National College of Ireland

Optimized Predictive Modelling to Unfold the Links of Crime with Education, Safety and Climate in Chicago

MSc Research Project
Data Analytics

Ratna Pillai
Student ID: x18134297

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Ratna Pillai
Student ID:	x18134297
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	12/12/2019
Project Title:	Optimized Predictive Modelling to Unfold the Links of Crime with Education, Safety and Climate in Chicago
Word Count:	5906
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th December 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimized Predictive Modelling to Unfold the Links of Crime with Education, Safety and Climate in Chicago

Ratna Pillai
x18134297

Abstract

Crime is one of the common issues faced by any country that impacts both society and economy. Although, the overall crime rate in Chicago began to decline since the early 20^{th} century, violent neighbourhoods still exist and continue to disrupt public peace. Moreover, the gun violence crimes aggravated the situation in Chicago in 2016, costing more than 700 lives and monetary costs estimated over 3 billion USD in 2018. Multiple studies are being carried out continuously to understand the cause of crime and violence in Chicago with a motive to improve public safety. Common census factors and ethnicity are studied enormously in this field to understand their relationships with crime. However, the complex nature of crime creates a wide scope to study several other factors which could possibly be a cause of crime. This research aims to identify whether violation and narcotics crimes in Chicago are linked to high schools, areas with surveillance cameras and climate. Using this link, crime occurrences are predicted at a geohash level rather than at a community level. To achieve this, machine learning models like multiple regression, XGBoost, random forest and artificial neural networks are used. Each model is optimized and evaluated using standard regression metrics such as R^2 statistic and RMSE (Root Mean Squared Error). XGBoost outperformed all the other models with a highest R^2 value of 88% and RMSE value of 2.57 crime counts.

Keywords: **Crime Prediction, Predictive Model, Machine Learning, Education, Climate, Safety**

1 Introduction

1.1 An overview of Predictive Policing

Predictive policing is a tool used by police departments in multiple countries, which derives insights about criminal activity from historical crime data using analytical and statistical approaches. Due to the abundant availability of crime data, scientists and analysts have started using the power of data mining and machine learning in this field. PredPol was the first predictive analytics tool used by the Los Angeles police department to prevent future crimes from happening (Brayne and Rosenblat; 2015). Such implementations help in bringing about changes to strategies in criminal justice, law enforcement and associated agencies.

As shown in Figure 1, this model considers three major aspects as stated by (Bachner; 2013), namely: geography, time at which the crime occurred and crime patterns based

on the historical data. The insights from the tool are then utilized by police departments or government to allocate resources, enable an efficient patrol strategy and introducing awareness programs in order to reduce crime and help the neighbourhood residents stay without any fear.

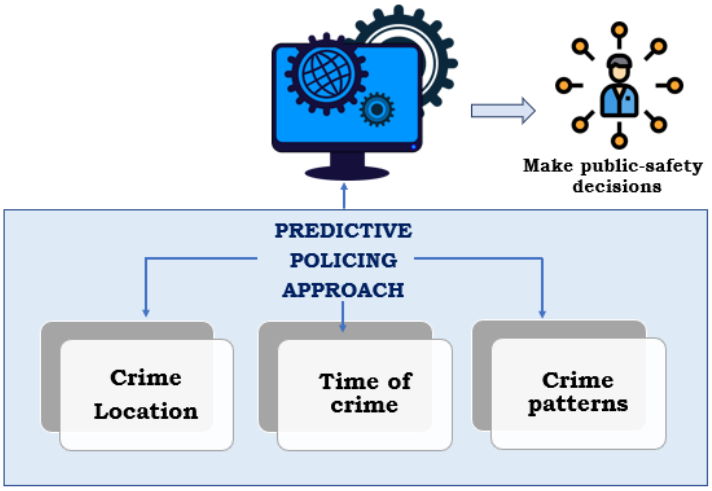

Figure 1: Predictive Policing (Bachner; 2013)

With the rise of big data in the crime field, machine learning and artificial intelligence has a great scope to perform stringent data analysis and derive interesting correlations of overlooked factors to assist the government and police departments to ensure public safety.

1.2 Background and Motivation

Chicago is a third largest city in the United States and is best known as the windy city. But recently, it is also being addressed as the capital of mass murders. As per the reports of High Intensity Drug Trafficking Area, this city is said to be a hub for conducting its money laundering for drugs activity.[1]

One key factor to note from theoretical evidences is that, both rural and urban communities face similar kind of crimes in the form of domestic and violent crimes. These involve murders, gun shootings and rapes, but somehow the link to these crimes are drugs in majority of the cases (Shukla et al.; 2019). In addition, drug consumption is seen mostly among the global youth population (Jiménez et al.; 2018; Burdick-will; 2018) and statistical studies have figured out that majority of the juveniles spent in detention centres of United States are high school drop outs.

Comparing these studies with the crime statistics in Chicago for the period 2015-2018, both assault and violation crimes have continued to increase, while homicide and narcotics crimes keep fluctuating which is shown in the Figure 2.

[1]https://www.justice.gov/archive/ndic/pubs/652/overview.htm

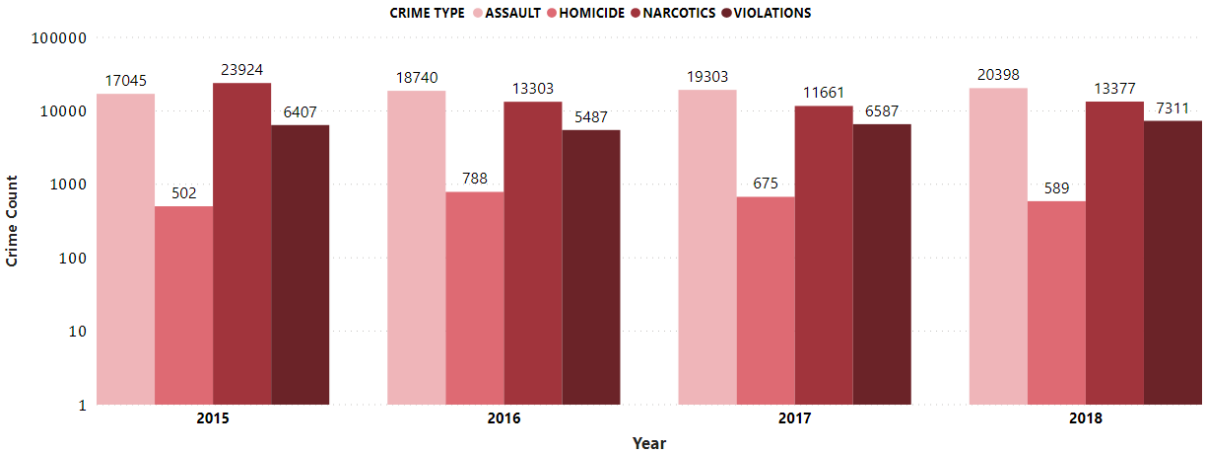

Figure 2: Distribution of Crimes in Chicago (2015-2018)

These theories that tie youths and crimes together, highly motivate the need to study the relationship between Chicago high schools and crimes. Moreover, combining high school factors with the presence of safety cameras in that area and climate might help understand crimes from a new viewpoint.

1.3 Research Question

"To what extent can machine learning approaches help in predicting Chicago crimes (such as assault, narcotics and violation crimes) using multiple factors such as high school performance, (enrollment, dropout, mobility, attendance, misconduct, suspension rates, etc.) camera surveillance areas and limited weather attributes?"

The motive of this research is to answer the above research question by merging multiple factors like attendance rates of students, teachers, misconducts, etc. with locations of surveillance cameras and limited climate factors like temperature, wind speed and precipitation with crime types like violations, assault and narcotics in Chicago. This data might help in deriving useful knowledge which will benefit Chicago police departments, government and high schools. Since the target variable to be predicted in this study is crime count (a continuous variable), regression algorithms are used in the implementation of this crime prediction model.

