University of East London

8/31/2023

DEVELOPING PREDICTIVE MODELS FOR CRIME USING LONDON-BASED TWITTER DATA

BY

DAMILOLA OLUWASENI OMISORE

2272153

A PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF MASTERS OF SCIENCE DEGREE IN DATA SCIENCE, SCHOOL OF ARCHITECTURE, COMPUTING AND ENGINEERING, UNIVERSITY OF EAST LONDON, UNITED KINGDOM.

SUPERVISED BY

DR. YANG LI

ACKNOWLEDGEMENT

I begin by extending my heartfelt gratitude to the Almighty God for His unwavering blessings and guidance that have enabled me to successfully complete this endeavour.

I wish to express my deep appreciation to my supervisor, Dr Yang Li, whose invaluable insights and constructive feedback significantly enriched and refined my work. Despite his demanding schedule, his contributions played a pivotal role in shaping the quality and focus of this study.

I am equally indebted to him as he is the esteemed program leader, whose guidance and mentorship throughout my master's journey have been instrumental in shaping my intellectual growth. I am truly grateful for his teachings and unwavering support.

My heartfelt thanks go to my family and friends for their unyielding encouragement and unwavering support throughout my pursuit of a master's degree. I am particularly grateful for the support, prayers, and uplifting words from Lanre and John during this phase.

Lastly, I extend my sincere appreciation to all the faculty members who imparted their knowledge and provided guidance during my M.Sc. program. Their dedication to teaching and mentorship has been invaluable to my academic journey.

ABSTRACT

The advent of social media has revolutionized communication patterns and interaction among people, leading to vast amounts of user-generated data. This data, particularly from platforms like Twitter, has become a valuable resource for scholars and data scientists seeking new analysis and forecasting methods. In this context, the study examines the feasibility of employing Twitter activity to predict crime in metropolitan areas, focusing on London as a case study. The research aims to uncover hidden relationships and early warning signs of criminal activity by analyzing patterns of retweets and likes associated with specific keywords related to crime. The objectives include collecting and preprocessing Twitter data, analyzing Twitter behaviors and crime patterns, and developing a crime prediction model.

The study compared the performance of four models: K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), and Naive Bayes. Each model's accuracy, Kappa value, sensitivity, specificity, positive predictive value, and negative predictive value were evaluated to determine their effectiveness in predicting crime. The Random Forest model emerged as the most promising, achieving an accuracy of 80.31% and demonstrating a strong ability to differentiate between low and high crime occurrences.

The study acknowledges potential limitations related to biases in social media data and data quality issues. While the findings showcase the potential of utilizing Twitter data for crime prediction, the study also highlights areas for future research. These include addressing data biases, exploring temporal dynamics, enhancing interpretability of models, and considering ethical implications related to privacy and bias. Ultimately, the research contributes valuable insights into the use of machine learning and social media data for crime prediction, with the Random Forest model showing promise for practical application in crime prevention and response strategies.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF BOXES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Nowadays, the way people interact and communicate has changed because of the growth of social media. A popular social media platform that is a very good crowdsourcing platform and also useful for retrieving a large number of human viewpoints and actions. For scholars and data scientists, the abundance of user-generated data on Twitter offers a once-in-a-lifetime chance to experiment with new analysis and forecasting methods. One such promising field is crime forecasting, where using Twitter activity might offer insightful information about the dynamics of criminal activity in metropolitan settings (Johannes Bendler *et al.*, 2014).

Even London, the United Kingdom's cosmopolitan and varied city, faces issues from crime. Traditional crime prediction techniques, such as demographic profiling and historical crime data analysis, have several shortcomings when it comes to capturing the rapidly changing nature of criminal episodes (Shamsuddin, Ali and Alwee, 2017). We may access a rich amount of information that represents public mood, social interactions, and current events by taking Twitter behaviours like retweets and likes into consideration as possible indications.

The purpose of the study is to look into the viability and effectiveness of using Twitter activity to forecast crime in London. We can find hidden relationships and possibly spot early warning indications of criminal activity by examining patterns of retweets and likes linked to particular keywords or phrases related to criminal situations. In order to provide light on the possible influence of social media on urban safety and security, the research will concentrate on examining the link between Twitter data and crime rates.

The three main objectives of this study are as follows:

1. Collecting and preprocessing Twitter data: The collection and treatment of Twitter data will result in the elimination of noise, spam, and extraneous information from a thorough dataset of tweets about London. The generated dataset will act as the starting point for further investigation.
2. Analyzing Twitter activities and crime patterns: Investigating the connection between Twitter activity, such as retweets and likes, and reported crime trends will include the use of statistical and machine learning approaches. The investigation will concentrate on finding trends, patterns, and connections between criminal conduct on social media and other types of behaviour.
3. Developing a crime prediction model: A predictive model will be created to anticipate crime rates in particular London neighbourhoods based on the knowledge acquired from the research. To produce precise and timely predictions, the algorithm will use information from Twitter, historical crime statistics, and other contextual variables.

The findings of this study may have important ramifications for policing organizations, decision-makers, and urban planners. Authorities may proactively allocate resources, improve situational awareness, and create specialised crime prevention plans by using the potential of social media data. Additionally, creating safer and more secure communities

can benefit from a knowledge of how social media affects crime dynamics. With a particular focus on the special context of London, the dissertation will add to the body of knowledge expanding in the area of crime prediction using social media data.

CHAPTER 2: LITERATURE REVIEW

Crime prediction is a mandatory duty for law enforcement agencies to carry out which they have to allocate resources and to prevent crime from occurring (Perry *et al.*, no date). As social media has become more popular and active, researchers have indulged in the use of Twitter activity as a potential for crime prediction. With a focus on the methodology, results, and problems in this area, this literature review intends to critically assess the existing work on crime prediction using machine learning algorithms and Twitter data.

2.1 Theoretical Framework

Crime prediction and machine learning are fields that have relatively worked to identify patterns and trends in criminal activities, ensuring actions are being taken and preventive measures are put in place. Criminological theories and ideas serve as the theoretical underpinnings of crime prediction, while machine learning offers the tools and methods for analysing massive volumes of data and making predictions based on trends. Several criminological theories, like the rational choice theory, the social disorganisation theory, and the theory of routine activity, provide insight into the variables affecting criminal behaviour. According to the notion of routine activity, crimes happen when suitable targets, motivated offenders, and the absence of qualified watchmen come together in location and time. The social disorganisation hypothesis places particular emphasis on how neighbourhood attributes affect crime rates. According to the rational choice theory, people weigh the costs and benefits of engaging in unlawful behaviour before choosing to do so (Siegel Larry J., 2015). These criminological ideas are enhanced by machine learning, which examines massive datasets and finds hidden links and patterns. In order to learn from past instances and forecast future criminal events, algorithms are trained on historical crime data. Traditional statistical approaches might not be able to detect complicated interactions and non-linear correlations, but machine learning algorithms can.

Machine learning algorithms can extract useful data for predicting crime by analysing Twitter data. For instance, analysing geolocation information from tweets might assist pinpoint crime areas and movement patterns. Sentiment analysis may indicate how people feel about safety, and it may even be related to crime statistics. Additionally, the early detection of possible threats might be helped by the identification of certain phrases and hashtags linked to criminal activity.

There are various advantages of integrating machine learning algorithms with Twitter data. The first benefit is that it makes it possible to use a huge quantity of real-time data, giving crime prediction a more rapid and dynamic approach. Second, by gathering information on crimes that are not reported to law enforcement organisations, Twitter data can enhance conventional sources of crime data, such as official crime statistics. Third, the analytical process may be automated with the use of machine learning algorithms, enabling scalable and effective crime prediction.

In conclusion, while machine learning offers tools for data analysis and prediction, the theoretical underpinnings of crime prediction are based on criminological ideas. With its real-time information and sociological insights, Twitter data has the potential to be a useful tool for predicting crime. However, it is important to carefully weigh the advantages and difficulties of combining machine

learning algorithms with Twitter data, including issues with data quality, representativeness, privacy, and ethics.

2.2 Crime Prediction Using Machine Learning

A vast amount of research and studies have explored using machine learning for crime prediction. A self-exciting point process model was proposed by Mohler G et al. to predict crime trends based on the hypothesis that one crime leads to other crimes (Mohler *et al.*, 2011). A series of machine learning and deep learning techniques such as logistic regression, support vector machine (SVM), k-nearest neighbors (KNN), Naïve Bayes, random forest, decision tree, time series and some other techniques were carried out by Safat W et al. on an empirical analysis for crime prediction and forecasting in Chicago and Los Angeles (Safat, Asghar and Gillani, 2021). These studies show how machine learning techniques are effective at predicting crime. Some of the machine learning algorithms commonly applied in crime prediction are Decision Tree, Random Forest, Support Vector Machines (SVM), Neural Networks, Bayesian networks, and Ensemble Methods.

