Intelligent Automation of SFO Crime Prediction using different AI methods

By
Praveen Kuruvangi Parameshwara

Project Summary

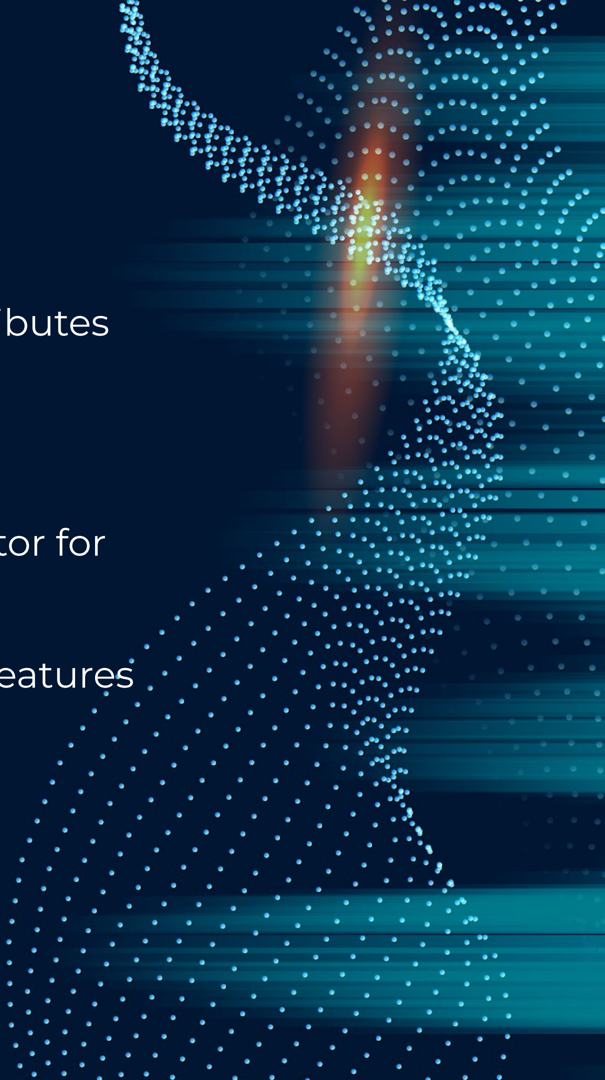
- Background
- Problem Definition
- Project Objective
- Data and Dataset
- Methodology
- Analysis
- Results
- Discussion
- Evaluation and Reflection
- Interactive Dashboard
- Web App
- Conclusion
- Limitation
- Future Scope

Project Background

- Crimes happens at certain areas, and it is illegal activities against society or someone else's property.
- 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 such as home, work, school, shopping, and entertainment areas.
- Crime analysis deals with investigating and detecting the crimes along with their connections to convict.

Problem Definition

- Different dataset has different attributes and some attributes are independent to the dataset.
- ML algorithms on tabular data is a common approach.
- Quality of the data in the dataset is most important factor for analysis.
- To build a model focus is on Identifying the important features and feature extraction are necessary.
- Focus on how to build complete project using python.

