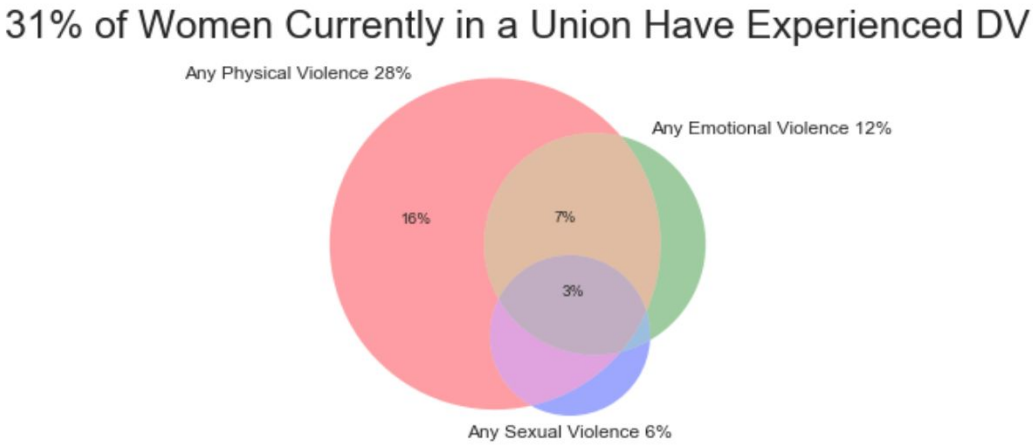
Using Customer Segmentation Techniques to Target Social Campaigns Against Domestic Violence in India

Mary Scott Sanders | October 2018

Introduction

With 1.2 billion people, India is the most populous democracy and the fastest growing economy in the world (Economist.) With 624 million males and 586 million females, it hosts one of the worlds largest systems of gender-based violence. If a female makes it past the sieves of sex-selective abortion, infanticide, and infant abandonment, there is a 30% chance she will go on to experience domestic violence ("dv".) Even though domestic violence has been a criminal offense since 1983, and "civil protections [have been] afforded to victims of domestic violence" since 2006, only 14% of women who have experienced domestic violence seek help, only 3% seek help from the police, and only 1% seek help "from a doctor or medical personnel, a lawyer, or a social service organization" (NFHS-4).



While much more work needs to be done in the police system to better protect women, in the legal system to more efficiently prosecute perpetrators, and in the healthcare system to better help women recover, there is also a significant amount of socio-cultural work that needs to be done to change the way men (and women) value and treat women. Men need to learn to value women, to protect them rather than mistreat them. Women need to learn that they are valuable, do not deserve to be mistreated, and when possible, to access the resources available. Abusers also need to be made aware that there are potentially legal consequences for their actions. In 2014, the World Bank "found that 38 percent of people surveyed [in India]

were not aware of laws on violence against women….Men who were unaware of the law were 1.5 times more likely to perpetrate intimate partner violence" (World Bank.)

The impact of social campaigns will be limited by their ability to communicate these ideas in a way that engages the intended audience and using media that is accessible to the intended audience. This project will use K-Means clustering, typically used by marketers to segment customer bases, to help India's government or relevant NGOs tailor social campaigns addressing domestic violence. This analysis would be particularly useful to an organisation like Bharatiya Grameen Mahila Sangh (BGMS,) which is an NGO that works through grass roots mechanisms toward community development and women's socioeconomic empowerment. This analysis attempts to understand what different groups of women exist when sorted by experience with domestic violence and investigates how that might inform the design of social campaigns directed towards those groups (and their partners.) With this research on hand, BGMS could generate a series of social campaigns specifically targeted to the subgroups identified in this analysis, with the relevant messages and medium for each group.

The Data

The data is from USAID's Demographic and Health Surveys, India: Standard DHS, 2015-16. The survey has 79,729 respondents for the domestic violence module and also covers many aspects of health and demographics. This analysis focuses on women currently in a union, of which there are 62,716 respondents in the original data set. There are over 4,000 columns in the original dataset. The selection used for this analysis are described below:

Variable	Description	Original Unique Values
v_sex_ttm	Anyone other than partner forced sexual acts	[nan, no, refused to answer/no response, yes]
v_phys_tt2m	Times hit by other than partner last 12 months	[nan, sometimes, not at all, often]
d113	Does partner drink alcohol	[no, yes]
d114	Times partner gets drunk	[nan, sometimes, often, never]
afraid	How often respondent is afraid of husband/partner	[never afraid, sometimes afraid, most of the time afraid]
v743a-d	Final say in household decisions: respondent's healthcare, large purchases, visits to family, husband's earnings	[husband/partner alone, respondent and husband, respondent alone, someone else]
v744a-e	Wife agrees with justifications for wife beating: wife going out, burning food, neglecting children, refusing sex,	[no, yes, don't know]

	argues	
d104	Ever any emotional violence	[no, yes]
dv_phys_less	Experienced any less severe physical violence	[no, yes]
dv_phys_more	Experienced any physical severe violence	[no, yes]
d108	Experienced any sexual violence	[no, yes]
control_issues	Number of control issues	[0,1,2,3,4,5,6]
worked_ttm	Whether the respondent worked in the last 12 months	[no, currently working, in the past year, have...
v102	De facto type of place of residence	[urban, rural]
v130	Religion	[hindu, christian, muslim, buddhist/neo-buddhi...
v190	Wealth Quintile	[middle, richest, richer, poorer, poorest]
can_read	Literacy	[able to read whole sentence, cannot read at a...
v157	Frequency of reading newspaper or magazine	[not at all, at least once a week, less than o...
v158	Frequency of listening to radio	[not at all, less than once a week, almost eve...
v159	Frequency of watching television	[almost every day, not at all, less than once ...
age	Respondent's age	[15, 16, 17, 18, 19, 20, 21, 22, 23...]
age_partner	Partner's age	[15, 16, 17, 18, 19, 20, 21, 22, 23...]
education	Respondent's years of education	[0,1,2,3,4,5,6]
education_partner	Partner's years of education	['do no know',1, 2, 3, 4, 5...]

