Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 01, January-2022

Leveraging ML to Predict Crime Against Women

Sonal Singh
Data Scientist
Fractal Analytics
Bangalore India

Abstract:- **In recent times, the crimes occurring against women is increasing at a rapid rate. It has become a major social issue not only in India, but across the world. Many attempts have been carried out with serious measures being taken in order to prevent such crimes. Every year, a massive amount of data is produced of different kinds of crimes being reported from different parts of the world. This knowledge can be very helpful in understanding and detecting violence, as well as assisting us in combating it to some degree. Analyzing such datasets can play a major role in identifying the crime patterns and occurrence of it. Here, data mining plays a huge role as it enables us to analyze, visualize and predict the different crimes which are occurring in a particular region. In this paper, we use Huber regression to analyze the data and visualize it using time series algorithm, based on different states of India and predicting out the particular crime occurring in a particular state.**

Keywords—Huber Regression, Time Series, Analyzing, Visualization

I. INTRODUCTION

In India, women are worshipped. They are given the status of goddesses. But the reality is completely opposite to it. With the passage of time, safety of women has started to become an alarming concern. With each passing day, the rate of crimes is increasing at a rapid rate. It is now considered to be a global issue, with many countries trying to bring in reforms, to curb down the crime rate. India too, is not far behind. With the data showed by the National Crime Records Bureau, the records of crime being reported against women has turned out to be higher than the past few years. Cases of rapes, murder, abduction and trafficking are occurring much frequent now. The government is trying to implement stricter laws and serious measures to prevent such crimes and ensure the safety of women. A huge amount of data is generated every year related to such different crimes in different regions of India. Analyzing such huge data records may seem a very tedious task. With the emerging technologies and methodologies, Data mining plays a huge role in analyzing of such large number of records with displaying accurate results and discovering out the patterns. The outcome of such analysis can’t be exactly equal to the perceived outcome; but it gives out an adequate rough figure of crimes which will be occurring in a particular state in the coming years. The main challenge in this prediction is to reduce out the losses and bring out the resultant number of a particular crime, say rape in a particular state like Andhra Pradesh in the coming years to the actual figure. The challenges that we are facing are:

- Analyzing data of different types of crimes based on each 28 states and 7 union territories.
- Reducing the loss to the minimum for each type of crime.
- Obtaining more datasets with more sets of crimes from various crime departments.

An ideal crime analysis tool should be able to identify crime patterns quickly and in an efficient manner for future crime pattern detection and action [efficient approach]. This will enable the law officials and the governments of various statesto enforce serious measures to reduce such crimes and to provide a safe place for women to live in. Previously, many data mining techniques like clustering have been used to predict a certain crime based on a single region with producing more accurate results. The present work proposed is to use Huber regression to analyze different types of crimes based on different states and territories and to predict out the number of that particular crime which will occur in that particular state/union territory and to visualize the data in a graphical form using time series to identify the number of particular crimes which will occur in that region. Also, this will indicate which type of crime is dominant in a certain region.