1.4 Research Objectives

Following objectives are defined in this study to meet the above research question:

- Data pre-processing and merging of multiple datasets with Chicago crime data.
- Selection of the best features contributing to crime prediction model using a combination of Random Forest and Recursive Feature Elimination models.
- Transformation and scaling of features using effective techniques such as one hot encoding and normalization.

- Implementation of machine learning models on the feature engineered data using Multiple Linear Regression, Random Forest, XGBoost (Extreme Gradient Boosting) and Artificial Neural Networks
- Cross validation and optimization of the models to perform predictions and evaluation.

Further, the paper is organised as Section 2 describes the related work carried out in this area, Section 3 explains the design Methodology followed for this research,Section 4 demonstrates the design flow implemented, Section 5 illustrates the machine learning model implementations, Section 6 shows the evaluation of results obtained and finally, Section 7 concludes the study.

2 Related Work

2.1 Relations of education and crime

Statistical studies done by (Lesneskie and Block; 2017) in the past state that crimes occurring in schools have negative side effects on the students such as creating a feel of low safety, drug use, depression and fear among the students. The author states that with an increasing level of safety and parent participation, these crime rates are meant to decline. Similarly, (Lockwood and Ph; n.d.) conducted a quantitative analysis which revealed that students who disobeyed the school rules often got caught up in possession of drugs, handguns or caused disturbance in the public areas like street or park. Another study by (Burdick-will; 2017) indicate that both racial attributes and socio-economic background of a neighbourhood in Chicago might have an impact on education and in turn crimes.

(Hardiman et al.; 2019) presented their findings on gun violence in America and stated that multiple causes like violence committed by the youth, school drop outs, drug and illegal weapons violation, family background, etc. might promote more gun violence. Influence of gun violence crimes on the birth outcomes by ethnicity was statistically analysed by (Matoba et al.; 2019) and the results suggest that low birth weights in infants are indirectly caused by violence in the neighbourhood. Consecutive studies by (Burdick-will; 2013), (Burdick-will; 2016) and (Burdick-will; 2018) presented the effects of neighbourhood and academics on crime in Chicago. The authors argue that both academic and justice system are connected and should be analysed together based on the past crime incidents in Chicago high schools.

2.2 Crime and other related factors

Many classification and regression problems in crime have been studied in the past. For a classification problem, evaluation is done using accuracy and for regression problems, R^2 statistic is used to define accuracy (good-fit) of the model. Below Table 1 represents the commonly addressed studies using identified potential factors influencing crime and have achieved maximum accuracy up to 97%:

Table 1: Classification and Regression studies in crime field

Author	**Models used**	**Selected features**	**Data Source**	**Best Accuracy**
(Alves et al.; 2018)	Random Forest Regressor	Urban census indicators	Brazil	80%
(Tayal et al.; 2014)	k-means clustering	Crime location, weapon used, age, sex and job of the victim	India	93%
(Kiran; 2018)	k-nearest neighbour, Naive Bayes classifier	Social, economical and environmental data	India	87%
(Gonzalez and Leboulluec; 2019)	Multiple linear regression, random forest regressor, naïve bayes regression, neural network	population, divorced male percent, poverty rates, etc.	California	96%
(Bogomolov et al.; 2008)	Random forest classifier	Mobile phone and demographic data	Europe	64%

Apart from the above associated factors, other factors have also contributed in the study of crimes. One such study reveals that the proximity of Chicago park played a major role in the reduction of crime in an area when compared to the days when the park was not implemented (Harris et al.; 2018).

Multiple factors were gathered and closely related to motor vehicle theft crimes in Manhattan by (Matijosaitiene, Mcdowald and Juneja; 2019), and with linear and boosting regression models, an accuracy of 77% was achieved. Similar analysis was performed by (Matijosaitiene, Zhao, Jaume and Jr; 2019) to study the effects of land use in Manhattan on crimes using classification models such as Logistic regression, kNN (k-nearest neighbours), Naïve Bayes and Random Forest. This work resulted in model accuracy ranging between 79% - 84%.

(Borowik et al.; 2018) examined the associations of social media and weather with theft related crimes in Poland and the results show that both seasonality and the distance of point-of-interest (a location that a person might find useful such as nearby restaurants or hospitals, etc.) can be used to determine future crime patterns. A similar research conducted by (Wang et al.; 2016) studied the impact of taxi flows and demographics on Chicago crime and with their study, they were able to reduce the error rate significantly by 17%.

(Chen et al.; 2015) obtained moderate results by analyzing the 6-hour interval weather data with twitter sentiment and theft crimes in Chicago which resulted in Area Under Curve (AUC) value of 67%.

Few notable studies which goes beyond typical census factors and explore innovative factors denote the state-of-the-art research in crime field which include the study of relationships of crime with factors such as neighbourhood traffic density (Rivera Ruiz and

Sawant; 2019), presence of urban trail parks (Harris et al.; 2018), land use (Matijosaitiene, Zhao, Jaume and Jr; 2019) and sport events occurrence (Copus and Laqueur; 2019).

2.3 Forecasting crimes using time series models

par Time series is another type of prediction that is carried out by many scientists to forecast the trends of crime patterns. Number of crimes in the future time space was forecasted on a 1-year and 2-year basis accurately using Autoregressive models in Chicago with a resulting accuracy of 84% and 80% respectively (Cesario et al.; 2016). However, the model over forecasts which is due to the shift in the residual errors towards negative scale. On the contrary, (Catlett et al.; 2019) eliminated the noise in New York and Chicago crime data using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for identifying high crime dense regions and then forecasted crime using ARIMA models. Compared to other models like RepTree and Random Forests, ARIMA models resulted in reduction of MAE (Mean Absolute Error) by 60% in this case.

The nature of the crime data could be highly complex with high non-linear relationships and dependencies with the corresponding datasets in consideration, and hence (Kang and Kang; 2017) proposed a deep neural network approach (DNN) to improve the crime prediction model rather than the traditional linear regression models. Considering spatial, temporal and environmental attributes were clustered and a deep neural network was implemented for predicting crimes in Chicago. The model performed exceptionally well with an accuracy of 84% and Area Under Curve value of 83% compared to SVM and Kernel Density Estimations (KDE).

Long short term memory (LSTM), Prophet and neural net models were applied on the daily crime data in San Francisco, Philadelphia and Chicago and the results indicated low RMSE up to 75.06 with neural net when optimized with an epoch of 300. (Feng et al.; 2019). Alternatively, an advanced approach followed by (Wang et al.; 2019) for San Francisco crimes using neural networks obtained significant forecasting results with low RMSE between 2-3 which were effective compared to the study by (Feng et al.; 2019).

Application of neural network with optimised epoch of 153 resulted in predicting the high-risk transportation stations in Chicago by studying the crime data and transportation stations data in Chicago (Kouziokas; 2017). This research also emphasizes highly on the spatial attributes and how they can affect the public safety.

2.4 Spatial and Temporal analysis of crimes

Past studies indicate that major crime activity revolves around the victim, criminal, spatio-temporal aspects and law (Brantingham and Brantingham; 1981). In addition, exploring both space and time of the crime helped in understanding the crime patterns such as behaviour of the criminal, location and time targeted for the crime based on the securities available during that time at a location.(Towers et al.; 2018) derived from their work that there exists a dependency of time, weather and crime in Chicago.

The spatial attributes of the crime were studied by (Andresen; n.d.) to predict crime clusters using census data and the census boundary units where the location attributes were geocoded, and 900 areas were finally selected for the analysis on which multinomial logistic regression was applied to identify whether a local crime cluster would occur or not. Results were significant with R^2 statistic of above 75%.

