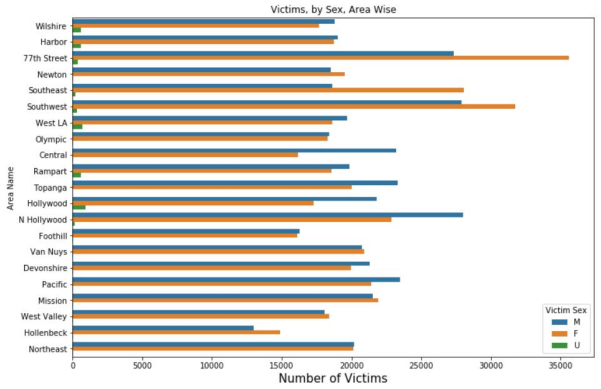
***Abstract*— LA is a very beautiful city, but one thing that scares a lot of people is the amount of crime that happens here. Our intention is to understand what makes this beautiful city to be infested with crime.To work upon this we have collected the dataset from LAPD official site.The dataset contains approximately 1.8 Million rows with around 28 attributes including DR Number, Date Reported, Date Occurred, Time Occurred, Area ID, Area Name Reporting District etc.Some of the questions which we are going to answer are Predicting the time of day, the days of a week, the days of a month, the days of a year in which different kinds of crimes are most likely to happen,to predict the hot spots for various crime categories, to predict the crime patterns, To predict the probability of an offender repeating the crime,Crimes correlation with age,Severity of crime by area (which crime is highest in each area) etc using the different machine learning algorithms like Decision Trees,Nave Bayes,Logistic Regression,K-Nearest Neighbours(K-NN),Support Vector Machines(SVM),Clustering Algorithms.The libraries which we are using include pandas, numpy, scikit-learn, pytorch, theano, etc.For the visualization of results the libraries being used are matplotlib, d3.js etc.**

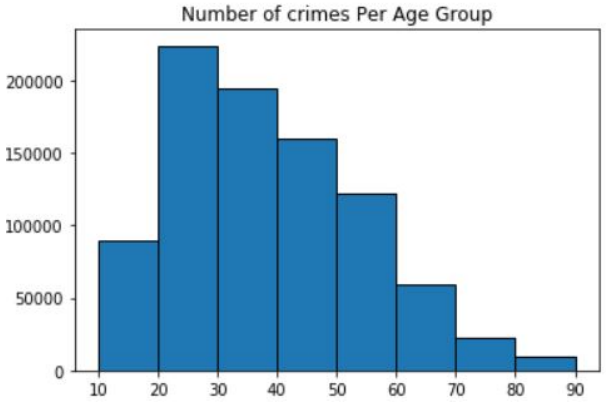
I. INTRODUCTION

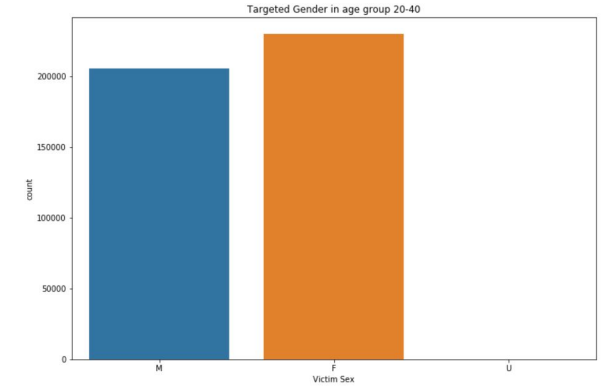
Beautiful and historic, Los Angeles is also infested with a high amount of crime. The total crimes reported to LAPD in 2017 were 129587(Violent: 29661, Property: 99926). The LAPD has made available the crime dataset from 2010-2018. In the Machine Learning experiments described below, we use the same crime dataset pre-processed from Kaggle containing records from 2010-2017. We performed some exploratory data analysis on the features and the results are as in the following section. In this project, we try to answer different questions based on the features. The questions and the algorithms used to evaluate answers to each question are described in the further section. We have used accuracy, F-score, mean squared error as our performance metrics. There have been multiple questions raised for the ethicality of such studies. The study in this project is only empirical and academic in nature. We do not intend to use the results and predictions on field or in conjunction with other predictive policing technologies in use today. Our aim is to apply various machine learning techniques to the selected features, keeping into account the literature survey, and evaluate answer to the selected questions. We intend to measure the accuracy of our results using the performance metric as described in the following section.

