Prediction of crime occurrence by category from crime and SES datasets

Mark. S. Kim

Sekwen Kim

Zhamila Klycheva

SPE 486

Claremont Graduate University

May 21, 2019

Abstract

Crime has been increasing at a steady pace threatening civilians' lives in the city of Los Angeles for the past few years. In order to reduce the crime rate and provide valuable policy implication for local authorities, this study combines crime dataset from LAPD and socio-economic status data from US Census in order to predict crime occurrence by category in the city of LA. By incorporating both supervised learning, namely Backwards Stepwise Regression, Multinomial Logistic Classification and Random Forest, as well as unsupervised learning through K-means clustering and Neural network, this paper seeks to suggest the most appropriate model for predicting crime based on data available on Los Angeles Police Department (LAPD) and US Census. After thorough assessment of models and testing, this paper suggests that random forest model assesses the crime analysis in the most efficient manner. This paper further recommends policies to control and prevent crime based on analysis which includes, educating the most targeted population group, controlling weapons and considering migration when corresponding crime.

Introduction

The prediction of crime occurrence has been paid extensive attention as the number of available literature suggests. However, the current study provides prediction of crime occurrence by category of eight different types of crime within the City of Los Angeles. Given crime dataset provided by LAPD and socio-economic status data by US Census on a zip code level, the current study aims to help policymakers address crime reduction by category of crime using machine learning techniques. On the initial stage of the research, Exploratory Data Analysis provides a comprehensive overview of the data and points out at important correlations between the variables in the dataset. The study deploys both supervised and unsupervised learning to meet the

goal. Thus, the first method used in EDA is K-means Clustering that provides a clustering analysis of the variables of interest, namely category of crime and draws important correlations used within the scope of the study. Further, the research considers supervised learning tools such as Stepwise Regression model to predict the total number of crimes unclassified by category as well as Multinomial Logistic Classification and Random Forest to predict crime by specific category. Finally, the study refers to unsupervised learning technique, namely Neural Network to predict crime by category. The best performing model is determined by the accuracy, sensitivity, specificity and AUC scores demonstrated for each model. Based on the evaluation results, the research will come up with the relevant policy implications to reduce crime rate and point out at the most robust model of performance.

Literature review

The number of past literature available on prediction of crime suggests a strong interest in the subject from the publicity and a sufficient support for policymakers. The current study explains two articles that serves as a basis and motivation for the research. In the first paper, Bogomolov, Oliver et.al. in their research, *Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data* forecast crime on a monthly basis using demographics and mobile data. Their model is built using training and testing data and a "5-fold cross validation strategy: logistic regression, support vector machines, neural networks, decision trees and ensemble tree classifiers" (Bogomolov et.al., p.430). To evaluate the model, the research used accuracy, F1 and AUC to determine the best performing one.

The second study, namely *Prediction of crime occurrence form multi-modal data using deep learning* that inspired the current research used deep learning to predict crime using images through Google Street View (Kang et.al. 2017). The Research Design considers 4 main layers of

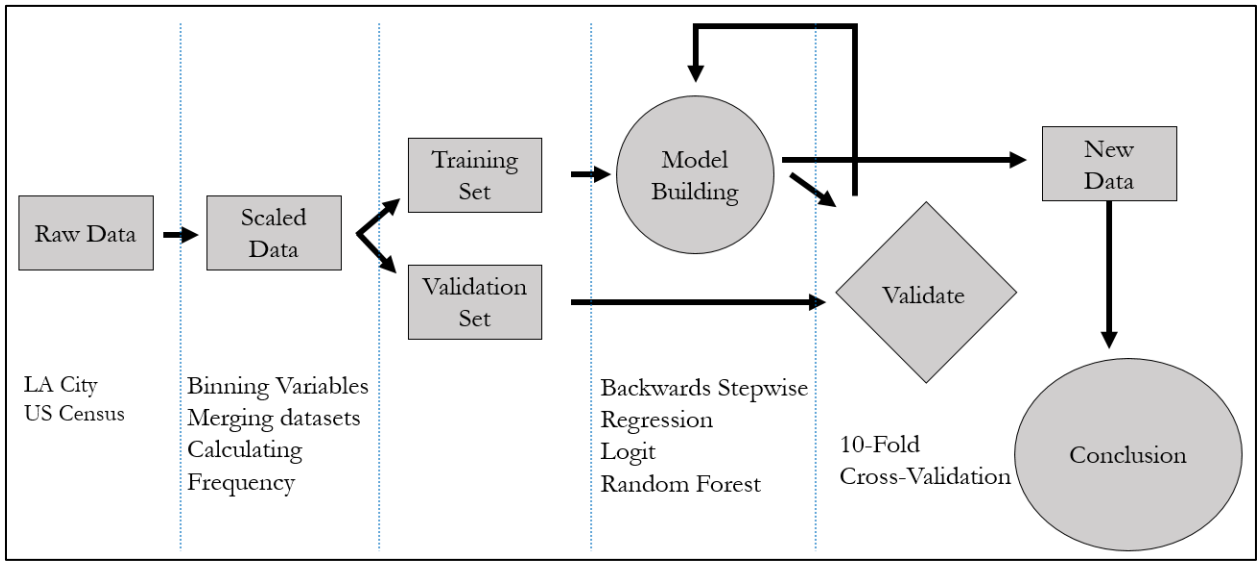
data: spatial, temporal, environmental and joint feature representation layers. Then, all four layers of data are unified into a single feature used in the analysis.

Research design

By incorporating both supervised and unsupervised learning, this study seeks to identify data arbitrage opportunity between crime and socio-economic status (SES) data to help policymakers address crime reduction by crime category(Seagal, ch.5). The main sources of data, namely crime dataset and SES data had different units of measurement, therefore had to be merged as a single dataset. Moreover, multiple variables of interest went through the cleansing process and were binned in fewer number of categories to ease the process. The study uses both Supervised and Unsupervised learning methods for a comprehensive analysis. In the Exploratory Data Analysis section, the paper deploys K-means Clustering technique to derive necessary information from existing data. The actual analysis starts with Supervised learning, introduced by Backwards Stepwise Regression Model, followed by Multinomial Logistic Classification and Random Forest. At a final stage, Unsupervised learning steps in with a Neural Network. Regardless of the research design, whether it is supervised or unsupervised, this study uses merged dataset from unclassified crime data from LAPD and classified SES data.