2.3 Twitter as a Data Source for Crime Prediction

Twitter data has been increasingly used and explored for crime prediction due to its real-time nature and abundance of information shared by users. A study to explore the potential of utilizing Twitter data in crime prediction was carried out by Matthew S. Gerber. The study was especially focused on the combination of Twitter data and a geographical analytic approach called kernel density estimation to identify regions where criminal behaviour was more likely to occur. The goal of the project was to develop a prediction model that may help identify places that are more likely to see criminal activity by examining geotagged tweets and using kernel density estimation. The study looks at the connection between Twitter usage and crime trends in an effort to improve crime prevention and prediction strategies by using social media data (Gerber, 2014).

2.4 Case Studies and Research Findings

Over the years, several studies have used machine learning and Twitter data to predict criminal activities. These studies have explored several algorithms, datasets, and assessment measures in an effort to get insightful knowledge on crime patterns. Here I will do a thorough analysis of some significant empirical studies. The first case study is by Vomfell L et al on improving crime count forecasts using Twitter and taxi data. The objective of this study was to improve the accuracy and reliability of crime count estimates in Chicago by integrating Twitter and taxi data into the forecasting process. The study's methodology involves data collection, data preprocessing, various constructed features from the data such as contextual features (e.g. weather, events, conditions) and temporal features (e.g. time of day, day of the week), machine learning algorithms, and some other statistical methods (such as rolling window prediction, count models). The researchers retrieved data from several sources such as crime reports from the Chicago Police Department, crime-related tweets in the Chicago area, and taxi GPS records. The machine learning algorithms used for this study were random forest (RF), gradient boosting machines (GBMs), and feed-forward artificial neural networks (ANNs)(Vomfell, Härdle and Lessmann, 2018).

The relationships and explanatory power of several characteristics for predicting violent crime and property crime were examined in this study. With the census data present in all settings, they used

eight possible feature combinations. Examining the regression coefficients, we concentrated on the important impacts that all models had found. In all non-exponential models for property crime, the vacancy rate had the biggest impact size and was highly correlated with property crime counts. New elements, such as the frequency of weekly taxi service and locations in the stores category, were also strongly linked to property crime. The number of residential locations and nightlife spots was linked to decreases in property crime. The number of property crimes was little yet significantly impacted by the Twitter function. Similar outcomes for the exponential models were seen. Social cohesiveness, as determined by the population's male ratio, had a significant impact on violent crime. Increases in violent crime counts were also correlated with rising female head of household rates and unoccupied dwelling rates. Ethnic diversity has a less evident impact on violent crime. Comparatively, the Twitter function had less of an impact on violent crime than on theft. Although the rise in violent crime caused by an extra food venue was very tiny, the category of food had the biggest effect size of all the new characteristics. Machine learning models were evaluated for variable relevance. Both the Twitter function and the cab feature were discovered to be crucial for both violent and property crime. The rankings of variable significance and the regression findings usually concurred. The accuracy of property crime prediction was increased, according to the predictive findings, when new data sources were used. The econometric models underperformed the machine learning models, with the Random Forest model with all features generating the least prediction error. The benefit of adopting innovative data sources was constrained for violent crime, though. The additional features had a minor positive impact on the econometric model's forecasts but had very little impact on the machine learning models. For the top-performing models, the results' robustness was examined. The fact that the property crime forecasts held up well against certain hyperparameter settings shows how good the new features are. The optimal hyperparameter combinations varied among windows, and the prediction errors for violent crime were more unpredictable(Vomfell, Härdle and Lessmann, 2018). Overall, the findings point to the potential for new variables, including taxi data and social media information, to increase the forecast accuracy for property crime but have limited potential for violent crime. Econometric models typically underperformed machine learning models, especially when it came to property crime.

The second case study is by Saraiva M et al on developing a crime prediction and monitoring system for the city of Porto, Portugal, by leveraging machine learning techniques, spatial analytics, and text analytics. The focus was on improving crime prevention and law enforcement strategies. In order to look at crime trends and forecast criminal incidents, the study used a multi-methodological approach that included geospatial analysis, machine learning modelling, and natural language processing (NLP). To analyse and depict localised crime trends, geospatial analysis employed ArcGIS software and kernel density estimation (KDE). Through conversation with law enforcement officers, the findings were confirmed. In order to find regions with persistent fluctuations in the incidence of crime, hot-spot analysis was also carried out. In order to investigate the impact of urban, morphological, and socioeconomic characteristics on crime, machine learning modelling approaches were used. Least absolute shrinkage and selection operator (LASSO) regression was used to find the variables that were the most important predictors. To forecast crime classes, four classification techniques—logistic regression, decision trees, random forests, and support vector machines—were used (Saraiva *et al.*, 2022).

In order to pinpoint the subset of variables that had the greatest impact on crime, the study used Lasso regression on the crime data from Porto. The prediction error was decreased, computing resources were optimised, and overfitting of the model was prevented by choosing fewer predictors with higher predictive power. By using an L1 penalty, Lasso regression enables the zeroing out of regression coefficients for uninteresting factors. For Lasso regression, the training and test sets were

split into 67% and 33%, respectively. According to positive regression coefficients, rising predictor factors increased the response variable (crime rate). Negative regression coefficients, on the other hand, indicated that a rise in the predictor variable was accompanied by a fall in the responder variable. In the study, other classification techniques were also used. Binary targets were identified using logistic regression with the L1 penalty for factors related to criminal activity. 70% of the data were used for training and 30% for testing the logistic regression model. Important factors including "Buildings with a wall structure in masonry with plate," "Buildings built before 1919," "Present population (male)," and "CCTV" exhibited negative coefficients, suggesting a decreased risk of crime occurring. On the other hand, factors like "Population with a low level of education" and "Classic family dwellings of usual residence with 1 or 2 rooms" exhibited positive coefficients, suggesting a higher risk of crime happening. SVM, decision trees, and random forests were among the other classification models that were created and fine-tuned using cross-fold validation (Saraiva *et al.*, 2022). The model with the highest accuracy of 0.832, recall of 0.99, precision of 0.79, and F1 score of 0.89 was random forest. Decision tree and random forest analysis highlighted "Classic Buildings," "Residents with the First Cycle of Basic Education," and "Present population (Male)" as significant factors.

Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression (L1 penalty = 0.151)	0.65	0.84	0.64	0.72
Decision Tree (criterion = entropy, max depth = 3)	0.61	0.56	0.70	0.63
Random Forest (max. features = 2, number of trees = 100, max depth = 5)	0.83	0.99	0.79	0.89
SVM (kernel = rbf, C = 1, gamma = 0.1)	0.80	0.87	0.82	0.91

Figure 1: Comparison of machine learning classification model performance *(Saraiva* et al.*, 2022).*

Overall, the findings are consistent with earlier studies, emphasising the role played by collective efficacy, gender, education, and density in crime dynamics. Big data interpretation and use should be done with caution since correlation does not indicate causation and the intricacy of micro-scale locations need a deeper comprehension than what is possible with standard algorithms and techniques.

CHAPTER 3: METHODOLOGY

The methodology used in this study to predict crime using machine learning on Twitter data and Crime data is presented in this chapter. This dissertation's multi-step technique will involve an overview of the research design, data gathering, preprocessing, analysis, and machine learning model creation and evaluation. The approach for performing the study on crime prediction in London using Twitter activity is outlined in the sections below:

3.1. Research Design

For this study, the research design adopted is a quantitative approach that integrates techniques of data science and machine learning. The project aims to forecast crime episodes using information from Twitter. Data collection, data preprocessing, feature extraction, model training, and model assessment are all processes in the study design that help to accomplish this.

3.2. Data Collection

The data collection process is done by accessing crime data and Twitter data with the use of Twitter API. The crime data is collected from the Metropolitan Police Service crime datastore which is updated monthly(*MPS Monthly Crime Dashboard Data - London Datastore*, no date). The data downloaded contained 33430 instances and was for the month of April and May 2023. The Twitter API allows programmatic access to the platform's large dataset of tweets related to crime incidents. To authenticate our request and gain access to the API endpoints, I created a developer account and acquired the required credentials, such as API keys and access tokens. I defined certain hashtags relevant to crime, law enforcement, and public safety to assure data relevance which are ["#LondonCrime", "#CrimeLondon", "#LondonSafety", "#SafeLondon", "#LondonViolence", "#CrimeWaveLondon", "#LondonPolice", "#LondonSecurity", "#LondonLawEnforcement", "#LondonCommunitySafety", "#LondonJustice", "#LondonCrimes", "#LondonSafetyAlert", "#LondonCrimeStats", "#LondonCrimePrevention", "#Crime + #London", "#truecrime + #London", "#detective + #London", "#knife + #London", "#knifecrime + #London", "#knifefree + #London", "#endknifecrime + #London", "#Arson + #London", "#Drug + #London", "#Burglary + #London", "#Assault + #London", "#homicide + #London", "#endknifecrime", "#putdowntheknives + #London", "#putdowntheknives", "MPSBarkDag", "MPSBarnet", "MPSBexley", "MPSBrent", "MPSBromley", "MPSCamden", "MPSCroydon", "ealingMPS", "MPSEnfield", "MPSGreenwich", "MPSHackneyCentr", "MPSHammFul", "MPSHaringey", "CrimeLdn"]. Over a one-month period, I gathered tweets from numerous cities and locales. I also took into account tweets with location-specific information or user-specified location information to narrow down to borough by borough and then area by area. To accomplish this, I made use of the 'rtweet' library in R and set a time period of a month. Kindly find the use of the library and link to the script in the Appendix. Finally, I also retrieve the population data from the London datastore (*Land Area and Population Density, Ward and Borough - London Datastore*, no date).

