A data mining logistic classifier approach for identifying the socio-economic determinants of early intention of antenatal care (ANC) in Bangladesh.

Group 7
Huda Abdaljalil and Tanjima Rahman
STAT 6440: Data Mining
Dr. Shuchismita Sarkar
April 21, 2024

Abstract

This study focused on finding the affecting characteristics of early and regular Antenatal Care (ANC) visits those reduce maternal and child mortality in Bangladesh, a developing country where such health factors are a concern. This research utilized data from the Bangladesh Demographic and Health Survey (BDHS) that was conducted in 2017–18 and employed several statistical analytical methods to identify factors that are influencing women's intention of early ANC services. The results revealed that 41.5% of women engaged in at least one ANC activity during the first trimester of their pregnancy, which, although higher than in other countries, falls below the global average. Both women's and their partner's education, had a significant impact on the intention to visit ANC early. Women from wealthy household and urban residential areas were found to be more likely to attend these early ANC visits rather than women from poor households and rural areas. Exposure to the mass media also showed great impact on their choice of having these ANC visits. These findings enhance our prospect that if these socio-economic characteristics can be improved for all the women in Bangladesh that will raise maternal and child health outcome at a great deal in Bangladesh. We also discovered that among several data mining approaches Random Forest is the best working method for these sorts of data analyzation. It usually gives us the more precise results than any other supervised algorithm.

Background of the study

Over the centuries, with the help of technology stillbirth may have been reduced a lot but complications related to pregnancy and maternal mortality are still major issues for the world. Most of these maternal deaths can be prevented by taking some important actions during the pregnancy and childbirth. Antenatal care (ANC) is one of those steps. It is a health care service that is given to pregnant women to prevent, diagnose and get treatment for any type of medical and pregnancy-related complexity. A quality full and improved ANC is a very important tool and key opportunity for every pregnant woman to improve the health of the mother and the child during pregnancy. Several studies have shown that those women who started having early ANC and attended them most frequently were more likely be healthy than those who started it later or never even got one (Yuba Raj Paudel, 2017).

Even though ANC is an important factor, the probability of attending these ANCs depend on different religious, cultural, geographical, socioeconomic and demographic factors. For developed countries these factors may not be as influential as in developing countries. In this study we tried to determine which socio-economic factors are affecting the intention of ANC among women in Bangladesh. We tried several data mining procedures to figure out those determinants. We found factor like education, residence status, wealth, exposure to mass media and health facilitated delivery are influential to change the early intention of ANC visits.

There have been few studies about this issue before, though most of them were about African women (Keolebogile M. Selebano, 2021) (Abdul-Aziz Seidu, 2022) (Michael Ekholuenetale, 2022) (Deogratius Bintabara, 2021). There are very few studies about Asian women (Yuba Raj Paudel, 2017) (Md. Ismail Hossain, 2024). We want to figure out the outcome variables for Bangladesh here.

Methodology

Source of the data

The data we used is from a cross-sectional national survey which was done by Bangladesh Demographic and Health Survey (BDHS), 2017-2018. Technical assistance was provided by the ICF and was funded by the United States Agency for International Development (USAID).

Sample design

A two-stage stratified sampling design was used to collect the data from both rural and urban areas of Bangladesh. In the first stage 675 areas of Bangladesh were selected, and in the second stage 30 households were sampled from these areas to collect the actual data. 20,127 reproductive women were interviewed who either received antenatal care or gave birth to live babies within the previous three years of the survey. Upon following our criteria and removing all the missing values we ended up with 4604 women, who's data was used for further analysis.

Response variable

Since our study was focusing on the early intention of antenatal care, we chose the early intention of antenatal care (ANC) visit at less than 4 months of the pregnancy or earlier as the response variable. For coding we assigned, a "1" if a woman reported attending ANC at 4 months of pregnancy or earlier, and a "0" if she either did not attend ANC or attended after 4 months.

Regressor variable

We treated several socio-economic factors as regressor variables and there were 16 of them. The variables we selected are respondent's age (15-24, 25-29, 30-49), respondent's education (No education, Primary education, Secondary education, Higher education), partner's education (No education, Primary education, Secondary education, Higher), wealth status (Poor, Middle, Rich), religion (Muslim, Non-Muslim), intended pregnancy (Yes, No), respondent's working status (Yes, No), exposure to mass media (Yes, No), residence status (Urban, Rural), region of residence (Eastern, Northern, Southern, Central), health facilitate delivery (Yes, No), age at first marriage (Less than 18, Greater or equal to 18), sex of the household head (Male, Female), age at first birth (less than 18, greater or equal to 18), healthcare decision maker (Respondent alone, Husband alone, Jointly, Other person) and husband's occupation status (Working, Not working).

