Analysis of Crime rate in India and forecasting

Basava Prabhu GK
Department of CSE
PES University
Bangalore, India
basavaprabhu1652002@gmail.com

Vijit Kumar
Department of CSE
PES University
Bangalore, India
vijitviku@gmail.com

Rhea Sudheer
Department of CSE
PES University
Bangalore, India
rheasudheer.19@gmail.com

Srinivas Katharguppe
Department of CSE
PES University
Bangalore, India
srinivas.katharguppe@gmail.com

***Abstract*—** The purpose of this paper is to analyse and predict different types of crimes against women in India. The prediction is done by following a systematic approach for identifying and analysing patterns and trends in the crime data set used.

Index - data, regression, prediction, forecast, time series, correlation, crime, women

INTRODUCTION

At present, women have become sex objects and are widely treated as interior to men in different spheres of life. In the rural areas, wife-beating, torture of unmarried daughters, sisters and other female relatives is common phenomenon. Girls are perceived as a burden on the family, because of the huge amounts of money required for their weddings. Girls are generally not encouraged to take up even middle or higher education. There is huge discrimination between men and women in the sphere of education and the reason attributed to such gender bias is the feeling of people that girls should be confined to the house.

In this project we aim to make a comparative analysis of crimes district-wise and state-wise and forecast the crime rate for each district for the future years using Random Forest.

I. RELATED WORK

Previous research on the spatial analysis of criminal activity has established that crimes are not randomly organized in space [2]. In [3], [4], [5], [6] researchers note that there are a lot of external factors that influence the formation and displacement of crime hotspots in a city. The main factors identified by scientists are Gross domestic product (GDP), population density, unemployment and number of homeless in the streets. In addition to social factors, there are also different spatial indicators. For example, in [7] Yanqing Xu et al. have postulated a link between street lights and spatial criminal patterns. In [8] the researchers have determined the relationship between traffic and crime. In [9], [10], [11] the authors determined that the weather also influence the criminal activity.

Several approaches to crime prediction have been proposed by researchers earlier. In [12] Hyeon-Woo Kang and Hang-Bong Kang proposed the predictive method based on deep neural network. With their model they achieved the accuracy of 74.35%. In [13] the scientists used the algorithm of random forest regressor to measure the impact of urban factors on homicides and predict the future number of crimes of this type. However, scientists note that the model developed in [13] works well only with small datasets. Renjie et al. [14] apply the Bayesian Learning Theory to implement the model, that could predict serial crime. Anneleen et al. [15] explored 3 approaches to predictive modeling: logistic regression, neural network, and ensemble model. The obtained models were tested on 3 types of crime: home burglary, street robbery and battery. The results with the highest accuracy have been achieved using the logistic regression. Several models, developed not only in the environmental but also in temporal context, were proposed by researchers [16], [17].

Cheng et al. [17] have implemented ARIMA (autoregressive integrated moving average), SES (simple exponential smoothing) and HES (Holt-Winters Exponential Smoothing) models to make a short-term forecasting of property crime. The comparison of prediction results shows that the ARIMA model has the highest accuracy.

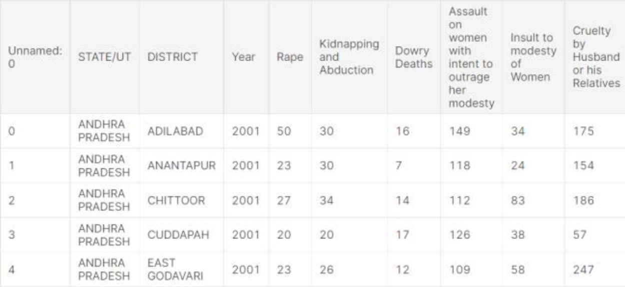
In the existing models the authors pay a little attention to the stage of the selection of factors that influence the criminal rate. However, this stage is one of the main ones in the procedure of modeling. The careful selection of features and their filtering improve the performance of predictive models and avoid their complexity and overfitting, is it was shown in the modeling tasks in other thematic areas [18], [19], [20], [21]. For example, in [19] authors use the feature selection method for prediction of solar radiation, Nicole et al. [20] implement this approach for handwritten character recognition. In the both of these studies the authors highlight that the use of this technique greatly improves the accuracy of the model.

