Tech Mahindra

PREDICTIVE POLICING:

LEVERAGING MACHINE LEARNING TO PREDICT CRIME HOT SPOTS

Anish Joseph[1] | Adarsha Murthy[2]
Tech Mahindra

Connected World.
Connected Experiences.

ABSTRACT

Crime prediction is a topic of special interest across the fields of Criminology, Smart City, Law Enforcement departments and Data Science. This article will give you insights on a Machine Learning enabled mechanism for identifying, parameterizing and predicting crime hotspots; by detecting the underlying patterns in the crime incident data. The objective is to give a bird's eye view of a critical project executed by Tech Mahindra Analytics Competency for the law enforcement departments, as a part of the Smart City initiative by the government; demonstrating the approach, outcome and the value proposition of predictive policing.

INTRODUCTION

The history of quantitative crime analysis goes back to the 19th century with a French Statistician designing the first visual map for crime data, revealing positive correlation between education and incidence of property crime. With the advent of affordable computation & data storage, accessibility to massive data sets, significant developments in the fields of data science and the availability of data scientists who can apply statistical analysis & machine learning algorithms to communicate actionable outcomes; data enabled crime prevention has recently come to be seen as its own discipline in the law enforcement field and has become a key component of modern policing.

With the increasing crime rates and its damaging effects on the economic & social fiber of the society, predictive policing is gaining popularity across the world. The U. S. Department of Justice and the National Institute of Justice had launched initiatives to support predictive policing. Police in Japan plan to introduce predictive policing into practice on a trial basis before the 2020 Tokyo Olympics and methods of predictive policing are being applied and tested in several European Countries.

Tech Mahindra has been a major player in the Smart City development, supporting critical projects across the globe. Staying true to the Core Value of Good Corporate Citizenship, TechMahindra is committed to contributing towards social, economic and environmental sustainability. One such initiative involved enabling the police department to address the increasing-crime rate in the province through predictive policing. This technique focuses primarily on identifying patterns in the Crime data and enabling inferences between different pieces of information; with the following key focus.

- Minimizing the crime occurrence through proactive police-patrolling
- Responding to the reported crime alerts in the shortest possible time
- Enabling the citizens to feel secure by being available for them in the vicinity
- Optimizing the police patrol vehicles based on changing crime hot spots for the respective day & time period
- Mark the Crime concentrated localities based on the historical & Futuristic Crime Rates
- Help understand factors and circumstances triggering crime in general
- Enable the effective planning of crime prevention measures like awareness campaigns etc.
- Reduce the overall efforts & costs in Police-Patrolling by enabling the patrol team to be more efficient