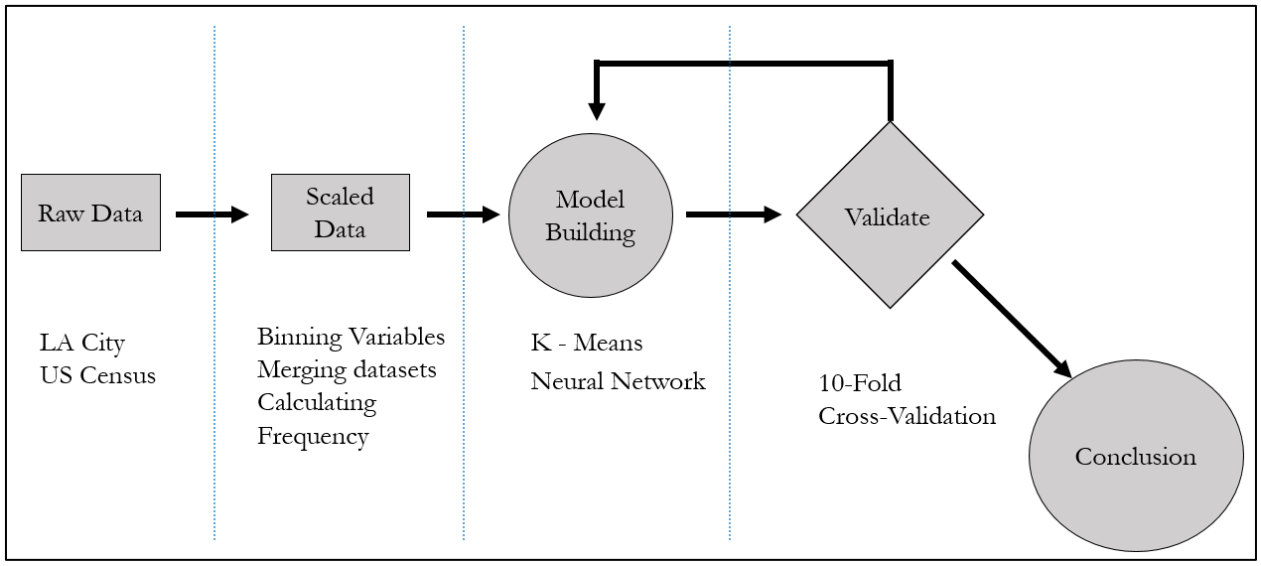
In order to build and validate supervised models, the paper partitioned the merged dataset into training and testing sets. It used 10-fold cross-validation to accurately evaluate the models and come up with less biased results. Figure 1 below summarizes the Supervised Learning flow.

Figure.1. Supervised Learning Research Design



Unsupervised learning has a slightly different research design. Instead of splitting the data in two sets, it builds the model based on the original dataset and further validates it. The initial data preprocessing follows the same exact pattern as Supervised method, since both designs rely on the same merged dataset. The paper uses K-means Clustering and Neural Network techniques under Unsupervised learning. The validation is performed using 10-fold cross validation to assess the performance of the model.

Figure. 2. Unsupervised Learning Research Design



Variable description

This study uses merged dataset of crime data from LAPD and SES data from US Census Bureau. Crime data from LAPD has 423,000 observations of detailed crime information including crime category, crime area, victim's age, weapons used and the time when crime occured. SES data from US Census Bureau contains economic and demographic information in LA city on a zip code level.

In order to provide a comprehensive picture of crime in general, this study uses crime data from LAPD as a core data set and adds corresponding SES information as a supplementary data set according to zip codes of crime data from LAPD. By doing so, we are not losing the richness of information and maintaining the competency in data volume for this study.

For random forest, logit, and neural network, this study treats crime category, which is a binned categorical variable with 8 categories, as dependent variable. The crime category is classified by the following categories: causing bodily harm, violent crime, involving property only, involving property and human, sex crime, against public order, fraud, deception &

corruption and other. The independent variables are area name, victim's age, weapon type, month and day of crime incidents, and victim's ethnicity from crime dataset, poverty, median household income, English proficiency, and education attainment from SES dataset. Within independent variables, weapon types and victim ethnicity are categorical variables and area name is a string variable. SES data set consists of continuous variables which are coded either in dollar value, for median household income and median home value, or percentage, for poverty, English proficiency, ethnicity and education attainment.

Table. 1. Variable description for Random Forest, Logit, and Neural Network.

Model	DV/IV	Variables	Description	Unit of analysis	Source
RF, Logit, & Neural Network	DV	Crime Category	Binned categorical var with 8 categories	Category	LA Crime data (423,000 obs)
	IV	Area Name	Name of crime spot	String	
	IV	Victim Age	Age of victims	Number	
	IV	Weapon Type	Category of weapon used in crime	Category	
	IV	Month	Month of crime	Number	
	IV	Day	Day of crime	Number	
	IV	Victim Ethnicity	Ethnicity of victim	Category	
	IV	Poverty	% of population under poverty line	Percentage	US Census
	IV	Med income	Median household income	USD	
	IV	Med Home Value	Median home value	USD	
	IV	English not Well	% of population with improficient English level	Percentage	
	IV	High School	% of population with high school diploma	Percentage	

For Stepwise Regression model, this study considers crime count,s which is the frequency of total crime of each zip codes in LA city, as a dependent variable. However, in order to better assess the socio-economic aspect of crime in LA, independent variables are victim's age from crime dataset, percentage of each Asian, Black, Native, Pacific, and white population, median

age, migration pattern, poverty, unemployment, median home value, higher education attainment and total male population of each zip codes from the Census dataset. The difference between the data for Stepwise Regression model and the data for classification models is that it only uses continuous variables, given either in count, percentage or dollar amounts.

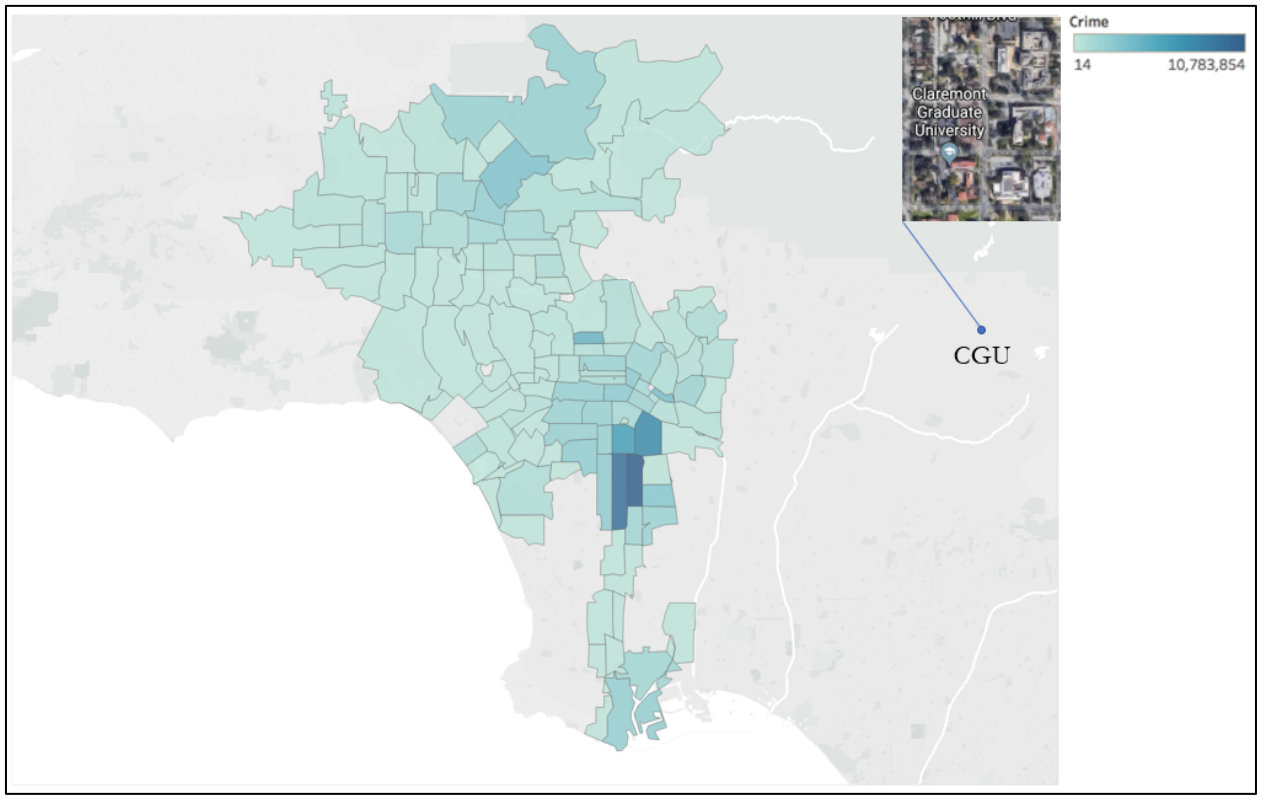
Table. 2. Variable description for Stepwise Regression model.

