Insights from Machine Learning Models on cases of Violence against Women in India

Sonali Mishra

09/05/2022

Abstract

The paper examines gender-based violence against women scenario in India across states, union territories and districts. We look at various factors such as presence of law enforcement specifically geared towards safety of women, socio-economic positioning of females in the society in terms of decision making, education and employment and current crime rate against women prevailing in the country. Perform clustering to understand similarities and resemblances between states and finally develop a predictive model to forecast crime rate.

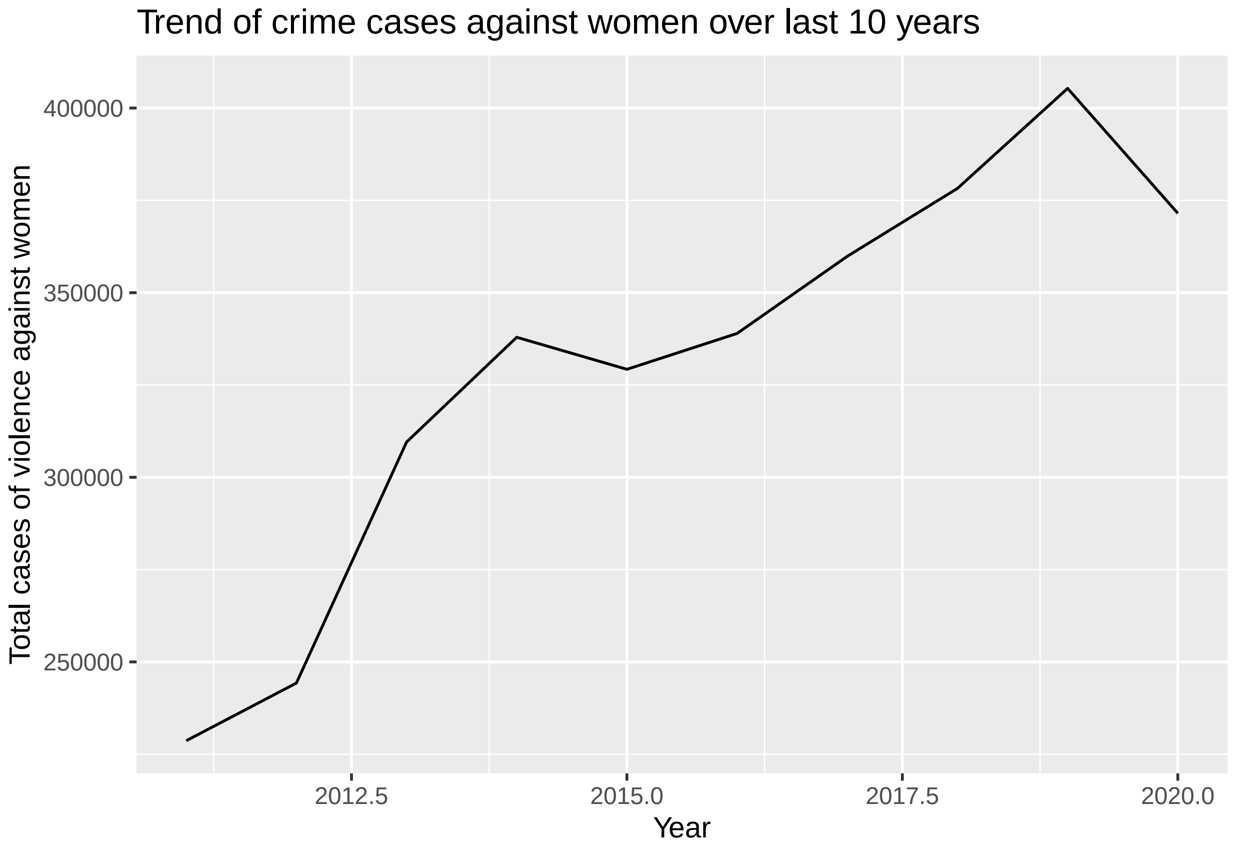
Overview of Crime against women in India

India ranks as one of the most unsafe countries for in terms of women safety. According to NCRB (National Crime Records Bureau) crime is committed against women every 3 minutes. Crime is committed in different forms such as domestic violence, rape, dowry, modesty defamation and more. Alarmingly 65% of the men believe that women "deserve to be beaten up". Recipients of the crime range from infants to elderly.

The literature on violence targeted at women highlights some key factors. Male dominance coupled with female submission is known to aggravate the problem. Multiple studies suggest that women feel more powerless than men given the same circumstances. Imposition by men who think they are entitled to supremacy makes the situation worse. Moreover, law enforcement institutions by executive (Police) and judiciary have not been efficient enough either. Huge number of women abuse cases are being reported as false, large number of cases are pending at court, the laws are not well defined, juvenile criminals are being excused and many more such gaps.

10 years ago, the famous "Nirbhaya" case led to an uproar across the nation leading to country-wide protest and condemnation from international organizations. This heat continued for next 5 years and in response to the same current Modi-led government began a one stop centre scheme wherein institutions are to be setup across the country to address all relevant challenges women face while tacking abuse and violence. The centre aims to be equipped to support with filing complaints, counselling, emergency services, medical assistance, legal aid, shelter, helpline and video conferencing facilities etc. My intention here is to propose a more machine learning driven approach to predict crime rate and in extension an indication of number of OSCs (One stop centre) needed at state level. The reason we compare at state level is because sanctions and funds are released at state level.

The below plot depicts a view of how crime rate has changed across states and union territories (UT) over the last 10 years.