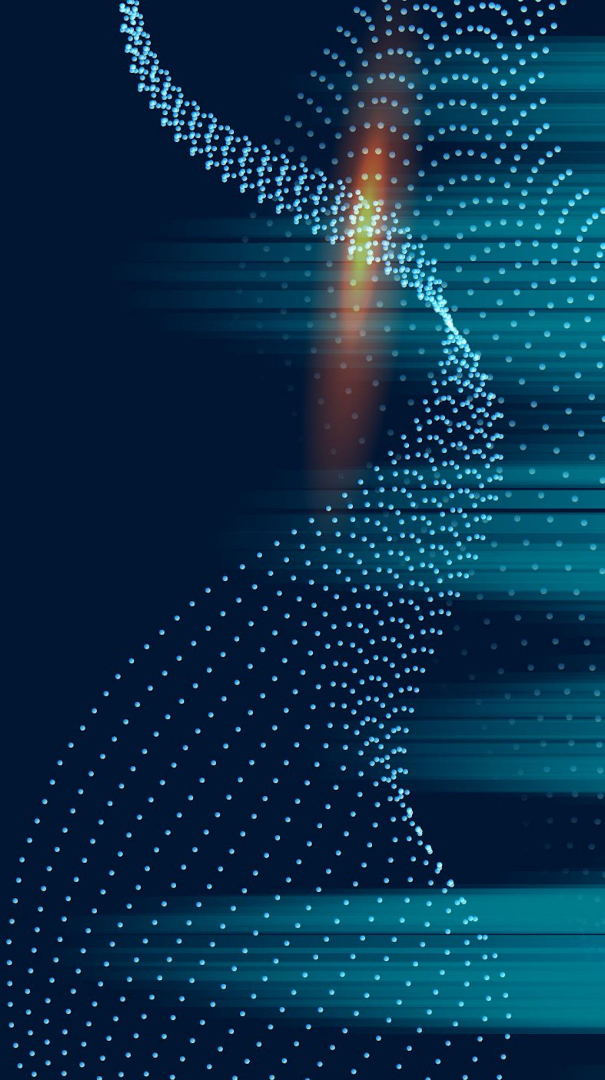


Project Objective

- The primary 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.
- The project address the limitation of existing paper and look towards the enhancement to explore on the usage of ANN methodology on the given SFO crime dataset.
- The project explore the use of python libraries for map and other visualization.
- The scope is extended to use different dataset or different cities' crime dataset to explore on the algorithm and its performance.
- Focus and deep dive into more insight of the crimes and its resolution and add the different metrics to provide more meaningful to the model and its architecture.
- Explore on ML algorithms, ANN, TabNet and Timeseries analysis.
- Create interactive dashboard using python.
- Create a web application using python.

Data and Dataset

- The Data is available on the SFO website. It has been derived from San Francisco Police Department Crime incident reporting system. The data contains the details of the crime from 2003 to present year.
- There are two csv files that are attached to the website. The first csv file contains 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.



Methodology

- Identify the required features and address the missing values
- Extracting the day, year, month, hour, season, timeofday from the timestamp attribute.
- Converting the character categorial to numeric categorical features.
- Converting the attributes values to numeric
- Split the data to ensure the required classification classes are present in train, validation and test dataset.
- Aggregate the data wherever necessary.
- Standardize the data for better modeling and classification.
- Identifying the required python libraries for the project support.

Analysis

- Incident date, time and location are more important factors.
- Time of the day and season of the year give more accurate information on the crime categories.
- Majority of the crimes are happening on Friday and more frequently occurring crimes are theft, burglary, robbery, missing person and drug.
- Highly committed crime areas are fall into North-east part of SFO.
- The cases like ROBBERY, Burglary and assault are fall into resolution status.
- Summer and Fall are the seasons that attract more crimes.
- The number of crimes had decreased during 2020 it might be tagged to Pandemic.

Analysis

- Afternoon hours are more prone to crimes and few are happening during evening hours.

- The trend of crimes has changed from fall to winter after pandemic.

- Timeofday, hour, latitude, longitude and police distinct are highly correlated attributes.

- Analysis of historical data from 2003 to 2018 gives more picture on the trend and model.

- Different models like RandomForest, K-nearest neighbor, ANN, Tabnet and time series analysis are created to understand the the behavior of the data with the model.

Analysis

- The complete method is applied to BOSTON crime data set only with initial dataframe code setup.

- The models are fine-tuned based on the hyperparameters, activation function, loss function, number of dense layers, optimizer and scheduler parameters.

- 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.

- Particularly intended for public safety awareness.

Result - Metrics

- Evaluation metrics include confusion metrics and classification report are more helpful to identify the performance of the models.

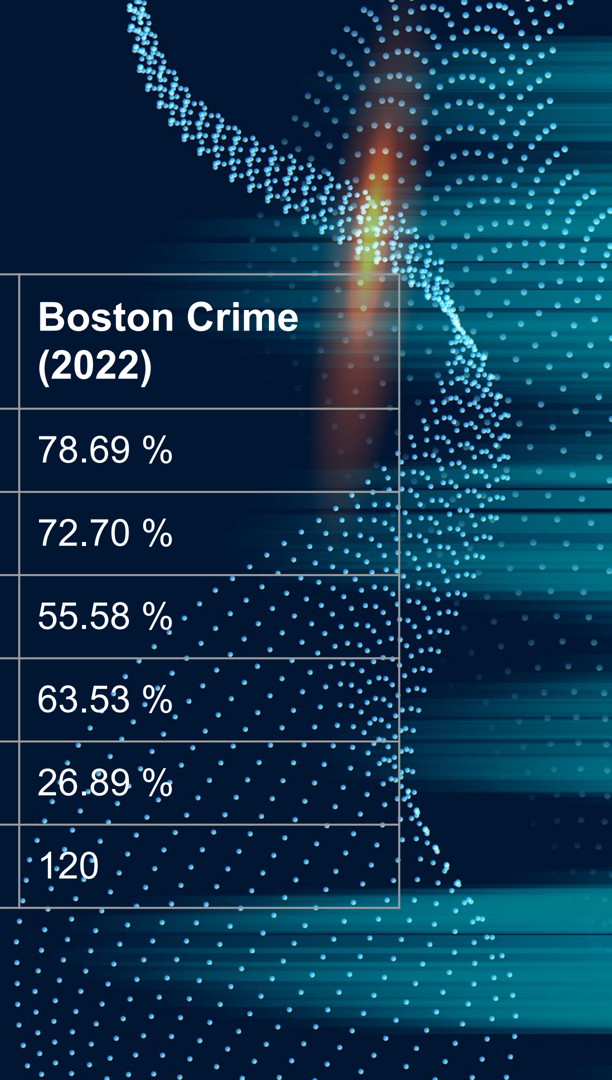
- Together, these metrics provide values for different errors that can be calculated and provides better understanding of classification.