Feature Engineering to Handle Categorical Features

- The first step taken was replacing all ‘yes’ and ‘no’ values with 1 and 0 respectively.
- The variable dv_phys_level was generated by scoring 2 if respondent replied ‘yes’ to dv_phys_more, 1 if respondent replied ‘yes’ to dv_phys_less, and 0 if respondent replied ‘no’ to both.

- V_sex_ttm and v_phys_tt2m were combined to create a new binary variable called ‘v,’ that indicates if the respondent reported abuse from someone other than the partner.
- The values for ‘afraid’ were replaced with 0 for ‘never afraid,’ 1 for ‘sometimes afraid,’ and 2 for ‘most of the time afraid.’
- V743a-d were combined by scoring each individual question with a 1 if the respondent indicated they had some say in the household decision, and then adding together the values from each individual question.
- V744a-e were combined by adding up the number of justifications a wife agreed with.
- V130 was transformed by one hot encoding.
- V157-9 were transformed by scoring each response in the following manner: ‘not at all’ = 0, ‘less than once a week’ = 1, ‘at least once a week’ = 2, ‘almost every day’ = 7.
- Wealth was transformed by ranking each response in the following manner: ‘poorest’ = 1, ‘poorer’ = 2, ‘middle' = 3, ‘richer’ = 4, and ‘richest’ = 5
- Can_read was transformed by scoring ‘able to read whole sentence’ as 1 and all others as 0.
- V102 was replaced with a variable called ‘urban’ which equals 1 if v102 equaled ‘urban.’

Data Cleaning

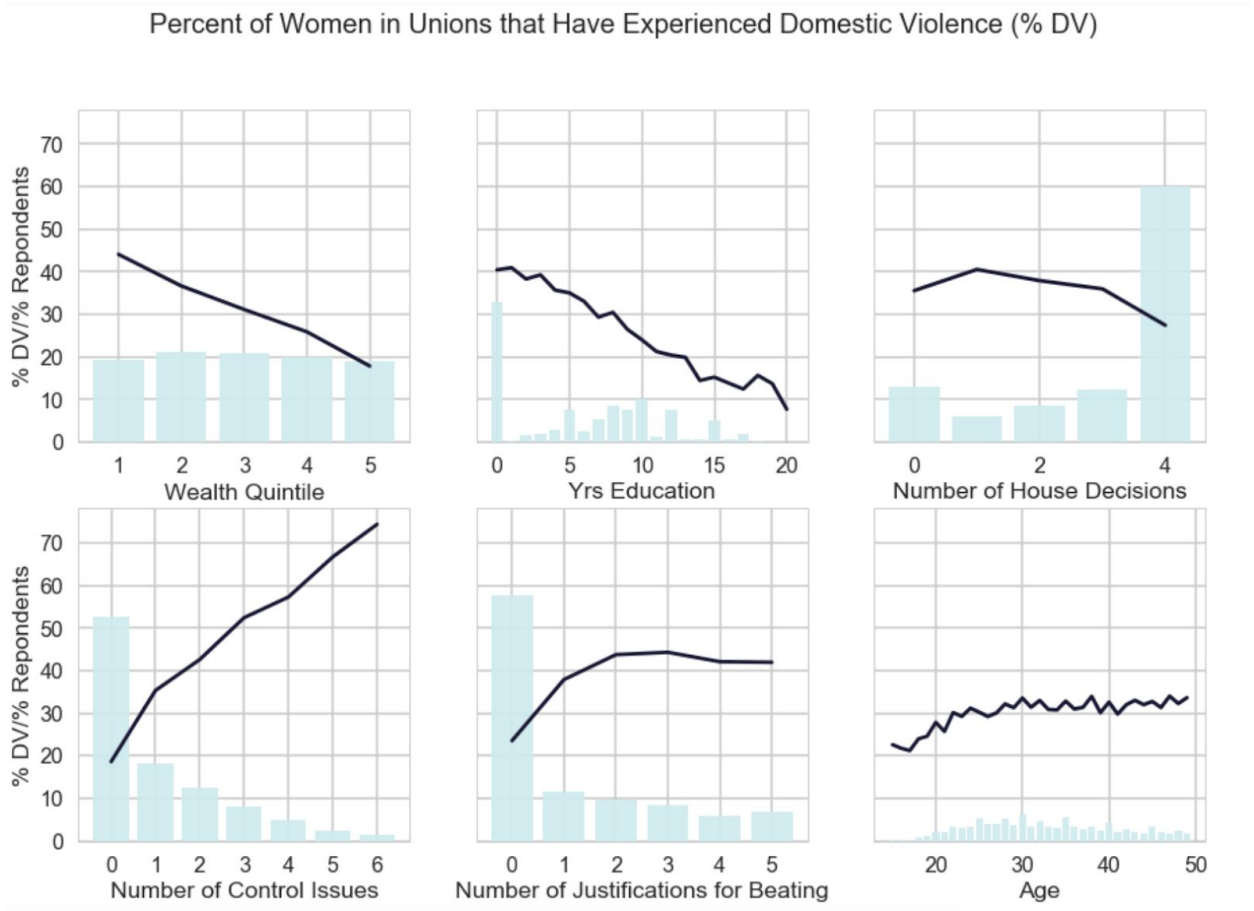
There were a couple of instances in which I decided to impute missing values. If a question relied on a previous question and was null due to the previous question’s answer, I filled the secondary question’s value with a zero. For example, a respondent whose husband does not drink (d113) was given a zero for the following question asking how often the respondent’s husband gets drunk (d114). For ‘education_partner,’ if the respondent replied that she did not know how many years of education her partner had, I replaced ‘do not know’ with the mean years of education for the rest of the column.