Compared to the studies conducted by (Andresen; n.d.) and (Towers et al.; 2018), hotspot detection of crimes using hyper-ensemble implemented by (Kadar; 2019) which cover both spatial and temporal aspects of crime by combining crime and low population density areas increased the hotspot prediction hit ratio by 7 percent. A novel evaluation framework was developed by (Adepeju et al.; 2016) which combined 4 different hotspot (a region which has high crime rates) prediction techniques based on repeated targets, effects of crime occurrence, clusters of crimes and predicting hotspots at random. Instead of just evaluating the accuracy of the model, the authors proposed 3 additional metrics i.e. compactness which is the ease at which the hotspots are predicted, variability which determines the variation to the hotspots with time, complementarity is the intersection rate of crimes.

From the above related work, it is significant to note that crimes are not influenced by one factor, rather there could be multiple indicators causing crime. Secondly, several work has based their ideas much on the census indicators and growth of the country and fail to notice potential factors such as educating the young minds.

3 Methodology

The various stages of this research implementation follows a design methodology resembling the stages of cross-industry process for data mining (CRISP-DM). Figure 3 represents the crime prediction design methodology covering the 6 stages namely: business understanding, data acquisition, data pre-processing, modelling, evaluation and decision making. The results generated could be then used for making effective decisions.

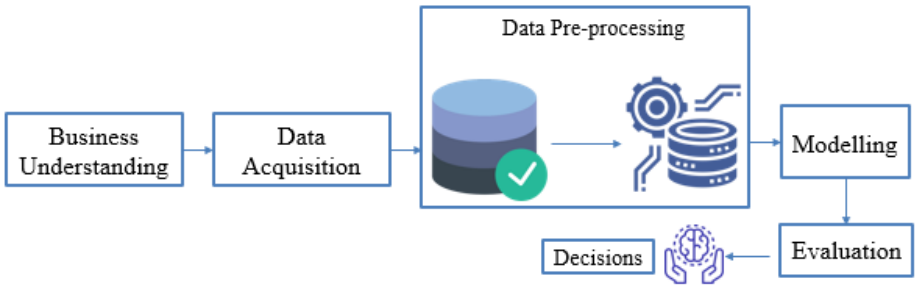

Figure 3: Crime Prediction - Design Methodology

3.1 Business Understanding

Common factors considered in the past crime prediction studies have remained more towards census features like poverty, unemployment, income, etc. or social media and environment factors like weather. However, the areas where there is no or limited access to social media or people not using social media when the crime occurs is an exception and may not contribute enough in such cases.

One practical solution to reduce Chicago crimes indicated by the theoretical studies is to consider a rigorous research on attaining insights about youth's way of thinking and assist them in taking improved choices of life by understanding the consequences of getting involved in a violent crime through education and awareness programs. This might possibly try to eliminate crimes from its roots.

3.2 Data Acquisition

Six datasets are gathered from multiple sources as shown in the Table 2, for this work. All the files except weather data, are downloaded as flat files in the form of csv (comma separated values) from Chicago Data Portal where the data is regularly updated and made available publicly for scientists and scholars to conduct studies. Limited weather data such as average temperature, wind speed and precipitation are extracted through Application Programming Interface (API) from Nation Centers for Environmental Information (NCEI) and saved as a csv file.

Out of several crime types (6.9 million records), only assault, narcotics, homicide and violation crimes comprising of 2,01,587 records are considered for this study, as these crimes are closely related to youth population based on the theoretical studies and statistical analysis. The high school report, location details (latitudes and longitudes) of speed cameras,red light cameras and police stations and weather data have been pre-processed first to merge it with crime data.

Table 2: Dataset Description

Dataset	**Record Count**	**Attribute Count**
Crimes(2015-2018)	2,01,587	30
High School Report(2015 - 2018)	752	106
Speed Camera Locations	161	7
Red Light Camera Locations	149	8
Police Station Locations	23	16
Daily Weather	1461	6

3.3 Data Pre-processing

Data pre-processing was one of the crucial phases of this study, where data manipulation activities like handling of missing values, merging of datasets and feature engineering activities such as scaling and transformation were done.

3.3.1 Handling of Missing values

- In crime dataset, all the missing values were related to geographical attributes such as latitudes, longitudes, zip codes and coordinates. Since this research focuses on merging all the datasets based on geography and time parameters, dropping these missing values would have resulted in loss of data, making the study ineffective.

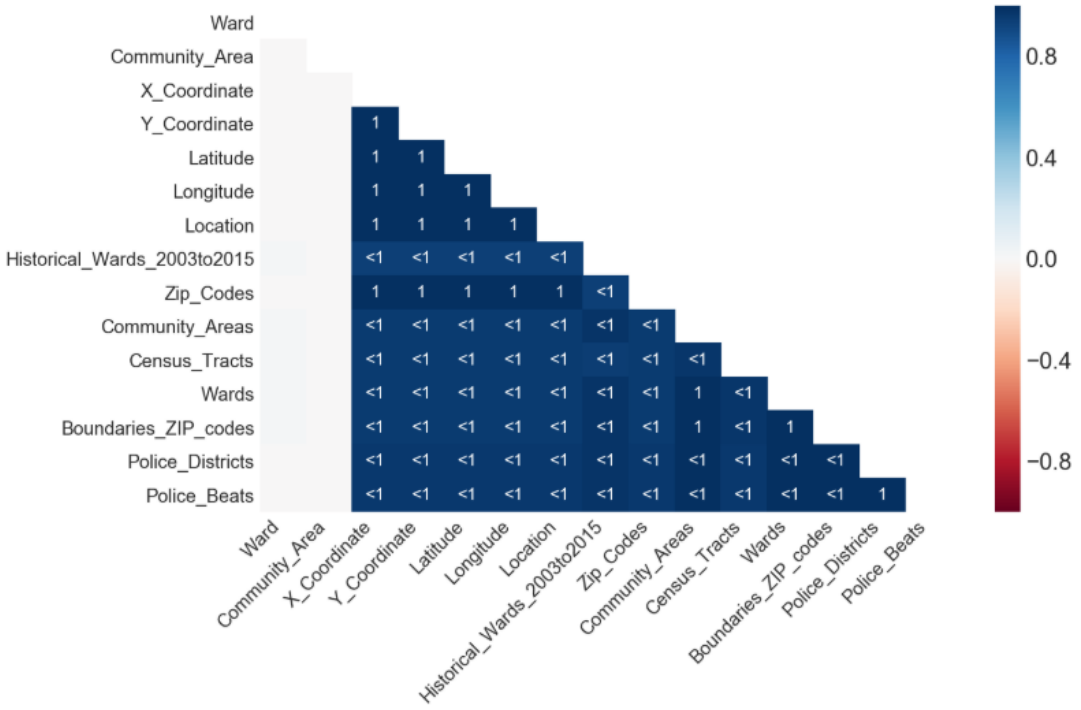

Figure 4: Missing Values in Crime Dataset

- Correlation matrix in Figure 4 clearly shows that one missing geography attribute is highly correlated to another, with high correlation coefficients close to 1. This indicates that a fix to one geography attribute will easily enable fixing other missing attributes.
- The block addresses were present for all the crime incidents reported and hence, was used in fetching the missing latitude and longitudes using Google maps API.
- After fixing, all the other geographical columns like community area, x_coordinate, y_coordinate, historical wards, zip codes and police beats were dropped, as they were not in scope for this study.
- Missing values check present in high school education data were imputed using median values, and in the remaining datasets, the location co-ordinates were already available.

3.3.2 Nearest Distance Computation

For each crime incident, distance between the latitude and longitude of the crime location and police station were computed first, to eventually get the smallest police station distance from crime location. Similarly, the nearest distance of presence of a red-light camera and speed camera were calculated using a user defined distance function.