Methodology and Data Gathering

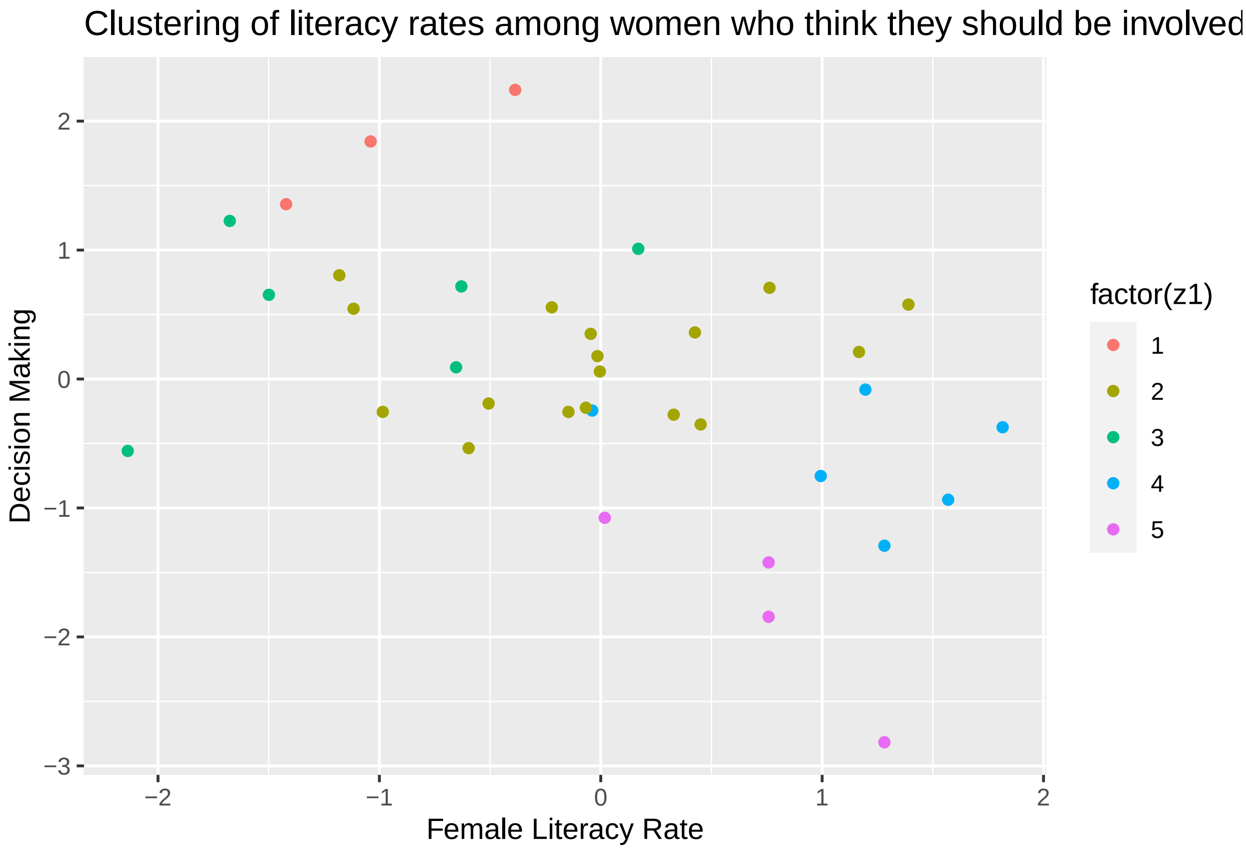
Data for this paper has been collected from multiple sources and is a combination of excel downloads and telephonic inputs. I first use agglomerative hierarchical clustering on two datasets – law enforcement and position of females. Law enforcement dataset consists of performance metrics at state level such as conviction rate and pendency rate. Position of females' dataset consists of interesting information. We look at whether female have tools to be independent in terms of employment, education and how are they involved in decision making of their day-to-day lives.

I am going to expand a little more on decision making attributes data. This is pulled from National Health Family Survey of India wherein women were asked multiple questions regarding their autonomous money expenditure, control over financial planning, participation in key household decisions, whether they jointly take decisions with their partners etc. This is to gauge how powerful do they feel when it comes to their own household. For example, in India 71% of women say that they take major decisions jointly with men but 56% men think that women should be allowed to make decisions jointly. The survey also covers a statement on earning parity. India only 40% of women earn nearly as much as their partners. Detailed description of variables used is given in the appendix along with data sources.

Given the smaller size of the dataset I observed better results from hierarchical clustering over kmeans. Post this I perform PCA to summarise 37 categories of crime against women. The idea is to compress these categories into a smaller number so I can further build a predictive model using these variables (and some more). Finally in the end I use random forest (over gradient boosting as rf provides lower out of sample error) to predict crime rate at district level. I then compare this to existing number of functional OSCs and analyse where we may need to augment efforts. The reason I don't directly predict the number of OSCs is because currently the structure of OSC is unknown to me – i.e number of employees dedicated to each wing of the centre, target audience magnitude, hiring plan etc.