II. STATE OF THE ART (LITERATURE SURVEY)

Researchers compared many data mining algorithms using a variety of real-world applications with some related works. Apart from those one of them is Crime against Women (CAW) Analysis and Prediction in Tamil Nadu Police Using Data Mining Techniques where they used Clustering in WEKA utensils, Euclidean distance calculation by S. Lavanyaa, D. Akila [1] which states that Clustering in WEKA utensils, Euclidean distance calculation gives improved exactness in the metropolitan urban areas violations rate to decrease and predict, according to the investigation. Analysis and Prediction of Crimes using Clustering, Classification and General algorithm by Rasoul Kiani, Siamak Mahdavi, Amin Keshavarz [2] where the main objective of occurrence frequency during different years. We used a theoretical model focused on data mining techniques including clustering and classification to analyse a real crime dataset collected by the police in England and Wales between 1990 and 2011. In this certain kind of weights are assigned so as to refine the model which is being used and eliminating the lesser values. Using the RapidMiner tool, the Genetic Algorithm (GA) is used to optimise the Outlier Detection operator parameters. Crime Analysis using K-Means Clustering on crime dataset using rapid miner tool by Jyoti Agarwal, Renuka Nagpal, Rajni Sehgal [3] where this project is based on the concept of crime analysis by using clustering algorithm on the obtained data set using a rapid miner tool. The analysis is done by taking in a particular type of crime that is homicide and presenting it in a graphical form, matching it with to the particular year it happened. The outcome derived was, that the rate of homicide is decreasing from year 1990 to 2011. Crimes Against Women in India: Analysis of Trends Using Regression and Visualization by R. Devakunchari, Bhowmick S, Bhutada S P, Shishodia Y [4] where Detection technologies improves identifying of the incidents and make use of public safety equipment as soon as possible. With faster identification of incidents, this helps to improve the response time and thus in this the accuracy and reliability of incident response and reporting, as well as the distribution of investigative resources, can all be improved with technology. This can also help in boosting up the clearing rates.

III. PROPOSED WORK

The data which used for doing the analysis plays a key role in finding out the patterns, especially in crime analysis. In this, the dataset is obtained from Kaggle which contains different types of crimes which are occurring against women such as ‘rape’,’ kidnapping’,’ dowry death’,’ assault on women’,’ cruelty against women’,’ importation of young girls’, ‘insult to modesty’ and ‘immoral traffic’. This time series-data is converted into supervised data, in order for predicting out the number of crimes that may take place in future. For analyzing these different crimes, Huber Regression is used. It determines the loss score, that is how much the difference is going to be between the actual and predicted value. It may be not hat accurate as the prediction is done based on the number of crimes that happened in the past years. There is no constant increase/ decrease in the crimes taking place. Based on that, the predicted values are brought closer to the actual value in order to reduce the loss. The predictions can be visualized in a bar-graph form, displaying the top states/union territories where the particular listed crimes are at the peak. It will also enable us to identify, which type of crime is dominant in a particular state.

IV. IMPLEMENTATION

A. *Methods Used*

Huber Regression

The Huber Regressor optimizes the squared loss for samples where |(y - X'w) / sigma| is less than epsilon, and the absolute loss for samples where |(y - X'w) / sigma| is greater than epsilon, where w and sigma are the parameters to be optimized. The sigma parameter ensures that if y is scaled up or down by a certain factor, epsilon does not need to be rescaled to maintain the same level of robustness.

```
sklearn.linear_model.HuberRegressor
```

Here in our project, we used Huber regressor as follows:

```
model rape = linear_model. HuberRegressor().fit(X_train_
rape,y_train_rape)
```

Time Series

Time series is a chain of data points which is mostly in chronological sequence, with gathering in regular intervals. This analysis can be applied to any number of variables that changes over time, generally data points that are closer together are more similar in nature than those further apart. Here we used time series algorithm to predict future number of crimes in each state that will happen. We can show the graph of rape in Andhra Pradesh as predicted in fig 1.

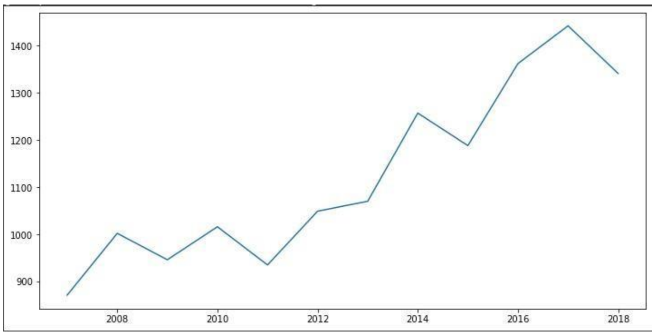
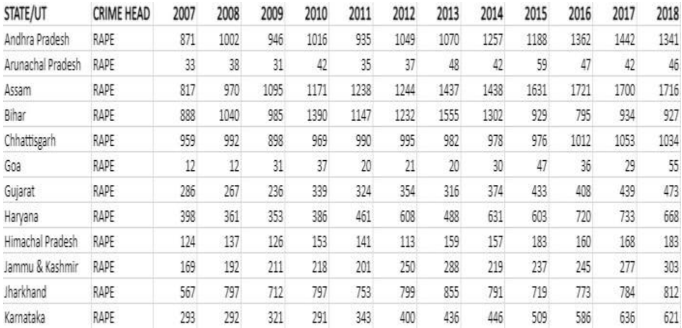