Model	DV/ IV	Variables	Description	Unit of analysis	Source
Stepwise Regression	DV	Crime Count	Frequency of total crimes	Number	LA Crime data (423,000 obs)
	IV	Ethnicity- Asian, Black, Native, Pacific, White	% of each ethnic group	Percentage	
	IV	Victim Age	Age of victims	Number	
	IV	Median Age	Median age of population per zip code	Number	US Census
	IV	Migration- abroad, diff county, diff state	% of each type of population	Percentage	
	IV	Poverty	% of population under poverty line	Percentage	
	IV	Unemployment	% of unemployment	Percentage	
	IV	Med Home Value	Median home value	USD	
	IV	Population with bachelors or higher	% of population with Bachelor's degree or higher	Number	
	IV	Total male	% of total male population	Percentage	

EDA

The exploratory data analysis serves as a starting point in order to identify key patterns in data, find significant correlations and guide through the analysis. The first model demonstrates the frequency of crime in LA by zip code. Based on the first visual, the crime is concentrated in the center of LA downtown and northern part of LA city. In Fig. 3, darker color indicates higher crime frequency, where downtown LA and northern part of LA city exhibit darker color indicating higher crime frequency.

Fig. 3. Crime frequency of LA city.



Before discussing the locality of crime frequency, this study performed K-means Clustering in order to identify multiple clusters of data by location and have a better understanding of inherent structure of the crime dataset. After performing K-means Clustering on gun incidents and violent crime, which is fatal death causing crime, it is found that both gun incidents and violent crime are prevalent throughout LA city area and exhibit similar pattern of clustering.

Fig. 4. K-means Clustering by gun incident

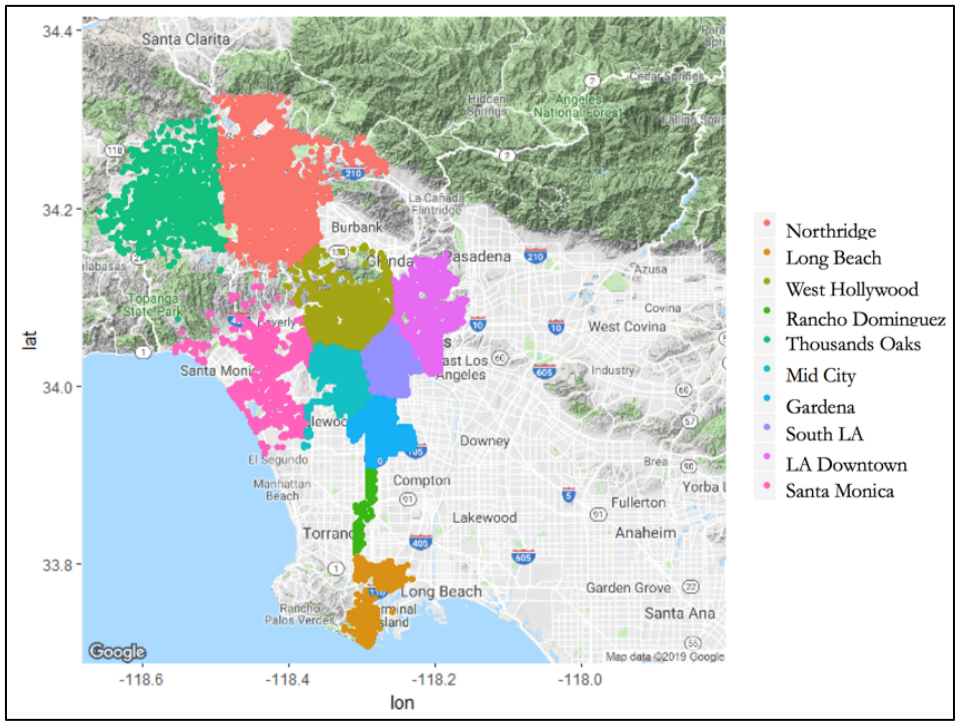
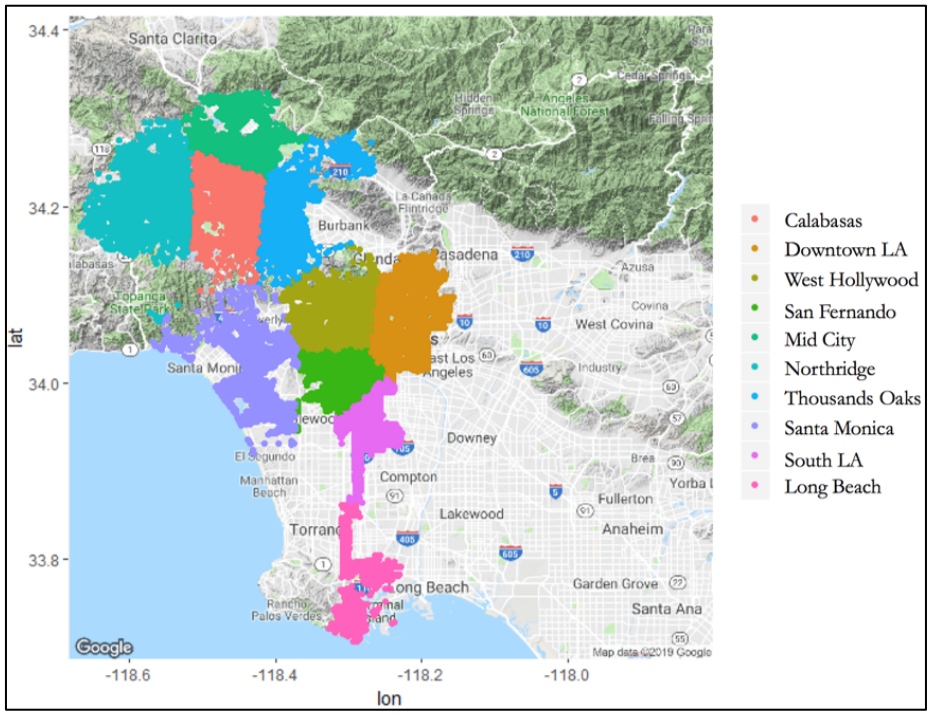


