Synthetic Data Generation for Modelling Sexual Harassment
ISM Final Report

Nandini Agrawal, ASP Student
nandini.agrawal_asp21@ashoka.edu.in

Advised by:
Prof. Debayan Gupta
Assistant Professor of Computer Science

I. Introduction

Massive amounts of data from survey results, census records,etc is stored in tabular form. This data can be analysed for patterns and used to make multiple decisions. Hence tabular data plays a pivotal role for every organisation, be it companies, researchers, etc, in decision making. But acquiring this data is not very easy. Despite the huge investment institutions put into collecting tabular data every day, real tabular data cannot always fulfill what is asked of it. Some of the issues are as follows:

a) Quantity of data might be insufficient
b) Quality of data is not very good. Example : Missing/ Incorrect values
c) Imbalance in data
d) Privacy concerns like access of sensitive information

All of these issues mentioned above are the reason we want to generate synthetic data. Synthetic data is data/attributes of humans synthesized by integrating a diverse range of data sources and using models to interpolate data. These data points are not a snapshot but a statistically accurate representation, they keep the actual data private and follow the same correlation as that. Machine learning models are heavily dependent on data and not having enough/having poor quality data can cause issues in training. Hence if we can generate synthetic data from the same underlying distribution it would help in the training phase. This also solves the issue of lack of access to data due to privacy concerns and govt regulations.

II. Synthetic Data for Sexual Harassment in India

Under the Indian Laws and guidelines laid down by the Hon'ble Supreme Court, sexual harassment is defined as:

- Inappropriate physical contact and advances
- A demand for sexual favors
- Making sexually colored remarks
- Showing pornography
- Others including any unwanted physical, verbal or non-verbal sexual remark

Punishment for this varies from fine, to three years of imprisonment. Employers in an organisations are obligated to report sexual harassment cases under the Indian Penal Code. Workplaces include government organisations, private organizations, trusts, NGOs, hospitals, sports facilities, places visited by an employee including the transportation taken, service providers, house,etc.

To model sexual harassment cases in India i.e have models to identify sexual harassment and if yes then what kind of harassment , if someone is likely to harass/be harassed,etc, we need to generate synthetic data since:

1. India lacks proper collection of data(not many instances are reported also)

2. Data is not accessible due to its sensitive nature.
3. Data that is available has imbalance as it is very gender skewed (men as victims to sexual harassment is not acknowledged much)

A. Proposed Columns in the Intermediate Dataset:

I shortlisted a few attributes that would be required to model sexual harassment. These columns are as follows:

- Gender: Many studies show that women are more likely to face sexual harassment than men. An online survey in 2018 by Stop Street Harassment showed that around 81% women have experienced sexual harassment whereas this number was just 43% for men.

- Age : Studies show young people to be more vulnerable to sexual assault as compared to older people due to attributes like sexual attractiveness, vulnerability and active social lives .

- State and District : A few places have reported more instances of sexual abuse as compared to others. Therefore, location does play an important role while modelling sexual harassment and abuse.

- Education, Job and Income : All these factors are interlinked in some way. People with lower education i.e. limited reading and writing skills struggle economically, are unemployed or have low paying jobs and this affects the chances of sexual violence. Lesser control over lives causes low self esteem and therefore higher risks of sexual assault. Women are at increased risk of sexual violence, as they are of physical violence by an intimate partner, when they become more educated and thus more empowered.

- Caste and Religion : Caste driven and religion driven rivalries are often major reasons for sexual violence in India. It is a tool of domination that the upper caste uses against the lower caste. Eg. Women of the Dalit community are very vulnerable to sexual violence by upper class Hindu men. According to the National Crime Records Bureau, more than four Dalit women are raped every day.

- Family size : The size and functioning of a family affects violent nature in adults. Child abuse is prevalent in India and studies show that 67.3% of the victims of child abuse come from nuclear families with family members between 4-10 members.

- Maritial Status : Marital Rape is a prevalent issue in India. Over one third of the women in India's battered women's shelters report being sexually assaulted by

their husbands. There is also a significant association between high percentage of divorced men, and incidence of reported sexual violence.

- Sexual Orientation : A study by Center for Disease Control and Prevention(CDC) says members of the LGBTQ community are likely to experience sexual violence at similar or higher rates than straight people. As a community, LGBTQ people face higher rates of stigma and marginalization, which puts them at greater risk for sexual assault. Studies also suggest that more than half of transgender people will face sexual violence at some point in life.