Fig. 1. Graph of Andhra Pradesh crime

A. *Dataset*

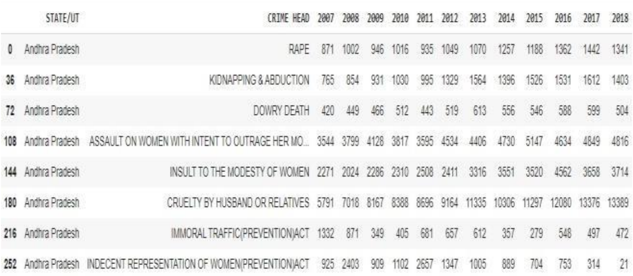
Here in dataset in fig 2, we have 8 different crimes such as rape, kidnapping and abduction, dowry death, insult to modesty of women, cruelty by husband or relatives, assault on women with intent to outrage her modesty, immoral traffic, indecent representation of women. These are depicted from 2007 to 2018 and we will be predicting crime number for the year 2019.



STATE/UT	CRIME HEAD	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Andhra Pradesh	RAPE	871	1002	946	1016	935	1049	1070	1257	1188	1362	1442	1341
Arunachal Pradesh	RAPE	33	38	31	42	35	37	48	42	59	47	42	46
Assam	RAPE	817	970	1095	1171	1238	1244	1437	1438	1631	1721	1700	1716
Bihar	RAPE	888	1040	985	1390	1147	1232	1555	1302	929	795	934	927
Chhattisgarh	RAPE	959	992	898	969	990	995	982	978	976	1012	1053	1034
Goa	RAPE	12	12	31	37	20	21	20	30	47	36	29	55
Gujarat	RAPE	286	267	236	339	324	354	316	374	433	408	439	473
Haryana	RAPE	398	361	353	386	461	608	488	631	603	720	733	668
Himachal Pradesh	RAPE	124	137	126	153	141	113	159	157	183	160	168	183
Jammu & Kashmir	RAPE	169	192	211	218	201	250	288	219	237	245	277	303
Jharkhand	RAPE	567	797	712	797	753	799	855	791	719	773	784	812
Karnataka	RAPE	293	292	321	291	343	400	436	446	509	586	636	621

Fig. 2. Crime Dataset

For example, for a particular state like Andhra Pradesh, the below dataset showcases all the different crimes in that particular region.



	STATE/UT	CRIME HEAD	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
0	Andhra Pradesh	RAPE	871	1002	946	1016	935	1049	1070	1257	1188	1362	1442	1341
36	Andhra Pradesh	KIDNAPPING & ABDUCTION	765	854	931	1030	995	1329	1564	1396	1526	1531	1612	1403
72	Andhra Pradesh	DOWRY DEATH	420	449	466	512	443	519	613	556	546	588	599	504
108	Andhra Pradesh	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MO...	3544	3799	4128	3817	3595	4534	4406	4730	5147	4634	4849	4816
144	Andhra Pradesh	INSULT TO THE MODESTY OF WOMEN	2271	2024	2286	2310	2508	2411	3316	3551	3520	4562	3658	3714
180	Andhra Pradesh	CRUELTY BY HUSBAND OR RELATIVES	5791	7018	8167	8388	8696	9164	11335	10306	11297	12080	13376	13389
216	Andhra Pradesh	IMMORAL TRAFFIC(PREVENTION)ACT	1332	871	349	405	681	657	612	357	279	548	497	472
252	Andhra Pradesh	INDECENT REPRESENTATION OF WOMEN(PREVENTION)ACT	925	2403	909	1102	2657	1347	1005	889	704	753	314	21