3.3. Data Preprocessing

The data contained 11 variables such as Month/Year, Category, Borough, Area name, Area code, Offence group, Offence subgroup, Measure, Financial year, FYIndex, and Crime count. To ensure quality and not the number of variables for the crime data, I remove irrelevant columns like financial year, area Code, category, measure, and FYIndex which had no significance to the research. To ensure the usability of the collected Twitter data, I performed a number of processes. First, I started off by removing duplicate tweets and retweets to reduce redundancy in the dataset. To accomplish this, I made use of Microsoft Excel data tools to take them out. Next, I filtered all tweets by crime and by boroughs to identify the total count of all the activities such as likes and retweets. This was

done using SQL queries. I leveraged on using the command “LIKE ‘%crime%’ AND LIKE ‘%location%’;”. Then, I performed data integration for the two datasets to result in one overall dataset. The integration was done by linking the areas of each borough and the crime category as primary keys. Next, was the integration of the population data retrieved and this was done using Microsoft Excel. I employed the use of a formula called ‘VLOOKUP’ using the area name as a primary key. Finally, to achieve my dependent variable, I calculated the crime count per thousand population by dividing the crime count by the population and then multiplying by a thousand. I then calculated the average and used it to create the crime level variable. For a crime count per thousand population that is less than the average, it is considered as low crime and if greater than average, it is considered high crime.

3.4. Data Description

The overall count of the instances was 33430 instances after preprocessing. The data initially had 11 variables but after preprocessing, it resulted in 12 variables due to the dropping off of irrelevant variables and the inclusion of Twitter data. Below are the final variables employed for this analysis.

Table 1: Dataset variables description.

Variables	Description	Type
Month_Year	This is of type character. Contains the date the data was created or formed.	Independent
Borough	This is of type character. This covers all 32 boroughs in London.	Independent
Area_name	This is of type character. This covers the areas under each London borough.	Independent
Offence_Group	This is of type character. This covers 12 groupings of crime	Independent
Offence_Subgroup	This is of type character. This covers subgroupings of the 12 groupings of crime.	Independent
Crime_Count	This is of type integer and contains linear or continuous data. This variable contains the total number of each crime for each area in each borough of London.	Independent
Tweets_Count	This is of type integer and contains linear or continuous data. Contains the total number of tweets retrieved on each respective crime and area.	Independent
Retweets_Count	This is of type integer and contains linear or continuous data. This is the total number of retweets on tweets on each respective crime and area.	Independent
Like_Count	This is of type integer and contains linear or continuous data. This is the total number of likes for tweets on each respective crime and area.	Independent

Population	This is of type integer and contains linear or continuous data. This is the total population of each area in each London borough.	Independent
Crime_per_TP	This is of type numeric and contains linear or continuous data. This is the standardized crime count per thousand population.	Independent
Crime_level	This is of type integer and contains categorical data. This is the crime level data divided into 0 (low crime) and 1 (high crime).	Dependent

3.5. Machine Learning Models

To accurately predict crime incidents, I used a number of machine learning algorithms. I tested many classifiers such as k-nearest neighbor (KNN), support vector machines (SVM), naïve bayes and random forest. These models were chosen because they performed well in similar prediction tasks in the literature and could handle both text and numerical information. The SVM is a powerful supervised learning algorithm that is effective in handling high-dimensional data and can handle both linear and non-linear classification problems using a kernel function. Random forest reduces overfitting and improves prediction accuracy by aggregating the results of individual trees. Random forests are particularly effective in handling high-dimensional datasets and capturing feature importance. Models were chosen after their performance was assessed using a variety of assessment criteria, such as accuracy, sensitivity, specificity, prediction values, and balanced accuracy In order to choose the best appropriate models for crime prediction based on Twitter activity data, I also took into account characteristics like computational efficiency and interpretability.

3.6. Model Training and Evaluation

With a fair distribution of criminal occurrences among the sets, I divided the dataset into training and testing sets in order to evaluate the performance of the machine learning models. Using the oversampled training data, I trained the machine learning models and tuned their hyperparameters to improve performance. The testing set was used as a separate dataset to assess the final model performance after the training set had been used for hyperparameter tuning and model selection. Based on the size and properties of the dataset, the splitting ratio such as 70% for training, and 30% for testing was chosen. The machine learning models had to be fitted to the training set during the training phase. In order to learn the model parameters and optimise the models based on the chosen loss or objective function, I employed the training set. After the models had been trained, I assessed the models' performance on the testing set. Using a variety of assessment criteria, such as accuracy, sensitivity, specificity, KAPPA, and both prediction values, I evaluated the models' prediction accuracy. These metrics revealed information on the algorithms' accuracy in identifying criminal incidences and non-incidents. I also evaluated the model's performance in terms of true positives, true negatives, false positives, and false negatives by examining the confusion matrix. Additionally, I used cross-validation methods to assess the generalizability of the models and performed statistical significance tests to verify the findings. I used statistical significance tests to confirm the data's significance. This required employing suitable statistical tests, such as t-tests or analysis of variance (ANOVA), to compare the performance of various models or variants. These tests enabled us to identify whether observed variations in performance were purely coincidental or

statistically significant. I wanted to choose the best-performing models for crime prediction based on Twitter activity data and deliver accurate and trustworthy predictions by using this technique for model training and assessment.

CHAPTER 4: RESULT OF ANALYSIS

In this chapter, I present the findings of our study on forecasting crime levels using machine learning algorithms based on Twitter activity. The main goal of this study was to see whether it was possible to forecast crime levels in London using publicly available Twitter data. Over the course of one month, I gathered Twitter data from a sample of London users for our study. I focused on phrases associated with criminal activity while retrieving tweets that were inside the city boundaries. I combined historical criminal records from the Metropolitan Police Service dashboard with Twitter data after cleaning the data to get rid of duplicates and irrelevant tweets.

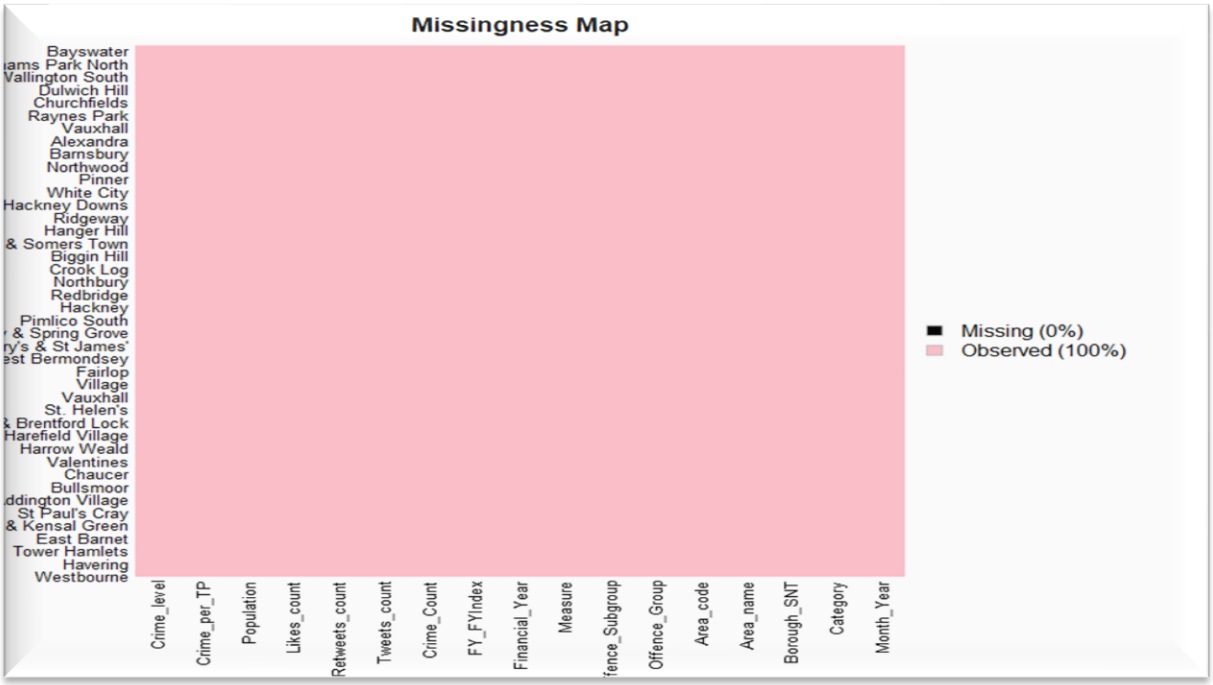

Figure 2: Missingness Map on the data

After carrying the method to identify if there are any missing data, it was observed that there were no missing data as shown in Figure 2. This ensures there is no loss of information, biases in results, and no effect on model performance.

