# HUMAN DEVELOPMENT IN INDIA

***By***

***Akshay Kulkarni***
***Anupam Gupta***
***Arpita Koundinya***
***Dhaval Mohandas***
***Saish Pai***

# *SUMMARY*

The development of a country can be analyzed based on some key factors like the social, economic and mental well-being of its subjects. It would be interesting to understand these factors for India by exploring "The India Human Development Survey(IHDS)" which is a nationally representative, multi-topic survey of 42,152 households in 1,503 villages and 971 urban neighborhoods across India for 2011-2012. The survey covers topics concerning health, education, employment, economic status, marriage, fertility, gender relations, social capital, village infrastructure, wage levels, and panchayat (village council) composition. It is a compilation of 14 data sets each comprising of a questionnaire concerning the aforementioned topics. The details for these datasets are provided in the appendix.
IHDS – II re-interviewed 83% of the households from the initial 2005 – 2006 (IHDS – I) survey, while still maintaining its diversity around urban and rural households across all the states and union territories of the countries. The questionnaires are similar across the two waves to enable comparisons over time. However, there are additions to the institutional modules to capture new programs and policies, which provide a quasi-experimental framework to test for their effectiveness. We propose leveraging these factors in accessing the variation in three primary topics mentioned below thus enabling us to gauge the development and the factors influencing it.
Our approach towards gaining insights from this dataset can be segmented into 3 parts:
1. Data cleaning and wrangling.
2. Explore data and build models using logistic regression and random forests to answer critical questions like -
    - *Women's LFPR (Labor Force Participation Rate)*
    - *Factors affecting Women's Employment*
    - *Income distribution and Inequality Gini coefficient)*
    - *Factors influencing Cultivated land*
3. Predict the factors influencing the human development index and growth rates utilizing logistic regression and random forests. Compare the pros and cons of the techniques utilized.

# *METHODS*

## *1 - Women's Labor Force Participation Rate(LFPR) and Factors Influencing Employment :*

After understanding and analyzing the dataset we thought there was an interesting scope to study women's LFPR and the gender disparity in India's growing employment sector.

Our Assumption: An increase in education is associated with an increase in women's employment.

Theories of human capital would imply that with more education, women acquire better skills and their earnings increase, resulting in higher labor force participation. However, it has been long observed that in India, women's education has an unusual relationship with labor force participation.
Cultural factors, such as norms restricting the mobility of women and structural factors, such as a lack of appropriate job opportunities for skilled women, play influential roles in determining the relationship between women's education and labor force participation in India. Measuring women's employment can be especially

tricky or challenging because often women are involved in part-time or seasonal jobs, or they could work from home, or they may participate in the labor market only in times of a family crisis

## Pre-Processing of data:

We took DS0003 dataset for eligible women and did the exploratory data analysis. The steps involved filtering the dataset for married women between the age of 25-59, dividing these women based on their education levels, filtering out any NA values for the level of education a woman has received. If a woman had an education level of none we categorized her as illiterate, class 1-4 as pre-primary, class 5-9 as primary and post-primary, class 10-12 as secondary with 12-14 as higher secondary and post-bachelors as a college graduate or higher. We then factored these results in increasing order of education levels from illiterate to pre-primary to further till college graduate or higher. We then implemented our first data exploratory analysis where we visualized working women by their education levels.

## Models:

The project proposal, Rmd code and this paper on github for your perusal.

### *Generalized linear regression model:*

We tried to fit a generalized linear regression model to perform logistic regression to predict women's employment based on a variety of factors all of which directly affect their employment status. We took variables such as spouse's education which directly correlated with the woman's education, their age, caste, religion, and other household income. Women tend to withdraw from the workforce as they age, while lower caste women tend to work more than upper caste women to support their household because they typically have lesser education and working opportunities in high paid jobs. We divided the dataset into 2 parts with the training dataset containing 80% of the data and test set containing 20% of the data. We converted the predictions gathered from the regressed model to match them with our response variable so that a confusion matrix could be built and applied the following formula:

```
fit1 <- glm(GR46 ~education + RSUNEARN + EW6 + SPED6 + ID13 + ID11 + GR48,
            data = women_par$train, family = binomial(link= "logit"))
```

Then afterwards we created a confusion matrix to get accuracy, sensitivity, specificity and subsequently the F-score.