Fig. 3. Andhra Pradesh dataset

B. Converting time series data into context-aware supervised learning

We'll use the Pandas library to import DataFrame and concat, and then convert using the function below.

```
def series_to_supervised (data, n_in=1, n_out=1, drop
nan=True)
```

where we define series to supervised (), a new Python function that converts a multivariate time series into a supervised learning dataset.

The role takes four arguments in this case:

- **data**: This is a set of observations in the form of a list or a 2D NumPy array.
- **n_in**: This is a collection of lag observations that is used as an input (X). Values will range from [1 to len(data)] and are completely optional. The default value can be set to 1
- **n_out**: This is a set of observations that is used as an output (y). Values will range from [0 to len(data)-1] and are completely optional. The default value can be set to 1.
- **dropnan**: This is a boolean form that indicates whether or not rows with NaN values should be dropped. It's also optional, and if it's present, it'll be set to True.

For getting scores and prediction we have divided the dataset into groups or context of years by specifying it into variables such as var1 for rape, var2 for kidnapping and abduction etc. We have done this for predicting number for the year 2019 by specifying var(t-2) as 2007 to 2016, var(t-1) as 2008 to 2017 and var(t) as 2018 and var(t+1) as 2019.

```
state_data = series_to_supervised (values,
2,2)state_rape =
state_data[['var1(t-2)','var1(t- 1)','var1(t)','var1(t+1)']]
```

Then we have done Train test split where for training it is 40% and testing is 60% and have done model fitting using linear model Huber regressor then from that we will get out scores and prediction.

C. Iterating to obtain scores and predictions for every state

Here we are predicting the scores and predict the rape, assault individually using this function and get our result where we will get scores in positive or negative value which is compared with the previous year and the predicted year if the crime number is more than the score goes positive and if the crime is less in number then the score goes negative. And also, we can find number and state of a particular crime and max number of a particular crime and in which state through which police department can take necessary steps to decrease that type of crime in their state.

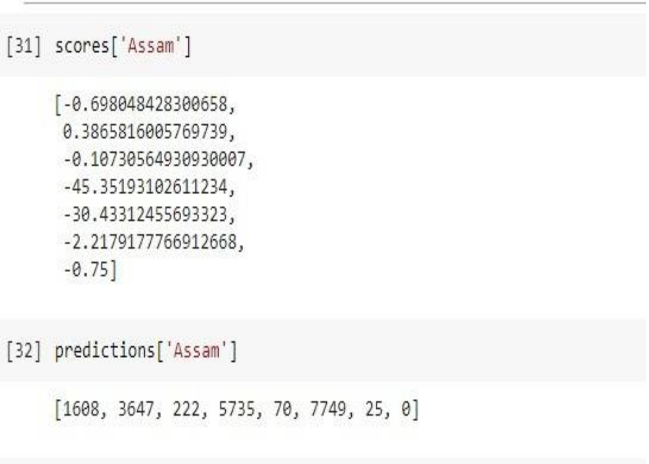
Such as for Maharashtra:

```
s, p = get_results_for_state('Maharashtra')
```

where s is score and p is prediction.

