Prediction of Domestic Violence

Proposal

Given the recent crime rate against women is increasing especially Domestic Violence, I decide to study the factors that could possibly explain the household domestic violence in India. In India, almost 55% of women undergo domestic violence, therefore, it will be helpful to understand what are the factors that may cause such incidences.
The Indian Human Development Survey(IDHS) is a multi topic survey of close to 1000 villages and neighbourhood in India. Interestingly, the survey also captured if the person was subject to domestic abuse. I decided to use this survey to try and model the economic, social and economical factors that influence domestic abuse.

Approach

To try and isolate factors which contributes to domestic abuse. The approach used is to frame this as a classification problem. Features were isolated using domain knowledge initially (like Day of birth is definitely not contributing to domestic violence but Year of birth is a potential candidate) and then manually combined to generate the feature representation for each user. Once this was done, several supervised models were trained to map the features of users to a binary classification decision indicating the presence or absence of domestic violence for that particular user. Feature Importance techniques were used to analyze the models and identify the factors that contributed to domestic violence.

Action

Data was obtained from the Indian Human Development Survey (IDHS) released by National Council of Applied Economic Research, New Delhi and the University of Maryland.
IDHS dataset generated 3 data file -

1. Eligible Women- Includes socio-economic details of individual women across 350 districts (county)
2. Household - Contains information on characteristics of the household's dwelling unit and characteristics of the residents and the usual visitors
3. Individual - Contains information about the social, political and economic aspects of an individual (men and women)

Since the aim of the project was to identify the factors that cause the domestic abuse for women in India the Eligible women dataset was the most informative.

Some key details of the Eligible Women dataset

- Number of observations - 39523 (records of data)
- Geography - 375 districts (districts represent County in USA)
- Variables in the dataset - 580
- The response variable "Beat if" was formed by consolidating 6 variables - beast if - disobedient, beat if- neglect on children, beat if neglect on elderly, beat if- suspicious on extramarital affair, beat if - no respect,

Initially close to 100+ variables were shortlisted out of the 580 variable using domain knowledge.

Variable like - Month of Birth, Date of birth, personal opinions like Contract HIV/AIDS: Mosquito bite, and other similar features were dropped instead I included features that were more informative and concise than the ones mentioned above.

The 100 + features were thus classified into 5 categories

The 5 categories were -

1. Decision-making situations/Scenarios - Variables are categorical in nature. This set of variables will gave a general indication of decision-making power given to a woman in a household. There were total of 85 original variables recording the above mentioned information. These variables were combined and consolidated to have total of 26 variables. Some of the example of variables are Cooking_Self, Purchase_Husband, If visit health center alone,If visit friend alone,If visit kirana shop alone, etc.
2. Freedom of choice - This set of variables include answers to many questions like does the woman in the house practice purdah or not, the amount of cash given to a woman, is she a part of Mahila Samiti, self-help group, etc. As it can be seen, the majority of variables are categorical in nature. There are total 24 variable in this set.
3. work-related variables- This set of variables decides whether a woman is allowed to work outside the house, wages given to a woman, etc. The variables are categorical as well as numerical in nature. There are around 6 variables relating to work. Eg. Allowed to work, Willing to work, Working:Who decides:Self, etc.
4. marital history of women - This set of variables gives information about the age at which woman was married, was she allowed to talk to her husband before marriage, whether she was shown her husband's photo, whether she knew her husband at all before marriage, etc. The variables in this set are majorly categorical in nature. There are 18 variable in this set.
5. variables related to the profile of women- This set of variables gives information about the level of education of the woman in the household, education level of other members of her house, etc. These variables are categorical in nature. There are 32 variables in this set.

Thus variables were consolidated into these categories and were reduced to around 100+ variables

Analysis:

Since the data is still high dimensional, I decided to do Principal component analysis for dimensionality reduction and decided to keep 15 Principal components
The transformed data was fit on supervised learning models like Logistic Regression and Random Forest. To reduce the effect of overfitting of the data I implemented L1 (Lasso) and L2 (Ridge) regularizer into my logistic regression. Ridge logistic model (Logistic regression with L2 regularizer) performed better than the other models by a marginal difference.

Feature importance plot for logistic regression, ridge Logistic Regression, Lasso Logistic regression and Random Forest were found. The subset of the top 10 features that all the models agreed on were:

1) Education Level of the family
2) Freedom to travel alone
3) Allowed to Socialize
4) Financial backing of woman
5) Husband known
6) Age at marriage
7) Decision making power of woman regarding children
8) Working freedom
9) Marriage choice
10) General Health of the individual

From the above features we get an idea about which features are major indicators to identify if a woman is undergoing domestic abuse. Thus keeping these indicators in mind the government can work towards improving the situation

The following are the few changes and initiatives can be taken to reduce the incidence of domestic violence based on the above factors:

1. Lack of education is one of the chief factors that cause domestic violence. Initiatives like scholarships and job guarantee will promote education and might improve the situation.
2. Freedom to travel is one of the important factors. However there can two reason that might put restriction on the movement of women - Lack of safety to travel alone, Restriction put by the family members. The government should provide or increase protection so that women could have a safer travel experience
3. Inorder for a woman to socialize the government could open free women centric clubs that teach cooking, sewing etc which the women may be allowed to participate.
4. Similarly for financial backing the government can open Self Help Groups for the women where the could work and have initiatives higher interest saving account so that the women have bank accounts in their names.