Intelligent Automation of SFO Crime Prediction using Different AI Methods

Background:

Crimes happen in certain areas, and it is illegal activities against society or someone else's property. [3] It is one of the dominant and alarming aspects of society. Crimes are increasing continuously due to the awareness in criminals of the technological enhancements, modern devices, and even social media. The crime offender rather than venturing unknown territories, frequently commit crimes in comfort zones or places where they are familiar with. The places such as home, work, school, shopping, and entertainment areas. The analysis of the crime is one of the organized approaches to identify the patterns and trends. [5] crime analysis deals with investigating and detecting the crimes along with their connections to convict. The article is doing spatial analysis of crimes and areas of higher concentration by understanding the past data. [4] It is difficult to analyze the huge volume of data that are part of crimes as well as criminals are a main task of law enforcing officers. [2] Highest accuracy in crime foresee tendency increases the crime prediction rate and pushes more security to the public and helps to reduce or prevent the amount of future crime. This will help the residents to feel safer in the cities and [2] recognize crime prone areas and crime zones to find their safe route to the destination.

It is important to study the problem to understand different attributes or features that are responsible for crimes based on the past data. Explore the background work associated with crime analysis and research done in the field of crime analysis using ML and different approaches. [3] It is observed that artificial intelligence has shown importance in almost all the fields of crime prediction. This gives us insight on data collection, pre-processing, classification, pattern identification and finally, the comparison of different data mining and [2] provides the path for the most agreeable techniques of machine learning classification for the prediction of crime rates. [3] It also highlights the worth and effectiveness of ML in prediction and it can be used by police to reduce crime rates in society. [4] It helps the organizations and users in identifying the hidden data from a massive database of crime records to investigate, control and prevent crime.

As per paper the crime characterizes is an "act of felony or grave offense against society or someone else's property which is prohibited by law. However, crime studies have revealed that crime does not happen evenly across all places [7] and that specific types of crime tend to occur more often in certain areas that are called crime hotspots [8]. Hence, the spatial analysis [9] of different types of crimes along with the areas of occurrence helps to predict the different types of crime occurs in future. It can help to predict the timing and the day of crime [10]. This shows higher percentage of crime are concentrated in hotspots [11] [12] and predicting them beforehand drvie towards the effective for law enforcement. This push towards law enforcement agencies to assign more resources to taken in-charge of the areas that are more concentrated in-terms of crime occurrence. This makes the residents to feel safer in their respective cities."

The paper has explored the different patterns, attributes and categories of crimes that are associated with data. The paper [2] deals with only a few categories of crime and limited parameters associated with the ML approach. The paper [3] came up with a little advanced portion of paper [2] and explored crime data based on weekdays and weekends but the number of

ML approaches and parameters remains almost the same. The paper [5] deals with the data of social media and offline sources and converts the emojis into text data. It identifies the user as suspect, normal, or criminal users but it is completely at the higher level and further investigation and other processes needs to be carried out. This is limited to specific types of data and text. The paper [6] gives a good overview of association mining, clustering, classification technique, correlation, and regression but not breakthrough in any of the methods.

The Scope of the paper contains many things like, the paper can predict crimes that happen at different places and predict the category of crime that happens at a place with highest accuracy. They best utilize and explore different ML algorithms to analyze and predict the crimes. It gives us a path on how to deal with huge amounts of data for analysis and model creation. Random Forest, Adaboost and Gradient boosting gave good accuracy for different classes. They also talk about KNN and how they tested by moving the K value from 100, 1000, 6000 and so on. The value 6000 gave good accuracy. They used a Decision tree with depth of 5 for better accuracy. They brief about different types of visualization to understand the data in a much better way.