DATA COLLECTION & INTEGRATION

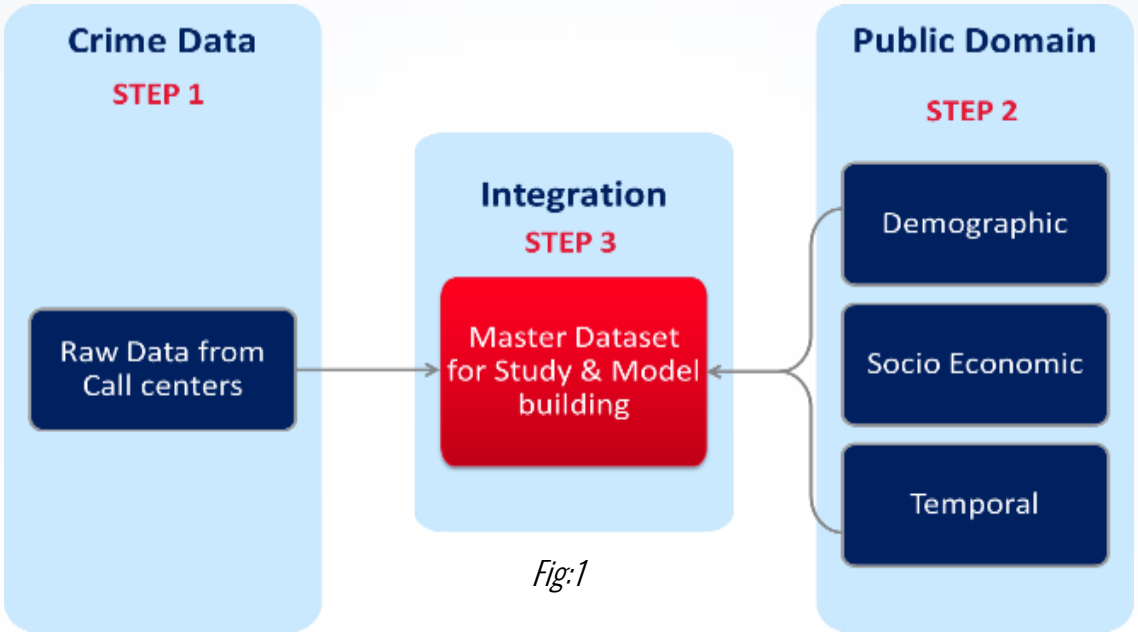


Fig:1

Crime patterns exist on a spatial, temporal, social and economic level and these patterns can be grouped and analyzed contextually based on the locality or town in which the crime occurs. Academic researchers in crime and human psychology points out to a relationship between the environmental characteristics and the incidence for crime. This led to the collection of data from multiple sources and integrating to generate the master data set **(Fig:1),** as demonstrated in the following three steps.

STEP 1: Collate the data from the police emergency call center managed by Tech Mahindra, which records close to a million calls on a daily basis from citizens who are reporting crime incidents or seeking help. This data involves the crime type, lat-long (latitude & longitude) of incident, crime time, the police action and several other variables that has been recorded for every reported incident.

STEP 2: Collate the Demographic, socio-Economic and temporal data (DST Data) data of the village & town in the province under consideration; which includes the lat-long coordinates (center point) of every village & town, social status of the citizens, life style, religious beliefs, possessions, nature of work, literacy rate, population ratio, living condition etc., in addition to the weather conditions in the given year, temperature and rain variations, key festivals, events and holidays.

STEP 3: Integrate this data with the police emergency call center data, which is available on a day-to-day basis for the reported crime incidents spread across the province. The challenge here is to map each of the police record with the corresponding DST data, which was accomplished by projecting the lat-long co-ordinates of the reported incident and the village & town lat-longs on to a planar co-ordinate system. Using the concept of Haversine distance, incidents are mapped to the respective village or town, which is falling within 5 km radius. The outcome is a master data set that contains all the variables and in this case, we had close to 180 of them.

EXPLORATORY DATA ANALYSIS (EDA)

It is important to understand the underlying patterns in the data, variable significance, correlations and hypothesis tests, which would help in defining the approach and identifying the independent, & dependent variables, before embarking on building a machine-learning algorithm. Statistical tests, multi-variant analysis and unsupervised algorithms like clustering and association rules were applied to the data, which revealed several interesting patterns, enabling the subject matter experts to guide the data scientist to pick the right variables for further statistical tests. Following are some of the interesting patterns that were observed and taken in to consideration, however not limited to these.