- Health & Nutrition levels: Columns like weight, mental illness, STDs fall under this large category. Under/over weight plays a role in the chances of being harassed. Nutrition plays a shielding factor against gender based violence. Also, sexual violence has been associated with a number of mental health and behavioural problems in adolescence and adulthood. Research on women in shelters has shown that women who experience both sexual and physical abuse are significantly more likely to have had sexually transmitted diseases.

- Alcohol/Drug consumption : Alcohol plays a significant role in certain kinds of sexual assault. Studies show that there are complex relations between alcohol and violence. Men are more likely to be violent when drunk as they feel they won't be accountable. Women's ability to defend themselves by interpreting and correctly responding on warning signals is affected when they use alcohol or drugs.

- History of Abuse : There is evidence that links experience of abuse in adolescence with patterns of victimization during adulthood. A national study of violence against women in the United States found that women who were raped before the age of 18 years were twice as likely to be raped as adults, compared with those who were not raped as children or adolescents (18.3% and 8.7%, respectively). Studies also show if a man has faced sexual abuse in his childhood, it might increase the chances of him comitting rape.

B. Original Data :

The data sources of the original data used for generating the synthetic population is as follows:

- India Human Development Survey for 2011(micro level data)
- Census Data for 2011 (aggregate level data)
- National Crime Record Bureau (aggregate level data)

C. Current Synthetic Population

Summarising the columns in the synthetic population dataset that I generated(current best version) along with its data type(Categorical/Range/Binary) is as follows:

Column Name	**Column Type**
State	Categorical
District	Categorical
PSUID	Categorical
HouseholdID	Categorical
HHSize	Range
Sex	Categorical
Age	Range
Marital_Status	Categorical
Education_Years	Categorical
Weight	Range
Religion	Categorical
Caste	Categorical
Mental_Illness	Categorical
STD	Categorical
Alcohol_Consumption	Categorical
JobLabel	Categorical
Income	Range
AbuseHistory	Binary (1 indicating yes)

III. Process of Synthetic Population Data Generation

The above section talked about the need for synthetic data to create models for sexual harassment. In this section we would talk about the methods used for generation of this synthetic data. There are multiple techniques that can be used to generate synthetic data such as:

- Models such as GANs and its variations(CTGAN)
- Models such as VAEs and its variations
- Expanding microdata from the surveys using iterative proportional fitting by using aggregate information and taking joint probability distributions into consideration.
- Sampling from a complex multidimensional distribution with unknown dependencies between the components of a random vector using Gibbs Sampling.

I explored the first and last methods proposed above. The second method could not be used fully as there wasn't aggregate information for a few columns present in the original data. Although it can be used to expand a subset of the original data, the remaining columns can be synthesized using other methods and combined to the final synthetic data.

A. Conditional Generative Adversarial Networks(CTGAN)

CTGAN generates high quality tabular synthetic data which outperforms other models like bayesian networks. The advantage of using a CTGAN model is that since the generator doesn't have access to the real data, it can be used in a privacy setting.I used the open source CTGAN library to implement this. This CTGAN uses a Variational Gaussian Mixture Model to detect modes of continuous columns and a fully connected network to generate synthetic data. Though, it requires a lot of data points to accurately generate samples that represent the original distribution as closely as possible. But one advantage of this over the MCMC Gibbs Sampler model is that it is much faster to generate these samples.

B. Monte Carlo Markov Chain using Gibbs Sampler

Gibbs sampling algorithm is a method for sampling from a complex multidimensional distribution with unknown dependencies between the components of a random vector. This works for sampling from a survey. It helps meet privacy requirements while staying close to the true distribution of the original data as it only requires conditional distributions and doesn't require any tuning parameters. It is more efficient than the Metropolis-Hastings algorithm (which is also used for a Monte Carlo Markov Chain) as it does not reject when drawing high dimensional samples. Implemented this with Kshitij's help by creating a Gibbs Class with functions using the same algorithm as mentioned in the paper given in the references section. One advantage of this method is that it can be used with less training points also, and yet it will give a very close approximation of the original distribution but the variance in each column shouldn't be too much. As we haven't made use of parallel computation, it is much more time taking than the CTGAN model.

IV. Statistics and Evaluation Metric

A. Evaluation of Synthetic data :

I used the internal evaluation method (sdv.evaluation.evaluate) to evaluate the model initially to see if the synthetic data and original data are similar. To do that, there are various evaluation techniques inbuilt:

- LogisticRegression Detection
- SVC Detection
- GaussianMixture Log Likelihood
- Chi-Squared (CSTest)
- Inverted Kolmogorov-Smirnov D statistic (KSTest)
- Inverted Kolmogorov-Smirnov D statistic
- Continuous Kullback–Leibler Divergence
- Discrete Kullback–Leibler Divergence

I used LogisticRegression Detection, CSTest, and SVC Detection. CSTest is a statistical metric whereas the other two are detection metrics. The CSTest metric uses the Chi-Squared test to compare the distributions of two discrete columns(string values also considered). The output for each column indicates the probability of the two columns having been sampled from the same distribution. The detection metrics evaluate how hard it is to distinguish the synthetic data from the real data by using a Machine Learning model. The machine learning models used in our two detection metrics are Logistic Regression Classifier and SVC Classifier.