4.1. Descriptive Statistics

I began the analysis with descriptive statistics to gain a comprehensive understanding of the data.

Table 2: Descriptive statistics on data.

	Crime Count	Tweets Count	Retweets Count	Likes Count	Population
Mean	9.75	0.02	0.08	0.10	42838
Median	3.00	0.00	0.00	0.00	13843

Standard Dev	39.13	0.39	2.11	2.06	85805.49
Minimum	1.00	0.00	0.00	0.00	3955
Maximum	2072.00	27.00	107.00	71.00	390719
1st Quantile	1.00	0.00	0.00	0.00	11028
3rd Quantile	7.00	0.00	0.00	0.00	16907

These descriptive statistics give a quick overview of the dataset and shed light on each variable's central tendency, variability, and distribution. The mean crime count is 9.75, which is an average of about 10 recorded offences each month. The distribution is right-skewed, as indicated by the median value of 3.00, with certain occurrences having extraordinarily high crime rates. The variety in crime incidences is highlighted by the standard deviation of 39.13. The average number of tweets, retweets, and likes on Twitter is 0.02 per tweet, 0.08 per retweet, and 0.10 per like, respectively. The fact that the median and first quantile are zero, showing that most cases have little to no social media interaction, supports these low results. For these Twitter indicators, the standard deviations are quite minimal, indicating that the data points are closely packed around the mean. In terms of population, the median is 13,843, while the mean is 42,838. The high standard deviation of 85,805.49 indicates that population numbers vary significantly. The variety of population counts is further illustrated by the low and high thresholds of 3,955 and 390,719 respectively. At 16,907, the third quantile shows that 75% of population counts are below this level. To further explore the data, visualisations such as box plot and a histogram is used.

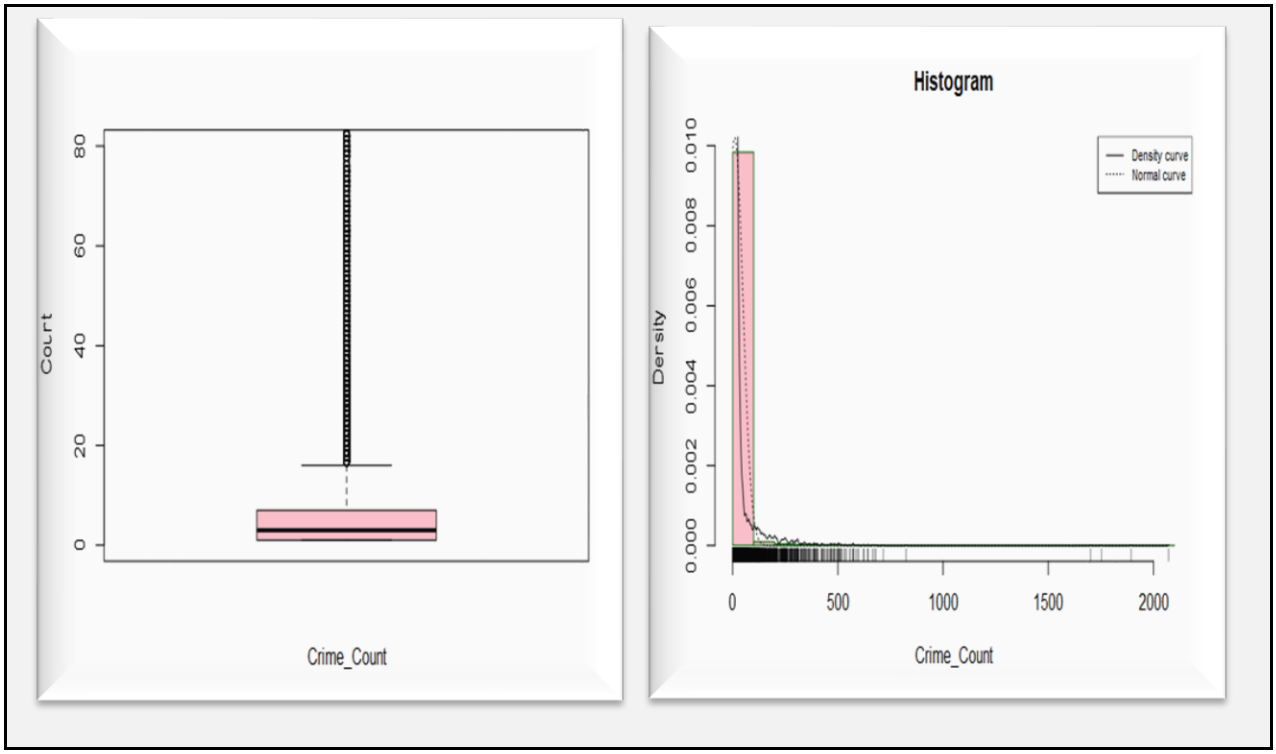

Figure 3: A box plot and histogram of the Crime Count variable.

From the box plot, the median's proximity to the box's bottom shows that the data are positively skewed. This indicates that the bulk of crime counts are concentrated at lower values, with a small number of extremely high crime counts (outliers) pushing the median upward. There is data anomaly or extreme values indicated by a large number of outliers and the outliers are extreme outliers which suggest that there are certain cases with disproportionately high crime rates that differ greatly from the majority of the data. These excessive figures may represent singular

occurrences or anomalies that are far greater than the normal distribution of crime counts. The majority of the observed places or time periods appear to have relatively low crime rates based on this trend, however, there are some cases where the crime rates are exceptionally high. These high crime rates may be related to particular incidents, places, or times where there was a very high level of criminal activity. Just as suggested from the box plot, The right-skewed histogram presents more proof in support of the first finding. Visual evidence supporting the claim that the data is actually skewed to the right comes from the histogram's right-skewed form, which has a long tail stretching towards higher crime counts. The form of the histogram, which is skew and has an extended tail on the right side, further shows that the crime count data does not have a normal distribution. . A smaller number of higher crime counts are responsible for the longer tail on the right, whereas the majority of crime counts are grouped together on the left side of the histogram.

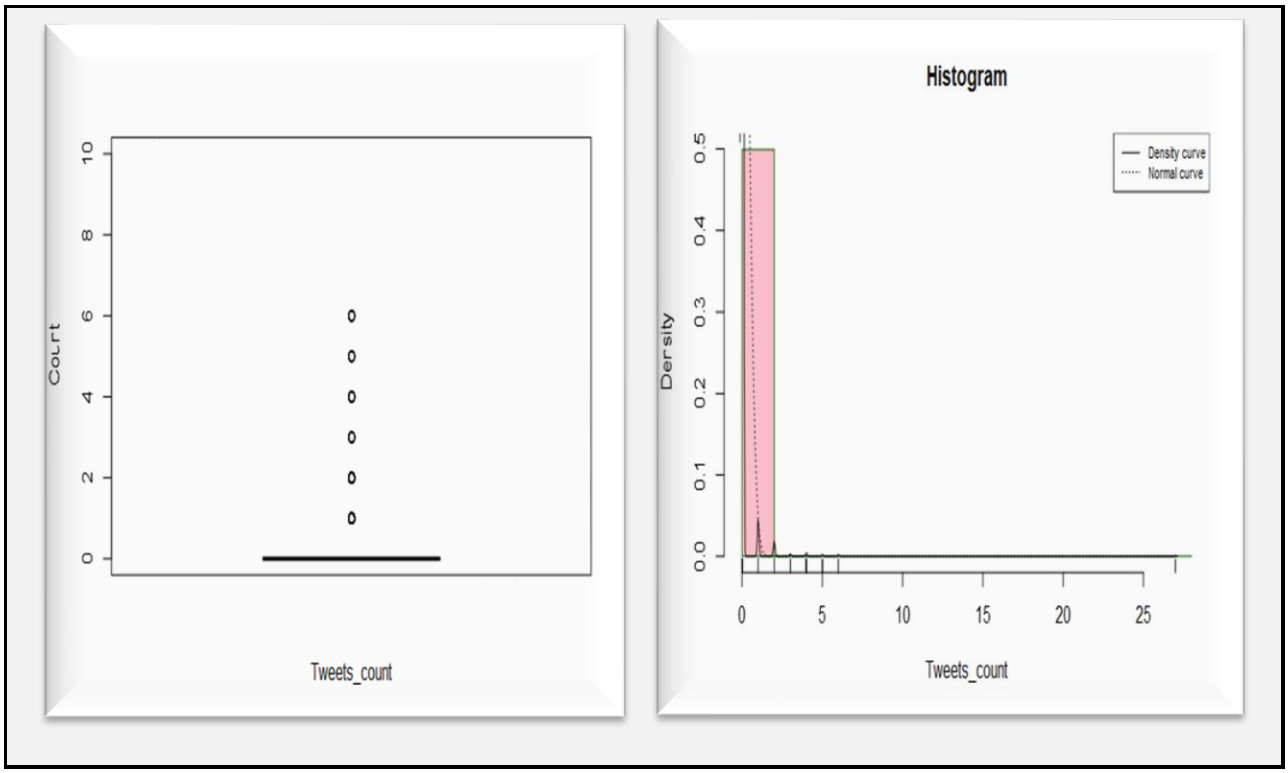

Figure 4: A box plot and histogram of the Tweets Count variable.