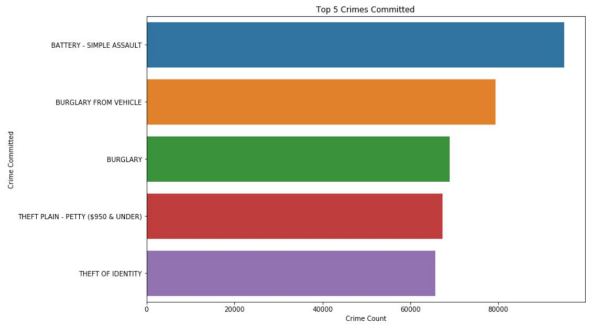
II. EXPLORATORY DATA ANALYSIS

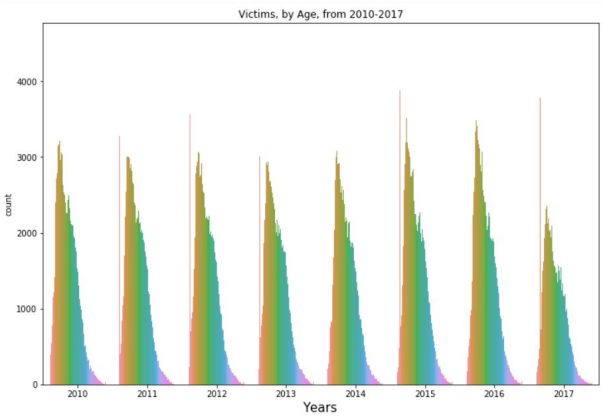


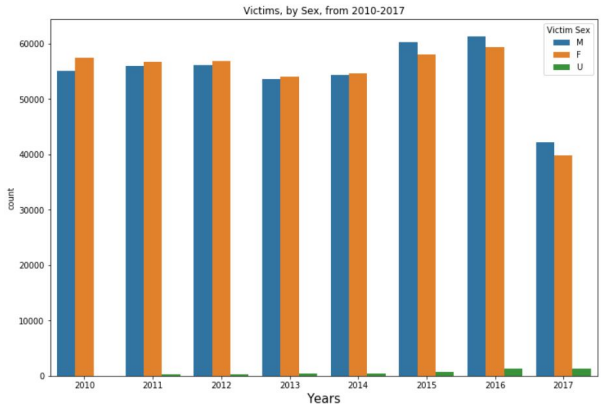












III. METHODOLOGY

A. Learning Algorithms

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations

- **SVM**: In this project we have used SVM with rbf kernel keeping the value of c to 1.0 and gamma to auto. Predict(X) method is used in predicting the crime code given victim age,victim sex.sklearn.svm has been used in importing the svm function.
- **Decision Trees**:Decision trees with a criteria of entropy and random state as 0 are taken.sklearn.tree.decision tree classifier is used to get the decision tree function.Predict(X) method has been used in predicting the crime code given victim age and victim sex
- **Logistic Regression**:Mutinomial logistic regression has been used in predicting the crime code given victim age and victim sex.Newton Solver has been used here. sklearn.linearmodel.logistic regression has been used for getting the regression function.Predict(X) method has been used for predicting the crime code.
- **Linear Regression**:Linear regression has been used in predicting the time occured given the weekdays.Here weekday is calculated from the date occurred using a python function.sklearn.linearmodel.linearregression has been used for getting the prediction function. predict(X) method has been used for predicting the linear regression.

B. Performance Metrics

- **Accuracy score**:From sklearn.metrics import accuracy score has been used in getting the accuracy score function.It is used for calculating the accuracy's of svm, decision trees and logistic regression in this project.
- **Root mean square error**:RMSE has been used in linear regression for measuring the accuracy of predicting time of the day given day and time. from sklearn.metrics import mean_squared_error has been used for getting the mean_squared_error function.

IV. VISUALIZATION FUNCTIONS

Packages seaborn and matplotlib in python have been used in visualization of the results.

V. Current Results

Question	Logistic/Linear Regression	Decision Trees	SVM	KNN
Prediction of crime code given victim age and victim sex	With grouping: ~33.12% Without grouping: ~11.55%	With grouping: ~19.96% Without grouping: ~13.32%	H/W unable to handle	With grouping: ~21.61% Without grouping: ~16%
Prediction of time occurred given weekday	rmse: 646.66	-	-	-
Prediction of area code given date occurred, time occurred and crime type	~7%	~6.2%	H/W unable to handle	~6.7%

VI. Relevant Work

The pioneer work in this field has been done by the team behind what is today known as PredPol. The introductory paper "Randomized Controlled Field Trials of Predictive Policing"[1] was first published in 2015. PredPol uses Crime type, Crime location, Crime date and time features and proprietary machine learning algorithms for forecasting. Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations [2] details the predictive policing strategies and their use in Law Enforcement. Predictive Policing: A Review of the Literature [3] from Portland State University reviews all the literature on predictive policing. Recent advancements include Partially Generative Neural Networks for Gang Crime Classification with Partial Information [4].

Acknowledgment

We thank Prof. Ion Muslea for guiding this project and USC DPS Officer Wyman Thomas for helping us with the understanding of the data and domain.

References

[1] G. O. Mohler, M. B. Short, Sean Malinowski, Mark Johnson, G. E. Tita, Andrea L. Bertozzi & P. J. Brantingham (2015) Randomized Controlled Field Trials of Predictive Policing, Journal of the American Statistical Association, 110:512, 1399-1411, DOI: 10.1080/01621459.2015.1077710

[2] Perry, Walter L., Brian McInnis, Carter C. Price, Susan Smith, and John S. Hollywood, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. Santa Monica, CA: RAND Corporation, 2013. https://www.rand.org/pubs/research_reports/RR233.html. Also available in print form.

[3] Portland State University. Criminology and Criminal Justice Senior Capstone, "Predictive Policing: A Review of the Literature" (2012). Criminology and Criminal Justice Senior Capstone Project. 5.

[4] Seo, Sungyong, Hau Kong Chan, P. Jeffrey Brantingham, Jorja Leap, Phebe Vayanos, Milind Tambe and Yan Liu. Partially Generative Neural Networks for Gang Crime Classification with Partial Information. (2017).

[5] http://michael-harmon.com/blog/crimetime.html

[6] https://machinelearningmastery.com/