Fig. 5. K-means Clustering by violent crime



However, after comparing the distribution of unemployment and higher education attainment across LA city and crime frequency, similar patterns are shared by crime frequency, unemployment and higher education attainment. Downtown LA and northern part of LA shows distinct characteristic of a higher rate in unemployment and lower rate in higher education attainment. That distribution goes along with the distribution of crime, demonstrating positive correlation between unemployment and crime and a negative correlation between education and crime.

Fig. 6. Unemployment rate by Zip codes

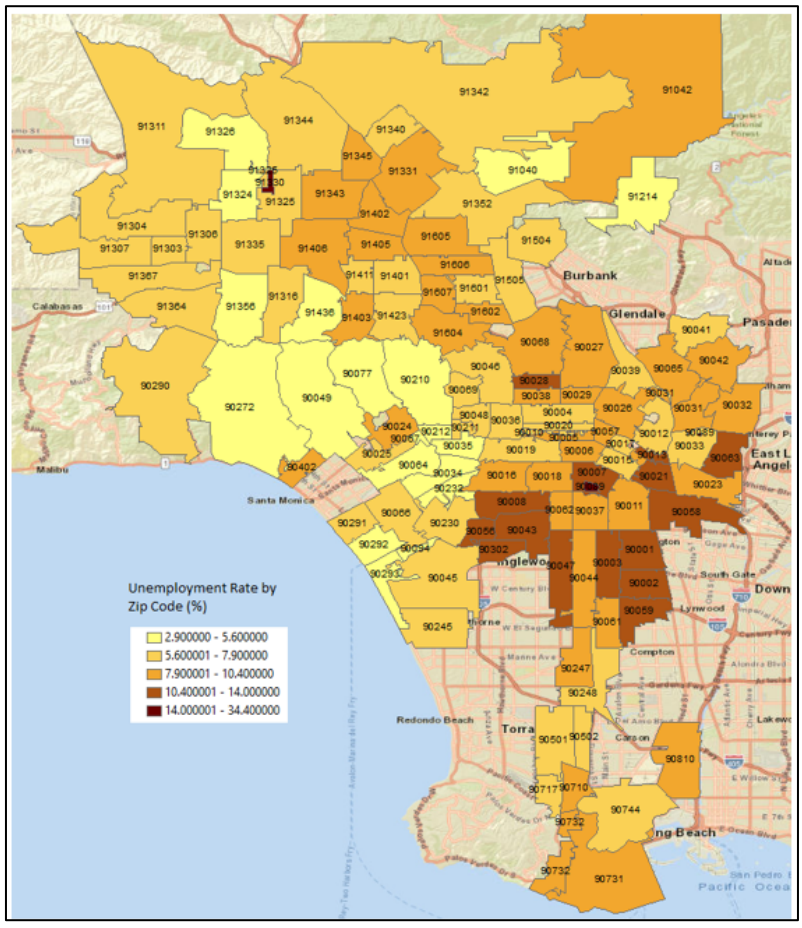
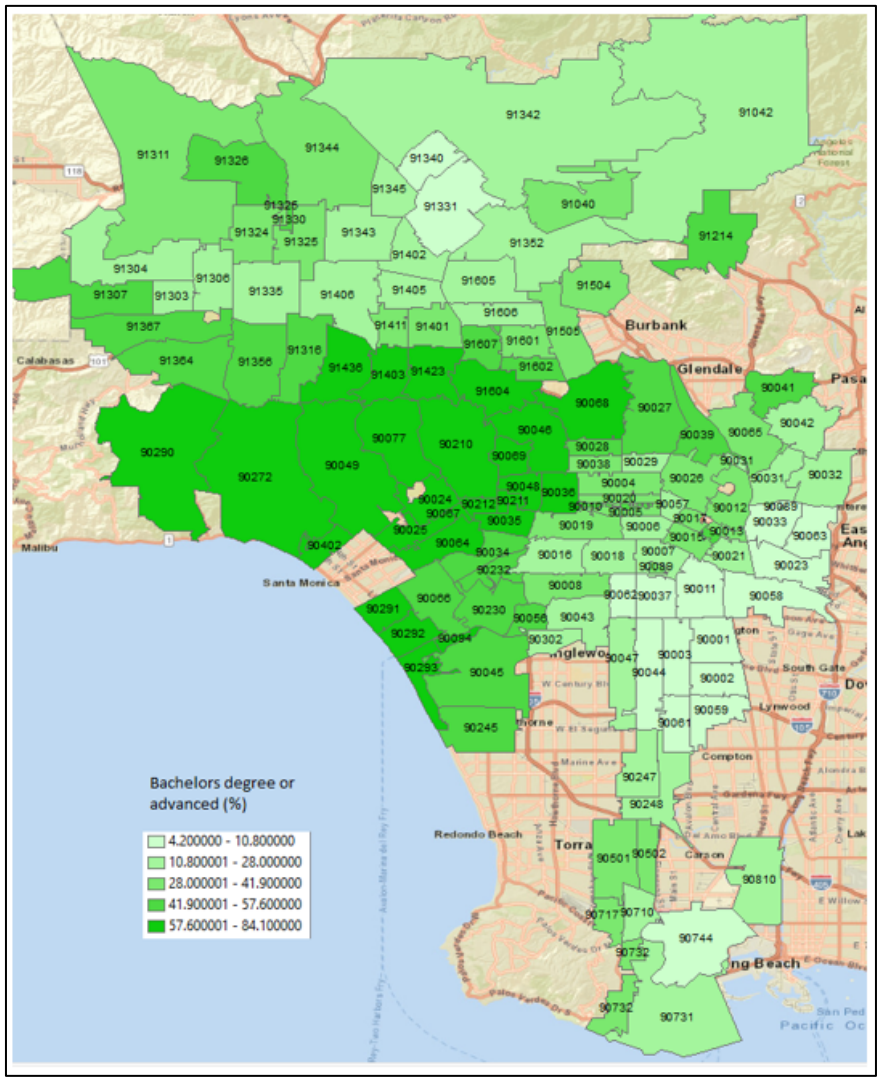