III. PROBLEM STATEMENT

Analysis of crime against women in India and forecasting the rate using different forecasting methods.

A. *Dataset*

We obtained our dataset from the official Government website National Crime Records Bureau, which provides data, documents, tools and applications for public use. The collected data for the work contains crime information of all the 29 states and 7 union territories over 14 years from 2001-2014. Initially, the dataset classified crimes under several different types like 'rape', 'kidnapping and abduction', 'dowry death', 'assault on women with intent to outrage her modesty', 'insult to the modesty of women', 'cruelty by husband or his relatives', 'importation of girls'. But importation of girls did not hold any significance, hence we scaled it down to 6 features.



Unnamed: 0	STATE/UT	DISTRICT	Year	Rape	Kidnapping and Abduction	Dowry Deaths	Assault on women with intent to outrage her modesty	Insult to modesty of Women	Cruelty by Husband or his Relatives
0	ANDHRA PRADESH	ADILABAD	2001	50	30	16	149	34	175
1	ANDHRA PRADESH	ANANTAPUR	2001	23	30	7	118	24	154
2	ANDHRA PRADESH	CHITTOOR	2001	27	34	14	112	83	186
3	ANDHRA PRADESH	CUDDAPAH	2001	20	20	17	126	38	57
4	ANDHRA PRADESH	EAST GODAVARI	2001	23	26	12	109	58	247

Fig 1. Snippet of the dataset used

B. *Exploratory Data Analysis*

It was found that all 9 of the attributes were of type – float64 Null values were found for all features, but the feature 'importation of girls' consisted of most null values due to which was ignored. We converted out dataset into label encoding which creates a tuple consisting of (year,state,district).

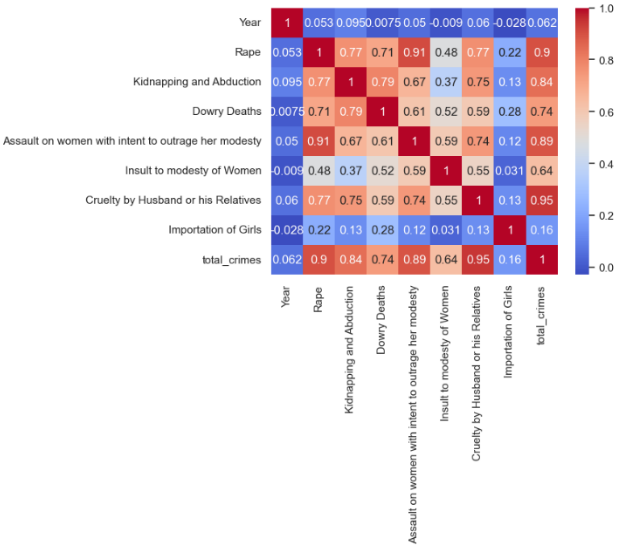


Fig. 2 Heatmaps of the 6 features.

It can also be inferred from Figure 2 that the correlation between rape and assault on women is high and there isn't great correlation between the others. Hence, we are keeping all the features, as they are important.

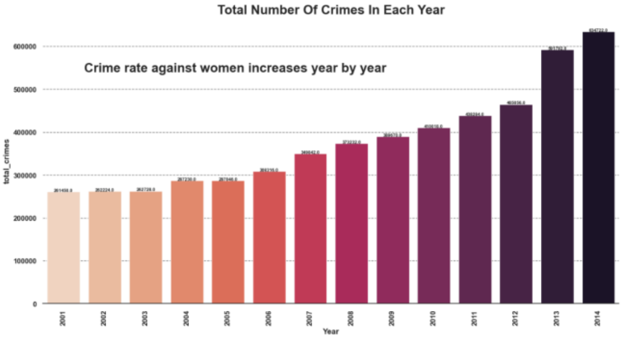


Fig. 3 Increase in crime rate over the years

From figure 3, we can infer that the crime rate has been increasing over the years