### *Stepwise Logistic Regression:*

After running the fitting our logistic regression model we decided to use a step-wise regression approach to figure out variables with the most impact on our response variable. Forward selection begins with no variables selected (the null model). In the first step, it adds the most significant variable. At each subsequent step, it adds the most significant variable of those not in the model, until there are no variables that meet the criterion set by the user. The model while running set the dataset ran into a row mismatch error which had to be resolved

using the na.roughfix parameter from the randomForest library which intuitively imputes missing values which resolved the error and gave us insight into which variable made the model performance better.

*Random Forest Model:*

After assessing the performance of our logistic regression model, it was clear that the model fit was not optimal and we needed a better algorithm to predict our response variable vs the data we had.

Random Forest is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier. It runs efficiently on large databases. It can handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification.
It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing. It checked a lot of the criteria that were crucial for us

And thus, taking into consideration the various advantages that the random forest presented us and given the categorical/classification nature of the response variable combined with our inconsistent voluminous dataset, we decided to employ it in our case study in hopes of yielding better results.

```
women.rf1<-randomForest(GR46 ~EW6+ education+ SPED6+ SPRO5+ INCOME+RSUNEARN,
                        data = train_partition, na.action = na.roughfix, mtry=6,ntree=200)
```

Starting with a wide variety of predictor variables and using the insight we had gathered from running a stepwise regression model, we had an idea of the key influencing factors for our response variable and therefore that made it significantly easier to choose and prune our predictor variables of our randomForest.
After deciding on fitting our parameters on the random forest model, we observed an increase in the accuracy of predicting women's employment which gave us more confidence on tinkering and improving our model to push its efficacy.

## *Income Inequality -*

In economics, the Gini coefficient sometimes called Gini index, or Gini ratio is a measure of statistical dispersion intended to represent the income or wealth distribution of a nation's residents and is the most commonly used measure of inequality. A Gini coefficient of zero expresses perfect equality, where all values are the same i.e. where everyone has the same income. A Gini coefficient of 1 (or 100%) expresses maximal inequality among values i.e. for a large number of people, where only one person has all the income or consumption, and all others have none, the Gini coefficient will be very nearly one.
The Lorenz curve is a graphical representation of the distribution of income or of wealth. For income inequality, it is a graph on which the cumulative percentage of total national income is plotted against the cumulative percentage of the corresponding population. The extent to which the curve sags below a straight diagonal line indicates the degree of inequality of distribution.

We had to explore all the data sets to find out the variables related to income and filter NA values. After this, we loaded the "ineq" package so that we can pass the income value to Lc(Lorenz curve). Since we need the proportion of income and population for Lc, we calculated those and stored them in a data frame for both 2004-2005 and 2011-2012. We then plotted the curves and edited the labels, fonts, and colors to make the plots readable.

## *2 - Cultivated Land:*

From the dataset, it was seen that the value of the crop residue being sold was less than the value of the crop residue being wasted.

| Value | Amount in Rupees |
|---|---|
| Total amount which was spent on cultivating land | 191,309,291 |
| Total value of crop residue | 41,992,026 |
| Total value of crop residue sold | 5,791,598 |
| Total value of crop residue wasted | 36,200,428 |

Looking at these figures, we wanted to find insights to enable farmers to maximize their income. However, no data was available which would give us the farmers' income. So, we have used the area of land a farmer was able to cultivate, given the amount of money they were spending, as a metric to gauge how successfully they were able to optimize their spending. We have predicted the area of land cultivated based on the amount of money being spent on factors like fertilizers, pesticides, irrigation, hiring labor, buying tractors, tillers, and other farm equipment.

## Pre-Processing of data:

We plotted the distribution of land and saw that most farmers had plots less than 30 acres. We filtered our dataset to remove outliers (land area more than 30 acres). We have imputed the missing values using the roughfix attribute of Random Forest in R. We split the data into training and test set, with 10% of data in the test set.