- 83 – 98 % accuracy achieved through different ML and AI approach.

- All metrics reported at the .2 decision threshold.

Result

Model and Classes	SFO Crime (2003 - 2018)	SFO Crime (2018 to Present)	Boston Crime (2022)
Random Forest	83.70 %	89.27 %	78.69 %
K Neighbour	98.79 %	98.59 %	72.70 %
ANN	86.07 %	81.63 %	55.58 %
TabNet	89.25 %	86.29 %	63.53 %
Time Series	37.33 %	42.56 %	26.89 %
Number of Classes	37	50	120



Discussion

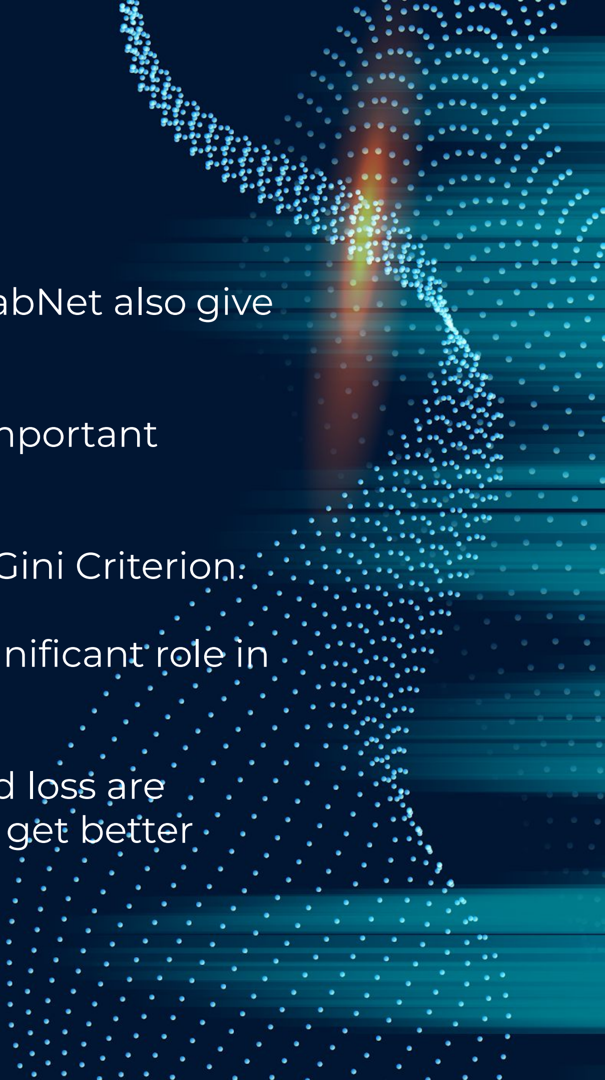
- Apart from ML algorithm, Deep neural network and TabNet also give good accuracy.

- The hyperparameters and criteria of algorithms are important factors to build the model.

- Entropy of RandomForest gives better accuracy than Gini Criterion.

- Number of estimator in K nearest neighbor plays a significant role in the algorithm.

- In ANN architecture, activation function, optimizer and loss are important factors and it should be carefully chosen to get better performance.



Discussion

- Learning ,decay rate and batch size plays major role in TabNet.
- Time series depends on how data is closely related to the previous trends.
- It is a good practice to keep less number of classes 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.
- Different model provide more confidence on the prediction.