Results reconciliation

When I perform hierarchical clustering on factors that affect empowering of women (appendix for more details) the output is quite satisfactory. For visualisation I have plotted the percentage of women of think they should be involved in major decision making as their husbands against the literacy rate of women. We definitely observe some form of grouping



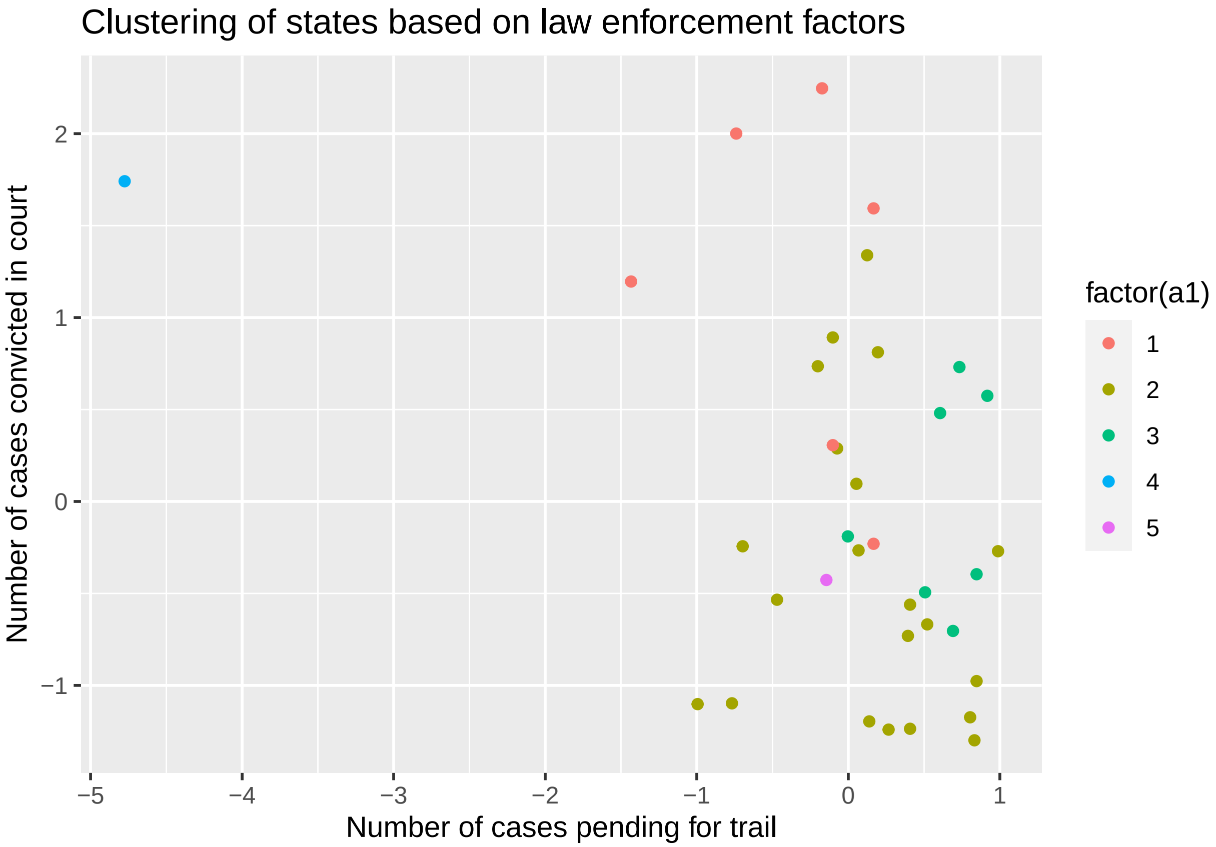
Below table represents state and cluster mapping

The plot shows that states in 4 and 5 cluster have high literacy rates but are less empowered in decision making and cluster 1 states have lower literacy but higher authority. Glancing at the table I can see cluster 4 states belong to the category with high crime rates and these expectedly appear in lower empowerment cluster. This is not the most balanced dendogram but still quite better than kmeans. Cluster 1 states belong to lower crime occurrence category. So, there is definitely some sort of grouping that is reflective of crime cases.

Now we do a similar analysis for law enforcement factors

Table 1: Clustering of states for women seeking equal role in decision making

Cluster	State/UT
1	Chandigarh
1	Mizoram
1	Nagaland
2	Delhi
2	Haryana
2	Punjab
2	Uttarakhand
2	Chhattisgarh
2	Jharkhand
2	Odisha
2	West Bengal
2	Arunachal pradesh
2	Assam
2	Manipur
2	Tripura
2	Gujarat
2	Maharashtra
2	Andaman & Nicobar Islands
2	Puducherry
2	Tamil Nadu
3	Himachal Pradesh
3	Meghalaya
3	Sikkim
3	Goa
3	Kerala
3	Lakshadweep
4	Jammu & Kashmir
4	Rajasthan
4	Madhya Pradesh
4	Uttar Pradesh
4	Bihar
4	Dadra & Nagar Damand & Diu
5	Ladakh
5	Andhra Pradesh
5	Karnataka
5	Telangana