## Model:

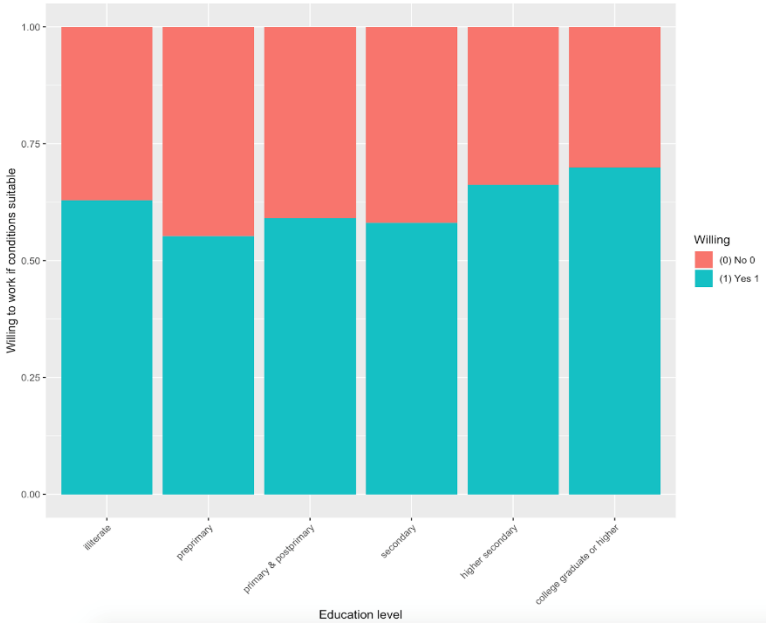
To decide which model to use, we plotted the response variable (area of land cultivated) vs each of the aforementioned variables. No linear relationship was found for any of the variables even after performing transformations like log, tanh, etc. So we employed random forest to build the model.

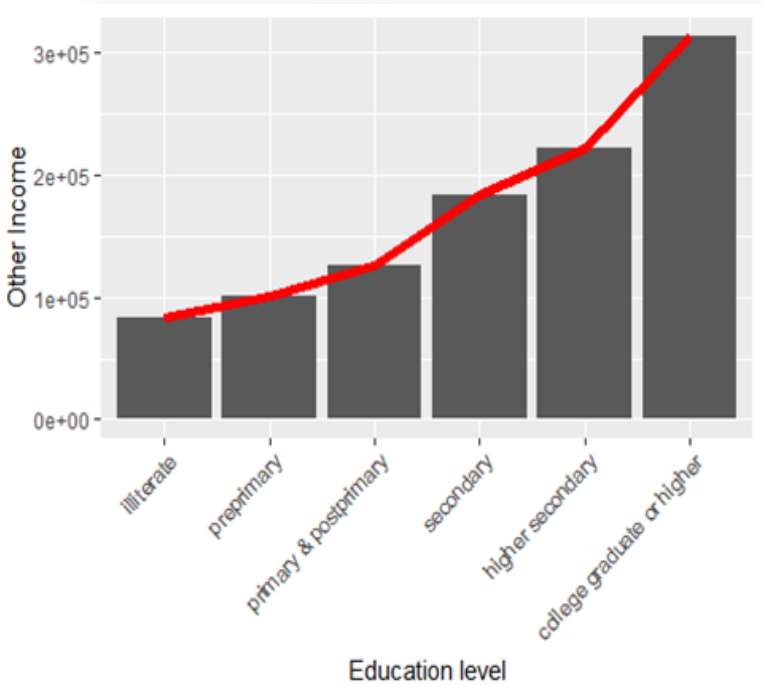
```
agri.rf1 <- randomForest(formula = cultivated_land ~ FM27B + FM29RS + FM30RS + FM31 + FM32 + FM34 + FM26A, data = agrarian, ntree = 200, subset = train, na.action = na.roughfix)
```