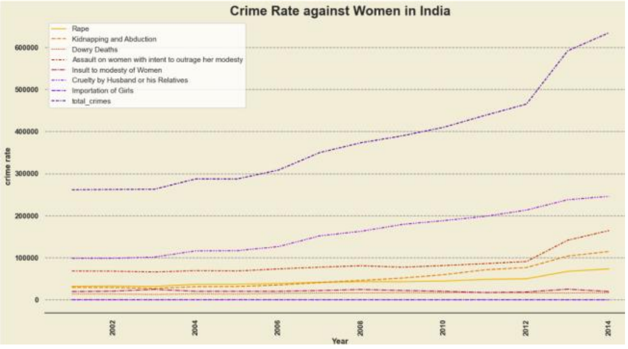


Fig 4. Crime variation for each type of crime over the years.

.

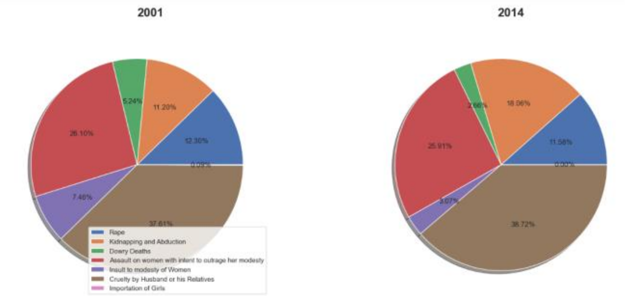


Fig 5. Comparison of crime in 2001 and 2014

Figure 5 compares the distribution of crime rate in the year 2001 and 2014. We can observe that the numbers of dowry deaths and insult to modesty has reduced by almost 50% and kidnapping and abduction has increased from 11% to 18%.

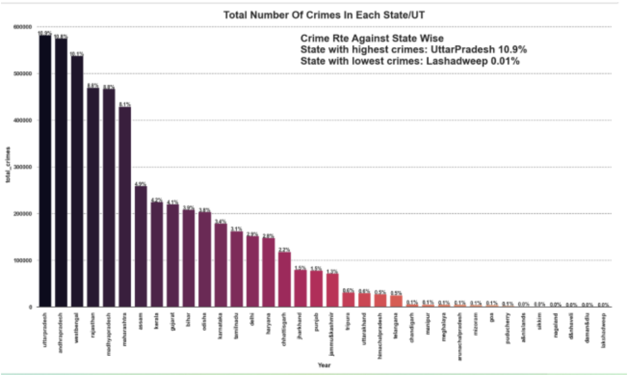


Figure 6. Comparison of crimes against women across the states.

From figure 6, we can infer that Uttar Pradesh has the highest crime rate and Lakshadweep has the least.

IV. PROPOSED MODELS FOR FORECASTING CRIME RATE

In this section we discuss the different models that we implemented to forecast crime rate. From EDA. Since we have to predict for all districts and most of the forecasting models predicts only one value, we are using Random Forests.

A. *Random Forests*

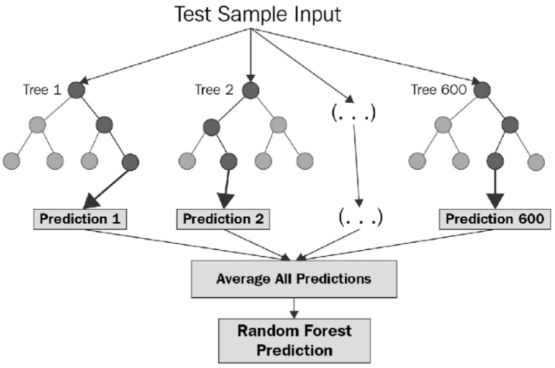


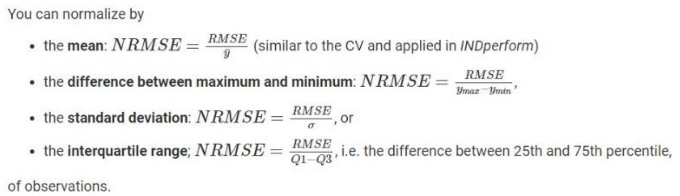
Figure 7. Structure of random forests

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. We use the mean or average prediction of the individual trees is returned.