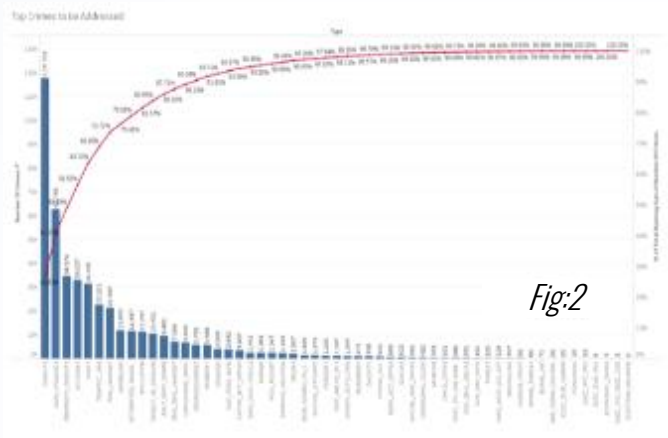


Fig:2

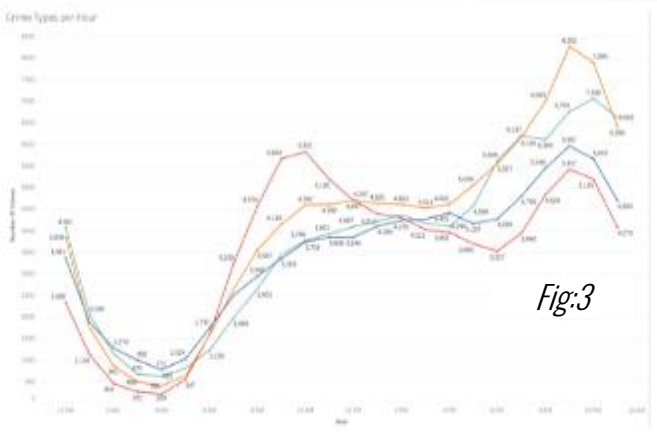


Fig:3

a) 80% of crimes are associated with seven crime types **(Fig: 2).** This was a good information to focus only on the top contributors and ensure that the data is more concentrated. There has been an interesting pattern on how the crime rate peaks between 04:00 PM and 11:00 PM and again between 06:00 AM to 11:00 AM **(Fig: 3),** based on this we can conclude the influence of the time of the day on the crime rate.

b) Some of the Crime types like dispute and other social crimes, including female harassment seems to be shooting up during some of the days **(Fig: 4 – red line)**

c) Crime rate shoots up during some of the festivals and holidays **(Fig: 5).**

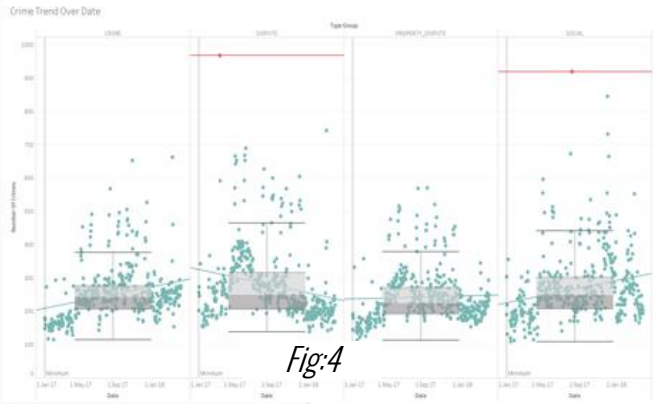


Fig:4

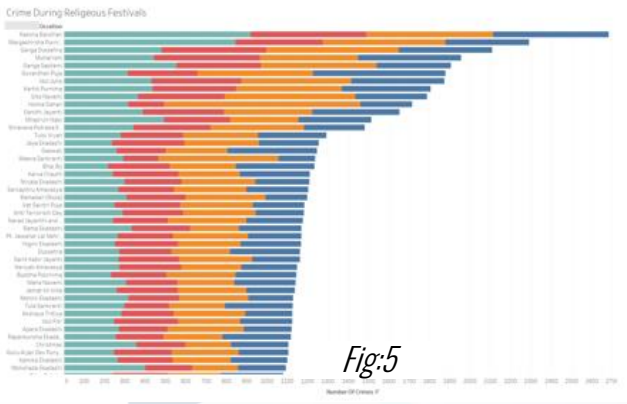


Fig:5

This information helped the team to retain the day of the week as well as festivals & holidays, in addition to other derived variables in the data set. The Exploratory data analysis helped the team to narrow down to 60 variables from 180, which was a significant reduction in the overall data size.

DIMENSIONALITY REDUCTION

The direct usage of the raw master-data variables could tend to yield the Machine Learning models with undesirable outcomes including coefficients biased away from zero, standard errors that are too small and confidence intervals that are too narrow indicating less test statistics and p-values, poor fit of the models and significant computation time. Hence it is imperative to consider only those key variables that eliminate redundancy & inter dependency and has the highest predictor properties.