There were about 1,000 observations dropped because they contained null values and therefore did not complete a full interview. There were about 1,500 dropped because the respondent was not a usual resident of the household. After handling missing values, there were 59,934 remaining observations (about 96% of the original in-union sample.)

There were several values used in the survey for error codes: ‘9,999’, ‘999’, ‘99’, ‘9, 9,998’, ‘998’, ‘98’, ‘8’, ‘9,997’, ‘997’, ‘97’, ‘7’. The only two columns that contained those values, ‘education’ and ‘education_partner,’ contained ‘7’, ‘8’, and ‘9’. These seemed plausible, so I did not manipulate them.

Exploratory Analysis

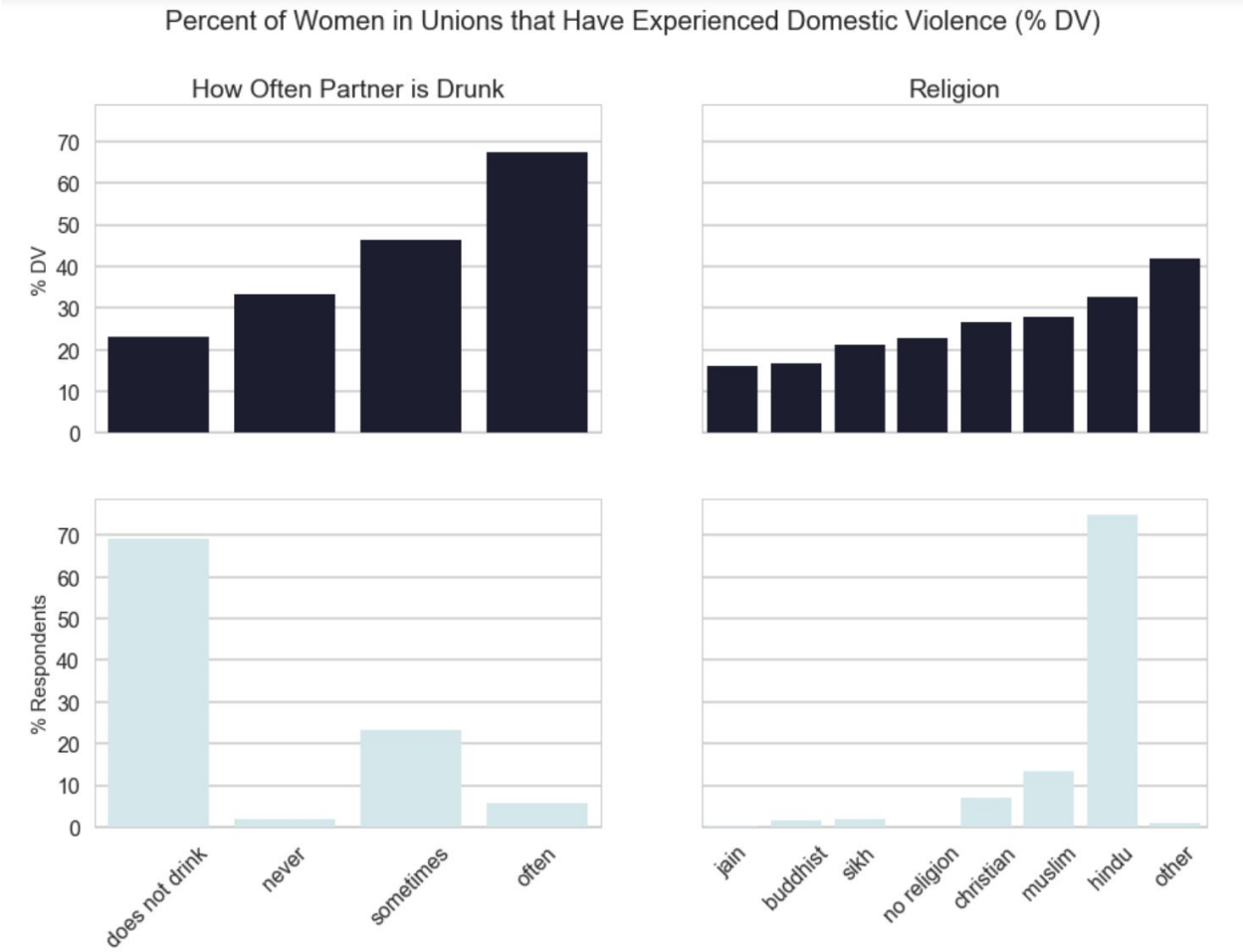
Of the in-union sample 31% reported having had experienced domestic violence. That can come in several forms. 28% experienced physical violence, 12% reported experiencing emotional violence, and 6% reported experiencing sexual violence. A significant portion reported experiencing more than one form of domestic violence. 3% reported experiencing all three.

