Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 13, Issue 1, January 2022: 212-216

Crime Prediction System: A Study on Machine Learning Approach for Crime Analysis and Prediction

Sonali Joshi [a], Robin Lobo [b], Joel Syrus Fernandes [c], Prof. Prajakta Bhangale [d]
[a,b,c] Student, [d] Assistant Professor, Fr. Conceicao Rodrigues College of Engineering (Affil. to University of Mumbai)
Email: sonalijoshi741963@gmail.com [a], robinlobo2001@gmail.com [b], joelfernandes711@gmail.com [c], prajakta.bhongale@frcrce.ac.in [d]

ABSTRACT

What is a crime? Crime is defined as an illicit act, or in simple words, it is an act performed against the law that poses a serious threat to society. Why does crime occur? The reasons for committing a crime are complex, it can vary from person to person. It may also depend on the situation of the individual, the various factors that contribute to guilt are as follows- some of them are longing, jealousy, vengeance, anger, on the other hand, some people commit crimes out of fear, or anger, therefore poverty and unemployment are major factors contributing to crime. All this raises the question, can crime be prevented? Or what can one do to prevent crime? Or how can crime be reduced? Nowadays when the world is achieving excellence in all fields whether it is development or technology nothing seems impossible now. The solution to the above problem is possible with some limitations. Crime Prediction System is a system developed for crime prediction and control. Along with this a question arises that what is a crime prediction system? Crime prediction is a structured approach to the discovery of patterns and directions of crime. This system helps police officers keep the law and keep track of places that are more prone to criminality.

Keywords: XGBOOST, data-preprocessing, Machine Learning, crime prediction system, crime analysis, NYPD, KNN.

.

1. INTRODUCTION

In modern times, many law-enforcement officers have been heightening their regular way of crime reporting with modern technological progressions to enhance their results by efficiently recording crimes to aid their investigation. Data is not only a record that holds pieces of information about the criminal activities undertaken over time but also comprises estimable insights to some sources and can help us get deeper into the investigation by providing various patterns in it. Previously, when there was less progress in technology, human intelligence was the only means to solve these devilish acts. But the rapidly increasing amount of data adds stress to this traditional method as the human thesis fails to deal with the gazillion records. That's where technology plays a prominent role.

Crime prediction and investigation have been studied and implemented by many researchers with varying success rates over the decade. But all these perspectives share a common fact that crime rate is directly proportional to location. To reduce the occurrence of such violations, the first step is to determine where and what types of offenses are prevalent. Machine Learning is a broad section of the study that automates the analytical model creation. It has revolutionized computer systems by giving them the sense to acquire knowledge from data, recognize patterns and produce decisions without the need for human intervention. A variety of systems, defined by machine learning, extend to solve problems where human intelligence fails. Its applications stretch to online fraud detection, medical diagnostics, email spam and malware filtering, search engine result purification, etc. Packs of machine learning algorithms act as brains to solve critical cases. Crime prediction is very essential and crucial for law enforcement as it has the potential to defend human lives and evade infringements in society.

2. WORKING

2.1. Data Collection

The crime data collection is achieved using the third-party API provided by New York City Police Department (NYPD) at the NYC Open Data portal which is reserved for free federal data to involve civilians in the reports generated and managed by the city administration. This dataset comprises all credible transgressions delivered to the New York City Police Department (NYPD). The dataset is updated every three months. The variables stored in the dataset comprise of the following:- the name of the borough in which the incident occurred, the date and time of occurrence for the reported event, an intimation of whether the crime was interrupted prematurely, attempted but failed, or completed successfully, the level of offense, the specific location of occurrence in or around the premises, the description of the crime, the date of reporting of the event, the victim and the suspect's age group, race description, and sex description, and the latitude and longitude of the crime incident.

2.2. Data Preprocessing

Data preprocessing includes reconstructing primary data to proper data sets since machines cannot use data that they cannot interpret. Primary data is usually deficient and has incongruous formatting. The adequacy or inadequacy of data preparation is associated with the success of every project that requires analysis or prediction of data. Data Preprocessing comprises both validation and imputation of data. The purpose of validation is to evaluate whether the data is both comprehensive and precise. The purpose of the imputation of data is to rectify errors and input missing values for the preprocessing of the dataset, we split it into two separate datasets, one dataset for analysis and the other for prediction. For the prediction dataset, there were quite a few missing values. Instead of dropping entire rows, we replaced missing values with "UNKNOWN" values to avoid the loss of data. We also used the date and time to add the year, month, day of the week, the part of the day, and the hour at which the crime took place for better analysis. For the analysis dataset, we took the date, time, latitude, longitude, category, and description of the crime. Since analysis prefers the data to be in numerical format, we created dummies for the crime categories and descriptions. We also used the date and time to add the year, month, and day of the week, part of the day, and the hour in a numerical format.