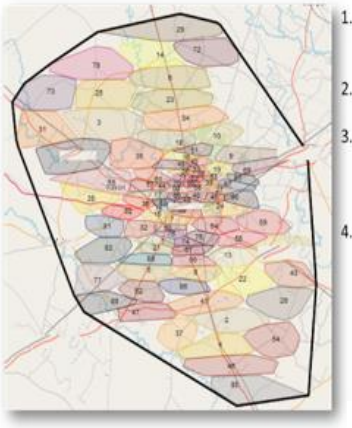
The team identified variable reduction techniques like "Conditional Random Forest technique", 'Principle Component Analysis (PCA)' and 'Factor Analysis' (FA). PCA Suggested around 25 principal components, however the linear relationship assumption failed to choose this model, as the data does not follow the Gaussian distribution. Before applying FA, KMO test was conducted to find that the MSA was very low **(Fig: 6)** because of which this approach had to be dropped. Conditional Random Forest will not assume the Gaussian distribution and works better when data has outliers and high degree of multicollinearity; and was considered a suitable technique in this case. This resulted in close to 20 variables that made to the final data set. Now we have a data set which reduced from 180 variables to 20 meaningful variables.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor_up100)
Overall MSA =  0.5
MSA for each item =
          district
               0.5
          priority
               0.5
       ps_category
               0.5
          latitude
               0.5
         longitude
               0.5
     day_midnight
               0.5
       day_morning
               0.5
   day_afternoon
```

Fig:6

DATA TRANSFORMATION

With a clear understanding of the data and its behavior, the next step was to define an approach to predict the crime rate across the province, which can be achieved by dividing them to clusters of manageable size. Hence, we have a classification problem wherein we would predict the probability of the crime occurrence in a cluster for a given day and time period within that day. Another approach that could be considered is the forecasting of the crime count in a cluster for a given day and time period. This led to the following tasks:



1. The graph shows cluster of geospatial co-ordinates with 5 km in radius
2. Each color depicts individual cluster
3. Now, we can study closely each clusters by applying Predictive Modelling Algorithms to predict / forecast the crimes in each cluster
4. There are around 80 Clusters for City A

Fig:7

a) Dividing the province in to Clusters based on the crime distribution, also considering the fact that the cluster size is small enough to ease the police patrolling **(Fig. 7)**. Clustering is an unsupervised machine learning technique where-in the lat-long coordinates of the reported crime incidents are statistically grouped together in such a way that it falls with-in a predefined radius, in this case 5Km. Clustering algorithms like DBSCAN, K-Means and K-Medoids were explored and the best output was considered. The lat-long coordinates representing the centroid of these clusters would be one of the predictor variables in the data set, enabling the selection of the cluster for crime prediction and hot spot identification.

b) Divide the day in to several time intervals – hourly intervals & 6 hour intervals, which was decided based on the way in which crime was distributed. The crime count is summed up to get the total crime occurring in these clusters for the given day and the time period. Hence we would be considering three data sets: daily crime rate, 6 hourly crime rate and hourly crime rate.

c) Define the independent variable for each data set: 1 = Probability of crime occurring is highest i.e. close to100% and 0= Probability of crime occurring is the least i.e. close to 0%. Each row in the three datasets are labeled as '1' or '0', which are defined based on the subject matter expertise and crime distribution, taking into consideration the count of minimum possible crime incidents which could occur in the given time period of the day. Additionally, 3 Class models have also been considered to assess the predictability.