The paper has some limitations to address such as, it is unable to predict the victims of crime or about the resolution that can be thought of for the given incident or category. The paper talks about ANN Initially but there is not much importance or usage of ANN. It provides details of accuracy of different models for reduction in classes but no details about any class. Even though the accuracy of naive Bayes and KNN is lower compared to random forest, Adaboost and gradient boosting, they have used those for prediction and are not sure about the reason behind the same. The model and other factors are specific to the San Francisco dataset and looks like data is skewed towards two places of east San Francisco. Analysis on different cities will give more clarity on the performance of the models. Effective visualization can still be done in different ways, but it is not defined, and no analysis done on resolution.

Project Objective

The goal of the project is to analyze and visualize the spatial and temporal relationship of crimes on various attributes and predict the category of crime in a particular location, address the limitation of existing paper and further look towards the enhancement to explore the usage of ANN methodology. Explore the different libraries of python to implement all the task of the project so that it limits the cost of infrastructure. The python libraries like folium, geopandas and widgets are used for map and other analysis as part of visualization. Extend the flexibility of code to use on different cities' crime dataset like BOSTON to showcase its analysis and performance. The focus is also on crimes and its resolution. Identify the different metrics to provide more meaning to the model and its architecture. Create an interactive dashboard and a web application using python. Explore the possibility to relate the data skewness to the population or poverty, or others based on the area using census data.

Data Collection

The Data is available on the SFO website, and it has been derived from the San Francisco Police Department Crime incident reporting system. The data contains the details of the crime from 2003 to present year. The data consists of two csv files, the first csv file has data from 2003 to

2018 and the second is from 2018 to present. The data contains 9 -10 required features with 37 - 56 incident categories.

Data Pre-processing

The quality of data helps in providing valuable information and helps in building the models. Some of the methods that are followed to improve the quality of the data are data cleaning which involve removing the invalid or unnecessary rows, addressing the missing values, data transformation involves extraction of new features from existing attributes and combining the required attributes to create the new features, data reduction involves discarding the null record or attributes that are not required or correlated and data conversion involves the conversion of character categorical to numeric categorical.

Analysis and Visualization

The important features that are responsible for analysis are incident date, time, and location and they are highly correlated attributes. The attributes such as day of the week, time of the day and seasons of the year are important factors for the crimes. It is observed that most of the crimes are happening on Friday and most frequently occurring crimes are theft, burglary, robbery, missing person, and drugs. More crimes are committed in the north-east part of San Francisco. There are quick resolutions for cases like robbery, burglary, and assault. The seasons like summer and fall are attracting more crimes. It is clearly visible that the number of crimes had decreased during 2020, it might be because of Pandemic. The crimes are mostly occurring during afternoon hours, and few are happening during evening hours. If we analyze the yearly trends, the occurrence of the crimes has changed from fall to winter after the pandemic. The attributes timeofday, hour, latitude, longitude, and police distinct are highly correlated attributes. Analysis of historical data from 2003 to 2018 gives more meaningful insights on the data pattern, trend and how to build the model. The different models like Random Forest, K-nearest neighbor, ANN, Tablet and time series analysis are created to understand the behavior of the data with the model.

Methods and Process

The methods that are followed to identify the required features and address the missing values are based on correlation and null values. The feature extraction such as day, year, month, hour, season, timeofday from the timestamp attribute are the most important factors to focus on the models. It is necessary to convert the character categorial to numeric categorical features and convert the attributes values to numeric. The data split requires more attention in proper distribution of all the classes of incident category in train, validation, and test dataset. The analysis of data needs aggregation which is one of the better methods. The creation of the model expects the standardization of the data. The libraries are important to support the project tasks. The models are fine-tuned based on the hyperparameters, activation function, loss function, number of hidden layers, neurons, optimizer, and scheduler parameters. The complete code is extended to apply the method to a different dataset which is the BOSTON crime data set. The code only needs initial dataframe setup and the rest of code needs only changes on the number of categories of the crime. It is easy to apply to any crime dataset in a very short time and in one go for all tasks such as analysis, visualization and modeling. It is better fit to all the different types

of visualization, analysis and model creation process and methods. This project is intended to be used for crime applications, such as assistance for the crime victims, police department, Victim service division, crime map and public safety awareness, Crime rates and statistics, Attorney, and legal advocacy. It is particularly intended for public safety awareness.