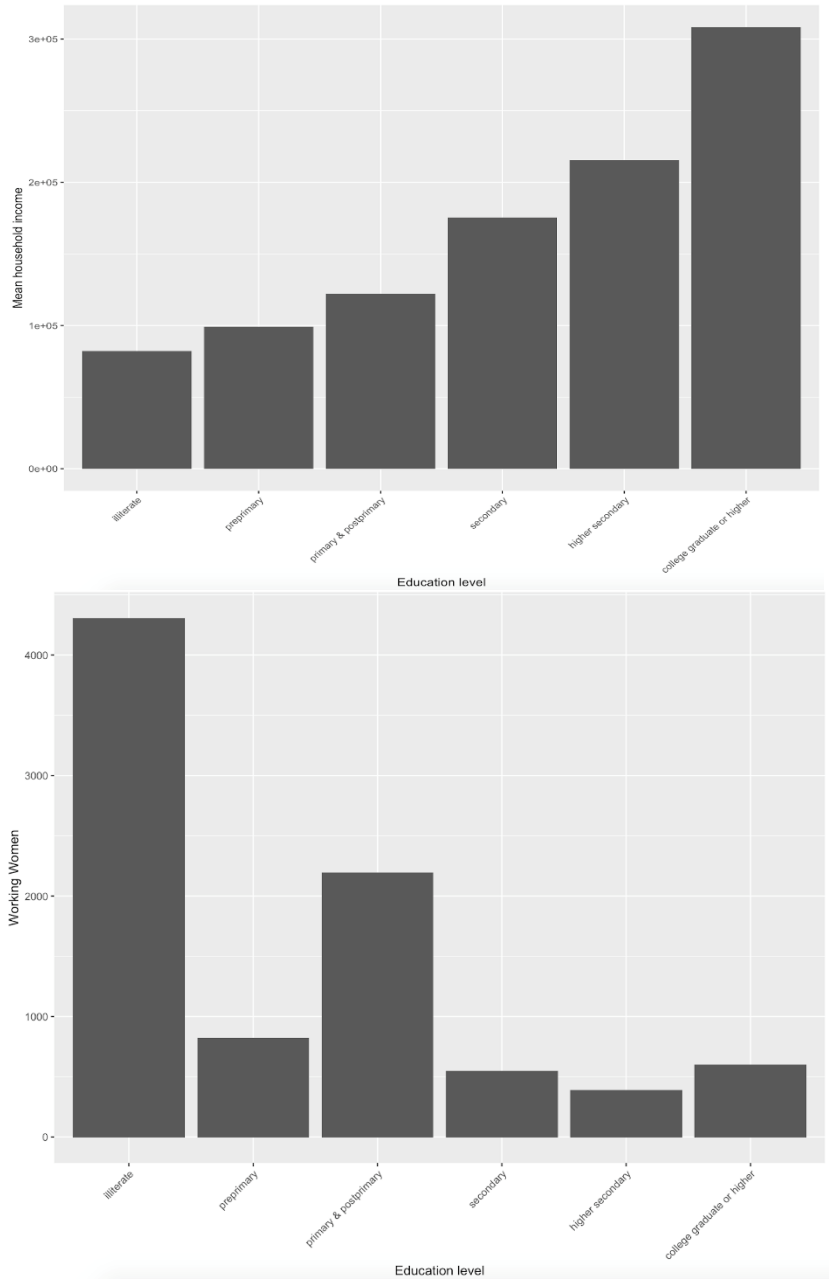
# *RESULTS*

## *Women's LFPR :*

We plotted count of women participating in labor force and found that with increasing education levels the count of working women decreases. We did not expect this trend and hence decided on exploring factors affecting this. A couple of which were other household income and willingness to work if conditions are suitable. We can see that higher educated women marry into richer families and although their willingness to work is higher than the rest it tells us that there are not enough suitable jobs for highly educated women.






## 1. *Logistic regression*

After applying logistic regression we saw that the accuracy of the model was only 67%. This was pretty low for a generalized linear model. So we explored random forest.