We used the RandomForestClassifier function from the sklearn.ensemble library and set the argument n_estimators as 125 after checking the model to give the best accuracy using brute force.

V. PERFORMANCE METRICS

We used normalized RMSE to check accuracy of the model. The formula for the same is shown in figure 8.



You can normalize by

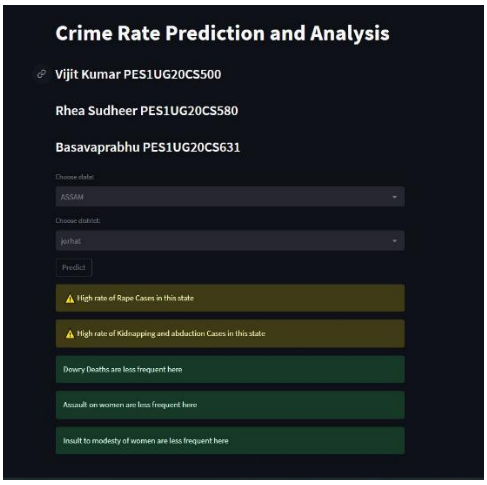
- the **mean**: $NRMSE = \frac{RMSE}{\bar{y}}$ (similar to the CV and applied in *INDperform*)
- the **difference between maximum and minimum**: $NRMSE = \frac{RMSE}{y_{max}-y_{min}}$,
- the **standard deviation**: $NRMSE = \frac{RMSE}{\sigma}$, or
- the **interquartile range**; $NRMSE = \frac{RMSE}{Q1-Q3}$, i.e. the difference between 25th and 75th percentile,

of observations.

Figure 9. Performance metrics used

VI. RESULTS

After trying various models for forecasting, random forests gave highest accuracy which we used to predict crime rates for future years based on the state and district name given by the user.



Crime Rate Prediction and Analysis

Vijit Kumar PES1UG20CS500

Rhea Sudheer PES1UG20CS580

Basavaprabhu PES1UG20CS631

Choose state:

ASSAM

Choose district:

jorhat

⚠ High rate of Rape Cases in this state

⚠ High rate of Kidnapping and abduction Cases in this state

Dowry Deaths are less frequent here

Assault on women are less frequent here

Insult to modesty of women are less frequent here

Figure 9. User interface to predict crime rate.

VII. CONCLUSIONS

As you can see through the visualisation, crime in India is increasing at an exponential rate and it's pertinent that apt measures are taken to curb this growth rate. However, in the meanwhile, ensuring the safety of women is the primary concern.

Through our solution, we hope to make the citizens more aware of the places they travel to and hope to make a meaningful contribution at prevention of crime.

ACKNOWLEDGEMENTS

We want to convey our sincere gratitude to Dr. Srinivas Katharguppe and the entire Data Analytics team for supporting us, giving us the chance to gain practical experience, and assisting us at every step of the road. We would also like to thank our class for giving us constructive feedback and suggestions to improve our project. This project has proven to be a great learning experience.

REFERENCES

P. A. C. Duijn, V. Kashirin, and P. M. A. Sloot, “The relative ineffectiveness of criminal network disruption,” Sci. Rep., vol. 4, 2014.

[2] S. Curtis-ham and D. Walton, “Mapping crime harm and priority locations in New Zealand : A comparison of spatial

analysis methods," Appl. Geogr., vol. 86, pp. 245–254, 2017. [3] O. K. Ha and M. A. Andresen, "Journal of Criminal Justice Unemployment and the specialization of criminal activity : A neighborhood analysis," vol. 48, pp. 1–8, 2017. [4] J. Phillips and K. C. Land, "The link between unemployment and crime rate fluctuations: An analysis at the county, state, and national levels," Soc. Sci. Res., vol. 41, no. 3, pp. 681–694, 2012. [5] J. J. Allen, C. A. Anderson, and B. J. Bushman, "The General Aggression Model," Curr. Opin. Psychol., vol. 19, pp. 75–80, 2018.