Statistical analysis

Data partition

Initially we partitioned the data set into two sets training and test. We built all our models on training data and evaluated the model performance on test data. Almost all our predictor variables were categorical, we made several dummy variables for them.

Logistic regression

At first logistic regression analysis was employed to check which variables are significant.

Let y_i be our response variable and π is the probability of response variable of being "Yes".

$$y_i = \begin{cases} 1; whenwoman\ reported\ attending\ ANC\ at\ 4\ months\ or\ earlier \\ 0;\ when\ woman\ did\ not\ attending\ ANC\ at\ 4\ months\ or\ earlier \end{cases}$$

$x_1, x_2, \dots, x_p$ are the regressor variables and $\beta_0,\ \beta_1,\ \beta_2, \dots, \beta_p$ be the corresponding regression coefficients. Then the logistic regression model is,

$$\log\left(\frac{\pi(y)}{1-\pi(y)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

Where ϵ is the random error term.

Stepwise

We also used several supervised algorithms to determine the significant variables.

Backward stepwise elimination with 10-fold cross validation approach was used to select the variables. In this method we began with the model where all the p regressors were present. Then for each regressor calculated its contribution and the regressor with smallest contribution was dropped from the model if the corresponding p-value was greater than α_{out}. We repeated the process until the p-value corresponding to the contribution at a particular step was less than α_{out}.

Lasso and Ridge regression

Two regularized regression method for variable selection were used. They are Lasso regression and Ridge regression. In regularized regression the regression coefficients are estimated by minimizing $SSE + P$, where P is the penalty factor.

For Lasso regression we minimized $SSE + \lambda \sum_{j=1}^{p} |\beta_j|$

For Ridge regression we minimized $SSE + \lambda \sum_{j=1}^{p} \beta_j^2$

Classification tree

We also used classification tree with cost complexity pruning. In this method the predictor variables were divided into nonoverlapping multidimensional rectangles in such way that each of those rectangles were as homogeneous as possible. A full-grown tree may lead us to overfitting. To overcome this problem, we employed cost complexity pruning to identify an optimal sized tree

that has minimum cost complexity. In this method the size of a full-grown tree was reduced by removing sections that has little power to classify instances. The cost complexity of the tree is given by,

$$CC_\alpha(T) = err(T) + \alpha L(T)$$

Here $err(T)$ is the fraction of misclassification of the tree,

$L(T)$ is the number of terminal nodes,

And α is the penalty factor.

Then we applied some ensemble methods too.

Bagging

The first ensemble method we used was the Bagging with 50 bootstrap samples. Model was fitted on each sample to create a simple predictive model. The final model was selected by choosing the majority class label among $T_1(x),\ T_2(x), \dots, T_{50}(x)$.

Random forest

The second one was the Random Forest with 25 bootstrap samples of $m = 2$ regressor variables. Decision trees were fitted on each bootstrap sample. The final prediction was selected by choosing the majority class label among $T_1(x),\ T_2(x), \dots, T_{25}(x)$ predictions.

Adaboost

And the last method we employed was Adaboost with 10-fold cross validation. At the beginning same weight was assigned to each observation. Then a decision stump was fitted, and the performance of the model was evaluated. After that the misclassified observations were assigned higher weight and the correct ones were assigned lower weight. A decision stump was fitted again on the newly weighted data. This process was repeated 50 times. The maximum depth of the trees was 2 and the learning rate was 0.1. The final prediction was the weighted average of the predictions of all 50 models.

For **Classification tree, Bagging, Random Forest** and **Adaboost** there was no direct way to represent whether a variable was significant or not. To remedy this, we used the variable importance term of these methods. Whichever variable had 50+ variable importance value we treated them as significant and the others as not significant.

Model comparison

In order to compare all these methods and to find out the best approach we compared them with their corresponding accuracy value, kappa value and lift chart.

After fitting the model on training data, we evaluated their performance on test data by creating confusion matrix for each of these methods. By doing that we got one accuracy value and kappa value for each of these methods. Then we chose the model as best performing model whichever had the highest accuracy and kappa value.

The other popular way we used to access the performance of these regression models was the Lift chart. We used the $gains()$ function and $gains$ from the R package to produce a lift chart for each of these models.