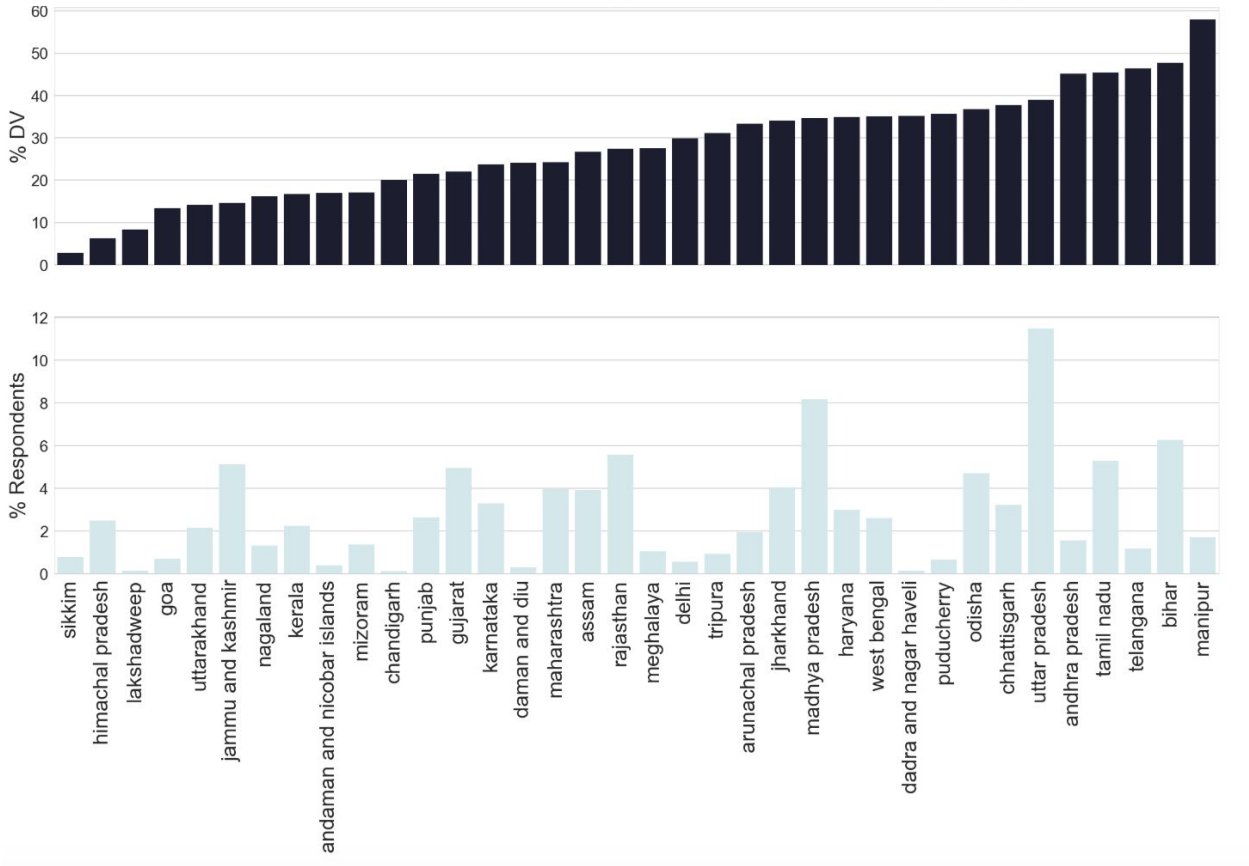


The charts above show the percent of women in unions that have reported experiencing at least one form of domestic violence with respective histograms.

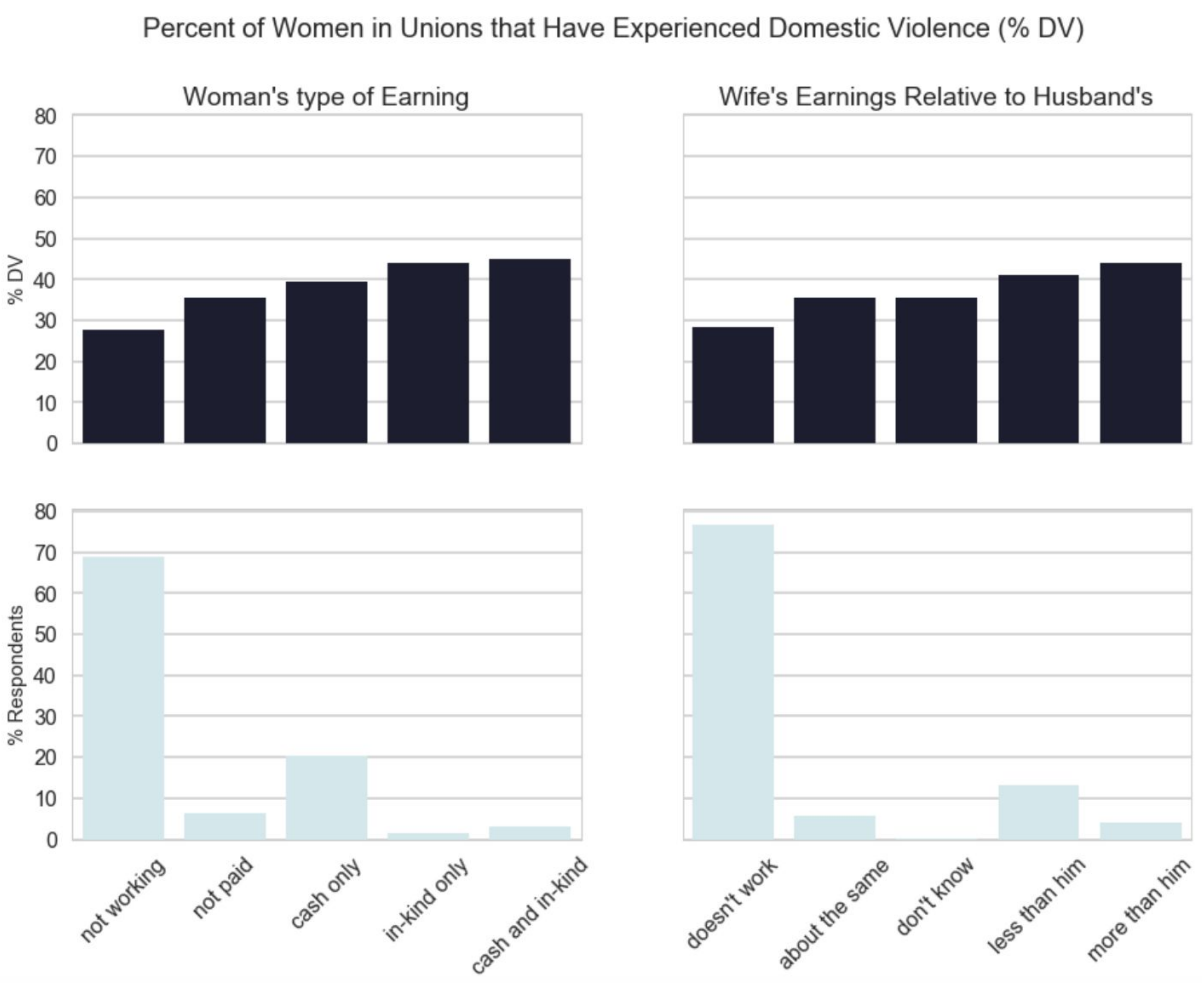
- Wealth is sharply negatively correlated with dv, while nearly 45% of women report dv in the poorest quintile report dv, only about 20% in the wealthiest quintile report dv.
- Years of respondents' education is also sharply negatively correlated with dv. The largest group of women at zero years of education, about a third of in-union women, report the highest rates of domestic violence at about 40%.
- Generally speaking, the number of household decisions a woman participates in is negatively correlated with reporting rate of dv, although not consistently so. 60% of women report to have some say in decisions related to her own health, household purchases, visits to her own family, and how to use husband's earnings, and also experience a dv rate slightly lower than overall "in union" group of 31%.
- The number of control issues a husband displays is positively correlated with dv, while there seems to be an exponential decrease in the number of women who reported each number of control issues.