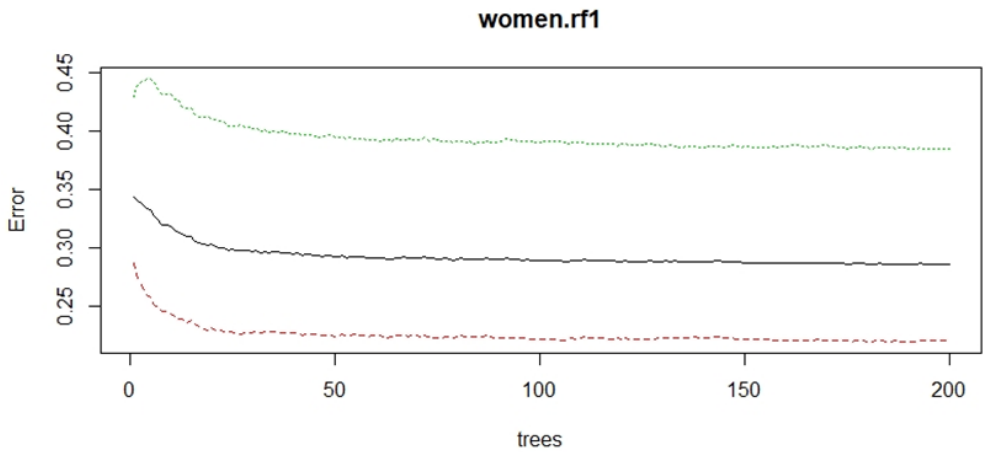
```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  (0) No 0 (1) Yes 1
##   (0) No 0      3365      1778
##   (1) Yes 1      319       883
##
##                Accuracy : 0.6695
##                  95% CI : (0.6578, 0.6811)
##     No Information Rate : 0.5806
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2654
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9134
##             Specificity : 0.3318
##          Pos Pred Value : 0.6543
##          Neg Pred Value : 0.7346
##              Prevalence : 0.5806
##          Detection Rate : 0.5303
##    Detection Prevalence : 0.8106
##       Balanced Accuracy : 0.6226
##
##        'Positive' Class : (0) No 0
##
```

```
cat("The F Score is",Cmatrix$byClass["F1"])
```

```
## The F Score is 0.7624334
```

## 2. *Random forest*

During optimizing our model, we ran several iterations with different permutation and combinations of predictor variables while also printing their individual importance values to ascertain their significance in the model. We also realised that the error rate plateaued after roughly 200 trees so the model was set to run according to that.




The confusion matrix was created and the information is displayed below.

```
Cmatrix1 <- confusionMatrix( p_cl,test_partition$GR46, dnn = c("Prediction",
"Reference"))
Cmatrix1
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  (0) No 0 (1) Yes 1
##   (0) No 0      2847      1011
##   (1) Yes 1      823      1641
##
##                Accuracy : 0.7199
##                  95% CI : (0.6985, 0.7211)
##     No Information Rate : 0.5805
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3984
##  Mcnemar's Test P-Value : 1.262e-05
##
##             Sensitivity : 0.7757
##             Specificity : 0.6188
##          Pos Pred Value : 0.7379
##          Neg Pred Value : 0.6660
##              Prevalence : 0.5805
##          Detection Rate : 0.4503
##    Detection Prevalence : 0.6102
##       Balanced Accuracy : 0.6973
##
##        'Positive' Class : (0) No 0
##
```

```
cat("The F Score is",Cmatrix1$byClass["F1"])
```

```
## The F Score is 0.7563762
```

```
importance(women.rf1)
```

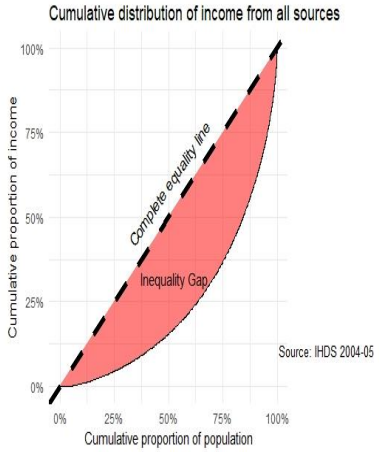
```
##           MeanDecreaseGini
## EW6               1511.9517
## education          633.4431
## SPED6             1589.5760
## SPRO5             1561.9494
## INCOME            3428.3759
## RSUNEARN          3621.4756
```

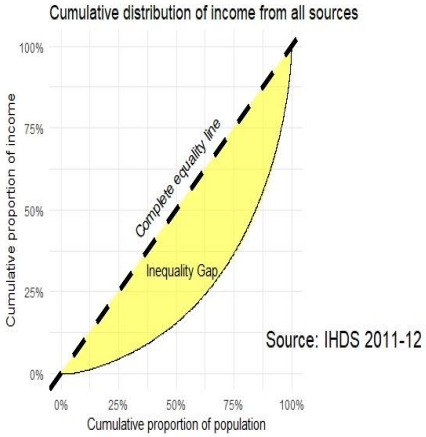
The importance values show a strong influence/correlation of women's employment with Total family income and other income sources. Which might suggest that the greater the income women's households have apart from their own earnings, the lower the chances of the woman being in the labor force.