Results

At first, we checked for any missing values, and there were many. We removed the missing values, as the total number of observations were much higher for our data and removing some missing values didn't affect much.

Then we partitioned the data into 80:20 partitions of training set and test set. The first method we ran on training data was logistic distribution.

Table-1: Variable selection in Logistic regression.

Variables	**Significant/ Not significant**
Age	Not significant
Respondent Education	Significant
Partner's Education	Significant
Wealth Index	Not significant
Religion	Not significant
Pregnancy intention	Significant
Respondent's working status	Significant
Mass media exposure	Significant
Type of residence	Significant
Region of residence	Significant
Health facilitated delivery	Significant

Age at first marriage	Not significant
Sex of the household head	Not significant
Age at first birth	Not significant
Health care decision maker	Significant
Husband's occupation status	Not significant

This Table1 shows us which of the variables we found significant through the logistic regression. As we can see here the age, wealth index and religion are not significant, but the education of both respondent and her husband are significant. According to this method women's intention of pregnancy, exposure to mass media, residence region, delivery process and who made the healthcare decision had impact on their choice of attending antenatal care visits before the first 4th month of their pregnancy period. While their age at first marriage and first birth did not have any effect on their choice.

Table-2: Variable selection for several supervised algorithms

Variables	**Stepwise**	**Lasso**	**Ridge**	**Classification tree**	**Bagging**	**Random forest**	**Adaboost**
Age	Not significant	Significant	Significant	Not significant	Significant	Not significant	Not significant
Respondent Education	Significant	Significant	Significant	Significant	Significant	Significant	Significant
Partner's Education	Significant	Significant	Significant	Significant	Significant	Significant	Significant
Wealth Index	Not significant	Significant	Significant	Significant	Significant	Significant	Significant
Religion	Not significant	Not significant	Significant	Not significant	Not significant	Not significant	Not significant
Pregnancy intention	Significant	Significant	Significant	Not significant	Significant	Not significant	Not significant

Respondent's working status	Significant	Significant	Significant	Not significant	Significant	Not significant	Not significant
Mass media exposure	Significant	Significant	Significant	Significant	Significant	Significant	Significant
Type of residence	Significant	Significant	Significant	Significant	Significant	Significant	Significant
Region of residence	Significant	Significant	Significant	Not significant	Significant	Not significant	Not significant
Health facilitated delivery	Significant	Significant	Significant	Significant	Significant	Significant	Significant
Age at first marriage	Not significant	Not significant	Significant	Not significant	Significant	Not significant	Significant
Sex of the household head	Not significant	Significant	Significant	Not significant	Significant	Not significant	Not significant
Age at first birth	Not significant	Significant	Significant	Significant	Significant	Not significant	Significant
Health care decision maker	Significant	Significant	Significant	Not significant	Significant	Not significant	Not significant
Husband's occupation status	Significant	Significant	Significant	Not significant	Not significant	Not significant	Not significant

Table-2 represents the significance of all the socio-economic factors for all seven methods that we used. Here we can see that the Ridge regression method interpreted all the variables as significant, and the Lasso didn't work that well too. Lasso only determined the religion and age at first marriage as not significant, the others were determined as significant. The stepwise worked well, it removed several variables and tried to find out the important ones. Among the ensemble methods