- The number of justifications of wife beating that a woman agrees with is positively correlated with the percent of women who report domestic violence. The largest increase in dv is seen between 0 and 1 reasons. Although it may be counter-intuitive, 36% of those who do not report domestic violence agree with at least one justification, while 56% of those who report dv agree with at least one reason.
- Reporting rate of dv also increases with age, with the steepest incline happening early on, between 15 and 25 years old.





- How often a partner is drunk is positively correlated with dv reporting rates. About 70% of partners do not drink, and that group has a slightly below average dv rate. While only about 5% of partners are reported as “often” drunk, this group has a dv rate almost twice the average.
- “Hindu”, which captures about 75% of the respondents, is the largest religious category, it also reports the second highest dv rate, which is slightly above average. “Jain,” “Buddhist,” and “Sikh” together account for less than 10% of women in unions and have a dv rate about two-thirds the average.
- Dv rate also varies by state, from less than 5% in Sikkim to over 55% in Manipur.



It would make sense to think that a woman's contribution to household earnings would increase her empowerment. While that may be true, it seems as if women who do not work experience the lowest rates of dv, compared to women who do work. In fact, women who make more than their partners have the highest rates of dv.

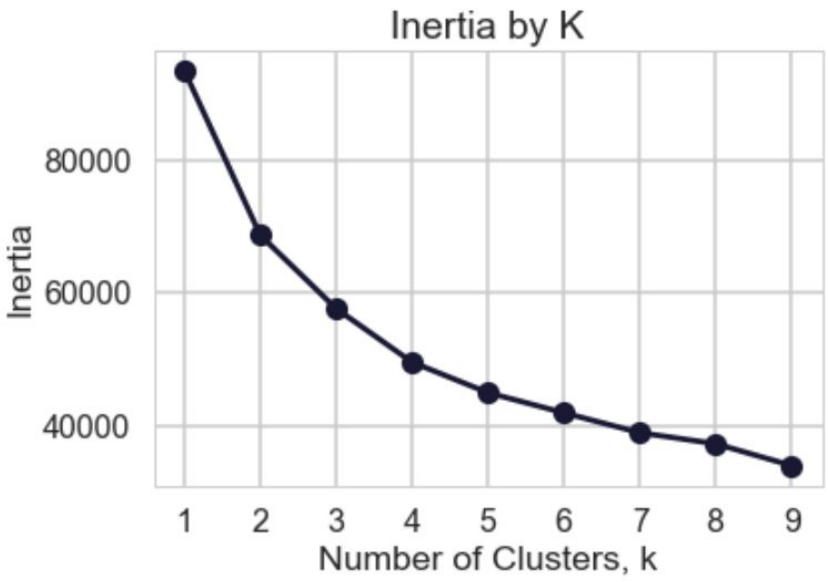
Clustering

The purpose of clustering in this analysis is to understand what the major groups of women are when segmented by experience with domestic violence. The variables used for segmentation were:

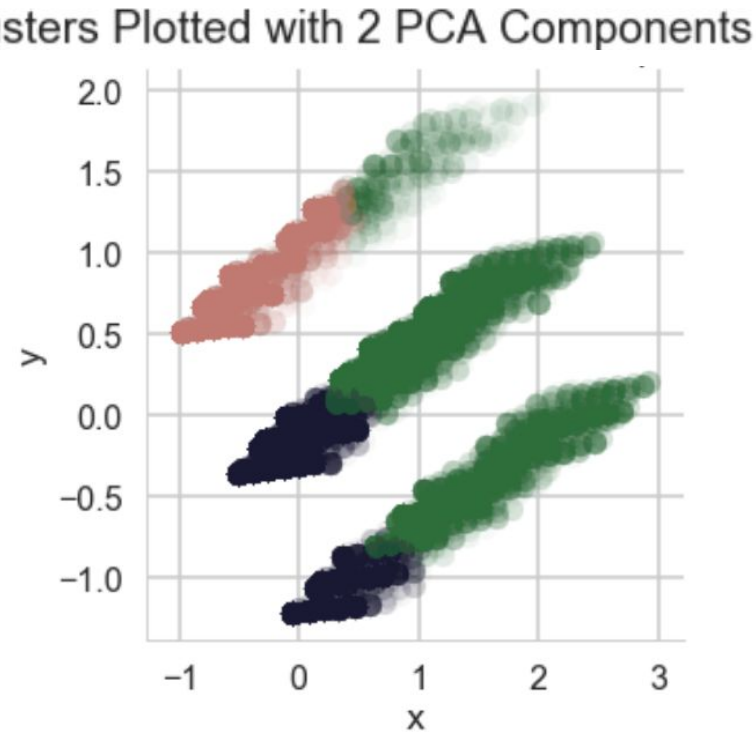
Segmentation Variables Used for Clustering