3.3.3 Merging of datasets

The six datasets in consideration had some relevant attributes based on which the merging has been carried out. These were primarily geographical attributes (latitude and longitude) or date-time attributes (date expressed as year, month, day and hour).

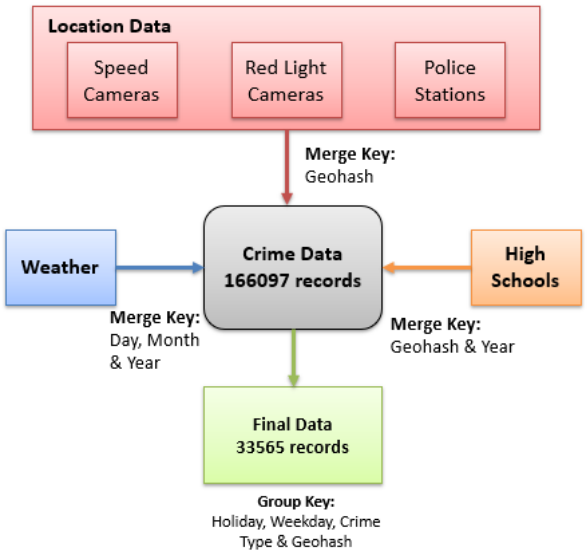

Figure 5: Merging of Datasets

- For better assessment of data, all the latitudes and longitudes in each dataset were converted to geohash with precision five, using the pygeohash library. This converts the location co-ordinates to a unique alphanumeric string. This helped in clustering multiple nearby co-ordinates into one area representing a geohash.
- As shown in the Figure 5, the location details of police station, speed cameras and red cameras were merged with crime data using geohash. Both weather and high school report had geohash and date attributes, based on which they were merged with crime data. Finally, flags indicating whether that particular day is a holiday or weekday was obtained and time of the day indicating morning, afternoon, evening and night were calculated based on the Day attribute against each crime record.
- Lastly, the columns latitude, longitude and day were dropped and the data was grouped using weekday, holiday and geohash attributes. The final dataset after merging and grouping resulted in 33565 records.
- It is interesting to note through the exploratory data analysis, that violation and drug related crimes tend to happen more during the afternoon and evening compared to morning or night time, which is also evident in the Figure 6.

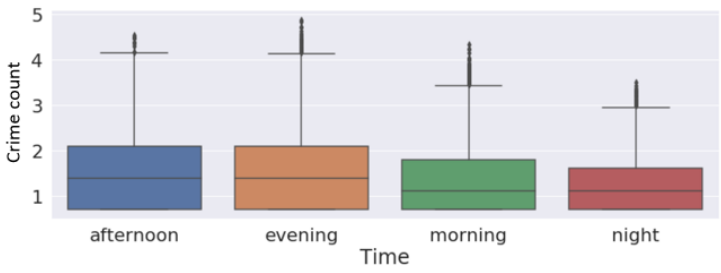

Figure 6: Time of the day at which crimes occurred (2015-2018)

- Correlation check was done on the merged dataset and very weak correlations were observed between the independent features and crime count. Below Figure 7 shows the matrix of features having a correlation co-efficient above 0.15.

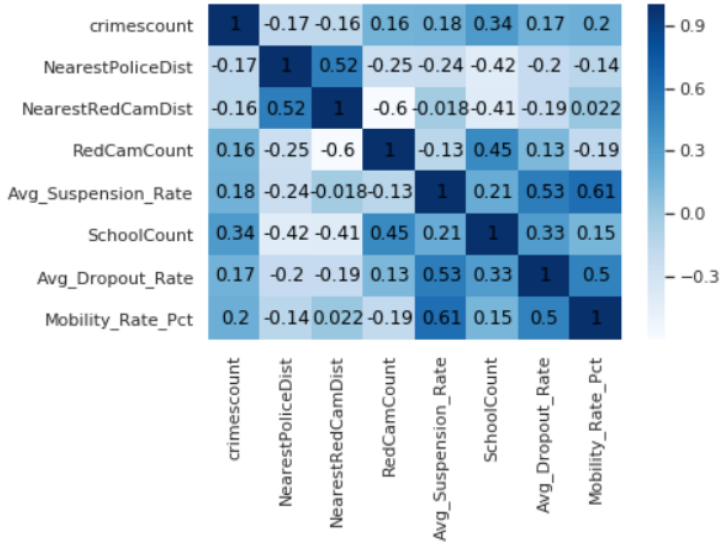

Figure 7: Crime correlation Matrix showing weak correlations

3.3.4 Feature Engineering

The merged dataset consists of both categorical and numerical features, and each require a different kind of engineering before the desired machine learning models are applied. Feature selection techniques are also implemented to decide on the best features from high school, safety enforcement and weather data defining the dependent variable, crime count.

3.3.5 One Hot Encoding

- The categorical variables present in the dataset include geohash, crime type, year, month, weekday, and holiday and these categories are encoded using One Hot Encoding technique.

- This type of feature engineering converts the categorical feature to its corresponding binary values by transforming each category to a feature in the dataset. For example, the attribute weekday having categories 0 and 1 is converted as weekday_0 and weekday_1. After this conversion, the total number of attributes in the dataset increased from 28 to 96 attributes.

3.3.6 Feature Scaling

- The numerical attributes in the dataset comprised of attendance, misconduct, suspension, school mobility features, etc. expressed as a percent rate, wind speed as miles per hour, average temperature as degree Celsius and precipitation measured in millimetres, the distances of police station, cameras are indicated in kilometres and the remaining features are expressed as counts.

- These different types of scales may highly affect certain machine learning models and their results. Hence, they were brought to a common scale using normalization. In addition, this also improves the performance of a model, in terms of speed (Bogomolov et al.; 2008).

- Normalization has not been carried out for tree-based models XGBoost and Random Forest as they build multiple decision trees using values that are absolute.

3.3.7 Feature Selection

For improving the crime prediction model accuracy and remove multicollinearity, it is essential to understand and obtain the best features from high school, safety camera locations and weather attributes, that describes the target variable crime count.

- In multiple crime studies, Pearson and spearman correlation coefficients are compared to understand the linear relationship of several factors with crime occurrences (Bogomolov et al.; 2008) and (Wang et al.; 2016). However, the data considered in this work does not possess linear relationship and hence an alternate method such as Recursive Feature Elimination using Random Forest is required.

- RFE (Recursive Feature Elimination) combined with Random Forest(RF) has been a successful feature selection in past studies (Granitto et al.; 2006). This technique ranks the best features defining crime by fitting a Random Forest model.

- Out of 20 numerical features, top 10 features namely: student attendance rate, teacher attendance rate, school mobility rate, nearest police distance, nearest red-light camera distance, red light camera count, school count, average temperature, average wind speed and precipitation with rank one are selected, as shown in Figure 8 based on feature rankings using RFE-RF method.