Metrics

Evaluation metrics include confusion metrics that contain the values of True positive, True Negative, False Positive and False Negative that helps in False Positive rate and False Negative rate to measure disproportionate model performance errors. The fraction of negative (not falling to the same category) and positive (same category) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. These metrics provide values for different errors that can be calculated and provide better understanding of classification. The accuracy of the models is between 83 – 98 % achieved through different ML and AI approaches. All metrics reported at the .2 decision threshold.

Discussion

We always approach ML algorithms for classification problems, but Deep learning models and TabNet are also good for classification problems on tabular data. To build the model's hyperparameter and different features are important factors for the algorithm. In the case of RandomForest using entropy criterion gives better accuracy than Gini Criterion. The number of estimators in K nearest neighbor plays a significant role in the algorithm. In ANN deep neural network architecture, activation, optimizer, and loss should be carefully chosen to get better performance. Learning, decay rate and batch size plays a major role in TabNet. Time series depends on how data is closely related to the previous trends. It is a good practice to keep a smaller number of classes or grouping similar classes as one to get better accuracy. High amount of data helps in fine tuning the model. Uniqueness and data consistency are important factors to build the model. If we build different models it provides more confidence and ideas to weigh to choose the model for prediction.

Evaluation and Reflection

Metrics like accuracy provide confidence in better performance of the model. Confusion matrix helps in visual representation of the prediction and its deviation. Classification reports provide the different metrics like precision, recall, f1-score for different classification. This gives better pictures on how the model is good enough for each classification. MSE and RMSE help in clear view of time series model performance. The loss function greatly influences the performance or accuracy of the model.

WebApp

To create the webapp start with YAML code as the first line of the jupyter notebook by providing the required filters as part of the code. Define the dashboard name as part of the WebAPP page. The webapp needs the mercury libraries hence install the libraries for python and use the

command such as jupyter trust, mercury add, and mercury watch on the created file. This will initiate the webapp in the link “http://127.0.0.1:8000” .

Interactive Dashboard

The interactive dashboard needs different python libraries for map and interactive display and methods. The folium library to display the map and ipywidgets library for interactive display are installed for python. The next step is to identify the important features for the dashboard and followed by initialization of the widgets for the features that are part of filter conditions.
The description and layout specify the display of the filter in specific format. The function should be defined to get the data and transform the data if necessary to required aggregation. Use folium method to display map and any other graph or chart if necessary. Use widget interactive method to invoke the function and required widget that are part of filters to display the interactive dashboard.

Conclusion

We always tend to move towards ML algorithms for classification problems as it is white box, but there are other models like deep learning, time series and TabNet which can better fit and are easy to implement. Even Though ML algorithms overcome the Deep learning and TabNet models, more data and fine tuning of any models perform better for the given data. This project deals with all the methods using python to make the users more friendly and reduce infrastructure cost. It is easy to use the same method to any crime dataset with little modification to the given data to fit to the required format and attributes.

Limitation

This project is to predict the incident categories. The number of categories may vary based on the data. It is not suitable for identification of person or thing responsible for crime; Crimes were categorized based on evidence produced and justified report. It is difficult to get the census data based on city and geographical location.

Future Scope

The crimes are concentrated towards the particular place and it might be tagged to population or poverty. In order to deep dive on this it is required to find the census data based on geographical location. Our future aim is to find the census data based on geography to identify and relate the crime patterns.

Existing Paper and its approach

The published paper is on the presentation to compare the different machine learning models and its performance. The paper used the algorithms like naive bayes, Bayesian networks, Gaussian naive bayes, decision tree, random forest, weighted k-nearest neighbors, multi-layer perceptron classifier, adaboost, gradient boosting, linear discriminant analysis and quadratic discriminant analysis to predict the crime category attribute. The paper result shows the better

performance of gradient boosting, random forest, decision tree and LDA in terms of accuracy. Some of the models perform better compared to others but the difference is less across the models. The model also dragged to use 10- fold cross-validation while calculating the training accuracy. This method can be used for different datasets, and this helps the law enforcement agencies to take advantage of machine learning models to maintain law and order and curb crime. This existing paper has more details and limitations. Hence, as part of our main paper we tried to address the majority of the aspects and some challenges. The details of the existing IEEE paper details are as below.