MODEL BUILDING

Multiple classification & forecasting algorithms have been considered for arriving at the best fit model (Fig. 8). These models were developed for each of the 3 data sets with the objective of selecting the most accurate model based on the validation output.

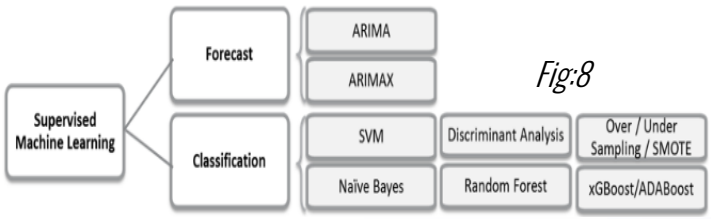


Fig:8

Classification Model: It was observed that the dataset with 6 hourly crime rate was performing better as compared to the hourly and daily crime rate data sets. Following is the output of various classification models that were developed.

Algorithms	Target Variable	Model Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Random forest	2 Class; Crime / No-Crime 0 – 0 to 1 Crime 1 – >=2 Crime	75.95	40.50	84.39	38.22	85.62
SVM	2 Class; Crime / No-Crime 0 – 0 to 1 Crime 1 – >=2 Crime	62.25	39.58	79.99	37.00	95.00
xGBoost	2 Class; Crime / No-Crime 0 – 0 to 1 Crime 1 – >=2 Crime	47.00	20.29	79.25	18.00	93.00
Bagging	2 Class; Crime / No-Crime 0 – 0 to 1 Crime 1 – >=2 Crime	61.00	33.50	77.45	19.00	89.00
Neural Network (ANN)	2 Class; Crime / No-Crime 0 – 0 to 1 Crime 1 – >=2 Crime	27.00	33.87	81.22	20.00	93.00

Fig:9

From **Fig. 9**, it can be noted that Random Forest has a comparatively high accuracy and has performed well on the validation data set with an accuracy of 75.95%. Fig.14 gives the details of the output of this model. Hence this has been selected as the best-fit model for predicting the probably of crime occurrence, over other models.

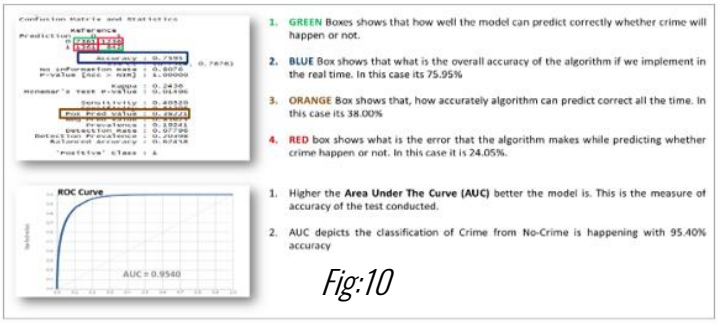


Fig:10

Forecasting Model: ARIMA and ARIMAX models were developed for the earlier mentioned 3 data sets. It was observed that 'Daily crime count' data set outperformed the others and ARIMAX gave the least "Root Mean Square Error (RMSE)" as compared to ARIMA and hence the same was selected

OUTCOME & VALUE

We have concluded on the following two models for predicting the crime occurrence, which is then transferred on to a GIS plot of the province, embedding the generated cluster with different color codes based on the predicted probability or count thresholds..

- Predict the probability of crime occurrence for the given day, time-interval and cluster
- Forecast the Crime Count for the given day and cluster

Following plot **(Fig. 11)** shows the visualization of crime hot spots (for comparison purpose prediction for two different days have been shown) with respect to each cluster and time period within the day. We can notice the Change in Crime pattern between Day A & Day B as well as within different time intervals in a day. The defined Threshold would determine the color of the cluster with RED being High probability of Crime and Green being the low probability (clusters below a certain threshold i.e. very close to zero crime probability has been eliminated in this plot). Similar plots were also developed for Forecasting where in the probability will be replaced with the potential crime count for each cluster, however only on a daily basis.

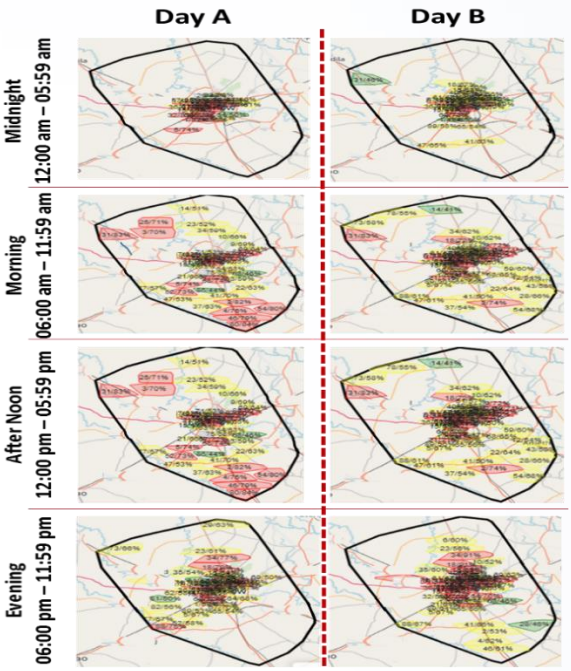


Fig:11