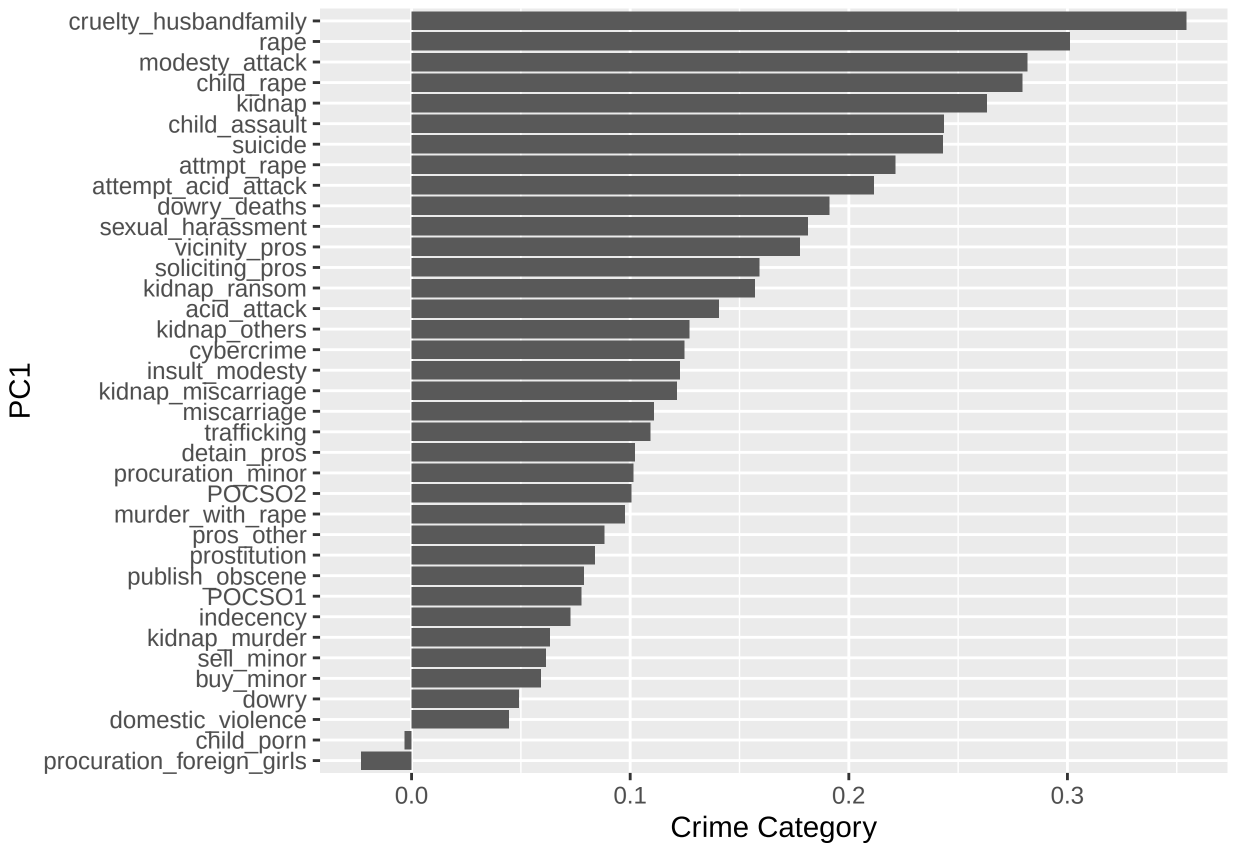
Clustering here was not very useful as most of the states were put in cluster 2. This is mainly because of outliers. While the outliers are not very intense but since the sample size is small their deviation is amplified. To improve this analysis it would be wise to move to a bigger sample size which means this information will have to be collected on district level which at the moment is not publicly available.

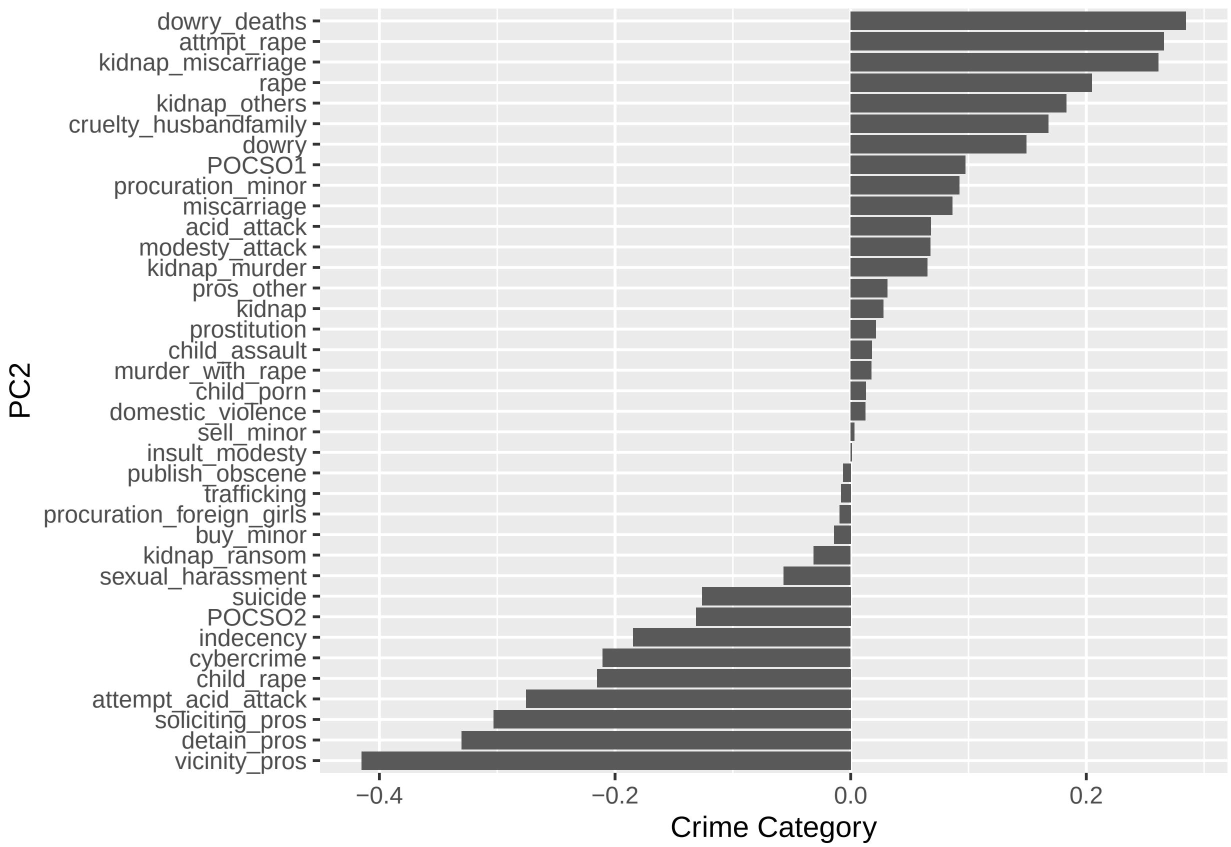
Next, we move on to the 3rd dataset which lists 36 categories of crime against women. Here the data is collected at district level. I bring this down to 18 categories of crime that explain 75% of the variation in the data. Detailed grid of 12 PCA components is attached in the Appendix.

```
## Importance of first k=18 (out of 37) components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1973 1.63373 1.45254 1.43995 1.34688 1.24205 1.18795
## Proportion of Variance 0.1305 0.07214 0.05702 0.05604 0.04903 0.04169 0.03814
## Cumulative Proportion  0.1305 0.20263 0.25965 0.31569 0.36472 0.40641 0.44455
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.10904 1.07777 1.07598 1.04921 1.03639 1.02846 1.01166
## Proportion of Variance 0.03324 0.03139 0.03129 0.02975 0.02903 0.02859 0.02766
## Cumulative Proportion  0.47779 0.50919 0.54048 0.57023 0.59926 0.62785 0.65551
##                           PC15   PC16    PC17    PC18
## Standard deviation     0.99900 0.9676 0.95147 0.94890
## Proportion of Variance 0.02697 0.0253 0.02447 0.02434
## Cumulative Proportion  0.68248 0.7078 0.73226 0.75659
```