For the Best Sample :

Model Evaluation Scores:
{'cstest': 0.7186,
 'logistic_detection': 0.9933,
 'svc_detection': 0.9975}

Overall Score (combining all evaluation techniques mentioned above) : 0.6773

B. Other Statistics:

For Range Columns:

Statistical Comparison	Age		Weight		AnnualIncome	
	Orig	Synth	Orig	Synth	Orig	Synth
mean	29.82	29.24	42.48	44.08	52608.23	28577.13
median	26.0	26.0	45.0	45.30	27892.5	660.0
std	20.36	20.28	18.72	21.73	80289.55	71352.50
min	0	0	1.2	1.2	0	0
max	99	99	150	150	2420000	1899999

The big difference between the mean and median of the original and synthetic data in the AnnualIncome column is due to the fact that the original data had Income missing for various data points, so this is an average over just 53,416 points whereas the synthetic data is an average of 1,00,000 points.

For Categorical Columns :

Sex Ratio	**Original Dataset**	**Synthetic Dataset**
Male	49.9%	41.87%
Female	50.1%	58.13%

Marital Status	**Original Dataset**	**Synthetic Dataset**
Unmarried	45.2%	47.02%
Married	46.3%	41.15%
Widowed	6.0%	8.21%
Separated/Divorced	0.5%	0.4%
Married, spouse absent	1.8%	2.83%
Married no gauna	0.2%	0.39%

Completed Education Years Distribution	Original Dataset	Synthetic Dataset
0 years	32.8%	28.94%
3 years	13.8%	16.53%
5 years	7.4%	5.17%
8 years	23.0%	25.5%
10 years	10.0%	12.75%
12 years	7.2%	5.67%
15 years	3.8%	3.08%
16 years	1.9%	2.36%

Religion	Original Dataset	Synthetic Dataset
Hindu	80.1%	81.15%
Muslim	13.6%	12.68%
Christian	2.6%	2.45%
Sikh	2.3%	1.53%
Buddhist	0.6%	0.5%
Jain	0.2%	0.6%
Tribal	0.4%	0.52%
Others	0.1%	0.51%
None	0.0%	0.06%

Caste	Original Dataset	Synthetic Dataset
Brahmin	5.0%	9.75%
General	22.9%	18.03%
OBC	41.0%	46.88%

SC	21.1%	13.55%
ST	8.5%	10.89%
Others	1.2%	0.91%

Mental Illness	**Original Dataset**	**Synthetic Dataset**
No	99.7%	%
Cured	0.0%	%
Yes	0.3%	%

STD	**Original Dataset**	**Synthetic Dataset**
No	100.0%	98.74%
Cured	0.0%	0.54%
Yes	0.0%	0.73%

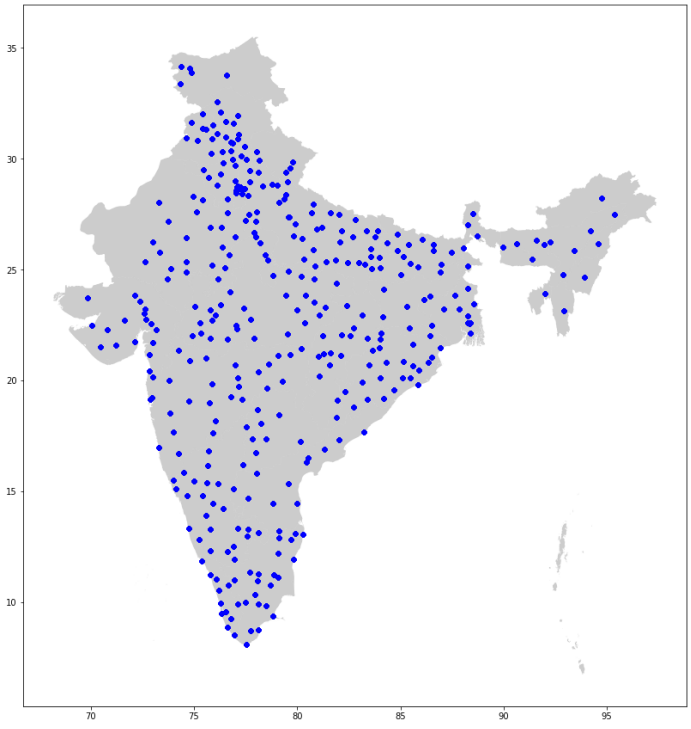
Alcohol Consumption	**Original Dataset**	**Synthetic Dataset**
Never	64.79%	76.49%
Sometimes	21.91%	14.46%
Daily	7.07%	4.70%
Rarely	6.23%	4.35%