It can be also specified as and can be shown in fig 4 and fig 5,

```
scores[i],predictions[i] = get_results_for_state(i)
number, state = get_final_results('cruelty')
```



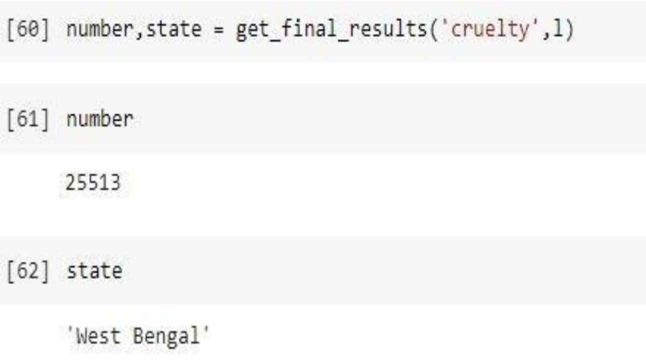
```
[31] scores['Assam']

     [-0.698048428300658,
      0.3865816005769739,
      -0.10730564930930007,
      -45.35193102611234,
      -30.43312455693323,
      -2.2179177766912668,
      -0.75]

[32] predictions['Assam']

     [1608, 3647, 222, 5735, 70, 7749, 25, 0]
```

Fig. 4. Score and Prediction of state wise



```
[60] number,state = get_final_results('cruelty',1)

[61] number

      25513

[62] state

      'West Bengal'
```

Fig. 5. Max cruelty number and state

D. Visualization

Here we get the prediction of states with most rapes, assault, dowry etc., as shown in following below figures 6,7,8.

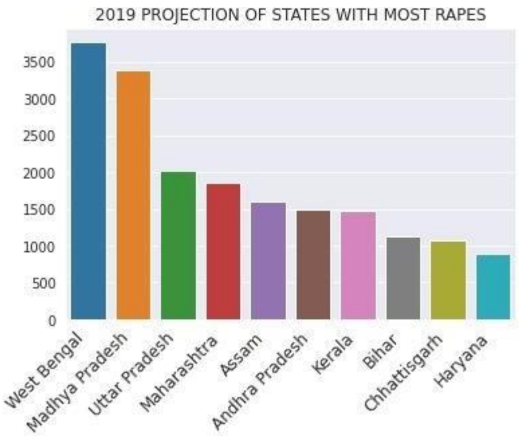


Fig.6. Graph with most rapes

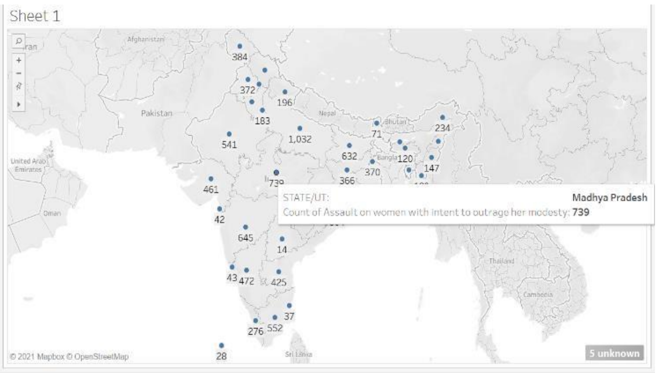


Fig. 7. Map of number of assaults in different states

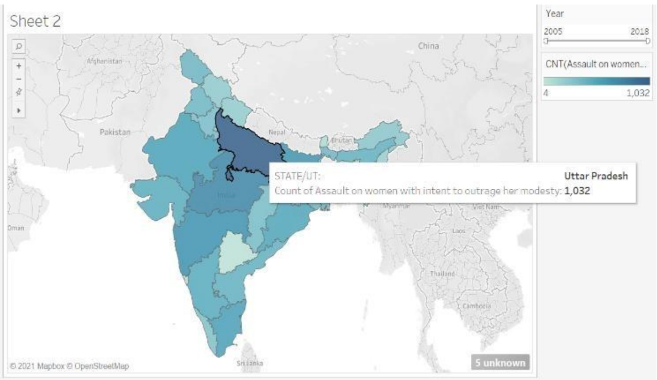


Fig. 8. Map with color to depict highest number of assaults in a state

V. CONCLUSION

Using Huber regression for analyzing different type of crimes for each state the predicted outcome is accurately close to the actual number of crimes which may occur in the frequent years. The predicted outcome is determined by the number of crimes taken from the previous years. Time series enabled the data to be visualized in a graphical form forecasting the increase/decrease of the crimes which may occur in a particular region in India. With the outcome of this whole program, the results help us to identify what type of crime is of dominance in a particular state and rough figures of different crimes in different states and union territories of India. This will help the law enforcement officials and government to implement stricter laws in order to curb down the crime rate and make a safer place for women to live in.