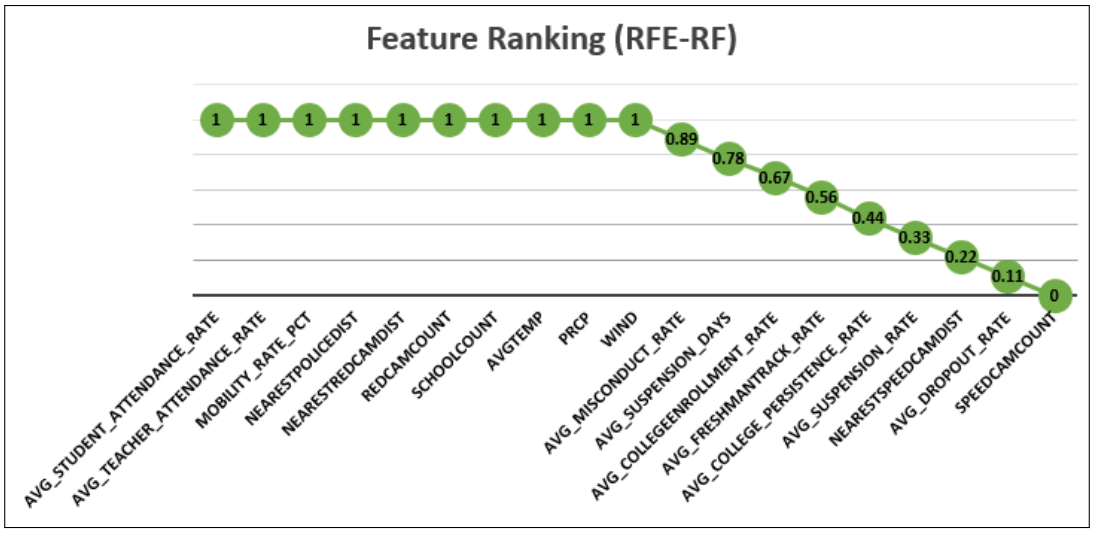

Figure 8: Feature Ranking using RFE-RF method

4 Design Specification

A three stage design flow has been followed, comprising of data, modelling and visualization stages as shown in Figure 9.

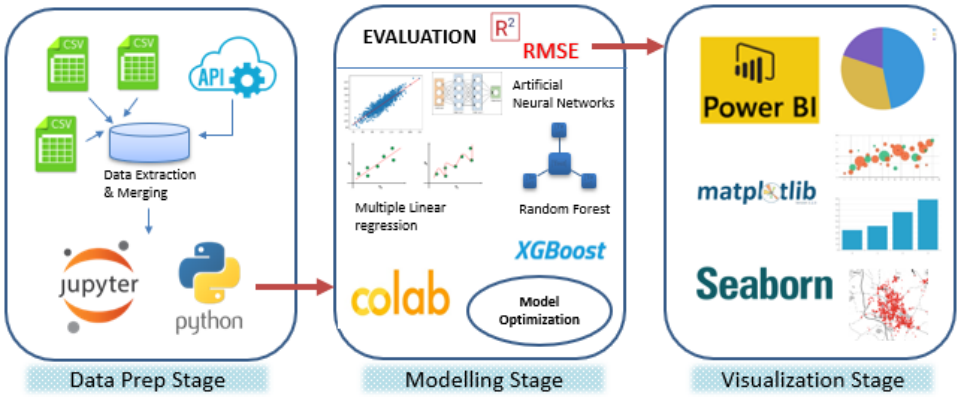

Figure 9: Crime Prediction - Design Flow

- The data preparation stage consists of all the stages carried out specific to data gathering, merging, exploratory data analysis, feature engineering and feature selection stages. Based on the data source, each data was downloaded as a csv (comma separated values) files or extracted using API connections using python programming on Jupyter Notebooks.
- Modelling stage deals with implementation of multiple models such as linear regression, xgboost, random forest and neural networks were done and evaluated using metrics such as R^2 statistic and RMSE (Root Mean Squared Error). Modelling and optimization have been done on Google Colaboratory.
- Lastly, the results obtained were presented in the form of plots and graphs as desired for visualization purpose.

Each model is tested for optimization, has undergone cross validation and evaluated using appropriate metrics which are explained in detail in the following sections.

5 Implementation

The implementation of this research is carried out in three phases, namely: sampling of data, cross validation and tuning of models with hyper parameters using Randomized Search CV (cross validation) techniques. The main reason for choosing this hyperparameter tuning method is the lower processing times associated with it.

5.1 Data Preparation for the models

The final dataset shape at the modelling stage was 33565 records with 86 columns after feature engineering and standardization. This data was then split into 80% training and

20% testing set. To achieve effective sampling of data and reduce over-fitting during the training phase, cross validation with a series of k-fold tests (3-30 folds) were performed. Figure 10 shows stability of mean squared error (MSE) and mean absolute error (MAE) between 10 and 20 folds for both Random Forest and XGBoost models. Also, as quoted in (Witten et al.; 2011), 10 folds are ideal for many datasets based on multiple experiments carried out in the past. Thus, data sampling was done using 10 fold cross validation was implemented across the models.

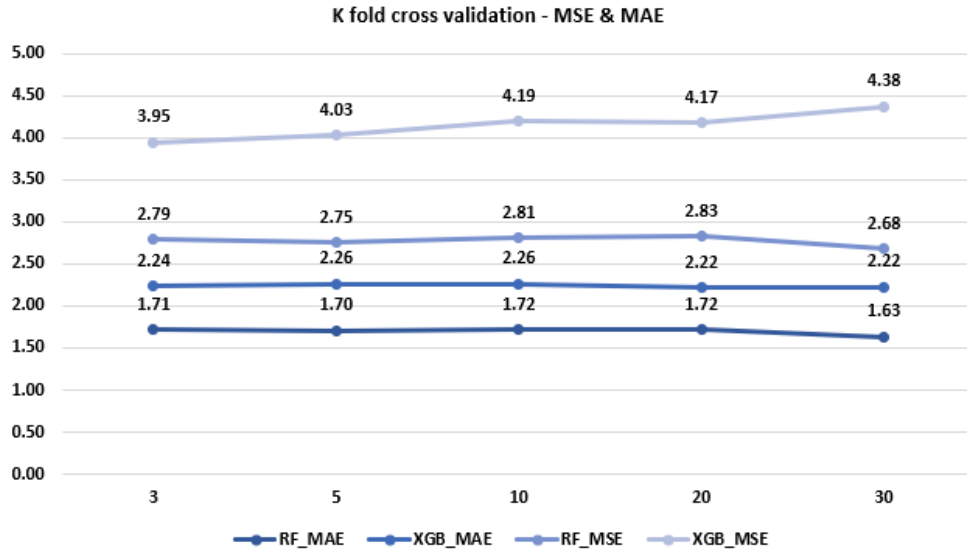

Figure 10: MSE vs. MAE for k fold (3-30 folds) cross validation in Random Forest and XGBoost models

5.2 Random Forest Regressor

Random Forest is robust to multicollinearity, scaling and normalization and this has made it very popular in machine learning (Alves et al.; 2018). Compared to linear models, this ensemble learning model builds multiple decision trees using bootstrap and the predicted output is determined across the trees by the majority votes or aggregating the predictions. This characteristic is often referred to bagging, as the features are bagged to obtain best prediction performance. In this study, random forest model was applied for default parameter values as well as the best parameters selected using optimization.

Optimization was done using Randomized search hyper parameter tuning and the best parameters selected are outlined in the below Figure 11 with max depth as 50, minimum sample split as 5, n estimators as 377 and minimum sample leaf as 4:

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=50,
                      max_features='auto', max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=4, min_samples_split=5,
                      min_weight_fraction_leaf=0.0, n_estimators=377, n_jobs=-1,
                      oob_score=False, random_state=20, verbose=0,
                      warm_start=False)
```

Figure 11: Random Forest using Python - Selected parameters

5.3 XGBoost Regression

Extreme gradient boosting (XGBoost) Regressor is the most popular algorithm used in machine learning to solve regression problems. As the name suggests, it is a boosting algorithm which implements gradient-boosted trees and the learning of the model occurs from the residuals of the former predictor variables (Matijosaitiene, Mcdowald and Juneja; 2019). Like random forest regressor, XGBoost regressor was also trained for both default settings and optimized using the best parameters.

With hyper parameter optimization using Randomized search cv, the best parameters were selected as outlined in the below Figure 12 with gamma value 1.5, n_estimators as 200 and learning rate 0.05:

```
XGBRegressor(base_score=0.5, booster='gbtree', bootstrap=True,
             colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1.0,
             gamma=1.5, importance_type='gain', learning_rate=0.05,
             max_delta_step=0, max_depth=8, min_child_weight=10, missing=None,
             n_estimators=200, n_jobs=-1, nthread=None, objective='reg:linear',
             random_state=20, reg_alpha=0, reg_lambda=1, scale_pos_weight=1,
             seed=None, silent=None, subsample=0.75, verbosity=1)
```

Figure 12: XGBoost using Python - Selected parameters

5.4 Artificial Neural Networks

Artificial Neural Networks (ANN) also referred as deep learning, is highly inspired by the working of neurons in a brain. The neural network architecture consists of three layers, namely: the input layer, hidden layer and output layer. Keras library in python is used to implement a sequential neural network model in this study. The number of layers and neurons are defined using Dense constructor. Iterative testing is done using multiple layers, different batch sizes and epochs. Error minimization is primarily done in order to achieve accurate predictions of the outcome (Kang and Kang; 2017) using a suitable optimizer.

As shown in the Figure 13, 21,377 total trainable parameters are produced by the ANN model for multiple dense layers, each having a set of neurons (128,64,32 respectively) in this case.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_163 (Dense)            (None, 128)               11008
_________________________________________________________________
dense_164 (Dense)            (None, 64)                8256
_________________________________________________________________
dense_165 (Dense)            (None, 32)                2080
_________________________________________________________________
dense_166 (Dense)            (None, 1)                 33
=================================================================
Total params: 21,377
Trainable params: 21,377
Non-trainable params: 0
```

Figure 13: ANN architecture built using Python with input, hidden and output layers

5.5 Multiple Linear Regression

The basic model to study the relationship of multiple features with the crime count is to use multiple regression models. This model is statistically represented as:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon_n \quad (1)$$

where represents coefficients, X_n represents the features is the residual and n is the total number of features used in the prediction of the dependent variable (Gonzalez and Leboulluec; 2019). This is a standard model applied for regression problems, but it specifically requires certain assumptions to be fulfilled such as the linearity of variables (which is absent in the data considered), homoscedasticity and normality of residuals. These assumptions are verified and evaluated in this study.

6 Evaluation

Evaluation of the models are done using R^2 statistic that determines the model fit, MSE and RMSE values that estimate the errors and are calculated based on actual and predicted values.

6.1 Feature Selection Experiments using RFE-RF

- Since the categorical variables were encoded using One-Hot Encoding, it resulted in more dimensions in the dataset. Using RFE-RF method, top 20 features with one hot encoded variables were selected.
- Table 3 shows the accuracy of the model with the selected numerical and categorical features which is around 55% with Random Forest and 57% with XGBoost models. Since much of the categorical variables were lost due to the order of importance, the remaining variables were not enough to define the target variable, and was making the model inefficient.

Table 3: Feature Selection (RFE-RF) Experiments

RFE-RF Experiment	**Random Forest (R^2)**	**XGBoost (R^2)**
Categorical+Numerical Features	55%	57%
Numerical Features	85%	65%

- Another round of feature selection was performed just with the numerical features, and the categorical features were retained as binary encoded itself. This approach considerably boosted the model accuracy upto 30% with random forest and 9% with xgboost models. Hence, the further experiments are carried based on RFE-RF features with best performing numerical features and all one-hot encoded categorical features.

6.2 Experiments with Random Forest

Random forest was experimented on the dataset with default and tuned parameters and evaluated using train and test accuracy with RMSE . As shown in Table 4, accuracy with default parameters of random forest led to over-fitting with 96% train accuracy, which means that the train model has learnt too well compared to the test model with 84% accuracy. This over-fitting was reduced using optimization and 10 fold cross validation, and the resulting train accuracy was brought down to 92% and test accuracy was improved overall by 2%.

Table 4: Random Forest Experiments

Experiment	**Train Accuracy (%)**	**Test Accuracy (%)**	**Overfitting**
Default Parameters	96.81	84.54	12%
Tuned Parameters	92.19	85.50	7%
10 fold + Tuned Parameters	92.29	86.85	5%

6.3 Experiments with XGBoost

Multiple experiments were conducted on XGBoost and evaluated using RMSE and train-test accuracy. Based on the figures shown in the Table 5, the model was training efficiently with an accuracy boost up to 23% using the optimized parameters obtained with hyper parameter tuning and 10 fold cross validation. There is also a notable reduction in the error rate (RMSE) by 2%, achieved through optimization and cross validation.

Table 5: XGBoost Experiments

Experiment	**Train Accuracy**	**Test Accuracy**	**Test RMSE**
Default Parameters	67.82	65.29	4.38
Tuned Parameters	92.19	86.46	2.7
10 fold + Tuned Parameters	92.29	88.14	2.52

6.4 Experiments with Artificial Neural Networks

Experiments with ANN is conducted by building two architectures, one with the default layers (input and output layers) and the other one with three layers respectively. Default layer architecture has an input layer with 128 neurons and the 3 layered architecture consists of (128, 64, 32) neurons in each layer respectively. With the rmsprop (Root Mean Squared Propogation) optimizer and reLU (Rectified Linear Unit) activation function, the models were trained for epochs through 10-50.

Table 6: ANN Experiments with default layers

Epochs	**max val MSE**	**min val MSE**
10	37.33	19.64
20	37.34	10.55
30	36.37	10.85
40	39.75	9.31
50	18.50	8.41

Initially, 128 neurons in the input layer was implemented for epochs (10-50) with batch size of 64. As shown in the Table 6, this model resulted in the minimum validation (val) loss i.e. MSE value to reduce up to 9 counts from 18 counts between 40^{th} and 50^{th} epochs, indicating the model is able to learn well.

Table 7: ANN Experiments with 3 layers

Epochs	**max val MSE**	**min val MSE**
10	38.75	10.49
20	40.06	11.20
30	27.99	10.12
40	41.88	8.07
50	23.51	7.36

The next round of experiment was done with 3 layers. Multiple rounds of epoch (10-50) were done, but using a batch size of 128 this time. As shown in the Table 7, the loss term (MSE) falls consistently. Compared to one layer model, this model was able to reduce loss term by 16 counts.

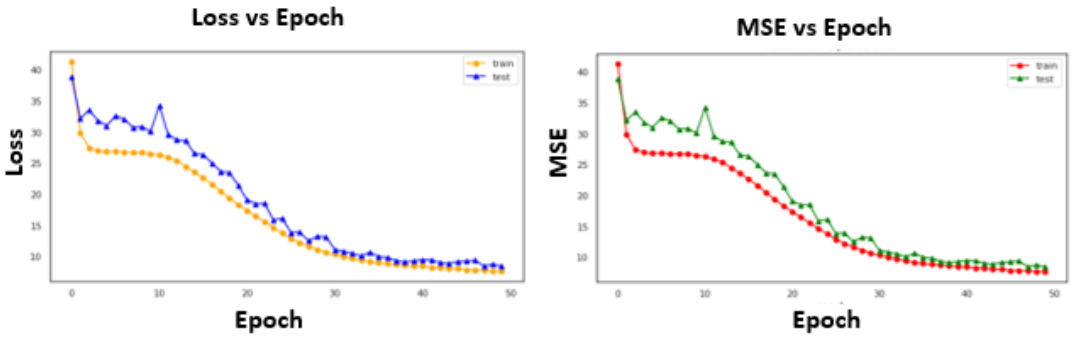

Figure 14: Loss and MSE Evaluation over epochs

As shown in Figure 14, initially the validation loss and mse seems to fluctuate, but the validation loss and mse (mean squared error) values tend to become stable after 36th epoch with the right choice of optimization.

Figure 15: ANN - crime count (actual values vs. predicted values)

With multi-layered architecture, R^2 obtained is 84% which is 1% higher when compared to a single layered ANN architecture. The resulting predictions have also been plotted against the actual crime count values in Figure 15, where the actual crime counts and predicted crime counts exhibit a strong correlation.