3. LITERATURE SURVEY

Several systems have been designed for the proposed problem. The proposed methods had given the desired results with some drawbacks. Most systems suffered from a lack of adequate datasets because they were not detailed [3,5,7]. Some did not include features such as population, habitat, and transport data. The datasets used were unbalanced [7]. The models also fail to confer the in-depth mapping of crime within the internal cities of the country [5]. The articles focus more on comparisons with different algorithms; however, they give a detailed description of the algorithm's performance [2]. Whereas, the model created using Multi-Linear Regression encountered a slight error while training the model. However, this research does not explicitly state the hindrances covered by other models the accuracy of KNN and decision tree classifier was observed to be highest but the analysis was done only for KNN and not for decision tree classifier [1]. The classification of violent/nonviolent crimes did not yield meaningful outcomes by the same classifiers. The initial data-set did not have enough predictability to achieve very high accuracy and found that a more meaningful approach was to divide crime categories into smaller, larger groups [15]. The model's accuracy is weak when prediction is concerned. The accuracy of both the approaches used was less than 50% [8]. Generalized errors were observed as a consequence of overestimation in regions with few reported crime incidents. The dataset used was a combination that ended in noise [9]. Some studies apply the latest equalization and boosting ensemble techniques from machine learning by analyzing crime at a granular level [13], while there is an inadequate comparison between models and no detailed description. Certain methods offered a solution which in real life may be too expensive to develop [11].

4. ALGORITHMS

4.1. Decision Tree

A Decision Tree algorithm creates prediction models by following non-parametric supervised learning approaches. Here, from the training data, simple decision rules are presumed for the creation of training models that can predict the classes and the values of actual variables. By slabbing down a dataset into more petite subsets, this algorithm progressively develops an associated decision tree. The resultant tree is the one with decision nodes and leaf nodes.

4.2. Random Forest

The Random Forest algorithm is a superintended algorithm for classification built from decision tree algorithms. In this, the accuracy and precision are directly proportional to the number of trees that the model builds. both classifications, as well as regression problems, are answered by the Random Forest Algorithm. It also helps to bypass overfitting and deals competently with the missing values in a dataset. It applies the bagging or bootstrap aggregation technique of ensemble learning methods to solve complex problems by consolidating several classifiers together taking the average or mean of the outputs.

4.3. XGBoost

XGBoost i.e., Extreme Gradient Boost is a Machine Learning algorithm that implements gradient boosting skeletons. Gradient boosting applies a gradient descent algorithm to lessen error in the sequential models. The decision tree and gradient boost are the key elements of this algorithm. It covers the boosting framework from the ensemble techniques. It is indicatively designed for optimal speed and performance that accompanies the models by generating more eminent results using scarcer computing resources in the quickest amount of time. Every tree is formed considering the estimates of the prior ones to get a final model with fewer deviations.

Boosting technique subdues the drawbacks of Bagging by dealing with bias or underfitting portions in a dataset. It braces various loss functions and operates excellently with interactions, just a fact that it requires scrupulous tuning of several hyper-parameters. Random Forest fails to provide a precise output since the latest prediction output relies on the mean of the predictions produced by subset trees. Boosting takes slow steps, making the predictors gradual rather than independent. It monotonously leverages the patterns in residuals, extends the model with low predictions, and makes it better. XGBoost has the potential to depreciate the error rate.

5. CONCLUSION

The offense rate in recent years has exponentially progressed with this, there arises a need for such a system that can help not only to analyze crime but also help predict it to some extent. The current prediction system has some loopholes and can be made accurate, multiple papers have targeted specific functionality like predicting the output without relating with other algorithms or just producing crime prediction for states rather than cities and places, or simply displaying the crime analysis, this can be advanced by using right prediction system and by using right dataset.

Hence, performing the research not only on a single crime type but also identifying multiple crime types will be beneficial in targeting the major crime and also be quite helpful for the officials to carry further investigation to reduce it to some extent. We will be using the XGBoost algorithm to predict the results of the outcomes as well as data analysis in order to display the entire summary of crime taking place.

REFERENCES

[1]. Crime Prediction and Analysis Using Machine Learning https://www.irjet.net/archives/V5/i9/IRJET-V5I9192.pdf

[2]. Predicting Crime Using Time and Location Data https://www.researchgate.net/publication/335854157_Predicting_Crime_Using_Time_and_Location_Data

[3]. Crime analysis in India using data mining techniques https://www.researchgate.net/publication/325117499_Crime_analysis_in_India_using_data_mining_techniques

[4]. A systematic review on spatial crime forecasting https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-020-00116-7

[5]. Safety App: Crime Prediction Using GIS https://ieeexplore.ieee.org/document/9137772

[6]. XGBoost: A Scalable Tree Boosting System https://arxiv.org/abs/1603.02754

[7]. Exploratory data analysis and crime prediction for smart cities https://www.researchgate.net/publication/334582376_Exploratory_data_analysis_and_crime_prediction_for_smart_cities

[8]. Crime Analysis Through Machine Learning https://www.researchgate.net/publication/330475412_Crime_Analysis_Through_Machine_Learning

[9]. Analyzing and Predicting Spatial Crime Distribution Using Crowdsourced and Open Data https://www.researchgate.net/publication/324107124_Analyzing_and_Predicting_Spatial_Crime_Distribution_Using_Crowdsourced_and_Open_Data

[10]. Crime Prediction & Monitoring Framework Based on Spatial Analysis https://www.sciencedirect.com/science/article/pii/S187705091830807X

[11]. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention https://www.researchgate.net/publication/351187525_Crime_forecasting_a_machine_learning_and_computer_vision_approach_to_crime_prediction_and_prevention

[12]. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest https://link.springer.com/article/10.1007/s42452-020-3060-1

[13]. Mining large-scale human mobility data for long-term crime prediction https://link.springer.com/article/10.1140/epjds/s13688-018-0150-z

[14]. Crime rate prediction in the urban environment using social factors https://drive.google.com/file/d/1BSnjpn134UGrL9JlUprzXo9lAKQiRtoW/view