Feature	Description	Values
dv_phys_level	If respondent's partner physically abused respondent (0 = "no" /"no partner", 1 = "less severe violence", 2 = "severe violence")	0 = "no", 1 = "less severe violence", 2 = "severe violence"
dv_emotional	If respondent's partner emotionally abused respondent	0/1
control	Number of control issues husband displays (divided by 3, ranges from 0 to 2)	6 possible controlling behaviors, total divided by 3, ranges from 0 to 2
dv_sex	If respondent's partner sexually abused respondent (binary)	0/1
v_between_parents	If respondent's father beat respondent's mother	0/1
v_sex_ttm	If someone other than respondent's husband forced sexual activity in the last 12 months (binary)	0/1
v_phys_ttm	If someone other than respondent's husband physically abused the respondent in the last 12 months (binary)	0/1
afraid	How often the respondent is afraid of spouse	2 = "often", 1 = "sometimes", 0 = "never"
justs	If the respondent agrees with any justifications for wife beating (binary)	0/1

These variables are intended to capture, to the extent possible, how the respondent has experienced domestic violence and the respondent's reaction to it. Rather than normalizing or standardizing the features, I used their relative scales as described above to weight their influence in the clustering process.



Using the “elbow” method, looking for the value of K at which the decrease in inertia slows down, the K was set to 3. This generated 3 clusters, each distinguishable when plotted using 2 PCA components.



Means by Cluster

Feature	Abused	In Danger	Less Harmed	Feature Description
cluster_perc	24%	55%	21%	Percent of sample in cluster
dv	100%	8%	14%	If respondents has reported dv
afraid_most	29%	12%	0%	If respondent is afraid of husband most of the time
afraid_sometimes	67%	88%	0%	If respondent is afraid of husband sometimes
afraid_never	4%	0%	100%	If respondent is never afraid of husband
dv_phys_less	98%	3%	11%	If respondent has reported less severe physical abuse
dv_phys_more	31%	0%	0%	If respondent has reported more severe physical abuse
dv_sex	21%	2%	2%	If respondent has reported sexual abuse
dv_emo	39%	3%	4%	If respondent has reported emotional abuse
v	9%	2%	2%	If respondent has reported violence from anyone other than husband
control_issues	2.01	0.83	0.61	Number of controlling behaviors the partner displays
justs	61%	37%	36%	If respondent agrees with one or more justifications for wife beating
can_read	41%	58%	66%	If respondent can read a full sentence in a major language
news_mags	0.62	1.13	1.43	How often each week respondent watches tv
tv	3.54	4.26	4.83	How often each week respondent watches tv
radio	0.4	0.43	0.52	How often each week respondent listens to the radio
news_mags_never	77%	65%	58%	Respondent never reads newspapers/magazines
tv_never	33%	25%	18%	Respondent never watches tv
radio_never	86%	85%	84%	Respondent never listens to the radio
wealth	2.5	3.1	3.3	Wealth quintile of respondent
age	33.2	32.6	33.3	Respondent’s age
age_partner	38.0	37.5	38.4	Partner's age
education	4.3	6.3	7.3	Respondent's years of education
education_partner	6.1	7.9	8.6	Partners years of education
v130_hindu	80%	74%	73%	If respondent identifies as Hindu

Every member in the “Abused” group has experienced domestic violence. Roughly 100% have experienced less severe physical violence, 30% experienced severe physical violence, 20% experienced sexual violence, and 40% experienced emotional violence by their partners. Correspondingly, almost all report being afraid of their husband either sometimes or most of the time. Members of this group are also the most likely to experience violence by others than just partners. Partners of this group have the largest average number of control issues. Members of this group are also the most likely to agree with at least one justification for wife beating. This group is the least likely to be able to read a sentence in a major language and has the fewest interaction with media throughout the week. In fact 77% never read a newspaper or magazine. However, 67% do watch tv, and the group averages watching tv 3.5 times throughout the week. This group has the lowest average wealth quintile. While this group is the least likely to have agency in decisions concerning their own health, still 72% report having some agency. In short, while this group is the one that needs the most support, it is likely the most difficult to reach with social ad campaigns.

The “In Danger” group, the largest group which represents 55% of the sample, was perhaps the most surprising group to come from the clustering process. While less than 10% of the “In Danger” group report domestic violence, 100% report being afraid of their husbands either sometimes (88%) or most of the time (12%.) This group would likely benefit from cultural progress in regard to domestic violence. Even though they have not experience dv yet (or are not reporting it,) their life is still impacted by the threat of it. On average, this group’s partners display less than half the control issues as the “Abused,” but about 35% more than the “Less Harmed.” This group is equally as likely as the “Less Harmed” group to agree with at least one justification for beating. This group tends to sit somewhere between the “Abused” and the “Less Harmed” group for most other features, like wealth, education, and access to media. If the “Abused” group is targeted effectively, it seems likely this group would be positively impacted as well.

100% of the “Less Harmed” group reports never being afraid of their partners. Partners average 0.61 control issues, 30% of the “Abused” group average. 36% of the women agree with at least one justification for wife beating, roughly equivalent to the value for the “In Danger” group. This group is the most likely to be able to have some say in decisions regarding their own health. This group has the highest literacy rate, with 66% of respondents able to read a sentence in a major language. This group, like the others, interacts with tv more frequently throughout the week than print media and radio. Only 18% never watch tv.