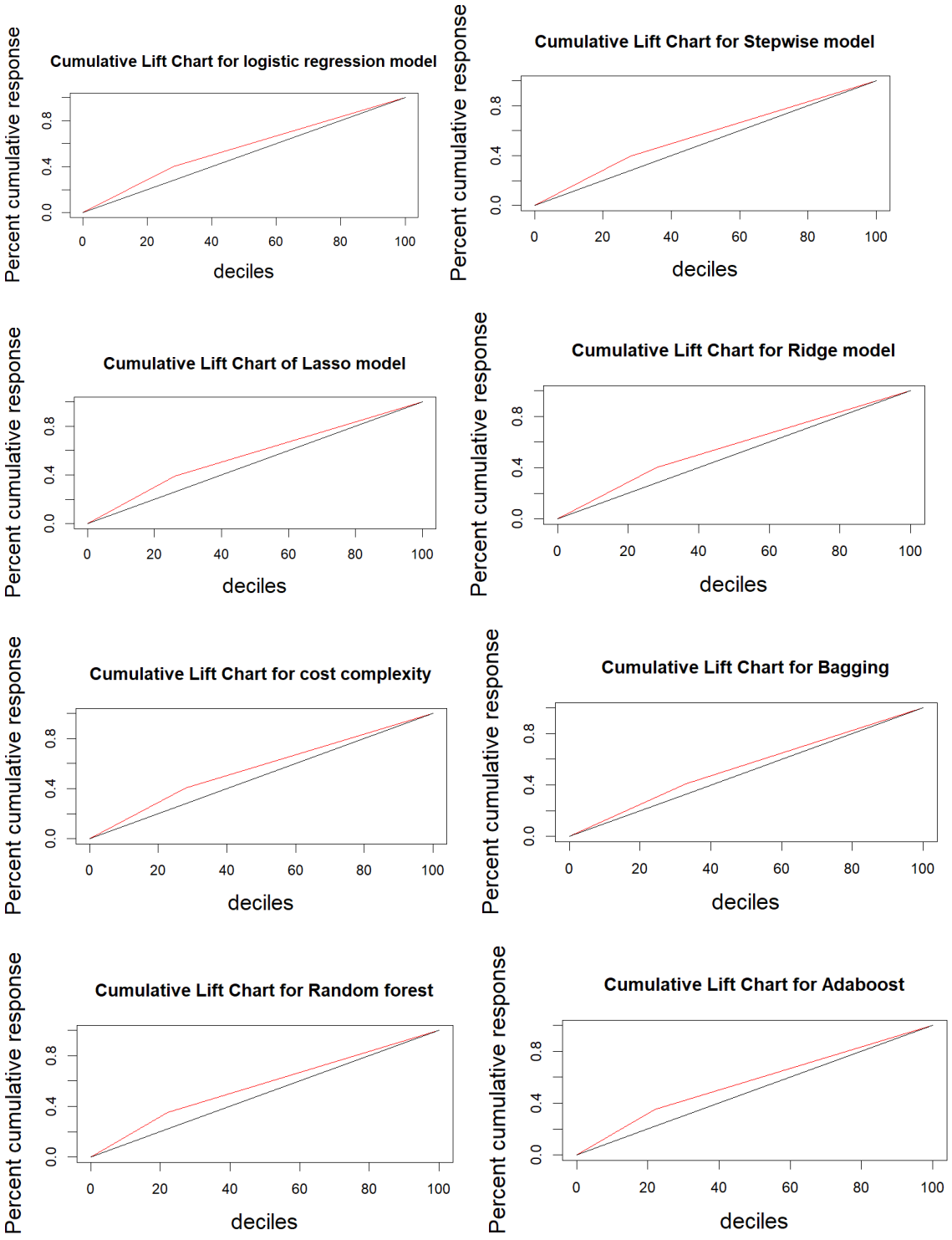
bagging performed the worst, as it did not remove many variables, while random forest and adaboost produced similar result. If we carefully look at this Table-2 we will see that the education for both respondent and her husband, exposure to mass media, residence type and how the previous baby was delivered were significant in each of these methods. So, these five factors have much more effect on early intention of ANC than others. Religion was only significant in ridge method, which implies that this factor has almost no effect on the early intention of ANC. The wealth index was significant in all the methods except for the stepwise, so it's obvious that this factor also has great impact on the choice of having early ANC.

Table-3: Model comparison with respect to Accuracy and kappa value

Variables	**Accuracy**	**kappa**
Logistic regression	0.6359	0.2169
Stepwise regression	0.6315	0.2074
Lasso regression	0.6458	0.2333
Ridge regression	0.6392	0.2225
Classification tree	0.6381	0.2222
Bagging	0.5941	0.1413
Random forest	0.6612	0.2547
Adaboost	0.6524	0.2377

Table-3 is presenting the 8 methods that we used for analyzing the data and their corresponding accuracy and kappa value. From this table it can be easily said that the performance of bagging has the lowest accuracy and kappa, which means that bagging performed the worst among all the methods. Random forest and adaboost have the quite similar values and we can conclude that random forest worked the best for this data set.

Lift charts



From the above lift charts, we can easily see that for this data the random forest has the highest prediction at the 20th percentile, while bagging has the lowest prediction. Lasso, Ridge, stepwise and logistic regression model's performances are almost the same.

After performing two types of comparison methods, we can conclude that the random forest worked the best among all the methods, while bagging turned out to be the least performing.

