Urban Crime Prediction with Machine Learning

Aakash Gajera
E-mail: aakashgajera9690@gmail.com

***Abstract*— Criminal activities in urban centres are prevalent in all urban areas around the world and a lot of resources; both personnel and money, are devoted to law enforcement so as to cope with these criminal activities. Tackling or eradicating urban crime completely has proven to be an impossible undertaking so much so that the main job of law enforcement in these urban areas is to just minimize its prevalence and spread. This hence implies that crime still exists but in isolated incidences and in lower numbers across various places within the urban centres. The law enforcement modus operandi is that more law enforcement officers/personnel are deployed so as to equalize or better the law enforcement to criminal ratio. This is expensive considering that each officer has to be facilitated throughout the course of his/her operations. Can Machine Learning help make law enforcement more efficient? The advent of Artificial Intelligence/Machine Learning coupled with the fast-paced research and innovation in these fields; have opened doors to innovative ways of solving various technical challenges, both complex like self-driving cars and less complex ones like binary image classification. Advancements in Machine Learning and Deep Learning have also made it possible to predict and map out urban crime and some models/solutions have been deployed in the real world environment by some police departments around the world. The results of this have brought mixed reactions more notably the staunch opposition from various activists especially in regards to bias in the deep learning models deployed, user privacy in the way they operate or are operated and also on ethical grounds.**

Whereas existing research has proposed some groundbreaking use cases of artificial intelligence and machine learning in crime prediction such as using video and image recognition tools, ethical considerations have largely been ignored as these methods are very much individual-centric. Little attention has been paid to methods that don't infringe on user privacy such as those that predict or map crime hot-spots partly because they aren't as trendy as their image/video recognition counterparts and also because most research until recently didn't take into account the ethical and privacy considerations of some of the proposed methods. In this study we present results from 2 different machine learning based crime rate prediction models relying on a curated tabular dataset with no reference to any personal information within it and an Artificial Neural Network model based on deep learning.

KEYWORDS: Artificial Intelligence (AI), Machine Learning (ML), Deep Learning, Urban Crime Prediciton,

I. Introduction

Criminal activities in urban centres around the world are a common occurrence in pretty much every country and of every continent and these have a long history. Criminal activities have evolved as much as society and technology have and they take on a new form with each new day that goes by. These criminal activities are a major social and societal problem as they affect directly and indirectly every aspect of human life, security and the economy.

Advancements and research in the data science and artificial intelligence/machine learning fields have opened a variety of use cases for these technologies such as in crime rate prediction; fueled by the increased data availability in various domains. In the case of crime prediction which is the main focus area in this study, some law enforcement/police departments and governments have taken steps to tap into the potential of this machine learning technology in combating crime by applying techniques such as computer vision based facial recognition to track individual's movement or suspicious action sequences and in some cases to predict which individuals in streamed CCTV footages are likely to commit crime based on their behaviour in the footages. This facial recognition mode of use of machine learning/deep learning technology has yielded catastrophic results as innocent civilians flagged as potential criminals or most likely to commit crime were arrested and in some cases sentenced to jail. For instance more black people were likely to be flagged as potential criminals by the AI algorithm deployed by the Chicago police department [1]. Furthermore, a lot of user data is mined without their consent in the name of law enforcement and this is privacy invasive and unethical as many critics have come out to explain. In this study we focus on the use of structured crime data to be able to predict future crime rates in given areas.

II. Methods

To predict the crime rate in particular areas, we train three different models on the same objective (crime rate) and same train data sample.i.e. an Xgboost, LightGBM and ANN models. Xgboost and LightGBM both produce gradient boosted models which are ensembles of weak learners/poor prediction decision tree models. Both Xgboost and LightGBM are highly optimized gradient boosting models capable of yielding very accurate results in short train times for various tasks. The main difference between both is the manner in which they grow their trees; Xgboost implements a level-wise/breadth-wise tree growth mechanism whereas LightGBM applies a leaf-wise/depth-wise tree growth mechanism. Performance-wise, LightGBM is faster than the Xgboost. The Artificial Neural Network (ANN) is a deep learning based model which consists of fully connected layers inspired by the neurons in the human brain. The first layer is usually an input layer that ingests input data and this

is followed by an n-number of hidden layers/neurons with activation functions after each and finally an output layer that generates the final prediction results.

A. ANN Model Architecture

While the Xgboost and LightGBM models have decision tree structures, the ANN is composed of fully connected layers as visualized below;