The majority of the data in the box plot falls from 0 to 0.0001, indicating that the great majority of tweets have very low counts. The fact that so many observations had identical, low tweet counts inside a condensed area of data suggests this. A right-skewed distribution is shown by the existence of severe outliers that extend well beyond the box and whiskers into significantly higher values (up to 6). These outliers shape the long tail of the distribution, which pushes the median and whiskers towards the distribution's lower end. This histogram also confirms the claims as shown in the figure above.

4.2. Hypothesis Testing

The hypothesis of this work is that there is a relationship between Twitter activity and the Crime rate in London.

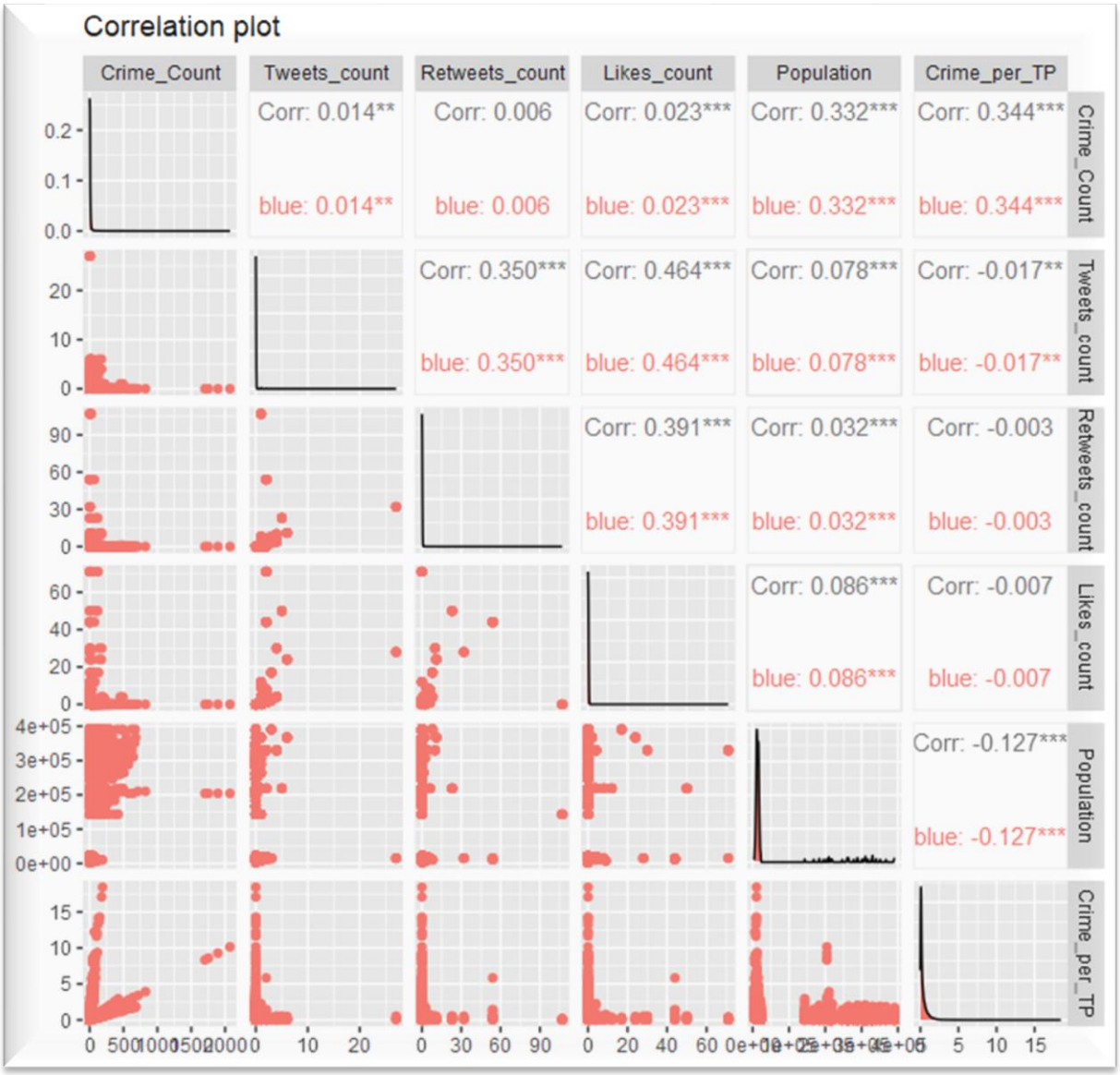

Figure 5: Correlation plot of all variables.

The graphic shows the pairwise correlation between the various variables in the dataset. The correlation coefficient between any two variables is shown in each matrix column and ranges from -1 to 1. A direct relationship is denoted by a positive correlation, whereas an inverse relationship is shown by a negative correlation.

I deduced from the correlation plot that there is a modest positive connection between the population and the Twitter activity metrics (Tweets_count, Retweets_count, and Likes_count). This correlation ranges from 0.01 to 0.46. Interestingly, the population and the crime-related variables (Crime_per_TP and Crime_Count) have a comparatively larger positive correlation (0.33 to 0.48). However, the relationship between crime-related characteristics and Twitter activity indicators is still only moderately strong (range from -0.02 to 0.18).

Additionally, there is a weak negative correlation between the population and the crime-related variables (Crime_level, Crime_per_TP, and Crime_Count), implying that places with larger populations have lower crime levels, crime per thousand people, and crime counts. A somewhat

negative association between the Crime_level variable and Twitter activity measures also exists, suggesting that lower crime rates may be correlated with increased Twitter participation.

4.3. K-Nearest Neighbour Modelling

Before creating the KNN model, all independent variables had to be scaled and this was carried out with the scale function in R. The function is generic for scaling and centring matrix-like objects. The reason for scaling is to avoid unequal weighting of variables, to have consistent distant measures, and to avoid uneven representation of data points in the space. By scaling the data, all variables are converted to a similar range, typically {0, 1} or {-1, 1}. This allows all variables to contribute equally to the distance computations. As a result, the KNN model improves in accuracy and becomes less sensitive to changes in variable scale.

Box 1: Evaluation of values of k.

	k	Accuracy	Kappa	AccuracySD	KappaSD
1	1	0.8006072	0.6012178	0.007803702	0.01560812
2	3	0.7935008	0.5869979	0.008926722	0.01785299
3	5	0.7951753	0.5903458	0.007529922	0.01505913
4	7	0.7965756	0.5931455	0.008596873	0.01719371
5	9	0.7974303	0.5948540	0.008530889	0.01706027
6	11	0.7974989	0.5949918	0.008331464	0.01666150
7	13	0.8005055	0.6010050	0.006746448	0.01349229
8	15	0.7996173	0.5992291	0.005225437	0.01045036
9	17	0.8008815	0.6017578	0.007109511	0.01421814
10	19	0.8003350	0.6006653	0.007324690	0.01464848

The presented KNN results show how the K-Nearest Neighbours algorithm performs on a classification problem with various values of k (the number of nearest neighbours evaluated). Accuracy and Kappa (Cohen's Kappa) scores, together with the corresponding standard deviations (SD), are the assessment metrics employed.

As the results were examined, I observed that the accuracy and Kappa scores generally increase as the value of k increases. With a Kappa value of 0.6012 for k = 1, the accuracy attained is roughly 80.06%. Although, the model might be sensitive to noise and outliers as it solely relies on the nearest neighbor. The accuracy steadily decreases as k gets bigger to 3, 5, and 7 and reaches about 79.35%, 79.52%, and 79.66%, respectively, with corresponding Kappa ratings of 0.587, 0.590, and 0.593. This suggests that taking into account more nearest neighbours improves predictions and raises concordance between predictions made by the model and the actual classes. Further examining the data, I noticed minor variations in accuracy and Kappa values as k increased from 9 to 19. The accuracy was fixed at around 79.74% for k = 9 and 11, with Kappa values at 0.595. Higher k values also lead to lower standard deviations for these measures, indicating improved model stability. The smaller standard deviations suggest that the model's forecasts become more consistent when

applied to several runs or datasets. Overall, the results suggest that a moderate k value, around 13 to 15, yields the best performance for this specific KNN model.

Using the best value of k, model testing with the test data is then carried out to evaluate the performance of the model. To examine the performance result, a confusion matrix containing the results is implemented.