[6] M. Coccia, "A Theory of general causes of violent crime: Homicides, income inequality and deficiencies of the heat hypothesis and of the model of CLASH," Aggress. Violent Behav., vol. 37, no. November 2016, pp. 190–200, 2017.

[7] Y. Xu, C. Fu, E. Kennedy, S. Jiang, and S. Owusu-Agyemang, "The impact of street lights on spatial-temporal patterns of crime in Detroit, Michigan," Cities, no. October 2017, pp. 0–1, 2018.

[8] L. P. Beland and D. A. Brent, "Traffic and crime," J. Public Econ., vol. 160, no. March, pp. 96–116, 2018.

[9] S. J. Michel et al., "Investigating the relationship between weather and violence in Baltimore, Maryland, USA," Injury, 2016.

[10] S. A. Salleh, N. S. Mansor, Z. Yusoff, and R. A. Nasir, "The Crime Ecology: Ambient Temperature vs. Spatial Setting of Crime (Burglary)," Procedia - Soc. Behav. Sci., vol. 42, no. July 2010, pp. 212–222, 2012.

[11] J. Tiihonen, P. Halonen, L. Tiihonen, H. Kautiainen, M. Storvik, and J. Callaway, "The Association of Ambient Temperature and Violent Crime," Sci. Rep., vol. 7, no. 1, pp. 1–7, 2017. -layer long short-term memory (LSTM) model with intermediate variables for weather forecasting." *Procedia Computer Science* 135 (2018): 89-98.

[12] H.-W. Kang and H.-B. Kang, "Prediction of crime occurrence from multi-modal data using deep learning," PLoS One, vol. 12, no. 4, p. e0176244, 2017.

[13] L. G. A. Alves, H. V Ribeiro, and F. A. Rodrigues, "Crime prediction through urban metrics and statistical learning," Physica A, vol. 505, pp. 435–443, 2018.

[14] R. Liao, X. Wang, L. Li, and Z. Qin, "A novel serial crime prediction model based on Bayesian learning theory," 2010 Int. Conf. Mach. Learn. Cybern., no. July, pp. 1757–1762, 2010.

[15] A. Rummens, W. Hardyns, and L. Pauwels, "The use of predictive analysis in spatiotemporal crime forecasting : Building and testing a model in an urban context," Appl. Geogr., vol. 86, pp. 255–261, 2017.

[16] W. Gorr, A. Olligschlaeger, and Y. Thompson, "Short-term forecasting of crime," Int. J. Forecast., vol. 19, no. 4, pp. 579–594, 2003.

[17] P. Chen, H. Yuan, and X. Shu, "Forecasting crime using the ARIMA model," Proc. - 5th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2008, vol. 5, no. January 2017, pp. 627–630, 2008.

[18] J. D. López-cabrera and J. V Lorenzo-ginori, "Feature selection for the classification of traced neurons," J. Neurosci. Methods, vol. 303, pp. 41–54, 2018.

[19] M. Almaraashi, "Investigating the impact of feature selection on the prediction of solar radiation in different locations in Saudi Arabia ಜ," Appl. Soft Comput. J., vol. 66, pp. 250–263, 2018.

[20] N. D. Cilia, C. De Stefano, F. Fontanella, and A. Scotto, "A ranking-based feature selection approach for handwritten character recognition," Pattern Recognit. Lett., vol. 0, pp. 1–10, 2018.

[21] P. Sciencedirect, M. Masila, A. Jalil, F. Mohd, N. Maizura, and M. Noor, "ScienceDirect ScienceDirect A Comparative Study to Evaluate Filtering Methods for Crime Data Feature Selection A Comparative Study to Evaluate Filtering Methods for Crime Data Feature Selection," Procedia Comput. Sci., vol. 116, pp. 113–120, 2017