Table 2: Cluster grouping of states according to law enforcement factors

Cluster	State/UT
1	Andaman & Nicobar islands
1	Lakshadweep
1	Mizoram
1	Nagaland
1	Puducherry
1	Sikkim
2	Andhra Pradesh
2	Arunachal Pradesh
2	Assam
2	Chhattisgarh
2	D&N Haveli and Daman & Diu
2	Goa
2	Gujarat
2	Haryana
2	Jammu & Kashmir
2	Jharkhand
2	Karnataka
2	Kerala
2	Madhya pradesh
2	Meghalaya
2	Odisha
2	Punjab
2	Rajasthan
2	Telangana
2	Tripura
2	Uttar Pradesh
2	West Bengal
3	Bihar
3	Chandigarh
3	Delhi
3	Himachal Pradesh
3	Maharashtra
3	Manipur
3	Uttarakhand
4	Ladakh
5	Tamil nadu

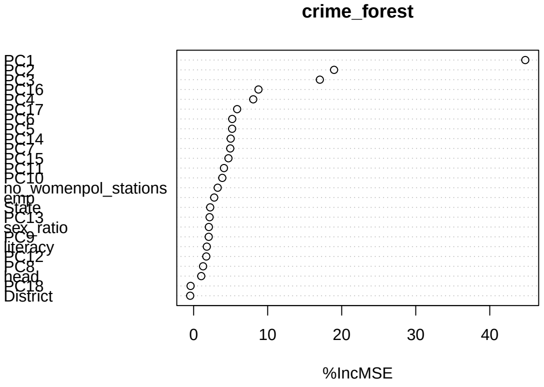

cruelty_husbandfamily
rape
modesty_attack
child_rape
kidnap
child_assault
suicide
attmpt_rape
attempt_acid_attack
dowry_deaths
sexual_harassment
vicinity_pros
soliciting_pros
kidnap_ransom
acid_attack
kidnap_others
cybercrime
insult_modesty
kidnap_miscarriage
miscarriage
trafficking
detain_pros
procuration_minor
POCSO2
murder_with_rape
pros_other
prostitution
publish_obscene
POCSO1
indecency
kidnap_murder
sell_minor
buy_minor
dowry
domestic_violence
child_porn
procuration_foreign_girls
PC1
0.0
0.1
0.2
0.3
Crime Category