Top 3 States in Original Dataset	**Top 3 States in Synthetic Dataset**
Uttar Pradesh - 10.5%	Maharashtra- 13.91%
Karnataka - 8.9%	Uttar Pradesh - 10.2%
Maharashtra - 7.8%	Karnataka - 8.01%

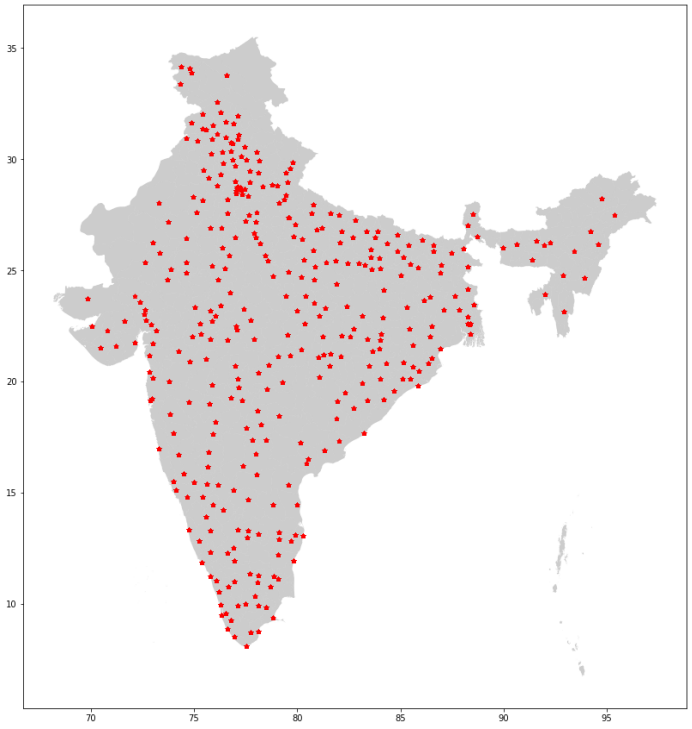
Bottom 2 States in Original Dataset	Bottom 3 States in Synthetic Dataset
Daman & Diu - 0.1%	Daman & Diu - 0.03%
Dadra + Nagar Haveli - 0.2%	Dadra+Nagar Haveli- 0.16%

Top 3 Jobs in Original Dataset	Top 3 Jobs in Synthetic Dataset
Ag labour - 28.65%	Construction - 21.36%
Construction - 25.53%	Ag labour - 16.48 %
Drivers - 4.11%	Teachers - 2.67 %

Bottom 2 Jobs in Original Dataset	Bottom 3 Jobs in Synthetic Dataset
Housewife - 0.002%	Housewife- 0.003%
Life science tech - 0.002%	Statisticians- 0.003%



Original Data (1,36,321 points)



Synthetic Data (1,00,000 points)

V. Roadblocks and Issues Faced

- The MCMC model gave samples that were close to the actual distribution when we didn't consider columns like AnnualIncome. Including that column makes the overall sample less accurate as the variance in the income is too much and hence sampling gets affected.

- A lot of the values for alcohol, job labels and income were missing, leaving me with a limited set of data points to train a CTGAN on. So, I trained separate CTGAN models(one for alcohol consumption's distribution and one for job+income's distribution) to generate synthetic data for these columns and combined it with the rest of the synthetic data. While combining a few things I kept in mind were:
    a) No individual with less than <18 years of age is given a job and thus has an income of 0. This might not be exactly similar to the original data though, as that has a few data points where agents whose age is less than 18 are working.
    b) No individual with age less than the overall minimum legal drinking age(18 is the minimum for all of India) has alcohol consumption as true. I chose the overall minimum and did not go with state wise legal drinking age as they can still have it in another state.

- Some data points in the synthetic data didn't really have a correlation between the age column and the marital status column i.e. children as young as 2 years of age were indicated as married. I changed these values to unmarried for all women less than 18 and all men less than 21 as per legal marriage age in India. This might not be exactly similar to the original data though, as that has a few data points where agents whose age is less than 18 (like 16/17 years old, mostly female) are married.