Evaluation and Reflection

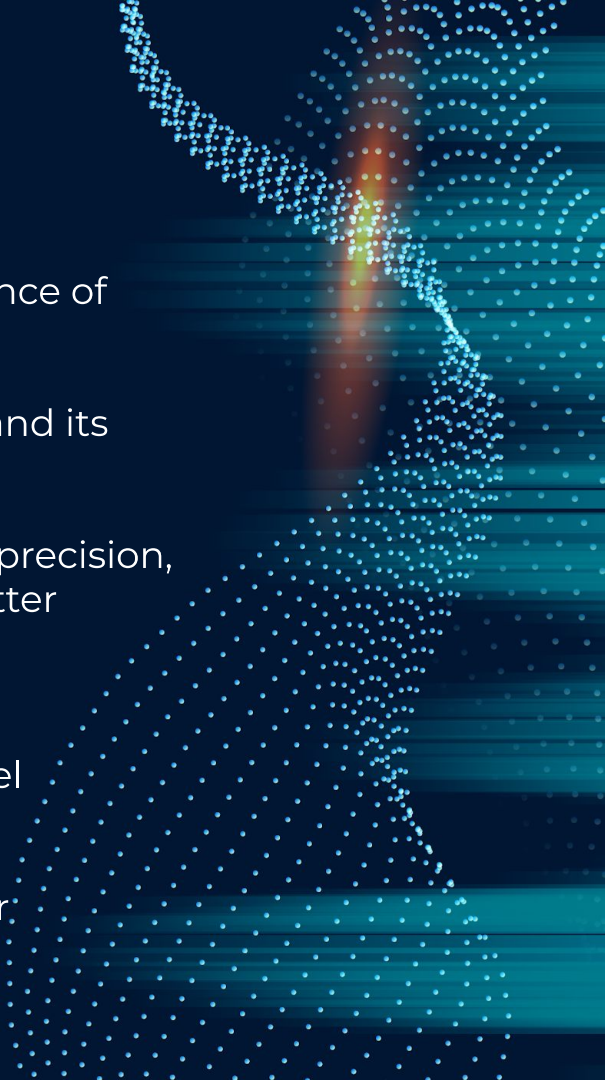
- Accuracy provides the confidence on better performance of the model.

- Confusion metrics helps in visualizing the prediction and its deviation.

- Classification report 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 helps in clear view of time series model performance.

- The loss function greatly influence the performance or accuracy of the model.



Interactive dashboard

- Install folium and ipywidgets libraries for python.
- Identify the important features for the dashboard.
- Create the widgets for the features that are part of filter conditions
- The description and layout specify the format.
- Define the function 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.

Webapp

- Create a YAML code as the first line of you jupyter notebook.
- Provide the required features as part of YAML code
- Provide the required name for the WebAPP page.
- Install Mercury libraries for python and execute below command
  - jupyter trust <filename>
  - mercury add <filename>
  - mercury watch <filename>
- The webapp will appear in the link “http://127.0.0.1:8000” .

Conclusion

- We always tend to move towards ML algorithms for the classification problems as it is white box, but there are other models like deep learning, time series and TabNet which can better fit and easy to implement.

- Even Though ML algorithm overcomes the Deep learning and TabNet models, more data and fine tuning of any models perform better for the given data.

- This project deals 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.

- Webapp and interactive map gives friendly and better visualization of the data.