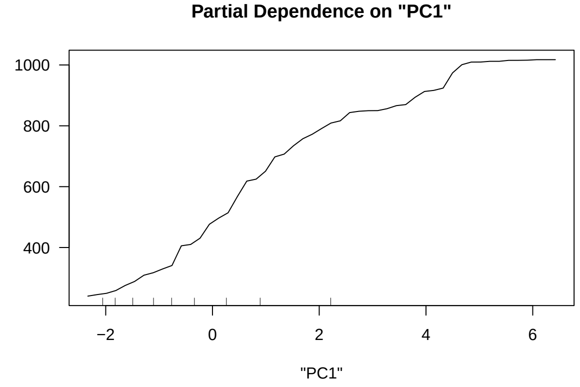


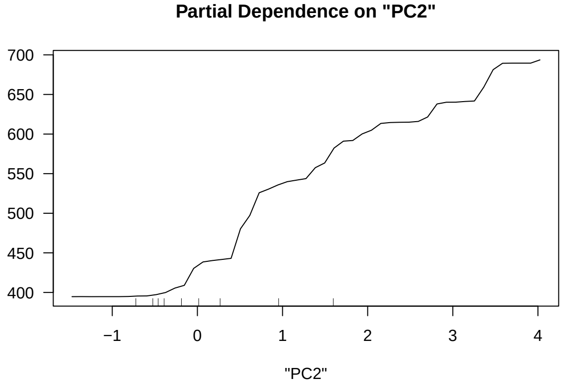
The first PCA component seems to be more or less an average (except for 2 categories which are lower in number). Later on when we perform random forest we will notice that this is the most important variable in the predictive modelling. The second PCA component is interesting in that it has contrasted the crime categories that are more prevalent versus less frequent. For example, rape, cruelty by family members of husband, dowry, coerced prostitution, child assault, acid attack etc occur (or at least reported) than kidnapping for ransom, insulting modesty etc. This is a good indication to policymakers that which area of crime and justice requires foremost focus. India is a huge country with million problems and therefore funnelling approach becomes very crucial. These stats can prove useful in those cases.

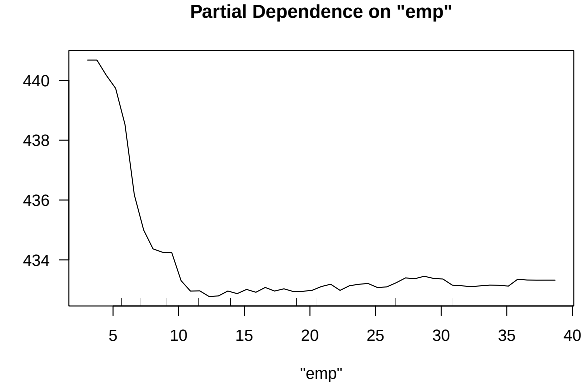
Now we predict number of cases using random forest. I used gradient boosting method as well but random forest gave a lower RMSE (116) when compared to boosting (160). This prediction is done on the testing set of the crime data. Also, this set is not comprehensive of all states. Since the district level information was collected by visiting individual websites and calling up police stations, I could not complete it for all 1500 districts and instead focused on districts with high number of cases.

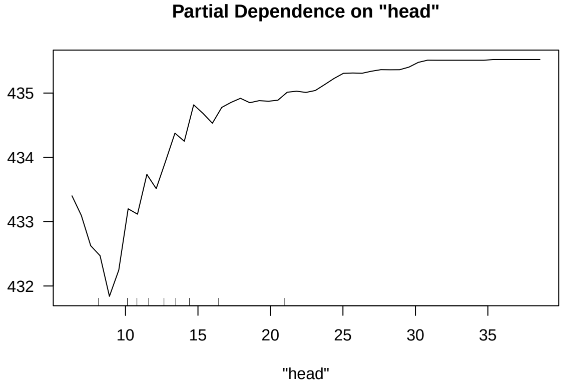
Looking at the variable importance graph, as expected PC1 is most important when it comes to modelling. Surprisingly number of women police stations is least important. Now this can mean either of two things: a) in fact women police stations don't make women safer which is highly unlikely or b) there aren't enough women police stations built to analyse the data which makes much more sense here. Most places don't have women police station or have just 1 as a result of which there is not much variation in this data for the algorithm to learn from. This is an interesting insight into how to interpret the variable importance graph.











Partial dependence plots of PC1 and PC2 are quite reasonable as they indicate crime rate, increasing those factors would increase likelihood of crime. Head graph depicts how crime rate changes as proportion of females heading household increases which is conceivable as heading household is empowering. What is surprising is the how crime rate changes with female employment. It reduces and then is pretty much ineffective. This goes to say that being employed is not sufficient enough to evade violence or feel empowered to evade crime. This is what various research literature also suggest.

Conclusion

Looking at the comparison between predicted crime rate and actual functional OSCs tells gross mismatch. Most centres need more centres. West Bengal has some districts topping the list in crime rate consistently and there is no functional centre yet. Whereas Arunachal Pradesh has excess of centres. Delhi and Rajasthan which have been infamous for rape cases are expected to handle over 100 cases. This explains that machine

Table 3: Number of cases to be handled by each OSC as per predictions

State	tot_pred_crime	OSCs	cases_per_esc
Andhra Pradesh	681.33	13	52
Arunachal Pradesh	98.75	24	4
Assam	9207.19	33	279
Bihar	1493.18	38	39
Chhattisgarh	1655.40	27	61
Delhi	605.56	11	55
Haryana	3758.66	22	171
Himachal Pradesh	371.79	12	31
Jammu & Kashmir	194.83	20	10
Jharkhand	887.11	24	37
Karnataka	555.58	30	19
Kerala	281.10	14	20
Madhya Pradesh	3129.87	52	60
Maharashtra	6163.96	37	167
Manipur	185.26	16	12
Mizoram	25.98	8	3
Nagaland	25.11	11	2
Odisha	5903.90	30	197
Punjab	812.35	22	37
Rajasthan	5640.77	33	171
Tamil Nadu	875.34	34	26
Uttar Pradesh	7404.43	75	99
Uttarakhand	32.09	13	2
West Bengal	1779.99	0	Inf

learning models are definitely superior to mere weighted average approach. The implications are huge here because based on their calculation, funds and contracts are sanctioned. Currently one of the major blockers this scheme has been insufficient funds. Therefore, judicious spending warrants a more sophisticated methodology of arriving at the number of centres.