Topic: Intelligent Automation of Crime Prediction using Data Mining

Paper Citation: A. H. Al-Ghushami, D. Syed, J. Sessa and A. Zainab, "Intelligent Automation of Crime Prediction using Data Mining," 2022 IEEE 31st International Symposium on Industrial Electronics (ISIE), Anchorage, AK, USA, 2022, pp. 245-252, doi: 10.1109/ISIE51582.2022.9831620.

IEEE Page Link:
https://ieeexplore-ieee-org.libaccess.sjlibrary.org/abstract/document/9831620?casa_token=Idj_SuNhppgAAAAA:2JA_HqKT5GADqLEM1l4IMkDQBbjosWVyTjqcQhGHu0xD348c5OlV9kq3rfCoVPIaG2Lv2BTQ3g

References:

1. A. H. Al-Ghushami, D. Syed, J. Sessa and A. Zainab, "Intelligent Automation of Crime Prediction using Data Mining," *2022 IEEE 31st International Symposium on Industrial Electronics (ISIE)*, Anchorage, AK, USA, 2022, pp. 245-252, doi: 10.1109/ISIE51582.2022.9831620.

2. D. M S and S. Shankaraiah, "Crime Analysis and Prediction using Machine Learning Algorithms," *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, India, 2022, pp. 1-7, doi: 10.1109/MysuruCon55714.2022.9971801.

3. Pratibha, A. Gahalot, Uprant, S. Dhiman and L. Chouhan, "Crime Prediction and Analysis," *2nd International Conference on Data, Engineering and Applications (IDEA)*, Bhopal, India, 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170731.

4. M. Menaka and B. Booba, "Analysis to Improve Classifier Accuracy in Crime Data Prediction," *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2022, pp. 721-725, doi: 10.1109/ICCMC53470.2022.9754066.

5. D. Dipakkumar Pandya, G. Amarawat, A. Jadeja, S. Degadwala and D. Vyas, "Analysis and Prediction of Location based Criminal Behaviors Through Machine Learning," *2022 International Conference on Edge Computing and Applications (ICECAA)*, Tamilnadu, India, 2022, pp. 1324-1332, doi: 10.1109/ICECAA55415.2022.9936498.

6. S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma and N. Yadav, "Crime pattern detection, analysis & prediction," *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2017, pp. 225-230, doi: 10.1109/ICECA.2017.8203676.

7. M.A. Tayebi, U. Glausser and P.L. Brantingham, "Learning where to inspect: Location learning for crime prediction", Intelligence and Security Informatics (ISI) 2015 IEEE International Conference on, pp. 25-30, May 2015.

8. W Sherman Lawrence, R Gartin Patrick and E Buerger Michael, "Hot spots of predatory crime: Routine activities and the criminology of place*", Criminology, vol. 27, no. 1, pp. 27-56, 1989.

9. A Tayebi Mohammad, Richard Frank and Uwe Glasser, "Understanding the link between social and spatial distance in the crime world", Proceedings of the 20th International Conference on Advances in Geographic Information Systems, pp. 550-553, 2012.

10. Patricia L Brantingham and Paul J Brantingham, "Nodes paths and edges: Considerations on the complexity of crime and the physical environment", Journal of Environmental Psychology, vol. 13, no. 1, pp. 3-28, 1993.

11. D. K. Rossmo, Geographic Profiling, CRC Press, 2000.

12. D. L., Groff E. R. Weisburd and S Yang, The Criminology of Place: Street Segments and Our Understanding of the Crime Problem, Oxford University Press, 2012.

13. https://builtin.com/data-science/random-forest-algorithm