Police department can now plan their patrolling based on these locations or clusters and the time period, leading to a proactive and predictive approach in minimizing the crime rate as well as swiftly addressing any crime incident. This data could also be correlated with the history of the nature of crime (Heinous or less critical) in respective clusters, to refine the patrolling decision-making. The results are formed into a User Interface which displays the information to a user through a web-based visualization tool, with features such as crime demographic breakdowns, cross-sections of crime types in a geographic region, cluster shape, GIS enabled Historical & predicted Crime Hot Spots. This information is made available at the police control center and the mobile devices carried by the patrolling vehicle allowing the user to get the real-time insight.

The police department has conceived the perspective that predictive policing tools are intended to empower police staff and need not replace them. However, this technology would enable the police department to optimize their forces and resources by enabling them to focus on critical geographies that needs their attention and swift action.

CONCLUSION

The prediction is not intended to override the cognitive thinking of the trained police officers but should augment their overall decision-making processes. Guided by the predicted hot-spot map and simply being in the right place at the right time, we believe that the patrolling police officers can efficiently plan their route as well as draw more rapid and accurate conclusions which will result in more optimized resource allocation, enhanced public safety, swift response to crime incidents and decrease in crime in general.

ABOUT AUTHORS:

[1]ANISH JOSEPH

Program Manager & Data Science Consultant - Tech Mahindra.

Anish is an Engineering Graduate with a Post Graduate Diploma in Data Science, accreditation in International Sales & Marketing and Global Leadership. He has over 16 years of experience in various leadership roles including Analytics Consulting & Sales Enablement, Account Management & Strategy, Alliance & Program Management and Business Development; with experience ranging across several Industry verticals.

[2]ADARSHA MURTHY

Data Scientist - Tech Mahindra.

Adarsha has over 14 years of experience in IT Services and Development. An Expert in statistical modelling and ML techniques with solution design exposure in healthcare and Industrial domains. Besides, he is a freelance Analytics trainer and visiting Faculty at Data Science University.

INTEGRATED ENGINEERING SOLUTIONS (IES)

Is a Connected Engineering Solutions business unit of Tech Mahindra. At Integrated Engineering Solutions, customers are at the core of every innovation. We align Technology, Businesses and Customers through innovative frameworks. We deliver future-ready digital convergence solutions across Aerospace and Defense, Automotive, Industrial Equipment, Transportation, Consumer Products, Energy and Utilities, Healthcare and Hi-Tech products. Our 'Connected' solutions are designed to be platform agnostic, scalable, flexible, modular and leverage emerging technologies like Networking, Mobility, Analytics, Cloud, Security, Social and Sensors, that enable launching of smart products and deliver unique connected consumer experiences, weaving a connected world. Coupled with this, our strong capabilities in Electronics, Mechatronics and Mechanical Engineering along with domain understanding and product knowledge, bring excellence to the entire lifecycle of these connected ecosystems.

CONTACT US AT

connect@techmahindra.com
www.techmahindra.com

SOCIAL MEDIA

www.youtube.com/user/techmahindra09
www.facebook.com/techmahindra
www.twitter.com/tech_mahindra
www.linkedin.com/company/tech-mahindra

Tech Mahindra