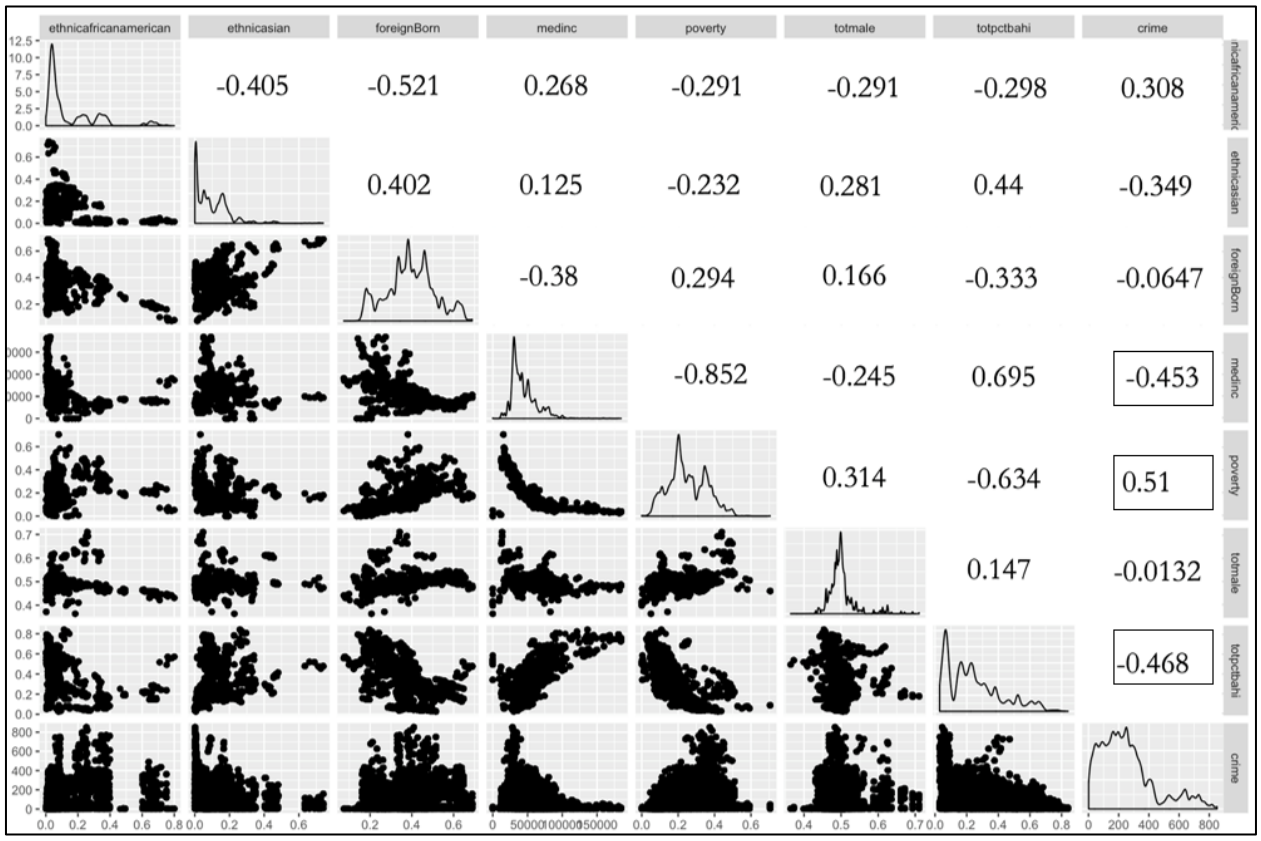
Fig. 7. Higher education attainment by Zip codes



In order to confirm the correlation, witnessed in the exploratory data analysis, this study performed the correlation plot. As a result, crime is significantly negatively correlated with median household income, positively correlated with the percentage of population under poverty, and negatively correlated with higher education attainment. It appears that crime is negatively correlated with median household income and higher education attainment, meaning zip codes with higher median household income and higher percentage of population with Bachelor's degree or higher would have lower crime frequency. However, crime frequency is positively

correlated with poverty which can be infers that areas with higher percentage of population under poverty would have higher crime frequency. Interestingly, the total male population appears to be insignificant in the relation to the total crime count as well as the percentage of people Foreign born in the US.

Fig. 8. Correlation plot



Model - Backwards stepwise regression

The first model we use in the study within the scope of supervised learning is Backwards Stepwise Regression model used predict total crime count. The baseline model includes a number of variables from SES data inspired by the existing literature. Prior to running the model, it goes through variable normalization process and then feature reduction based on the AIC number (Hastie, ch.3). The baseline model includes a full list of variables, step-by-step

eliminating features and decreasing the AIC. At the end, the algorithm comes up with the model that has the smallest AIC for predicting crime frequency given the initial set of variables. The following model below shows the baseline model with the original set of variables before the feature elimination took place. Crime count is a dependent variable which stands for the total number of crimes reported from 2011 to 2017 in the city of Los Angeles. This number is not differentiated between types of crime, but rather presents an aggregate number of crime. The first 5 variables present the percentage of people of a given ethnic group from total population per each zip code in Los Angeles:

$$Crime_{it} = \beta_0 + \beta_1 African\ American_{it} + \beta_2 Asian_{it} + \beta_3 White_{it} + \beta_4 Pacific_{it} + \beta_5 Native_{it} + \beta_6 Poverty_{it} + \beta_7 Foreignborn_{it} + \beta_8 medage_{it} + \beta_9 Male\ with\ BA\ or\ higher\ degree_{it} + \beta_{10} Migrants\ from\ abroad_{it} + \beta_{11} Migrants\ from\ dif\ county_{it} + \beta_{12} Migrants\ from\ different\ statesit + \beta_{13} Unemp_{it} + \beta_{14} Median\ hoval_{it} + \beta_{15} Med\ income_{it} + \beta_{16} Total\ male_{it} + \varepsilon$$

Based on the algorithm results, only two variables get eliminated from the initial model. Interestingly we find that Victim's Age and percent of people Native Born are not considered in the final model. Despite a common thought of an importance of Victim's age in predicting crime number, we realize that it doesn`t matter as much as other variables in the model do.

Below is the model outputs table. As we can see form the table, every single variable is significant at a 1% level except for the percentage of people from total with Bachelor's degree or higher. Based on the results, we can conclude that the area demographic data such as ethnicity, percentage of people foreign born, median age, migrated population, total male population as well as economic data, namely poverty level, unemployment rate and median house value are important in predicting the number of crimes occurring in LA.

Table. 1. Stepwise Regression Results

Dependent Variable	Crime Count	
African American	0.260***	(0.003)
Asian	-0.171***	(0.003)
White	-0.977***	(0.007)
Pacific	-0.369***	(0.007)
Native	-0.061***	(0.003)
Poverty	0.490***	(0.004)
Foreign Born	0.059***	(0.003)
Median Age	-0.277***	(0.005)
Male with Bach or higher	-0.144***	(0.017)
Mig abroad	-0.359***	(0.004)
Mig dif county	0.154***	(0.010)
Mig dif state	0.722***	(0.009)
Unemployment	-0.242***	(0.004)
Tot pop with bach or higher	-0.029	(0.018)
Median house value	0.054***	(0.004)
Total male	0.257***	(0.005)
Constant	0.181***	(0.003)
Observations	423,340	
Adjusted R2	0.404	
Residual Std. Error	0.167 (df = 423323)	
F Statistic	17,943.910*** (df = 16; 423323)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

To evaluate the model, we additionally estimate Root Mean Squared Error, which is equal to 0.17.