REFERENCES

[1] S. Lavanyaa, D. Akila Crime against Women (CAW) Analysis and Prediction in Tamilnadu Police Using Data Mining Techniques International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5C, February 2019

[2] Tahani Almanie, Rosha Mirza and Elizabeth Lor Crime Prediction Based on Criminal Types And Using Spatial and Temporal Criminal Hotspots International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.4, July 2015

[3] Mehmet Sati Vural, Mustafa Goku Criminal prediction using Naive Bayes theory September 2017

[4] Shiju Sathya Devan, Devan M. S, Surya S Gangadhara Crime Analysis and Prediction Using Data Mining (ICNSC), 2014

[5] Yamuna.S, Sudha Bhubaneswar D, Data mining Techniques toAnalyze and Predict Crimes, International Journal of Engineering and Science (IJES) Vol-1, Issue-2, PP 243-247

[6] Sylvia Welby, Improving the statistics on violence against women, Statistical Journal of the united nation ECE 22(2005)193-216.

[7] Bewley, J. Friend and G. Mosey, Eds, Violence against Women, London: Royal College of Abstrictions and Gynecologists', 1997.

[8] Malathi.A and Dir.'s. Santhosh Baboo. Article: an enhanced algorithm to predict a future crime using data mining. International Journal of Computer Applications,21(1):1-6, May2011.Published by foundation of Computer Science.

[9] Azbuka and Gifford, 'Fuzzy association rule mining for community crime pattern discovery', in ACM SIGKDD workshop on intelligence and security Informatics, Washington, D.C., 2010, PP.1-10.

[10] Web Crime Data using Data Mining, International Journal of Engineering and Innovative Technology (Ajeet)2(3)

[11] Devendra Kumar Tayal et al., Crime detection and criminal identification in India using data mining techniques, AI & Soc (2015) 30,pp.117-127.

[12] Roslin V. Husk and Hsinchu Chen., Colin: A Case of Intelligent Analysis and Knowledge Management, Proceedings of International Conference on Information Systems, 1999, pp.15-28.

[13] Prajakta Yarded, Vaishnavi Guldur Predictive Modelling of Crime Dataset using Data Mining (IJDKP) Vol.7, No.4, July 2017.

[14] Deepika K.K, Smitha Vinod Crime analysis in India using data mining techniques 2018.

[15] Rasoul kiang, Siamak Mahdavi, Amin Keshavarz, Analysis and Prediction of Crimes by clustering and Classification. (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol-4, No:8,2015

[16] Amaranthine, L.C. (2003) Technological Advancement: Implications for the Crime, the Indian Police Journal, AprilJune.

[17] Mayank Motwani, Paratha Pawar, Rachit Mathur, Aatif J Mashed An Efficient Approach towards crime gainst women using time series algorithm International Journal of Computer Applications (0975 – 8887) Volume 179 – No.34, April 2018.

[18] Dr.D. Akila, Ms. Vidya and Mrs. Rajesh, "Optimization Based Information Retrieval with the Enhancement of Annotator in WordNet Application", Journal of Advanced Research in Dynamical & Control systems Volume 10, Issue 2, pp. 318- 323,2018.

[19] D.Akila, S.Sathya, G.Suseendran, "Survey on Query Expansion Techniques in Word Net Application", Journal of Advanced Research in Dynamical and Control Systems, Vol.10(4), pp.119- 124, 2018.