6.5 Experiments with Multiple Linear Regression

Achieved R^2 with multiple linear regression was around 41% and as suspected due to the presence of weak linear relationship between the predictors and target feature crime count, this model did not perform well. Other assumptions of multiple linear regression such as homoscedasticity and residual linearity were also checked, however these assumptions were not satisfied with this type of data selection.

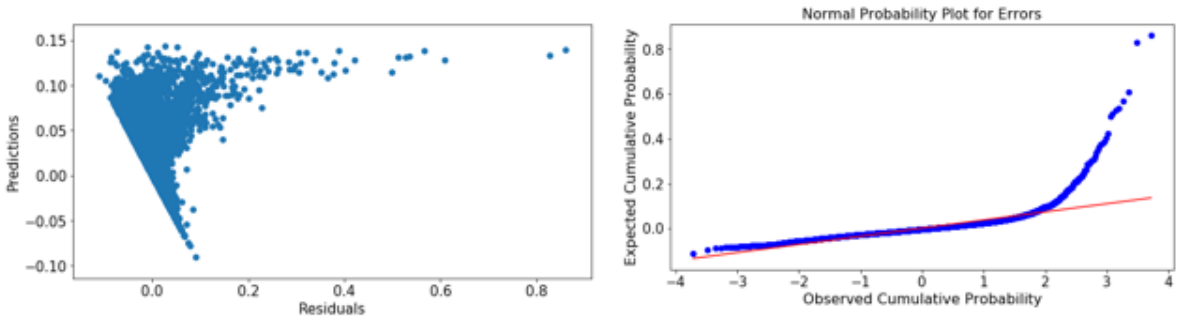

Figure 16: Multiple Linear Regression - Assumptions check

In the Figure 16, the left plot shows a cone shaped scatter of error denoting that there is no random distribution of error term (heteroscedastic), whereas the right plot shows the quantile plot where the distribution of the error is slight curved towards the end instead of being linear.

6.6 Discussion

This study is done with an idea of exploring the relationships of multi variate factors from high schools, traffic surveillance camera areas and climate on the crimes occurrences in Chicago. Unlike the traditional linear models implemented by (Gonzalez and Leboulluec; 2019), where the socio-economic data like poverty, illiteracy, unemployment rates tends to have strong relationships with crime rate, the major challenge faced was the complex independent variables considered in this research which had weak correlations with the crime count. One possible theory behind strong correlations in these studies is because, the crime rate itself is computed using population and the socio-economic indicators are also derived based on population and this derivation might possibly explain strong linear relations between them. However, due to the unavailability of population data geohash wise for the period 2015-2018, crime rate was excluded and this study had to rely on crime count as a dependent variable.

Four machine learning models were implemented and evaluated in this study. Various techniques to improve the model performance were adopted, such as grouping the dataset based on multiple features, one-hot encoding and feature selection techniques. In addition, significant performance improvement are brought to the work, by optimizing the models using randomized search cv. All the models except multiple linear regression, performed well with an accuracy value greater than 80%. As shown in Figure 17, XGBoost is the best performing model which resulted in highest R^2 value and is 2% to 4% higher than accuracy obtained using random forest models and ANN. The error rates (RMSE) is 2.52 crime counts which is considerably less than the errors generated by ANN (8.83 counts). Both xgboost and random forest are ensemble learners, however the approach of building decision trees in both the cases is different. On the contrary, XGboost handles the bias-variance trade-off effectively, however random forest trees use bagging approach which just works on reducing the variance.

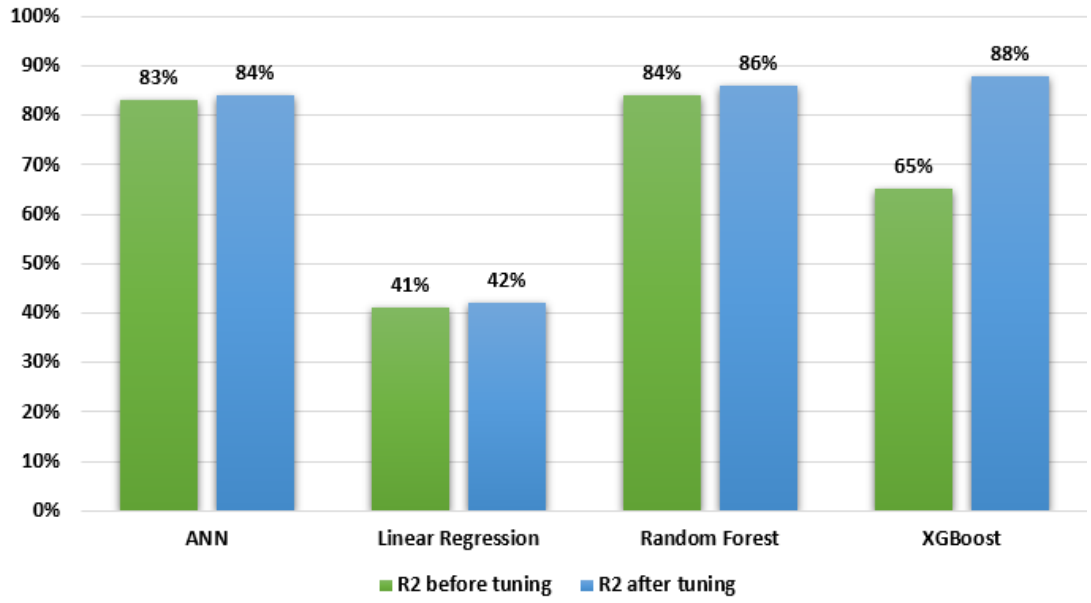

Figure 17: Performance comparison of models

Based on the results, it is observed that boosting models perform well with this type

of data compared to bagging. This result is remarkable as both the bias and variance is handled effectively. While comparing the results obtained in this research with the past studies, gradient boosting algorithm opted in the study (Matijosaitiene, Mcdowald and Juneja; 2019) resulted in a mean R^2 value of 61%, while this study was able to produce the highest R^2 of 88% by opting extreme gradient boosting technique. Crime studies using homicide crimes done by (Alves et al.; 2018) resulted in 80% R^2 value and (Bogomolov et al.; 2008) achieved 64% accuracy using random forest regressor, but this implementation attained R^2 statistic around 86%. Further, an accuracy value of 84% is obtained with artificial neural network, which is similar to the accuracy obtained with the deep neural network implemented to predict the crime occurrence (classification problem) in Chicago (Kang and Kang; 2017).

The performance of multiple linear regression was very poor with a resulting accuracy of 42% which is due to a lack of fulfillment of the linear model assumptions like linearity, homoscedasticity and normality. These problems have impacted several other crime studies in the past where these linear models resulted in predicting negative values, especially in cases that deal with crime count as a target variable (Alves et al.; 2018) and (Ingilevich and Ivanov; 2018). One permissible solution to tackle these problems is to test using models like Poisson or Negative Binomial regression (Wang et al.; 2016) and (Mares and Moffett; 2019).

7 Conclusion and Future Work

In this research, links of Chicago high schools, areas surveilled by cameras and climate were studied on assault, narcotics and violation crime types. The major high school factors identified to be having an effect on crime occurrences were school mobility rates and attendance rates of students and teachers. Even though the impact of these factors on crime cannot be determined due to a lack of measurable relationship, important decisions still can be taken by high schools to improve these rates and create more awareness among the youth to prevent them from committing crimes.