- Some data points in the original data don't really have a correlation between age and weight. The approaches considered to fix this issue was:
    a) Clean source data and then generate synthetic data from that.
    b) Regression on age and gender and then generate weight.

  This isn't working fullyl and synthetic data still has anomalies. Discussed some approaches with Bhavesh but haven't got an accurate model. Will continue to work on that.

- Very few data points in the synthetic data didn't really have a correlation between the age column and the highest level of education column. These were fixed by dropping those points(not sure if that's the best way but I did not lose many data points). The basis on which I picked out such rows is as follows: if the age of the agent is less than the number of education years or the age of the agent is less than the number of education years + 2, those rows were dropped. I chose 2 as that's the age people start schooling. Eg: An agent is 4 years old and has 3 years of education, this isn't possible, hence dropped. Similarly if age is 7 years old then and they have 8 years of education, that also isn't possible.

- Data on sexual orientation of the people of India not available(for 2011 as microdata is from that year) hence could not add to the synthetic data. Only statistics of the number of transgender people is available statewise but the the IHDS survey does not consider the 'Other' gender.

- Faced issue understanding methods of combining aggregate data with microdata. The authors of that paper did not reply back. I will continue to look into this. For the time being, the history of abuse(in this case rape) data was in the aggregate form but I added it to the rest of the data(microdata) temporarily using some basic granularity like Age and State (on the basis of the NCHB report of reported rape cases in 2011). One of the big issues here is most of the data is targeted at women, i.e. could not find data about the abuse of men.

- The census reports and data from the IHDS surveys have some gaps (in statistics). Example: Comorbidities like STD, IHDS survey indicates 0% population has been infected with that but that's not true.

- Lastly, one of the biggest roadblock was mine and my family's health which hindered my progress immensely. I, along with my father and extended family contracted COVID.

VI. Other Work : Adding Essential Worker and Public Transport column in TW dataset

To add a boolean value as to whether an individual uses public transport or not, we make use of their jobLabel. There is a list which mentions all the possible well paying jobs and if an individual's jobLabel matches any job in that list: we mark a 0 for them indicating that they don't use public transport, else we mark a 1 i.e. they use public transport. These well paying jobs are defined subjectively based on my understanding from the india wage report 2011. Another possible way of doing this is to make use of the Annual Income column within the IHDS survey (WSEARN). Any individual with a salary less than 5000/10000 a month can be categorised in the low/mid income category and hence we mark a 1 i.e. they use public transport, else we mark a 0. Similarly, essential workers were also defined based on the agent's jobLabel.

References and Papers

1. For Columns of the Intermediate Dataset:
    a) https://www.who.int/violence_injury_prevention/violence/global_campaign/en/chap6.pdf
    b) https://web.uri.edu/iaics/files/15RebeccaSMerkin.pdf
    c) https://www.npr.org/sections/thetwo-way/2018/02/21/587671849/a-new-survey-finds-eighty-percent-of-women-have-experienced-sexual-harassment
    d) https://www.unwomen.org/en/what-we-do/ending-violence-against-women/facts-and-figures
    e) https://news.psu.edu/story/278094/2013/05/30/research/young-people-are-overwhelmingly-victims-sexual-assaults
    f) https://www.dw.com/en/caste-dynamics-behind-sexual-violence-in-india/a-43732012
    g) https://bprd.nic.in/WriteReadData/userfiles/file/201609221217334612798Summary.pdf
    h) https://www.hrc.org/resources/sexual-assault-and-the-lgbt-community
    i) https://www.statista.com/statistics/705970/india-number-of-transgender-people-by-state/
2. Original Data Sources :
    a) https://ihds.umd.edu/data/ihds-2
    b) https://censusindia.gov.in/2011-common/censusdata2011.html
    c) https://ncrb.gov.in/sites/default/files/crime_in_india_table_additional_table_chapter_reports/Table%205.3_2011.pdf

3. Models and Evaluation:
    a) Modeling Tabular data using Conditional GAN: https://arxiv.org/abs/1907.00503
    b) Monte Carlo Markov Chain using a Gibbs Sampler: https://sci-hub.do/https://ieeexplore.ieee.org/abstract/document/6680524
    c) Open source CTGAN library: https://github.com/sdv-dev/CTGAN
    d) Evaluation Metric : https://github.com/sdv-dev/SDV/blob/master/EVALUATION.md
4. Others:
    a) Synthetic Data Vault : https://dai.lids.mit.edu/wp-content/uploads/2018/03/SDV.pdf
    b) Combining aggregate and individual data: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3347777/