Model - Logit

The next model under supervised learning is the Multinomial Logistic Classification (Hastie, ch.4). This model along with the remaining models answers a different question. It predicts a crime by each class or category, therefore presents a more sophisticated version of the Binary Logistic Classification. The dependent variable is crime category, broken down into 8

types mentioned in the data description part. The model below lists all independent variables as a function of Crime Category.

Crime Category *~ Area Name, Victim Age, Weapon Type, Month, Day, Victim Ethnicity, Poverty, Med income, Med Home Value, Unemployment, Total Population, Ethnic African American, English not Well, Foreign Born, High School, Male Percent from Total with Bachelors*

We run a prediction model based on the set of listed variables. The model is built on the training set of data and assessed through a 10-fold cross-validation. In order to assess the model we consider 4 main of measurements, namely accuracy, sensitivity, specificity and AUC.

In order to define the key drivers of the dependent variable, we run a Variable Importance test in R. Based on the results (see Model X in Appendix), the following is the list of the most important variables sorted from the most important to the least (top to bottom)

1. Weapon Type
2. Area Name
3. Percentage of African American from the population total
4. Victim Ethnicity
5. Percentage of people below poverty line
6. Percentage of people not proficient in English
7. Unemployment

Now we proceed to evaluating the model (Runkler, Chapter 8). Based on the results, the model accuracy is 43.51%, which indicates the percentage of correctly identified classes either true positive or true negative as a proportion of total. However, given the unbalanced nature of the data, we infer to sensitivity and specificity provided in the Table 2. Based on the table results,

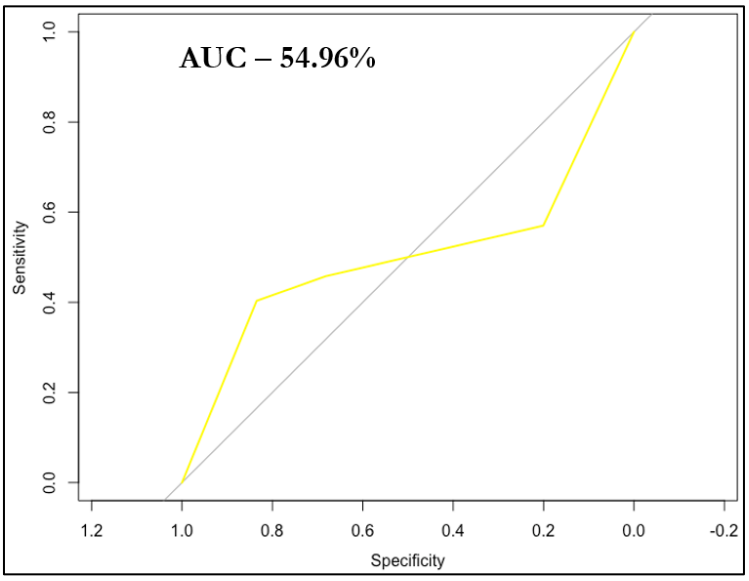
the most balanced crime category is Violent Crime, whereas the model does a poorer job on classifying Fraud, Deception and Corruption as well as Property crime.

Table. 2. Logit Model Evaluation

Accuracy – 43.51%		
Crime Category	**Sensitivity**	**Specificity**
Against Public Order	0.0001	1.00
Causing bodily harm	0.33	0.81
Violent Crime	**0.70**	**0.81**
Fraud, Deception, Corruption	0.00	1.00
Involving property only	0.26	0.91
Involving property and human	0.00	1.00
Sex crimes	0.57	0.72
Other	0.00	1.00

The last measure of assessment is AUC curve, presented in the Figure 8 below. The model demonstrates a trade-off between the two indicators and proves that there is balance between sensitivity and specificity in certain cases.

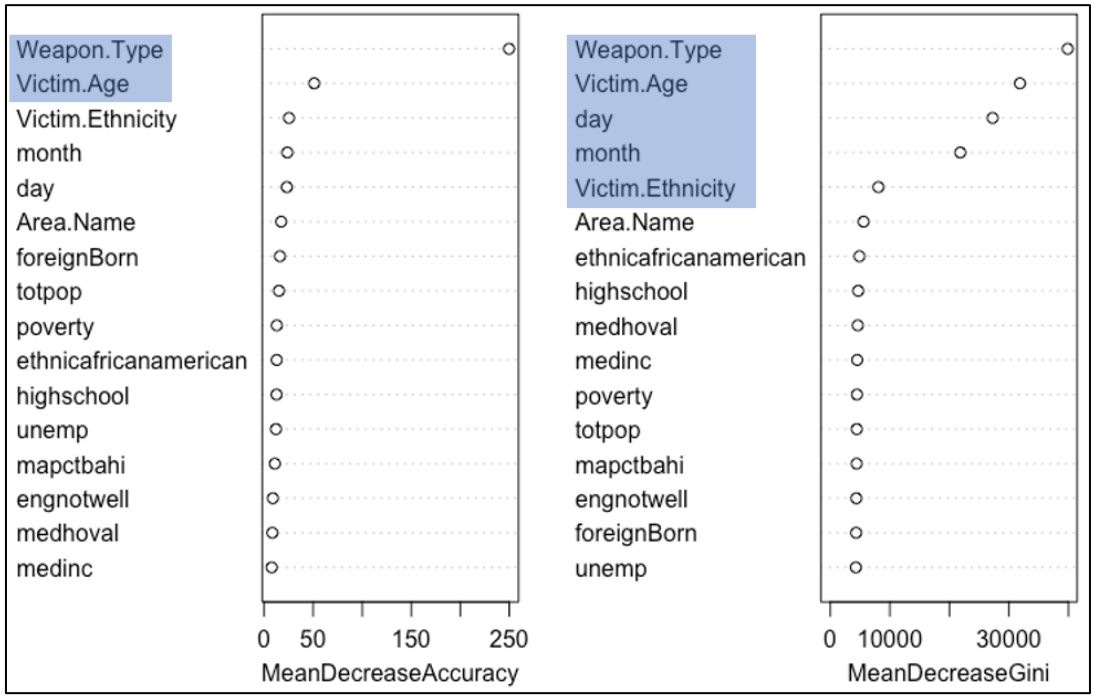
Fig. 8. Trade-off between Sensitivity and Specificity



Model - Random forest

The next model of supervised learning is Random Forest, which is built on the same set of variables as Logistic Classification. The model was built based on the training dataset using 100 trees. Based on the varImp function in R, we find important variables for the model, which slightly diverge from the logistic classification. However based on this model, we find slightly different variables as important using the "varImp" function. Based on the Figure 9, the most important variables of the model are Weapon Type, Victim Age, day, month and victim ethnicity. Thus, the key variables are taken from the crime dataset.