Box 2: Confusion Matrix and Statistics of KNN Model after prediction.

```
Confusion Matrix and Statistics


          Reference
Prediction    0    1
         0 4889 1152
         1 1412 5186


               Accuracy : 0.7971
                 95% CI : (0.79, 0.8041)
    No Information Rate : 0.5015
    P-Value [Acc > NIR] : < 2.2e-16


                  Kappa : 0.5942


 Mcnemar's Test P-Value : 3.138e-07


            Sensitivity : 0.7759
            Specificity : 0.8182
         Pos Pred Value : 0.8093
         Neg Pred Value : 0.7860
             Prevalence : 0.4985
         Detection Rate : 0.3868
   Detection Prevalence : 0.4780
      Balanced Accuracy : 0.7971


       'Positive' Class : 0
```

The outcomes of a binary classification model are shown in the KNN confusion matrix, where the predicted classes are contrasted with the true or reference classes. The number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions is shown in the confusion matrix. The reference and anticipated classes are denoted by the labels 0 and 1, respectively.

The KNN model's accuracy is computed as 0.7971 (79.71%), indicating the proportion of correctly classified instances out of all instances. 4889 instances of class 0 (low crime) are accurately classified by the model as true positive, as evidenced by the sensitivity (true positive rate) of 0.7759 (77.59%). As for the specificity (true negative rate), it had 0.8182 (81.82%), which indicates a very good capacity to accurately detect occurrences of class 1 (high crime).

The Kappa statistic indicates the degree of agreement between the model's predictions and the actual classes, accounting for any possible coincidences in agreement. The Kappa value in this instance is 0.5942, which suggests that there is a moderate level of agreement, meaning the model is performing significantly better than random chance. A Kappa score close to 0 indicates that the model's predictions are not considerably more accurate than chance. There is a statistically significant difference between the model's error rates for class 0 and class 1, as shown by the Mcnemar's Test P-Value of less than 3.138e-07. This underlines how crucial it is to solve the problem of class inequality.

The majority of incidents of class 0 (low crime) are accurately identified by the model, according to the sensitivity, or true positive rate, of 0.7759 (77.59%). Also, the specificity (true negative rate) for the model is 0.8182 (81.82%), which indicates that it has a very strong capacity to identify incidents of class 1 (high crime). The percentage of occurrences classified as class 0 that were properly predicted, or the positive predictive value (Pos Pred Value), was 0.8093 (81.93%). The model's class 1 (high crime) predictions are reasonably good, according to the negative predictive value (Neg Pred Value), which is 0.7860 (78.60%). The reported balanced accuracy is 0.7971 (79.71%). The average of sensitivity and specificity is used to compute balanced accuracy, which accounts for class imbalance. A balanced accuracy of about 79% indicates that the model is successfully differentiating between the two groups and is not considerably more accurate than random guessing.

In conclusion, the statistics and confusion matrix for the KNN given here provide key insights into how well the model predicts crime levels. The high specificity illustrates the model's strength in properly identifying cases of high crime and the high sensitivity shows that the algorithm successfully recognises instances of low crime.

4.4. Random Forest Modelling

The outcome of the Random Forest model sheds important light on a classification model created to forecast the degree of crime using a variety of input data such as Crime count, number of tweets, likes and retweets, population and crime count per thousand population. The model was built using the “randomForest” function, and as it is a classification type, it serves the role of categorising instances into various classes.

Box 3: Result of the Random Forest Model with training data.

```
Call:
 randomForest(formula = Crime_level ~ ., data = train_data3)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2
```

```
        OOB estimate of  error rate: 19.23%
Confusion matrix:
      0     1 class.error
0 11560  3066   0.2096267
1  2562 12081   0.1749641
```

The random forest ensemble consists of 500 decision trees, which were constructed from bootstrapped samples of the training data. Only two randomly chosen variables from the dataset were taken into consideration during the creation of each decision tree in the forest at each split. This characteristic aids in adding variety and unpredictability to the ensemble, improving its capacity for prediction and lowering overfitting. The Out-Of-Bag (OOB) estimated error rate utilising the data not included in the training bootstrap samples resulted to be 19.23%. When the model is tested on hypothetical data, the error rate reflects the percentage of cases that are incorrectly categorised.

The confusion matrix also includes the class error rates. The percentage of cases of class 0 that were incorrectly categorised out of all instances of class 0 is determined as 0.2096267 (20.96%). Likely indicating the percentage of cases of class 1 that was incorrectly categorised out of all occurrences of class 1, the class error for class 1 is 0.1749641 (17.49%). Overall, the Random Forest model achieves an OOB error rate of 19.23%, indicating that it is capable of decent generalisation. The larger class error for class 0 than for class 1 indicates that there is still potential for improvement, particularly in accurately identifying instances of class 0.

Box 4: Confusion Matrix and Statistics of Random Forest Model after prediction.

```
Confusion Matrix and Statistics


          Reference
Prediction    0    1
         0 4911 1099
         1 1390 5239


               Accuracy : 0.8031
                 95% CI : (0.796, 0.81)
    No Information Rate : 0.5015
    P-Value [Acc > NIR] : < 2.2e-16


                  Kappa : 0.6061


 Mcnemar's Test P-Value : 6.144e-09
```

```
            Sensitivity : 0.7794
            Specificity : 0.8266
         Pos Pred Value : 0.8171
         Neg Pred Value : 0.7903
             Prevalence : 0.4985
         Detection Rate : 0.3886
   Detection Prevalence : 0.4755
      Balanced Accuracy : 0.8030

       'Positive' Class : 0
```

Class 0 (low crime) had 4,911 true positives, according to the confusion matrix, showing that the model accurately identified occurrences of low crime as class 0. The algorithm properly classified cases of high crime as class 1 since there are 5,239 true positives for class 1 (high crime). However, there are 1,099 false positives (instances predicted as low crime but were actually high crime) and 1390 false negatives ((instances predicted as high crime but were actually low crime). The percentage of occurrences that were successfully categorised out of all instances is given as 0.8031 (80.31%). The accuracy's 95% confidence interval (CI) is (0.796, 0.81), and it indicates the range in which the accuracy's actual value is most likely to fall.

The accuracy attained by consistently foreseeing the majority class is known as the No Information Rate (NIR). The NIR in this instance is 0.5015 (50.15%), which is much less accurate than the model's accuracy. The model performs statistically substantially better than a straightforward majority-class predictor, as shown by the accuracy versus NIR p-value (2.2e-16), which shows that the difference between the model's performance and the NIR is statistically significant. The agreement between the model's predictions and the actual classes is measured using the Kappa statistic, which takes into account the possibility that the agreement might also happen by coincidence. The Kappa value in this instance is 0.6061, which is much greater than zero. A Kappa value that is closer to 1 suggests that there is stronger agreement than would be predicted by chance, indicating that the model is working effectively. According to the McNemar's Test P-Value of 6.144e-09, there is a statistically significant difference between the model's error rates for classes 0 and 1. This implies that the model's performance is not the same for the two classes, which may call for more research and optimisation.

The model's ability to accurately detect incidents of class 0 (low crime) is demonstrated by the sensitivity (true positive rate), which is given as 0.7794 (77.94%). The model's capacity to accurately detect incidents of class 1 (high criminality) is indicated by the specificity, or true negative rate, of 0.8266 (82.66%). The average of sensitivity and specificity, or the balanced accuracy, is 0.8030 (80.30%). A model that performs well and successfully distinguishes between the two groups has a balanced accuracy close to 1. The percentage of occurrences classified as class 0 that were properly predicted, or the positive predictive value (Pos Pred Value), was 0.8171 (81.71%). The percentage of class 1 cases accurately anticipated among all instances correctly forecasted as class 1 is represented by the negative predictive value (Neg Pred Value), which is 0.7903 (79.03%).

Finally, the model successfully distinguishes between occurrences with low and high crime levels with high levels of accuracy, sensitivity, and specificity. The model's overall dependability is highlighted by the Kappa value, which denotes a considerable agreement above and beyond chance. To enhance the model's performance on both classes, additional research may be necessary given the disparity between the error rates for class 0 and class 1. Overall, the model shows potential in estimating crime rates based on Twitter activity, providing insightful information for tactics for crime prevention and response.

4.5. Support Vector Machine Modelling

The output of the support vector machine (SVM) model used to forecast crime levels using input variables from the dataset is shown in the box below. Its important parameters and performance metrics are displayed. With the cost parameter set to 1, the radial kernel was used to train the model.

Box 5: Result of the SVM Model with training data.

```
Call:

svm(formula = Crime_level ~ ., data = train_data3, kernel = "radial",
cost = 1)

Parameters:

   SVM-Type:  C-classification

 SVM-Kernel:  radial

       cost:  1


Number of Support Vectors:  13996

 ( 6979 7017 )


Number of Classes:  2

Levels:

 0 1
```

During training, the SVM model found 13,996 support vectors. The training set's data points known as support vectors are essential for establishing the decision boundary or hyperplane that divides the two classes. The SVM model was created for binary classification, identifying the dataset's two classes as (0 and 1). 6,979 support vectors belong to class 0 (low crime) and 7,017 in class 1 (high crime) out of a total of 13,996 support vectors. The support vector's even distribution implies that both classes were fairly represented throughout training.