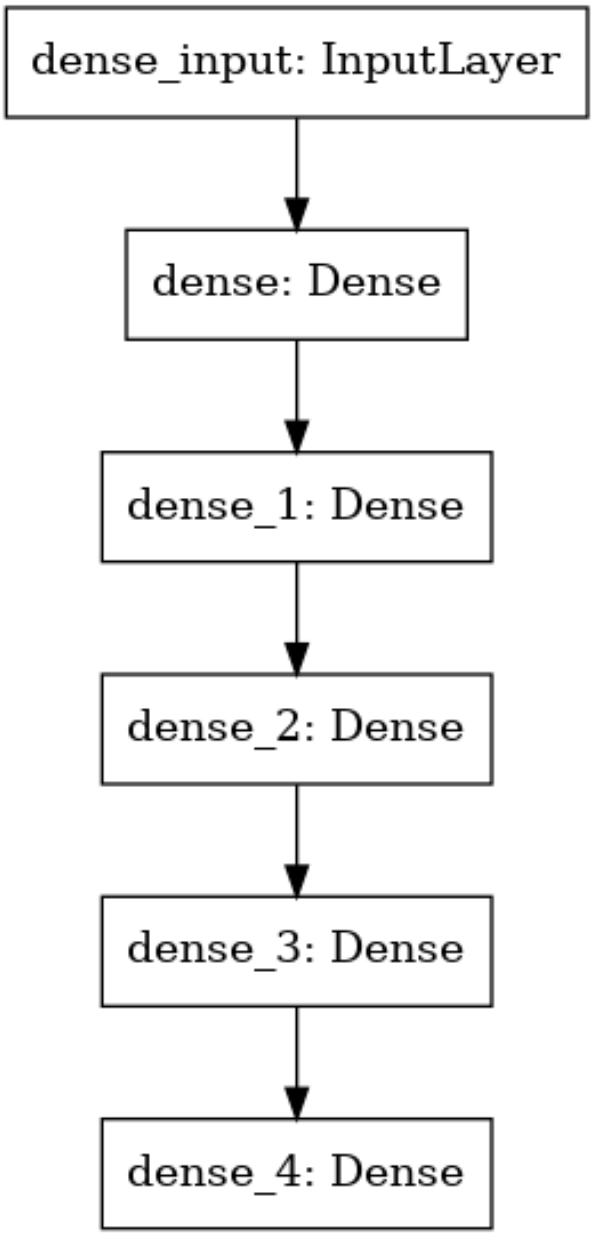

Fig. 1. ANN Model Architecture.

III. Dataset

The crime dataset [2] sourced from online public datasets comprised of criminal activity levels in various communities and was already cleansed so it didn't require much more data cleaning except for formulating the target variable (crime rate). The dataset consists 1993 records which is quite small for any generalizable performance to be achieved for any model to be of real world use.

The crime rate in our study is defined relative to the violent crimes per population of an area in the dataset and thus a binary target variable is formulated to represent areas with high violent crimes per population as 1 (high crime rate) and those with a low violent crimes per population as 0 (low crime rate). The distribution of the final target variable (crime rate) is visualized below;

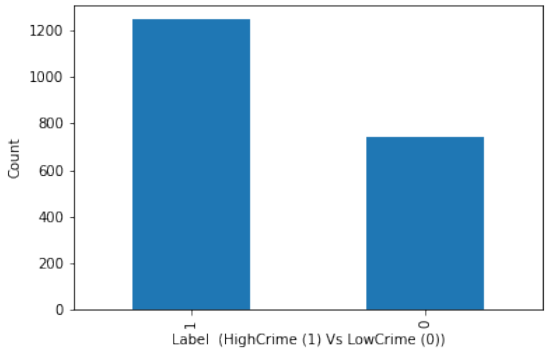

Fig. 2. Target Variable Distribution .

IV. Results

Given that there is an imbalance in the dataset due to the higher number of samples of high crime rate and the relatively lower number of low crime rate samples, the Xgboost and LightGBM models are trained with KFold cross validation technique to ensure representativeness of training samples hence reducing bias in model predictions of the trained models achieve an accuracy of above 80%. The ANN model input data however rely on KFold split sets but rather a single train-test-split set. The 5 layer ANN is trained for 8 epochs with a batch size of 8 and achieves a slightly higher accuracy on the hold out set than the Xgboost and LightGBM models. The ANN model train and validation loss and accuracy trends over 8 epochs for the ANN are visualized below;

Fig. 3. ANN Accuracy Vs Loss Plots over 8 epochs.

The accuracy results of the 3 models on the holdout set are presented as below;

MODEL	ACCURACY SCORE
Xgboost	81.183%
LightGBM	81.183%
ANN	84.649%

Fig. 4. Accuracy Results.

V. Discussion

The three models trained all achieve high accuracy scores but if only one model is to be chosen, then technically the one that achieves a higher test score would be chosen. So in this case the final model would be the ANN model. However, because the dataset size trained on is significantly small and that the variance between the best score and the next best score is small, we we'll select the Xgboost model as the final model. This is also in part because it is trained with cross validation to enhance generalization and reduce model bias.

VI. Conclusion

The crime rate prediction models trained do achieve good scores but given that the dataset used for training is small in size, the results obtained can only be used in a research environment and thus the models are no where near real world use.

VII. Acknowledgements

Credits to https://github.com/tina31726/Crime-Prediction/ from where the data used for model training was obtained and a little insight into the problem relative to this dataset. and to [3] which was the reference research paper for this study.

VIII. Challenges

The implementation as described in the sections above does deviate from what was stated in the proposal because of some technical obstacles faced in trying to achieve what was proposed. It was quite a challenge to find a dataset that very much suited the main objective as outlined in the proposal so much so that the first solution developed for this study had little to no correlation with the objective as outlined in the proposal. Fortunately a much better structured dataset was found and it's what is used to build the solutions as described above. Another blocker was the implementation of the hybrid model as stated in the proposal. After some research, implementing this seemed much more un-achievable and got no leads to prior practical work thus concluded that the proposal may have been too 'optimistic'.

References

[1] Renata M. O'donnell. Challenging Racist Predictive Policing Algorithms Under The Equal Protection Clause.
[2] https://github.com/tina31726/Crime-Prediction/
[3] José Ribeiro, Lair Meneses, Denis Costa, Wando Miranda, and Ronnie Alves. Prediction of Homicides in Urban Centers: A Machine Learning Approach.