Fig. 9. Random Forest – Variable Importance Plot



In order to assess the model we rely on the same set of measures. Accuracy based on the testing set is 48.92%. For sensitivity and specificity we refer to Table 3 below. Interestingly, violent crime appears of the same importance for Random Forest and performs great prediction on this type given the set of important variables, including weapon type. Another important observation

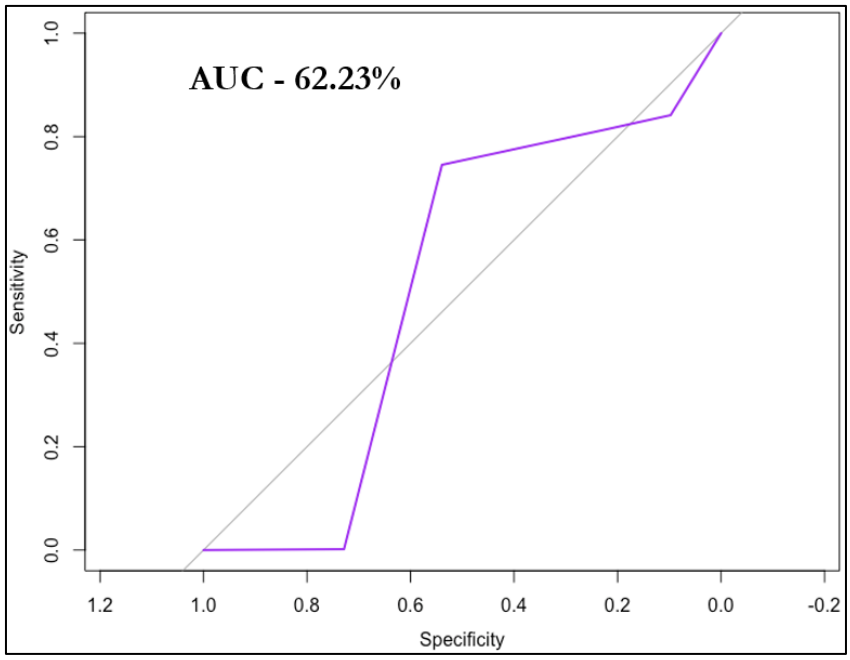
from the table is high specificity values, which stand for the true negative values classified as correctly identified by the algorithm.

Table. 3. Random Forest Model Evaluation

Crime Category	Sensitivity	Specificity
Against Public Order	0.53	0.99
Causing bodily harm	0.85	0.85
Violent Crime	**0.90**	**0.95**
Fraud, Deception, Corruption	0.47	0.99
Involving property only	0.67	0.98
Involving property and human	0.26	0.99
Sex crimes	0.74	0.92
Other	0.49	0.99

Figure 10 below shows the ROC curve with AUC equal to 62.23%, which is notably higher than Logistic Regression. Here, the shape of ROC is slightly different where the optimal point has higher sensitivity and specificity.

Fig. 10. Random Forest ROC Curve



Model – Neural Network

The final model we built to predicting the variable 'Crime Type' was using a Neural Network from the 'NNET' package in R(A beginner's guide, Skymind). NNET is design for a feed-forward neutral networks with a single hidden layer. Given the limitation of using a single hidden layer on multi-classification across numerous variables for Crime Type prediction - we anticipated less than stellar performance. The time to compute exceeded 24 hours before the algorithm converged to a solution. The Confusion Matrix and Statistics for the Neural Net are shown in Table 4. As expected the performance was marginal. Accuracy of the model was 32.8%. The model had a high bias towards selecting three classes: Causing bodily harm, Causing Death, and Sex Crime.

And ignored the remaining classes. The NN model did not label any prediction in the categories: Against public order, Fraud, Deception & Corruption, Involving property and human, and Other which are reflected in it poor performance. Future research should attempt to run the model

again on a Neural Network centric platform that will enable the algorithm to expand beyond a single hidden layer (Why deep learning, 2018). It is possible 5-6 hidden layers would greatly improve the prediction accuracy of the Neural Net model. The Neural Net was worst performing model among models tested.

Table. 4. Neural Network Results

Confusion Matrix and Statistics

Prediction	Reference: Against public order	Causing bodily harm	Causing Death	Fraud, Deception & Corruption	Involving propertly only
Against public order	0	0	0	0	0
Causing bodily harm	4028	34021	23249	82	17682
Causing Death	119	455	367	3	324
Fraud, Deception & Corruption	0	0	0	0	0
Involving propertly only	30	220	110	0	222
Involving property and human	0	0	0	0	0
Other	0	0	0	0	0
Sex Crimes	18	116	92	0	45

Prediction	Reference: Involving property and human	Other	Sex Crimes
Against public order	0	0	0
Causing bodily harm	1208	708	22144
Causing Death	21	12	326
Fraud, Deception & Corruption	0	0	0
Involving propertly only	10	4	91
Involving property and human	0	0	0
Other	0	0	0
Sex Crimes	2	2	121

Overall Statistics

Accuracy : 0.3282
95% CI : (0.3253, 0.331)
No Information Rate : 0.3289
P-Value [Acc > NIR] : 0.703

Kappa : 0.0034

Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: Against public order	Class: Causing bodily harm	Class: Causing Death	Class: Fraud, Deception & Corruption
Sensitivity	0.00000	0.97728	0.015409	0.0000000
Specificity	1.00000	0.02702	0.984637	1.0000000
Pos Pred Value	NaN	0.32991	0.225569	NaN
Neg Pred Value	0.96036	0.70812	0.774953	0.9991968
Prevalence	0.03964	0.32894	0.225055	0.0008032
Detection Rate	0.00000	0.32146	0.003468	0.0000000
Detection Prevalence	0.00000	0.97439	0.015373	0.0000000
Balanced Accuracy	0.50000	0.50215	0.500023	0.5000000

	Class: Involving propertly only	Class: Involving property and human	Class: Other	Class: Sex Crimes
Sensitivity	0.012149	0.00000	0.00000	0.005335
Specificity	0.994689	1.00000	1.00000	0.996693
Pos Pred Value	0.323144	NaN	NaN	0.305556
Neg Pred Value	0.828323	0.98827	0.99314	0.786022
Prevalence	0.172660	0.01173	0.00686	0.214321
Detection Rate	0.002098	0.00000	0.00000	0.001143
Detection Prevalence	0.006491	0.00000	0.00000	0.003742
Balanced Accuracy	0.503419	0.50000	0.50000	0.501014

Confusion Matrix and Statistics

Model – Ensemble Model

The Table 5 summarizes the key variables and performance of the models used in the research project. Overall performance varied across model type. The significant variables were extracted using the variable importance function from the package "varImp" are listed in the table. The Random Forest algorithm performed best among the models tested accurately predicting Crime Type at 48.92%. Like Neural Network, Random Forest was limited in the number of trees used. The original attempt to select 1000 tree caused the computer to seize. The number of tree was decreased incrementally until the algorithm could finish within 24 hours. Given the large dataset 100 trees was the optimal number to complete the computations. The "CARET" short for Classification and Regression Training package was extremely useful for feature selection, model tuning and estimating model performance from the training set.