## Income Inequality:

For this analysis, we used the 2004-2005 survey as well as the 2011-2012 data. Although the Gini index is calculated for a nation's income, we wanted to implement it on our data set and see the results. Below are the

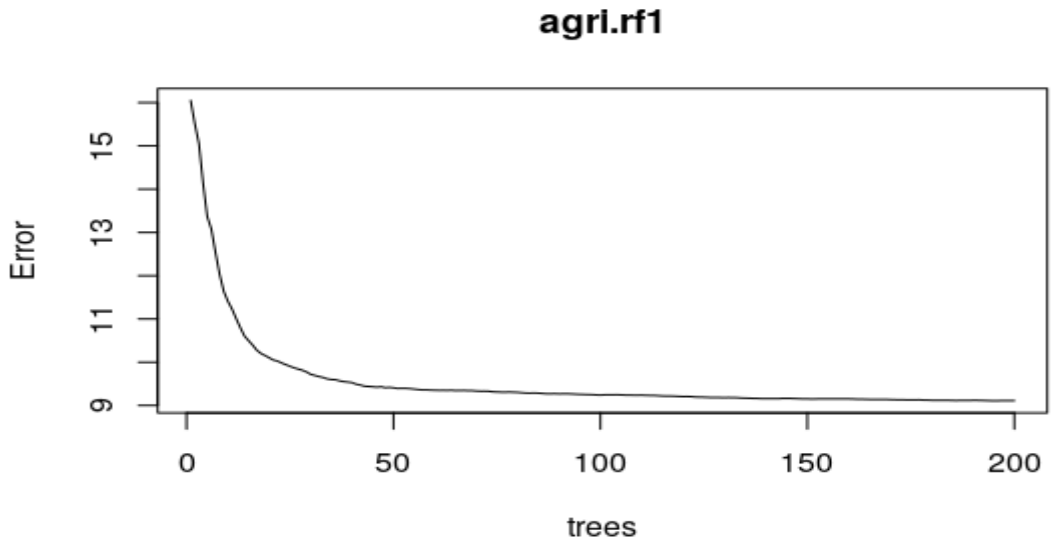
graphs for the same.
From our calculations, we see that the Gini coefficient in 2005 was 0.53 and 0.54 in 2012. There is a very marginal increase considering that our data frame does not cover the whole nation's population and income. Looking at real-world data, the Gini coefficient in 2005 was 0.45 and 0.51 in 2012.




## Cultivated Land:

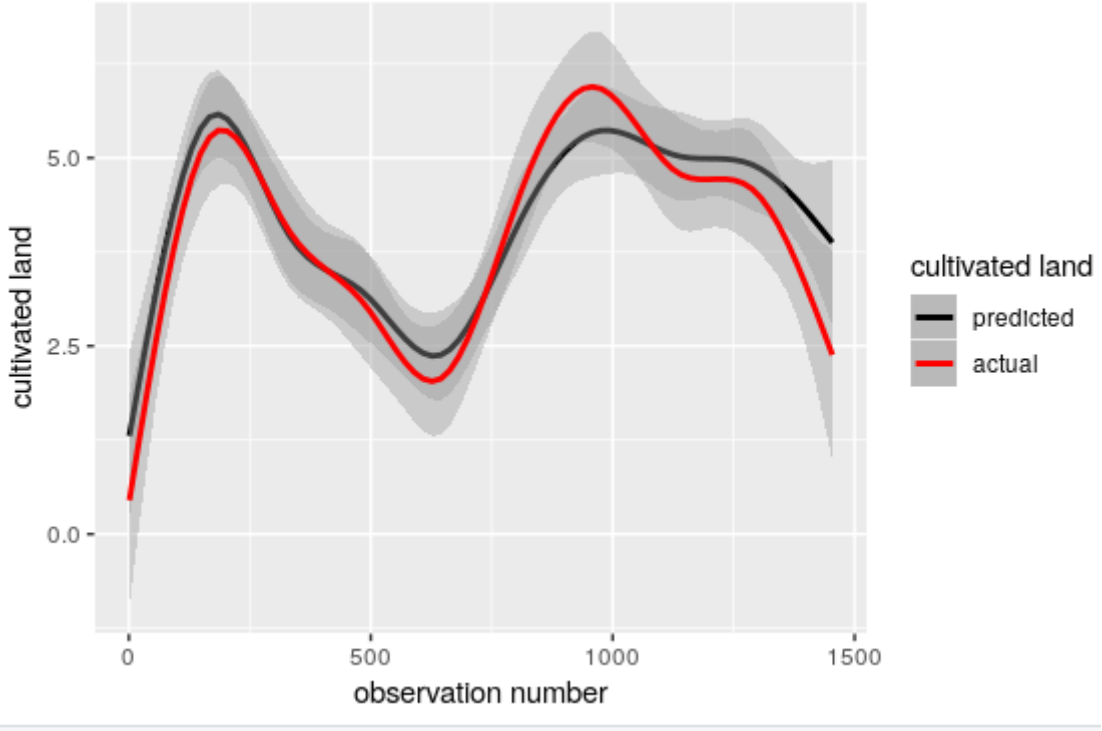
Random forest was used to predict the area of land cultivated. We saw that the error reduces very fast and plateaus beyond 200 trees. Thus we have kept the number of trees in our model to 200.