A future development of this project would be to somehow factor in victim blaming sentiment in the realm of violence. India is plagued by victim targeting as is suggested by multiple surveys however there is no official public data available on this. Such factors influence whether women/girls are willing to report these cases. Even today thousands of females don't report sexual abuse because of the fear of it being backfired.

Clustering has some very useful insights with respect to policy. Discerning similarities between states can help exchange policies that have worked for one them into the others. Implementing policy without having to reinvent the wheel cannot be done naively but certainly is a low-hanging fruit.

A further improvement would be to gather data at district level in order to obtain bigger samples. Personally, I feel the samples size for clustering were not big enough for the algorithms to learn. While I had a large sample for random forest but I think one could benefit from rich data of Indian demography.

I still have no explanation for why factors such as female employment, female heads, literacy rate have lower importance in the forest modelling but that is not what I am trying to do here. Random forest gave the best and least out of sample error and therefore the its predictive powers are promising.

Appendix

Description of variables:

I1 = Alone or jointly with their husband decide how their own earnings are used I2 = Alone or jointly with their husband decide how their husband's earnings are used I3 = Earn more or about the same as their husband I4 = Percentage of women who usually make decisions alone or jointly with their husband I5 = Percentage of men who say that a wife should have an equal or greater say alone or jointly with her husband in Emp = female employment with respect to female population Head = Percentage of households headed by women Literacy = Percentage of women aged seven and above who can read and write in any language with respect to female population Pol_chrgsheet = Percentage of cases charge sheeted by police Pol_pending = Percentage of cases pending investigation by the police Pol_convic = Percentage of convictions by police Court_pending = Percentage of cases pending trial in court Court_convic = Percentage of cases convicted by the court Pol_perlakh = Number of police officers per 100000 population Pol_100sqkm = Number of police officers per 100 km sq Perc_womenpol = Percentage of women police officers No_womenpol_stations = Number of women police stations in the area Sex_ratio = Proportion of women to men Tot_crime = Total number of crime against women Murder_with_rape = Sum of Murder with Rape/Gang Rape Dowry_deaths = Sum of Dowry Deaths (Sec. 304B IPC) Suicide = Sum of Abetment to Suicide of Women (Sec. 305/306 IPC) Miscarriage = Sum of Miscarriage (Sec. 313 & 314 IPC) Acid_attack = Sum of Acid Attack (Sec. 326A IPC) Attempt_acid_attack = Sum of Attempt to Acid Attack (Sec. 326B IPC) Cruelty_husbandfamily = Sum of Cruelty by Husband or his relatives (Sec. 498 A IPC) Kidnap = Sum of Kidnapping & Abduction (Sec. 336 IPC) Kidnap_murder = Sum of Kidnapping & Abduction in order to Murder (Sec. 364 IPC) Kidnap_ransom = Sum of Kidnapping for Ransom (Sec. 364A IPC) Kidnap_miscarriage = Sum of Kidnapping & Abduction of Women to compel her for marriage (Sec. 366 IPC) Procuration_minor = Sum of Procuration of Minor Girls (Sec. 366A IPC) Procuration_foreign_girls = Sum of Importation of Girls from Foreign Country (Sec. 366B IPC) Kidnap_others = Sum of Kidnapping and Abduction of Women – Others (Secs.363A, 365, 367, 368, 369 IPC) Trafficking = Sum of Human Trafficking (Sec. 370 & 370A IPC)
Sell_minor = Sum of Selling of Minor Girls (Sec. 372 IPC) Buy_minor = Sum of Buying of Minor Girls (Sec. 373 IPC) Rape = Sum of Rape Attmpt_rape = Sum of Attempt to Rape Modesty_attack = Sum of Assault on Women with Intent to Outrage her Modesty Insult_modesty = Sum of Insult to the Modesty of Women Dowry = Sum of Dowry Prohibition Act, 1961 Prostitution = Sum of Procuring, inducing Children for the sake of prostitution (Section 5) Detain_pros = Sum of Detaining a person in premises where prostitution is carried on (Section 6) Vicinity_pros = Sum of Prostitution in or in the vicinity of public places (Section 7) Soliciting_pros = Sum of Seducing or soliciting for purpose of prostitution (Section 8) Pros_others = Sum of Other Sections under ITP Act Domestic_violence = Sum of Protection of Women from Domestic Violence Act Publish_obscene = Sum of Publishing or Transmitting of Sexually Explicit Material (Sec. 67A/67B (Girls) IT Act) Cybercrime = Sum of Other Women Centric Cyber Crimes (Ex. Blackmailing/ Defamation/Morphing/ Fake Profile) Child_rape = Sum of Child Rape (Sec. 4 & 6 of POCSO Act) / Sec. 376 IPC) Child_assault = Sum of Sexual Assault of Children (Sec. 8 & 10 of POCSO Act) / Sec. 354 IPC) Sexual_harassment = Sum of Sexual Harassment (Sec. 12 of POCSO Act) / Sec. 509 IPC) Child_porn = Sum of Use of Child for Pornography/Storing Child Pornography Material (Sec. 14 & 15 of POCSO Act) POCSO1 = Sum of POCSO Act (Sections 17 to 22) / Other offences of POCSO Act POCSO2 = Sum of POCSO Act r/w Section 377 IPC / Unnatural Offences Indecency = Sum of Indecent Representation of Women (Prohibition) Act, 1986

Data Sources:

Ministry of Women and Child Development National Crime Records Bureau
Census of India 2011 District Handbook, Census of India Telephonic conversations with Police Districts in India