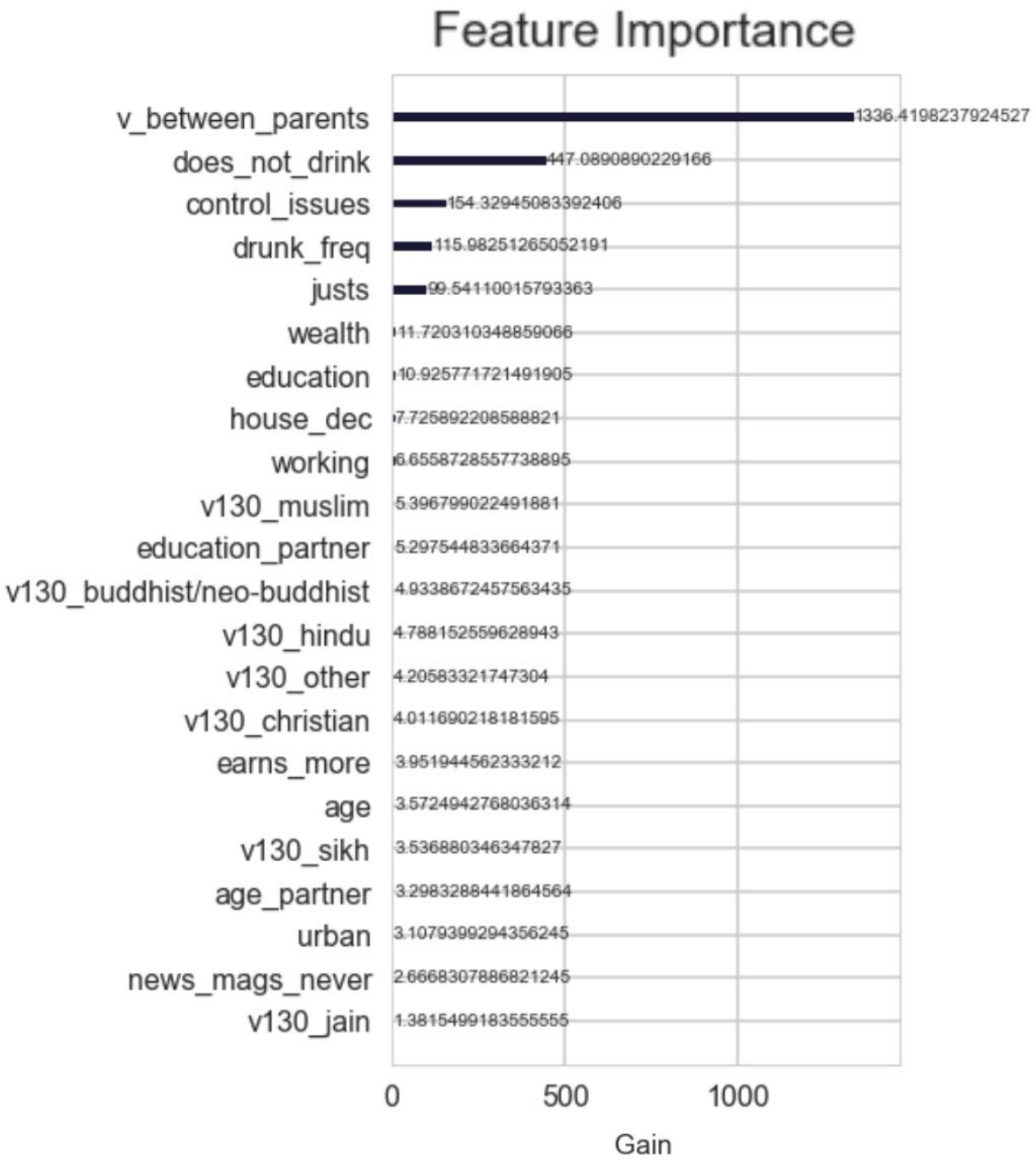
Gradient Boosting

In order to understand what the most important factors are in domestic violence, I used an XGBClassifier to predict whether or not a woman reported any type of violence (any_v.) I set aside 20% of the data as test data. I set the learning rate at 0.02, the number of estimators at 500, and the number or rouns to stop at if there was no improvement at 10 (“early stopping rounds”.) In order to tune the model, I used GridSearchCV with three-fold cross validation to