Another interesting point to note from this research is that the presence of red-light cameras in a location also contributed to the crime prediction. However, the presence of a speed camera did not contribute to the types crimes selected in this research. Although, all the weather attributes considered impact these crimes in some way, effective patterns could not be generated. However, the findings from this study might potentially benefit multiple stakeholders. Chicago police department would be able allocate appropriate resources in a geohash locality based on the presence of surveillance cameras. With an adequate consultation and awareness programs, consequences of crime and violence can be taught in the schools to guide the young minds to progress in life instead of ending up in detention centres for committing crimes.

Consideration of data for a period of four years limits the capability of the models and adding more data could result in better insights. Alternatively, effective use of clustering techniques like DBSCAN or Hierarchical DBSCAN can be done for eliminating noise in crime data and to improve crime predictions.

Acknowledgement

Firstly, I would like to thank my research supervisor Dr. Muhammad Iqbal for relentlessly supporting and motivating me throughout the research work. For a period of 13 weeks, my supervisor assisted me with my queries and guided towards right direction. I must sincerely thank my family for their exceptional love, constant understanding and support. Lastly, thanks to my friends who have been an inspiring energy source to me.

References

Adepeju, M., Rosser, G. and Cheng, T. (2016). Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study, *International Journal of Geographical Information Science* **30**(11): 2133–2154.
URL: *http://dx.doi.org/10.1080/13658816.2016.1159684*

Alves, L. G. A., Ribeiro, H. V. and Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning, *Physica A* **505**: 435–443.
URL: *https://doi.org/10.1016/j.physa.2018.03.084*

Andresen, M. A. (n.d.). Predicting Local Crime Clusters Using (Multinomial) Logistic Regression, **17**(3): 249–262.

Bachner, B. J. (2013). Predictive Policing : Preventing Crime with Data and Analytics, pp. 86–90.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F. and Pentland, A. (2008). Once Upon a Crime : Towards Crime Prediction from Demographics and Mobile Data.

Borowik, G., Wawrzyniak, Z. M. and Cichosz, P. (2018). Time series analysis for crime forecasting.

Brantingham, P. and Brantingham, P. (1981). *Environmental criminology*, Sage Publications.
URL: *https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=87681*

Brayne, S. and Rosenblat, A. (2015). Predictive Policing.

Burdick-will, J. (2013). NIH Public Access, **86**(4).

Burdick-will, J. (2016). Neighborhood Violent Crime and Academic Growth in Chicago: Lasting Effects of Early Exposure, **95**(September): 133–157.

Burdick-will, J. (2017). Neighbors but Not Classmates : Neighborhood Disadvantage , Local Violent Crime , and the Heterogeneity of Educational Experiences in Chicago, **124**(NOVEMBER).

Burdick-will, J. (2018). Neighborhood Violence , Peer Effects , and Academic Achievement in Chicago.

Catlett, C., Cesario, E., Talia, D. and Vinci, A. (2019). Spatio-temporal crime predictions in smart cities : A data-driven approach and experiments, *Pervasive and Mobile Computing* **53**: 62–74.
URL: *https://doi.org/10.1016/j.pmcj.2019.01.003*

Cesario, E., Catlett, C. and Talia, D. (2016). Forecasting Crimes using Autoregressive Models.

Chen, X., Cho, Y. and Jang, S. Y. (2015). Crime prediction using Twitter sentiment and weather, *2015 Systems and Information Engineering Design Symposium, SIEDS 2015* **00**(c): 63–68.

Copus, R. and Laqueur, H. (2019). Entertainment as Crime Prevention: Evidence From Chicago Sports Games, *Journal of Sports Economics* **20**(3): 344–370.

Feng, M., Zheng, J., Ren, J., Hussain, A. and Qiaoyuan, L. (2019). Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data, pp. 106111–106123.

Gonzalez, J. J. and Leboulluec, A. (2019). Crime Prediction and Socio-Demographic Factors : A Comparative Study of Machine Learning Regression-Based Algorithms, **13**(1): 13–18.

Granitto, P. M., Furlanello, C., Biasioli, F. and Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products, **83**: 83–90.

Hardiman, E. R., Jones, L. V. and Cestone (2019). Social Work in Public Health Neighborhood Perceptions of Gun Violence and Safety : Findings from a Public Health-Social Work , **1918**.

Harris, B., Larson, L. and Ogletree, S. (2018). Different Views From The 606 : Examining the Impacts of an Urban Greenway on Crime in Chicago.

Ingilevich, V. and Ivanov, S. (2018). Crime rate prediction in the urban environment using social factors, *Procedia Computer Science* **136**: 472–478.
URL: *https://doi.org/10.1016/j.procs.2018.08.261*

Jiménez, R., Anupol, J., Cajal, B. and Gervilla, E. (2018). PT NU Authors names and affiliations, *Addictive Behaviors Reports* .
URL: *https://doi.org/10.1016/j.abrep.2018.09.005*

Kadar, C. (2019). Public decision support for low population density areas : An imbalance- aware hyper-ensemble for spatio-temporal crime prediction, *Decision Support Systems* **119**(September 2018): 107–117.
URL: *https://doi.org/10.1016/j.dss.2019.03.001*

Kang, H.-w. and Kang, H.-b. (2017). Prediction of crime occurrence from multi- modal data using deep learning, pp. 1–19.

Kiran, J. (2018). Prediction Analysis of Crime in India Using a Hybrid Clustering Approach, *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on* pp. 520–523.

Kouziokas, G. N. (2017). ScienceDirect The application of artificial intelligence in public administration for forecasting high crime risk transportation areas in urban environment, *Transportation Research Procedia* **24**: 467–473.
URL: *http://dx.doi.org/10.1016/j.trpro.2017.05.083*

Lesneskie, E. and Block, S. (2017). School Violence : The Role of Parental and Community Involvement, **16**(4): 426–444.

Lockwood, D. and Ph, D. (n.d.). National Institute of Justice.

Mares, D. M. and Moffett, K. W. (2019). Climate Change and Crime Revisited: An Exploration of Monthly Temperature Anomalies and UCR Crime Data, *Environment and Behavior* **51**(5): 502–529.

Matijosaitiene, I., Mcdowald, A. and Juneja, V. (2019). Predicting Safe Parking Spaces : A Machine Learning Approach to Geospatial Urban and Crime Data, pp. 1–15.

Matijosaitiene, I., Zhao, P., Jaume, S. and Jr, J. W. G. (2019). Prediction of Hourly Effect of Land Use on Crime, pp. 1–13.

Matoba, N., Reina, M., Prachand, N., Davis, M. M. and Collins, J. W. (2019). Neighborhood Gun Violence and Birth Outcomes in Chicago, (June): 1251–1259.

Rivera Ruiz, D. and Sawant, A. (2019). Quantitative Analysis Of Crime Incidents In Chicago Using Data Analytics Techniques, *Computers, Materials & Continua* **58**(2): 389–396.

Shukla, R. K., Stoneberg, D., Lockwood, K., Copple, P., Dorman, A. and Jones, F. M. (2019). University of Central Oklahoma, (0123456789).

Tayal, D. K., Jain, A., Arora, S., Agarwal, S., Gupta, T. and Tyagi, N. (2014). Crime detection and criminal identification in India using data mining techniques, *AI and Society* **30**(1): 117–127.

Towers, S., Chen, S., Malik, A. and Ebert, D. (2018). Factors influencing temporal patterns in crime in a large American city: A predictive analytics perspective, *PLoS ONE* **13**(10): 1–27.

Wang, H., Kifer, D., Graif, C. and Li, Z. (2016). Crime Rate Inference with Big Data, pp. 635–644.

Wang, Q., Jin, G., Zhao, X., Feng, Y. and Huang, J. (2019). Knowledge-Based Systems CSAN : A neural network benchmark model for crime forecasting in, *Knowledge-Based Systems* (xxxx): 105120.
URL: *https://doi.org/10.1016/j.knosys.2019.105120*

Witten, I. H., Frank, E. and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.