Table. 5. Model Summary

Model	**Stepwise Regression**	**Model**	**Logit**	**Random Forest**	**Neural Network**
Significant Variables	Median Age	**Important Variables**	Weapon Type	Weapon Type	Median Age
	Unemployment		Area Name	Victim Age	Unemployment
	Migration different county		% of Black population	Day	Migration different county
	Median House Val		Victim Ethnicity	Month	Median House Val
	Median Income		Poverty	Victim Ethnicity	Median Income
	% of Pacific population		% of population not proficient in English	Area Name	% of Pacific population
	% of Black population		Unemployment		% of Black population
	% of Migration	**Accuracy**	43.51%	**48.92%**	32.82%
Adj. R-Squared	40.4%	**Precision**	23.25%	**58.11%**	25.10%
RMSE	0.17	**Recall**	90.63%	**96.29%**	90.13%
AIC	-311887.5	**AUC**	54.96%	**62.23%**	49.98%

Extra Research – Ensemble Learning

The research team attempted to seek additional machine learning tools to investigate if model performance could be further enhanced. The next step was to experiment with an Ensemble model. An Ensemble model attempts to combine several predictive models into one single model that is effectively a simple weighted linear combination of models. Utilizing the 'caretEnsemble package' the caretList function created a list of models to be ensembled together: Decision Tree, General Linear Model, K-Nearest Neighbors and SVM. The ensemble model was cross-validated using 10 folds and 3 repeating cycles as the earlier models. Each of the caret models are sequentially trained on the same training data with the same re-sampling parameters. Correlation analysis was run to determine the relationships between models. Conceptually, combining low correlated models are desirable to maximize "triangulation". The correlation matrix did not show any pair-wise correlations above 0.60 (decision tree and GLM being the most positively correlated). The remaining relationships were below 0.22.

After cross validation is completed the caret Ensemble function was called to perform predictions on the test dataset. When comparing models, ROC is the best determinant of classification performance among compared models. The Ensemble model ROC score is 0.6957 with the next best performing model, GLM with an ROC of 0.6828.

Table. 6. Ensemble Model

```
Models: rpart, glm, knn, svmRadial
Number of resamples: 30

ROC
              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
rpart     0.5532787 0.6372439 0.6596889 0.6611706 0.6851843 0.7423287    0
glm       0.5266393 0.6455772 0.6686283 0.6828798 0.7300000 0.7767970    0
knn       0.4674233 0.4944704 0.5318082 0.5381992 0.5825255 0.6250525    0
svmRadial 0.4035309 0.4626042 0.5047068 0.5025949 0.5295923 0.5927350    0

Sens
              Min.   1st Qu.    Median        Mean   3rd Qu.       Max. NA's
rpart     0.1250000 0.2750000 0.4169872 0.445042735 0.6602564 0.79487179    0
glm       0.2750000 0.4679487 0.5064103 0.514914530 0.5860577 0.69230769    0
knn       0.1282051 0.2516026 0.3000000 0.303739316 0.3458333 0.50000000    0
svmRadial 0.0000000 0.0000000 0.0000000 0.001688034 0.0000000 0.02564103    0

Spec
              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
rpart     0.6065574 0.6721311 0.7868852 0.7799636 0.8833333 0.9508197    0
glm       0.6393443 0.7012295 0.7622951 0.7486612 0.8024590 0.8500000    0
knn       0.5833333 0.6721311 0.7704918 0.7362750 0.7868852 0.8524590    0
svmRadial 0.9180328 1.0000000 1.0000000 0.9961566 1.0000000 1.0000000    0
```

Conclusion

The current research assessed and evaluated different methods and approaches to answer the question of a data arbitrage opportunity between crime and SES data to help policy makers address crime reduction. As a result, this study concludes that weapon types and victim ages and ethnicity are significant variables in terms of assessing and predicting the crime frequency. It is also found that the random forest predicts with the highest accuracy, precision and recall among multiple models and approaches we performed. It is found that violent crime is the most accurately predicted crime category consistently within models. However, there is no such thing as a perfect model and so as this study. Despite the volume of data set this study uses, this study would perform better with data set from additional sources, such as mobile phone uses, Google street view images, and other visual translated dataset. Incorporating beyond systematic, socio-economic data, such as crimes of passions and psychological data set, would enhance the strength of the study as well.

Policy implication

After the thorough assessment of different models and approaches, this study would like to predict the crime count in Beverly Hills, California since Beverly Hills, which exhibits highest median household income, median home value and considered to be one of the ideal places to live. In order to do so, the data obtained from the Beverly Hills zip code is plugged in the Final Stepwise Model output. Based on the result, it is predicted that even residents in Beverly Hills are exposed to 41 violent crimes per year. This predicted exposure to violent crime somewhat coincides with the status quo of crime vulnerability and prediction which this paper proposed in the beginning of this study.

With results from both supervised and unsupervised learning, this study recommends policy makers to focus on weapon control, targeting area, considering migration and ages of population when designing or implementing policy on controlling crime. As the result of unsupervised learning, the crime frequency is highly predictable with the variable category of weapons type. In this sense, this study recommends to implement the policy on gun control, especially focusing on specific types of gun and reducing amount of guns. Further, the research suggests to focus on crime frequent areas with consideration of median household income and median home value. Lastly, this study recommends policy makers to consider and focus on the volume of in-bound migration and crime prevention program focusing on average victim age and median age of victim which is 33 years old.

References

1. *A Beginner's guide to Neural Network and Deep Learning*(pp. 1-15, Rep.). (n.d.). Skymind.
2. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014, November). Once upon a crime: towards crime prediction from demographics and mobile data. In Proceedings of the 16th international conference on multimodal interaction (pp. 427-434). ACM.
3. Hastie, T., Tibshirani, R., & Friedman, J. (n.d.). Linear Methods for Regression. In *The elements for Statistical Learning*(2nd ed.). Springer.
4. Hastie, T., R. T., & Friedman, J. (n.d.). Linear Methods for Classification. In *The Elements of statistical learning*. Springer.
5. Kang, H. W., & Kang, H. B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. PloS one, 12(4), e0176244.
6. Mahapatra, S. (2018). *Why Deep learning over Traditional Learning?*(pp. 1-7, Rep.).
7. Runkler, T. A. (n.d.). *Data Analytics: Models and Algorithmsfor Intelligent Data Analysis*(Vol. Chapter 8).
8. Siegel, E. (n.d.). The Ensemble Effect. In *Predictive Analytics*(pp. 133-143).