Overall, utilising the radial kernel and cost parameter of 1, the SVM model offers a solid basis for crime level prediction. An appropriate classification boundary is the consequence of the model having learnt complicated correlations between the input characteristics and the target variable, which is implied by the existence of a large number of support vectors. To guarantee the model's

generalizability and usefulness in real-world settings, I further assess the model's performance using other metrics including accuracy, sensitivity, specificity and other results on a separate test dataset.

Box 6: Confusion Matrix and Statistics of SVM Model after prediction.

```
Confusion Matrix and Statistics


          Reference
Prediction    0    1
         0 5108 1377
         1 1193 4961


               Accuracy : 0.7967
                 95% CI : (0.7895, 0.8036)
    No Information Rate : 0.5015
    P-Value [Acc > NIR] : < 2.2e-16


                  Kappa : 0.5934


 Mcnemar's Test P-Value : 0.0003064


            Sensitivity : 0.8107
            Specificity : 0.7827
         Pos Pred Value : 0.7877
         Neg Pred Value : 0.8061
             Prevalence : 0.4985
         Detection Rate : 0.4041
   Detection Prevalence : 0.5131
      Balanced Accuracy : 0.7967


       'Positive' Class : 0
```

The SVM confusion matrix and statistics presented enable a thorough assessment of the model's performance. It is observed that the model classified 5108 instances of class 0 (low crime) as true positives and classified 1377 instances of low crime as false positives. Then as of class 1 which represents high crime, 4961 instances are classified as true negatives and 1193 instances of high crime as false negatives.

The accuracy of the model is given as 0.7967 (79.67%) which represents the percentage of cases that were successfully classified out of all instances. The accuracy achieved by consistently predicting the majority class (low crime) is shown by the No Information Rate (NIR), which is 0.5015 (50.15%). The accuracy against NIR p-value (2.2e-16) indicates that the model performs considerably better than predicting the majority class. The Kappa score in this instance is 0.5934, which denotes moderate agreement. With a sensitivity of 0.8107 (81.07%), the model's ability to accurately detect instances of low crime (class 0) is high. The specificity for the model on the other hand is 0.7827 (78.27%), indicating that it has a slightly lesser capacity to accurately detect instances of high crime (class 1). The Positive Predictive Value, which is the percentage of class 0 cases properly predicted among all instances classified as such, is 0.7877 (78.77%). The percentage of class 1 cases accurately anticipated among all instances correctly forecasted as class 1 is represented by the negative predictive value, which is 0.8061 (80.61%). The average of sensitivity and specificity is 0.7967 (79.67%). A model that performs well and successfully distinguishes between the two groups has a balanced accuracy close to 1.

The SVM confusion matrix and data shown show that the model does a fair job of predicting crime levels. It has great sensitivity and accuracy for detecting low crime levels (class 0). However, it seems to have difficulty recognising high crime levels (class 1) events with sufficient specificity.

4.6. Naïve Bayes Modelling

Naive Bayes modelling is a straightforward yet efficient classification approach that uses the Bayes theorem and the premise of feature independence. Below is the result of the model built.

Box 7: Result of the Naïve Bayes Model with training data.

```
Naive Bayes

29269 samples
    6 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 26342, 26342, 26342, 26342, 26342, 26342, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.7308759  0.4618721
   TRUE      0.7980794  0.5961444

Tuning parameter 'laplace' was held constant at a value of 0
```

```
Tuning parameter 'adjust' was held constant at a value of 1

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were laplace = 0, usekernel = TRUE and
adjust = 1.
```

Cross-validation with 10 folds was used to evaluate the model. The dataset was partitioned into about equal-sized subsets for examination, according to the summary of sample sizes for each fold. The resampling results for various tuning settings are shown, with the tuning parameter 'usekernel' having two possible values: FALSE and TRUE. The kernel density estimator in the model is probably controlled by the 'usekernel' option.

The model attained an accuracy of about 73.08% and a Kappa value of about 0.46 with the 'usekernel' option set to FALSE. While the model performed better when the 'usekernel' option was set to TRUE, producing results with an accuracy of about 79.80% and a Kappa value of roughly 0.59. This suggests that adopting a kernel density estimator improved predictions and increased agreement above and beyond chance. 'Laplace' and 'Adjust', two tuning parameters, were maintained throughout the analyses. The 'laplace' parameter is utilised in Laplace smoothing, a method for avoiding zero probability for characteristics that are not visible. Its value was set to 0 in this instance, indicating that no Laplace smoothing was utilised. The class priors were not updated during model training, as shown by the value of the 'adjust' parameter, which was set to 1. The final settings for the model were laplace being 0, usekernel set to TRUE, and adjust as 1, which generated the maximum accuracy among the examined choices. Accuracy was the selection criteria for the model.

Box 8: Confusion Matrix and Statistics of Naïve Bayes Model after prediction.

```
Confusion Matrix and Statistics


          Reference
Prediction    0    1
         0 4794 1075
         1 1507 5263


               Accuracy : 0.7957
                 95% CI : (0.7886, 0.8027)
    No Information Rate : 0.5015
    P-Value [Acc > NIR] : < 2.2e-16


                  Kappa : 0.5913


 Mcnemar's Test P-Value : < 2.2e-16
```

```
            Sensitivity : 0.7608
            Specificity : 0.8304
         Pos Pred Value : 0.8168
         Neg Pred Value : 0.7774
             Prevalence : 0.4985
         Detection Rate : 0.3793
   Detection Prevalence : 0.4644
      Balanced Accuracy : 0.7956

       'Positive' Class : 0
```

For instances of low crime (class 0), the model was able to accurately predict 4794 instances as true positives and 1075 instances were incorrectly classified as low crime (false positives) but should have actually been high crime (class 1). The model predicted 5263 instances of high crime (class 1) correctly as true negatives and predicted 1507 instances that should have been classed as low crime (class 0) by the model instead it fell into the high crime category.

The model's accuracy in this case is 79.57% (0.7957). The percentage of cases that were successfully categorised out of all instances is represented by this number. The No Information Rate (NIR), which measures accuracy by consistently predicting the majority class (low crime), is 50.15% (0.5015). The accuracy against the NIR p-value (2.2e-16) indicates that the model performs considerably better than predicting the majority class. Also disclosed is the Kappa statistic, which has a value of 0.5913. A Kappa score greater than 0.5 denotes moderate to significant agreement.

The reported value for sensitivity, also known as the true positive rate, is 76.08% (0.7608). It shows how well the model can recognise occurrences of class 0 (low crime). The model's specificity, also known as true negative rate, is stated as 83.04%, indicating that its accuracy in detecting occurrences of class 1 (high crime) is likewise rather high. Prevalence in the dataset (proportion of the positive class) is 49.85% (0.4985), which represents the percentage of incidents in class 1 (high crime). The percentage of occurrences expected to be class 0 out of all instances, or the detection prevalence, is 46.44% (0.4644). The average of sensitivity and specificity, or the balanced accuracy, is 79.56% (0.7956). A model that performs well and successfully distinguishes between the two groups has a balanced accuracy close to 1. The McNemar's Test P-Value, which measures the difference in error rates between class 0 and class 1 predictions, is very low (2.2e-16).

Overall, the accuracy of the Naive Bayes model is decent, and its balanced accuracy indicates a strong model. Scores for sensitivity and specificity show that the model successfully classifies occurrences of both types. Given the strong Positive Predictive Value, it is quite probable that the model will be accurate when classifying a given case as class 0. However, the significantly smaller Negative Predictive Value indicates that there may be a greater risk of error when the model forecasts an instance as class 1.

4.7. Model Performance Evaluation

Comparing the results of the models, the best model can be identified based on their performance metrics and characteristics.

Table 3: All model's performance.

	KNN Model	RF Model	SVM Model	Naïve Bayes Model
Accuracy	79.71%	80.31%	79.67%	79.57%
Sensitivity	77.59%	77.94%	81.07%	76.08%
Specificity	81.82%	82.66%	78.27%	83.04%
Pos Pred Value	80.93%	81.71%	78.77%	81.68%
Neg Pred Value	78.60%	79.03%	80.61%	77.74%
Balance Accuracy	79.71%	80.30%	79.67%	79.56%

The accuracy of the KNN model ranged from 80.06% (k=1) to 79.75% (k=11). After predicting with the scaled test data, the accuracy resulted in 79.71% which is moderately good. The accuracy is 80.06% and the Kappa value is 60.12% for the best result when k=1. As k rises, the accuracy and Kappa values somewhat decline but generally stay rather high. The KNN model performs well in differentiating between low- and high-crime occurrences, with a balanced accuracy of 79.71%. But the model might be sensitive to noise and outliers as it solely relies on the nearest neighbor. Due to this, the KNN model would not be the ideal option for this particular situation. The Random Forest model fared ok, scoring a Kappa of 0.6061 and an accuracy of 80.31%. Because of its greater balanced accuracy of 80.30%, it is better able to differentiate between the classes. There is definitely room for improvement, particularly in terms of cutting down on false positives and false negatives. The SVM model had an accuracy of 79.67% and a Kappa score of 0.5934. The sensitivity for categorising low crime levels (class 0) was good at 81.07%. On the other hand, the specificity of 78.27% is slightly less (ability to detect high crime incidents, class 1). 79.67% is the balanced accuracy. The SVM model has promise, but more work has to be done on it to increase specificity and performance as a whole. The accuracy and Kappa score of the Naive Bayes model with kernel density estimation were 79.57% and 0.5913, respectively. It demonstrated well-balanced specificity of 83.04% and sensitivity of 76.08%, which resulted in a better-balanced accuracy of 79.56%. The performance of the model is encouraging, but there is still a tiny class imbalance as seen by the negative predictive value being less than the positive predictive value.