For the random forest model being fit, the following are the parameters

| Parameter | Value |
|---|---|
| Type of random forest | Regression |
| Number of trees | 200 |
| No. of variables tried at each split | 2 |
| Mean of squared residuals | 9.111223 |
| % Var explained | 57.58 |

We obtained the following plot to see how our model was performing on the test set.




We obtained residual plots of the fitted model vs each of the predictor variables and saw that the residues were randomly scattered indicating there was no bias.

## *CONCLUSIONS*

1. While fitting our model for predicting cultivated land, we saw that the money being spent on fertilizers, manure, and pesticides in the previous year were the most influential factors affecting the area of land a farmer was able to cultivate. This indicates that a farmer should spend more on fertilizers and hiring farm labor than other factors like irrigation, buying tractors, etc in order to cultivate more area of land.
2. Since we had memory constraints, we were not able to join data sets to answer cross-functional questions. To resolve this issue, a cloud computing service like AWS can be used to store the data. Another alternative can be to store the data as relational databases using MS SQL Server or MySQL and retrieve the data via R.
3. Since our data sets are more biased towards rural than urban areas and contain huge variations within the variables, we can use Lasso and Ridge techniques for normalization and regularization.
4. To improve the accuracy and efficiency of our models, advanced machine learning models like SVM and neural networks could be tested against the results of our models.

## *CONTRIBUTIONS*

Akshay Kulkarni - Women's Data Analysis (LFPR) and Modelling to predict Women employment (randomForest) and identifying its key influencing factors.

Anupam Gupta - Women's LFPR analysis and modelling (generalised linear model) with exploring gini index for income inequality.

Arpita Koundinya - Data exploration, cleaning and preparing it for the models.

Dhaval Mohandas - Modelling area of cultivated land, exploratory data analysis and conclusion for the same.

Saish Pai - Data exploration, cleaning and preparing it for the models. Identifying the machine learning model that could be used.

## *REFERENCES*


1. https://en.wikipedia.org/wiki/Gini_coefficient
2. https://en.wikipedia.org/wiki/Lorenz_curve
3. https://www.demographic-research.org/volumes/vol38/31/
4. https://onlinelibrary-wiley-com.ezproxy.neu.edu/doi/full/10.1111/j.1728-4457.2016.00149.x
5. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2723560

## *APPENDIX*

| Data Set | Variables | Observations |
|---|---|---|
| DS1 - Individual | 337 | 204,569 |
| DS2 - Household | 758 | 42,152 |
| DS3 - Eligible Women | 580 | 39,523 |
| DS4 - Birth History | 21 | 111,193 |
| DS5 - Medical Staff | 18 | 23,327 |
| DS6 - Medical Facilities | 158 | 4,447 |
| DS7 - Non Resident | 21 | 14,248 |
| DS8 - School Staff | 17 | 31,994 |
| DS9 - School Facilities | 141 | 4,267 |
| DS10 - Wage & Salary | 19 | 64,289 |
| DS11 - Tracking | 41 | 215,754 |
| DS12 - Village | 545 | 1410 |
| DS13 - Village Panchayat | 17 | 14,305 |
| DS 14 - Village Respondent | 10 | 11,024 |

## Cultivated Land Plots

Residual Plots