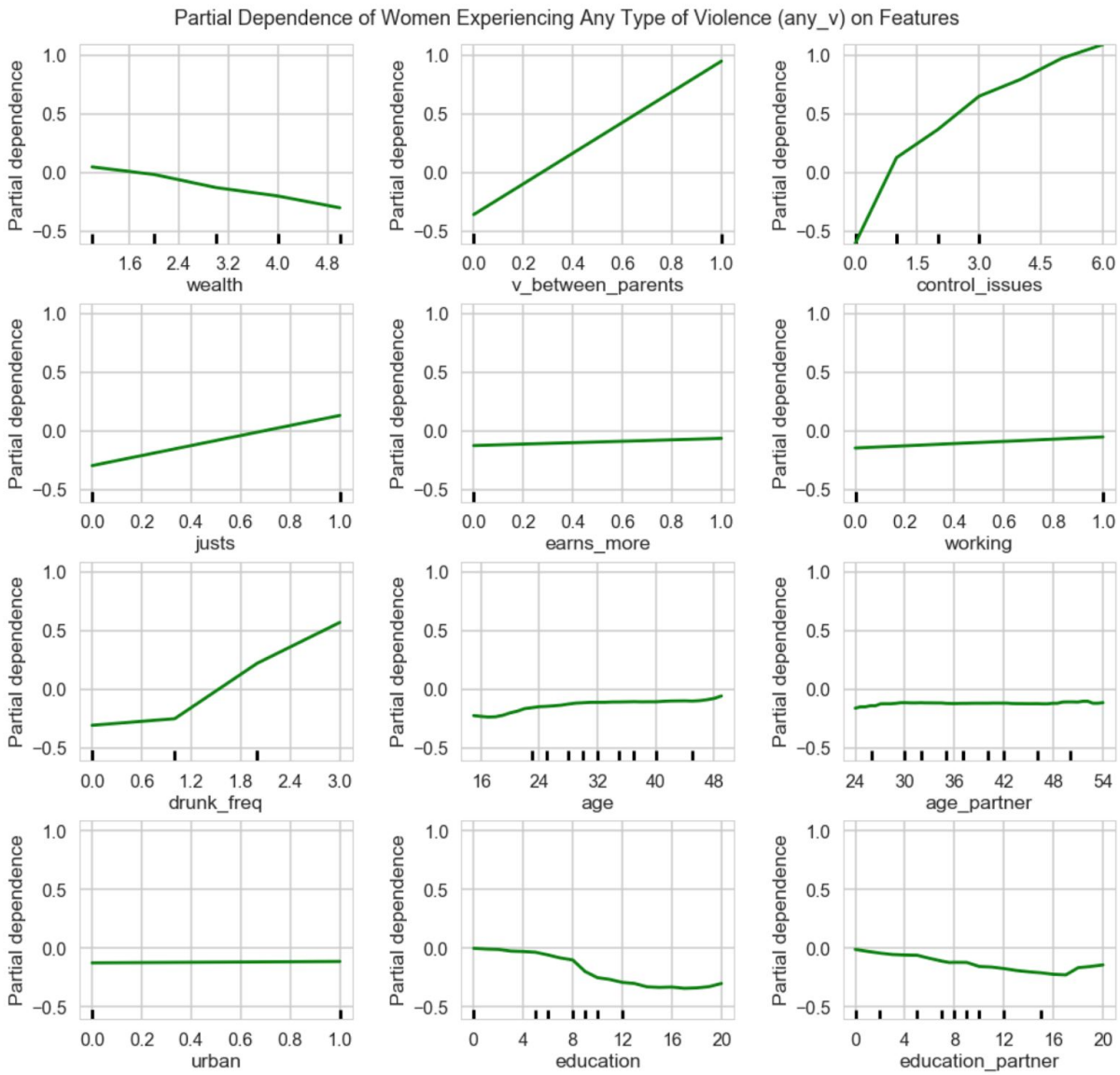
discover what portion of the parameter grid, with max tree depths between 1-10, and columns samples by tree between 10%-75%, was most advantageous.

Results from Gradient Boosting Model with GridSearchCV Tuning	
AUC	67.2%
Accuracy	75.0%
Recall	44.9%
Precision	67.7%
Best colsample_by_tree	30.0%
Best max_depth	5

In the end, the model results were not impressive, but the primary purpose of use classification in this analysis was to understand the relative influence of individual features. Below is a chart of features arranged by their importance in terms of “gain,” or how valuable a feature is in reducing impurity during when splitting.



Above, the behavioral factors like drinking, control issues, agreeing with justifications for wife beating and violence between parents swamp socioeconomic factors that showed up more prominently in the univariate charts in the data exploration section. This would imply that the two groups of factors are correlated, resulting in the overemphasis of wealth and education in the univariate plots.



While the univariate plots in the data exploration section are impacted by covariates, partial dependence plots attempt to separate out a single feature impact on the target variable. “Partial dependence plots show the dependence between the target function and a set of ‘target’ features, marginalizing over the values of all other features (the complement features)...For classification you can think of it as the regression score before the link function.” (scikit-learn) For example, if there were a major difference between a univariate plot for a specific feature and its partial dependence plot, it would be worth investigating what other variable might have been influencing the univariate plot. In general, the partial dependence plots seem less severe although relatively similar in shape to their corresponding univariate plots. Socio-economic features like wealth and education seem to have less impact, relatively speaking, than it appeared originally, while behavioral features like control issues, violence between parents, drunkenness really pop out. It is interesting to see that the difference between education plots

for women and their partners especially around 8 years of education. It looks like whether or not a woman goes to highschool is a larger determiner or whether or not she will experience dv.

Recommendations

Mechanisms

While the “Abused” group needs the most outreach, it is also the most out-of-reach of traditional media channels. However, out of tv, radio, and newspapers/magazine, TV is the most equitable across groups and the most likely channel to reach the “Abused” group, which averages 3.5 views a week.Tv would also be best for this group since most people in this group cannot read a sentence in a major language and campaigns could rely mainly on audio/visual presentations, rather than text. Further research should be done to investigate the most advantageous ways to incorporate anti-domestic violence messages into this demographic’s regular tv consumption. While it seems that behavioral factors are more influential in dv outcomes, socioeconomic information is still useful for targeting high risk groups. What shows are the lower-income, less-educated, more-rural demographic watching? What services are they using to watch television? Would commercials be feasible on these services for those shows? Could those shows directly incorporate anti-domestic violence into their programming?

Further study should be devoted to if and how education can reduce domestic violence. Findings in this report suggest that in addition to promoting messages that relate directly to domestic violence, curriculum should also address the associated behavioral factors, like control issues and drunkenness. It is important to note that those in the “Abused” category are least educated, so using education as a tool to address domestic violence should start early. On average, women in this group have 4 years of education and men have 6.

Messages

The two main categories of messages that should be furthered are those that address potential/actual perpetrators and those that address potential/actual victims. Certain types of messaging should target both boys and men and should aim to teach them to respect women, even to protect women. About 50% of women are afraid of their partners, even though they report no domestic violence. There should be messaging to convert these partners to becoming less antagonistic. About 20% of women report no dv and that they are never afraid of their partners. There should be messaging geared to this group to convert them to advocates for women. As about 30% of women reported some form of domestic violence, there should also be messaging to make them aware that most violence against women is actually illegal and what the potential consequences of their actions are. As 42% of women in unions agreed with at least one reason for wife beating, women also need to be assured that they should not be physically abused and their value should be reaffirmed. As only 11% of those in this sample who have experienced domestic violence had ever told anyone but the interviewer about what they experienced, further research should go into reasons why services are not accessed when

available, and women should be made aware of how to access any public services that are available.