Discussion

In our data we have seen that 42% of women attended the ANC early, while more than half of these women either did not attend or attended late. We know that antenatal care is a very important factor for pregnant women and their unborn children. Because by attending them they can easily detect if there is any complexity with the child or mother. At the initial stage of pregnancy many complications can be resolved, or further steps can be taken to prevent them. But if you do not know anything about these complexity at all or diagnose them late that can be harmful for both mother and child. In this study we have determined that factors like education, wealth, residence status, exposure to media, previously health facilitate delivery and so on usually affect the choice of Bangladeshi women to attend the early ANC visit. The education, wealth and residence status have been found as key socio-economic determinants in other studies too (Keolebogile M. Selebano, 2021) (Gebretsadik Shibre, 2019). Women's education plays the most important role in affecting the intention of having early ANC (Michael Ekholuenetale, 2022). In developing countries like Nepal wealth and transportation were found greatly affecting the decision of having early ANC (Yuba Raj Paudel, 2017).

In our study we employed several variable selection data mining approaches to determine the factors that are affecting the choice of attending early ANC in Bangladesh. We used 4 regression methods, 1 decision tree and 3 ensemble methods. After analyzing the data with each of these methods we found our significant variables as well as the best method. By comparing these methods performance, we came to a decision that random forest performed the best while bagging was the least. Random forest was capturing the highest number of predicted values than any other methods. We have noticed that all the regression methods were performing similarly, though Ridge regression selected all the variables as important. The performance of decision tree was also good.

Limitations and strengths of this study

Using a nationally representative data was our first strength, because using this data gave us a good picture of the whole population of Bangladesh. The second strength was using several supervised algorithms. By employing these methods, we were able to analyze the data in several ways and

that made our result more accurate and dependable. Despite the strong points we also came along with some limitations. Firstly because of data constraints we could not include some other important factors that may have some influence on early intention of ANC among women. And secondly the cross-sectional nature of the data did not allow us to establish any link between the factors and the early intention of ANC.

Conclusion

In conclusion, this study focused on figuring out the key determinants that have significant influence on early intention of ANC visit. Respondent's and her partner's education background, type of residence, mass media, wealth status and previously health facilitated delivery were found as crucial factors of impacting intention of early ANC. Our findings emphasize that improvement in these factors may increase the amount of attending ANC visit during the early stage of pregnancy, which ultimately will improve maternal and neonatal mortality. We also tried to find the better approach for feature selection, and we came up with random forest as our best model. For future study one can employ several other methods like (KNN, Neural network) and may find more appropriate result. They should also overcome the limitations that we faced. Longitudinal studies can be considered for further studies to establish a linkage between early antenatal care and improved health outcomes for pregnant women and their children. This study provided a more nuanced understanding of the impact of timely healthcare access on maternal well-being. Finally, we found some foundations for improving maternal and neonatal health in Bangladesh along with some best data mining methods.

References:

Abdul-Aziz Seidu, J. O. (2022). Inequalities in antenatal care in Ghana, 1998–2014. *BMC Pregnancy and Childbirth*, 1-6. doi:https://doi.org/10.1186/s12884-022-04803-y

Deogratius Bintabara, N. B. (2021). Twelve-year persistence of inequalities in antenatal care utilisation among women in Tanzania: a decomposition analysis of population-based cross-sectional surveys. *BMJ Open*, 1-10. doi:10.1136/bmjopen-2020-040450

Gebretsadik Shibre, W. M. (2019). Socio-economic inequalities in ANC attendance among mothers who gave birth in the past 12 months in Debre Brehan town and surrounding rural areas, North East Ethiopia: a community-based survey. *Reproductive Health*, 1-14. doi:https://doi.org/10.1186/s12978-019-0768-8

Keolebogile M. Selebano, J. E. (2021). Decomposing socio-economic inequalities in antenatal care utilisation in 12 Southern African Development Community countries. *SSM - Population Health*, 1-8. doi:https://doi.org/10.1016/j.ssmph.2021.101004

Md. Ismail Hossain, T. R. (2024). Survival analysis of early intention of antenatal care among women in Bangladesh. *Scientifc Reports*, 1-10. doi:https://doi.org/10.1038/s41598-024-55443-5

Michael Ekholuenetale, C. I. (2022). Effects of socioeconomic factors and booking time on the WHO recommended eight antenatal care contacts in Liberia. *PLOS GLOBAL PUBLIC HEALTH*, 1-14. doi:https://doi.org/10.1371/journal.pgph.0000136

Yuba Raj Paudel, T. J. (2017). Timing of First Antenatal Care (ANC) and Inequalities in Early Initiation of ANC in Nepal. *ORIGINAL RESEARCH, 5*, 1-5. doi:https://doi.org/10.3389/fpubh.2017.00242