Table 4: Principal Components of crime categories

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
murder_with_rape	0.10	0.02	-0.03	0.05	-0.22	0.05	-0.22	0.11	-0.09	-0.36	0.23	0.36
dowry_deaths	0.19	0.28	-0.20	-0.30	-0.20	0.08	0.14	-0.05	-0.08	-0.06	-0.14	-0.03
suicide	0.24	-0.13	0.27	-0.15	0.13	0.01	-0.05	-0.16	0.08	-0.03	0.03	0.15
miscarriage	0.11	0.09	-0.05	-0.10	-0.12	0.00	-0.53	-0.23	0.09	0.17	-0.09	-0.10
acid_attack	0.14	0.07	0.06	-0.13	0.04	-0.11	-0.15	0.13	-0.30	0.40	-0.23	0.22
attempt_acid_attack	0.21	-0.28	-0.09	-0.15	0.21	0.08	-0.16	0.03	-0.18	0.02	0.12	-0.21
cruelty_husbandfamily	0.35	0.17	0.06	0.11	0.15	-0.04	-0.01	0.04	-0.03	-0.03	0.00	-0.03
kidnap	0.26	0.03	0.32	0.04	-0.15	0.14	-0.01	-0.02	-0.07	0.12	0.03	0.03
kidnap_murder	0.06	0.07	-0.15	-0.09	-0.09	0.06	0.10	-0.07	0.25	0.31	0.39	0.10
kidnap_ransom	0.16	-0.03	-0.32	0.22	-0.01	0.05	-0.27	0.01	0.20	0.15	0.04	-0.07
kidnap_miscarriage	0.12	0.26	-0.33	-0.27	-0.10	-0.02	0.12	0.12	0.01	-0.20	-0.09	-0.06
procuration_minor	0.10	0.09	-0.12	0.12	0.16	-0.12	0.08	0.30	0.40	0.00	-0.04	0.00
procuration_foreign_girls	-0.02	-0.01	-0.01	0.02	0.02	0.00	0.00	0.03	0.05	0.01	-0.02	-0.07
kidnap_others	0.13	0.18	-0.20	0.20	0.17	-0.03	0.08	0.11	-0.07	-0.20	0.14	0.26
trafficking	0.11	-0.01	0.12	0.16	-0.02	-0.07	0.10	0.28	0.25	0.06	0.04	-0.29
sell_minor	0.06	0.00	-0.04	0.00	0.15	-0.03	-0.09	0.50	-0.15	0.34	0.00	0.16
buy_minor	0.06	-0.01	0.09	-0.03	-0.02	0.14	0.03	0.06	-0.03	-0.08	-0.01	-0.58
rape	0.30	0.20	0.05	0.19	0.04	0.15	0.11	-0.06	-0.06	-0.13	0.08	-0.09
attmpt_rape	0.22	0.27	-0.08	0.29	0.17	0.08	0.02	-0.12	-0.15	0.14	-0.04	-0.07
modesty_attack	0.28	0.07	0.21	0.15	-0.20	-0.08	0.02	0.10	0.01	-0.12	0.14	-0.02
insult_modesty	0.12	0.00	0.20	0.03	0.02	-0.57	-0.02	-0.21	-0.03	0.00	0.01	-0.04
dowry	0.05	0.15	-0.21	-0.27	-0.14	-0.26	0.09	-0.08	-0.13	0.08	0.10	-0.17
prostitution	0.08	0.02	0.11	-0.09	0.01	0.17	0.11	-0.12	0.03	-0.18	0.12	0.12
detain_pros	0.10	-0.33	-0.29	0.07	-0.15	-0.05	0.15	-0.12	0.05	0.08	-0.19	0.06
vicinity_pros	0.18	-0.42	-0.23	-0.02	0.14	-0.02	0.05	-0.06	-0.02	-0.11	0.05	-0.08
soliciting_pros	0.16	-0.30	-0.04	-0.13	0.23	0.08	-0.04	0.00	-0.11	-0.16	0.20	-0.03
pros_other	0.09	0.03	0.13	0.03	0.08	-0.48	0.05	-0.24	0.17	-0.04	-0.03	0.02
domestic_violence	0.04	0.01	0.10	-0.03	-0.04	0.13	0.14	0.05	0.26	0.16	-0.05	-0.15
publish_obscene	0.08	-0.01	-0.12	0.07	-0.27	-0.06	-0.53	0.03	0.21	-0.16	0.02	-0.06
cybercrime	0.12	-0.21	0.00	0.23	-0.43	-0.08	0.14	0.10	-0.20	0.02	-0.08	-0.06
child_rape	0.28	-0.22	0.13	-0.16	0.01	0.09	0.00	0.11	0.13	0.10	-0.05	0.10
child_assault	0.24	0.02	0.11	-0.35	-0.15	0.20	0.10	-0.05	0.16	0.03	-0.18	0.10
sexual_harassment	0.18	-0.06	-0.19	-0.18	0.17	-0.31	0.06	0.12	-0.07	-0.10	-0.09	-0.02
child_porn	0.00	0.01	-0.08	-0.07	-0.12	-0.07	0.13	-0.11	-0.03	0.32	0.65	0.00
POCSO1	0.08	0.10	-0.15	0.25	0.18	0.18	0.00	-0.40	-0.19	0.12	-0.10	-0.07
POCSO2	0.10	-0.13	-0.14	0.14	0.00	0.05	0.19	-0.21	0.32	0.07	-0.21	0.28
indecency	0.07	-0.18	-0.06	0.21	-0.37	-0.05	0.16	0.05	-0.25	0.06	-0.06	0.02