Limitation

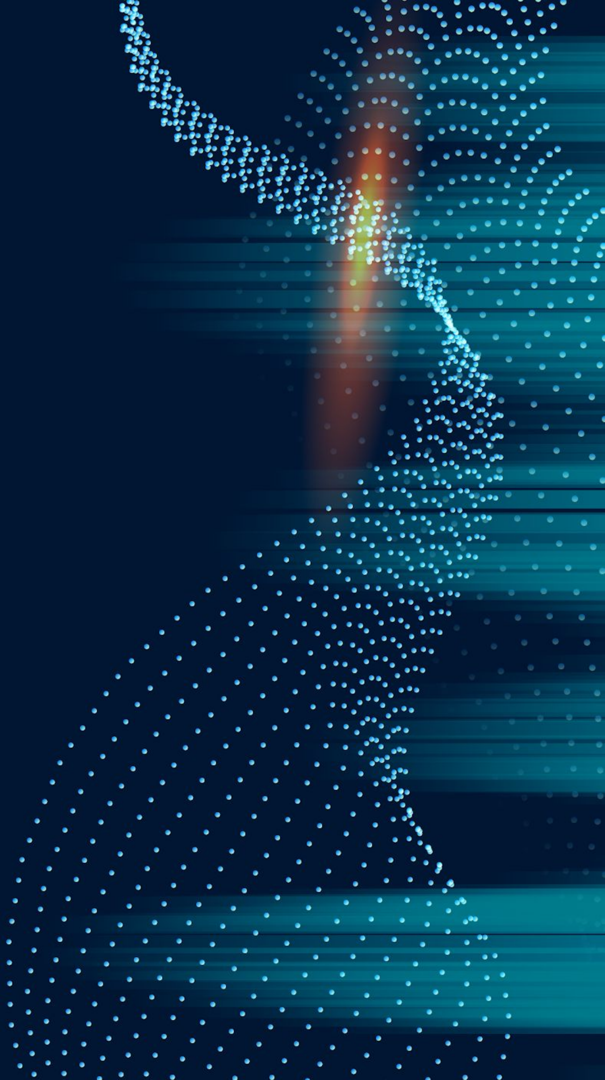
- This project is to predict the incident categories. The number of categories may vary based on the data.

- It is not suitable for identification of person or thing responsible for crime. Crimes were categorized based on evidence produced by the justified report.

- It is difficult to get the census data based on city and geographical location hence linking the crime to specific location is bit challenging task.

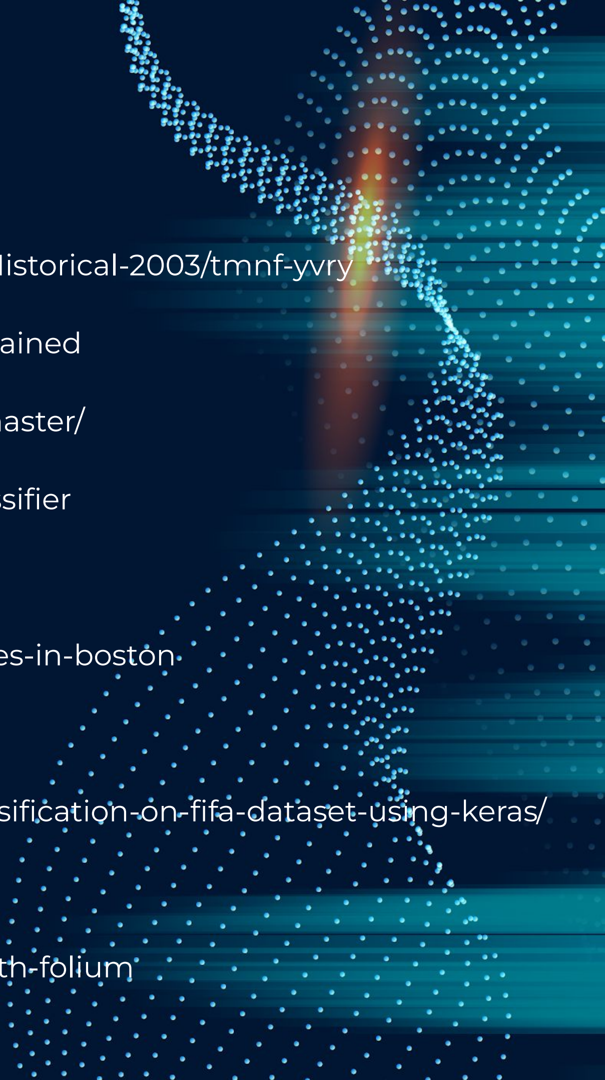
Future Scope

Find the census data based on geography to identify and relate the crime patterns.



References

- https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry
- https://www.kaggle.com/code/flafuji/sf-crime-eda-visualization-model-explained
- https://github.com/joshlingy/SF-crime-data-analysis-and-modeling/blob/master/
- https://www.kaggle.com/code/klyushnik/san-francisco-crimes-catboostclassifier
- https://towardsdatascience.com/deep-dive-into-sf-crime-cb8f5870a9f6
- https://www.kaggle.com/code/abhishekr7/time-series-forecasting-on-crimes-in-boston
- https://jovian.com/msameeruddin/00-cs1-eda-mv-tsa-bow-tfidf-final#C159
- https://www.analyticsvidhya.com/blog/2021/07/performing-multi-class-classification-on-fifa-dataset-using-keras/
- https://github.com/marcellusruben/sf-crime-voila
- https://www.dominodatalab.com/blog/creating-interactive-crime-maps-with-folium



Thank You