Considering the given metrics, the Random Forest model stands out among the ones that are offered as the best model. It scored the greatest balanced accuracy, demonstrating a more accurate capacity to differentiate between the two groups. The Kappa value also shows a moderate to large agreement between the model's predictions and the actual classes. It also demonstrated a respectable sensitivity and specificity for both classes. Although there is certainly space for development, it offers a strong foundation for estimating crime levels based on Twitter activity.

CHAPTER 5: DISCUSSION

K-Nearest Neighbours (KNN), Random Forest, Support Vector Machine (SVM), and Naive Bayes were four machine learning models that were compared in the study. A number of performance indicators, including accuracy, Kappa, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), were used to evaluate each model. The Random Forest algorithm outperformed the other four models in terms of predicting crime levels. It outperformed the other

models with an accuracy of 80.31% demonstrating its capability to distinguish between cases of low and high criminality. The results also align with previous studies as discussed in the literature review.

In comparison, the accuracy ranged from 79.57% to 79.71% for the KNN, SVM, and Naive Bayes models, which similarly displayed reasonable performances. They were somewhat more accurate and balanced than the Random Forest model, but not quite as accurate. Although simple to grasp and intuitive, the KNN model may have been constrained by the choice of the number of neighbours (k), and larger k values may have produced more accurate findings. The SVM model with a radial kernel demonstrated competitive performance, but more hyperparameter tweaking may be necessary for it to operate to its fullest ability. Despite its simplicity, the Naive Bayes model performed pretty well, but its performance may be improved by looking at other text processing strategies.

The use of Twitter data and the inherent biases in social media data were two of the study's many drawbacks. Sample bias might affect the data gathered from Twitter. The demographics and socioeconomic backgrounds of Twitter users may not be entirely representative of the whole population, and certain groups may be over or underrepresented. This bias may restrict the application of the established models to larger populations and alter the generalizability of the findings. Twitter data is highly susceptible to noise, including spam, irrelevant content, sarcasm, and false information. Even after preprocessing, there could still be some noise in the dataset, which could affect the precision and dependability of the prediction models. Future research should focus on enhancing data quality by creating more advanced noise detection and filtering methods. Twitter data is also multilingual and includes slang, cultural allusions, and subtleties unique to each language. The models created in this study might not fully capture these subtleties, which could have an impact on how well they predict outcomes. To address these issues, future research can investigate sentiment analysis, cross-lingual transfer learning, or language-specific modelling approaches. Criminal behaviour on Twitter and crime patterns may both display temporal dynamics, such as seasonality, periodicity, or trends over time. The temporal features of crime prediction using Twitter activity were not fully examined in this study. Future research might use temporal characteristics and time-series analysis methods to capture and take advantage of temporal dynamics for more precise predictions. It may also be difficult to comprehend the fundamental causes influencing crime forecasts due to the interpretability and explainability issues with the machine learning models used in this work. Future research should focus on making the models easier to understand and provide clear justifications for their forecasts. Techniques like rule extraction, attention processes, and model-agnostic interpretability methodologies can be investigated. The use of social media data to forecast crime concerns issues of privacy, permission, and possible biases. The study and application of the produced models must be carefully monitored to make sure that they don't violate people's privacy rights or worsen pre-existing prejudices in law enforcement procedures. Future research should carefully analyse these moral issues and provide the necessary protections and regulations.

In summary, I've described the process used to predict crime using machine learning methods using data from Twitter activity. I went over the steps of data collecting, data preparation strategies, feature extraction approaches, machine learning models, and model building and testing. I also recognised the study's limits and noted possible areas for advancement and investigation in the future. This study offers important information on the potential of machine learning on Twitter activity data for crime prediction. This chapter's methodology lays the groundwork for creating reliable and accurate crime prediction models. We can open the door for more efficient crime

prevention methods and progress in the field by taking into account the drawbacks and unfinished business mentioned.

CHAPTER 6: CONCLUSION

In summary, this study has successfully delved into the realm of crime rate prediction in London through the utilization of machine learning techniques based on Twitter activity. Among the four models investigated, the Random Forest model emerged as the most robust performer, boasting an impressive accuracy of 80.31%. Such outcomes hold considerable significance for London's law enforcement agencies, policymakers, and urban planners. By harnessing the power of Twitter data, particularly metrics like tweet counts, likes, and retweets, authorities can gain invaluable insights into the intricate dynamics of criminal behaviour and potential hotspots. The precision exhibited by the Random Forest model in predicting crime levels offers an invaluable tool for resource allocation and the proactive formulation of crime prevention strategies.

While the Random Forest model shines, it is important to acknowledge that no single model is devoid of limitations. To further enhance the prediction accuracy, future studies could explore the amalgamation of multiple models or the utilization of ensemble methods, capitalizing on the unique strengths of each approach. Furthermore, the potential for improvement lies in the inclusion of additional attributes or a more in-depth analysis of the textual content within tweets. This could potentially yield more accurate predictions of crime levels by considering the nuances of language and sentiment expressed in social media.

This dissertation contributes not only to the field of crime prediction but also to the broader application of machine learning in addressing real-world challenges. It provides a solid foundation for future research endeavours focused on crime prediction leveraging social media data. The insights garnered from this study can pave the way for proactive, data-driven crime prevention strategies, fostering safer neighbourhoods and enhancing the overall quality of urban life. As we move forward, it is imperative to recognize the potential ethical considerations surrounding the use of social media data for crime prediction. Safeguarding individual privacy and ensuring that the predictive models do not exacerbate existing biases in law enforcement practices are paramount concerns that demand careful attention.

In conclusion, this study demonstrates the power of machine learning to harness the vast potential of social media data for addressing complex urban issues. It bridges the gap between technology and societal challenges, offering a glimpse into a future where data-driven solutions contribute to the creation of safer and more secure urban environments.

REFERENCES

Gerber, M.S. (2014) 'Predicting crime using Twitter and kernel density estimation', *Decision Support Systems*, 61(1), pp. 115–125. Available at: https://doi.org/10.1016/j.dss.2014.02.003.

Johannes Bendler *et al.* (2014) *INVESTIGATING CRIME-TO-TWITTER RELATIONSHIPS IN URBAN ENVIRONMENTS - FACILITATING A VIRTUAL NEIGHBORHOOD WATCH*. AISeL.

Land Area and Population Density, Ward and Borough - London Datastore (no date). Available at: https://data.london.gov.uk/dataset/land-area-and-population-density-ward-and-borough (Accessed: 9 August 2023).

Mohler, G.O. *et al.* (2011) 'Self-exciting point process modeling of crime', *Journal of the American Statistical Association*, 106(493), pp. 100–108. Available at: https://doi.org/10.1198/jasa.2011.ap09546.

MPS Monthly Crime Dashboard Data - London Datastore (no date). Available at: https://data.london.gov.uk/dataset/mps-monthly-crime-dahboard-data (Accessed: 9 August 2023).

Perry, W.L. *et al.* (no date) *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Available at: www.rand.org.

Safat, W., Asghar, S. and Gillani, S.A. (2021) 'Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques', *IEEE Access*, 9, pp. 70080–70094. Available at: https://doi.org/10.1109/ACCESS.2021.3078117.

Saraiva, M. *et al.* (2022) 'Crime Prediction and Monitoring in Porto, Portugal, Using Machine Learning, Spatial and Text Analytics', *ISPRS International Journal of Geo-Information*, 11(7). Available at: https://doi.org/10.3390/ijgi11070400.

Shamsuddin, N.H.M., Ali, N.A. and Alwee, R. (2017) 'An overview on crime prediction methods', in *6th ICT International Student Project Conference: Elevating Community Through ICT, ICT-ISPC 2017*. Institute of Electrical and Electronics Engineers Inc., pp. 1–5. Available at: https://doi.org/10.1109/ICT-ISPC.2017.8075335.

Siegel Larry J. (2015) 'Criminology_Theories_Patterns_and_Typolo'.

Vomfell, L., Härdle, W.K. and Lessmann, S. (2018) 'Improving crime count forecasts using Twitter and taxi data', *Decision Support Systems*, 113, pp. 73–85. Available at: https://doi.org/10.1016/j.dss.2018.07.003.

APPENDIX

DIS.R