Targeting of social transfers: Are India's poor older people left behind?

Project Report for
ECON F241: Econometric Methods

Shreyash Bhardwaj: 2020A7PS2066H
Suraj R Nair: 2020A7PS0051H
Krishn Parasar: 2020A7PS2093H
Mandar Naphade: 2021B3A72734H
Aditya Jain: 2021B3A82573H

Birla Institute of Technology and Science, Pilani, Hyderabad Campus

Under the supervision of
Prof. Bheemeshwar Reddy A

Contents

1. Introduction

2. Approach
    a. Pre-Processing techniques used
    b. Analysis

3. Table for Descriptive Statistics

4. Interpretation of coefficients for the regressions given in the paper

5. Table for Regression Results

6. Logistic Regression - Model Suggestion

Introduction

The changing demographics, a large informal sector, and weakening family support for the elderly are significant factors contributing to old-age poverty in developing countries. As migration and declining fertility lead to a reduction in the traditional multi-generational household model that supported older people, those working in the informal sector are at increased risk of old-age poverty due to the lack of social protection coverage.

To address this issue, social pensions, in the form of cash transfers, have been introduced to provide income security for the elderly who lack social protection coverage. The Indian government implemented the National Old Age Pension Scheme in 1995 to support poor older people as part of the National Social Assistance Programme.

However, the targeting performance of the social pension scheme for poor older people remains an under-researched topic in India. Existing studies have limitations and do not focus on the effectiveness of the social pension scheme in reducing old-age poverty among the poor older people.

Therefore, this study focuses on evaluating the targeting performance of the social pension scheme for poor older people. The success of the social pension scheme in reducing old-age poverty depends on whether it reaches the poor older people, which needs to be assessed accurately.

Firstly, we try to understand the extent to which social pensions are reaching the poor elderly population, which is a crucial requirement to evaluate the effectiveness of social pension schemes in developing countries like India, which face similar targeting challenges.

Secondly, from a methodological standpoint, this study contributes to existing targeting research by comparing targeting performance indicators to a hypothetical random distribution of social pensions. Furthermore, by utilizing panel data to investigate the factors that affect access to social pension benefits, this study aims to minimize potential omitted variable bias.

Dataset used : The IHDS dataset by the National Council of Applied Economic Research and the University of Maryland (Desai et al., 2007, 2015).

Approach

- First we import the necessary packages and create a subset of the original dataset.
- Then we create a subset of the original dataset for training the model.

```
# CREATING SUBSET OF THE ORIGINAL DATASET FOR OUR USAGE
df2=OGData %>% select(HHID.y,IDHH,PERSONID,IDPERSON,ME13, ME7, MM3W, MM3M,FM3, FM4A,FM4B,FM4C,MM2W, MM2M,CO22, HHEDUC.y, NP
NADULTM.y, NADULTF.y, RC1B1, ID18A, RO5, IN11S1.x, ASSETS2005.x, RO4, RO6, URBAN2011.x, ID13.x, RO3, ED2,ID11.x,ME4,ME5,ME6
ASSETS.x,ED6,WS4,WS14,WS13,WKSALARY,WKAGLAB,WKNONAG,WKNREGA,WKANY5,SN2H1,SN2H2,FU1,SN2E1,SN2E2,SN2F1,SN2F2,SN2G1,SN2G2,SN2I
IN11S4.x,IN11S2.x,RC1B3)
```

The columns shown here are selected and stored in the df2 dataframe.

Preprocessing Techniques used -

- First step is to remove ‘useless’ columns from the dataset ,i.e , there are some variables such as time of interview, type of roof in household , animals kept in the house ,etc.
  Since these variables are not directly affecting the pension earned by an individual we remove these from the dataset.

- Next up we impute the missing values in the resulting dataset since these can interfere with our regression results . The

following code snippet is how we are imputing missing values.

```r
# Define a function to impute missing values with mode
impute_mode <- function(x) {
  if(is.numeric(x)) {
    mode_val <- names(sort(table(x[!is.na(x)]), decreasing = TRUE))[1]
    mode_val <- as.numeric(mode_val)
  } else {
    mode_val <- as.character(names(sort(table(x[!is.na(x)]), decreasing = TRUE))[1])
  }
  x[is.na(x)] <- mode_val
  return(x)
}

# Apply the impute_mode function to every column in the dataset
df2_imputed <- as.data.frame(lapply(df2, impute_mode))
```

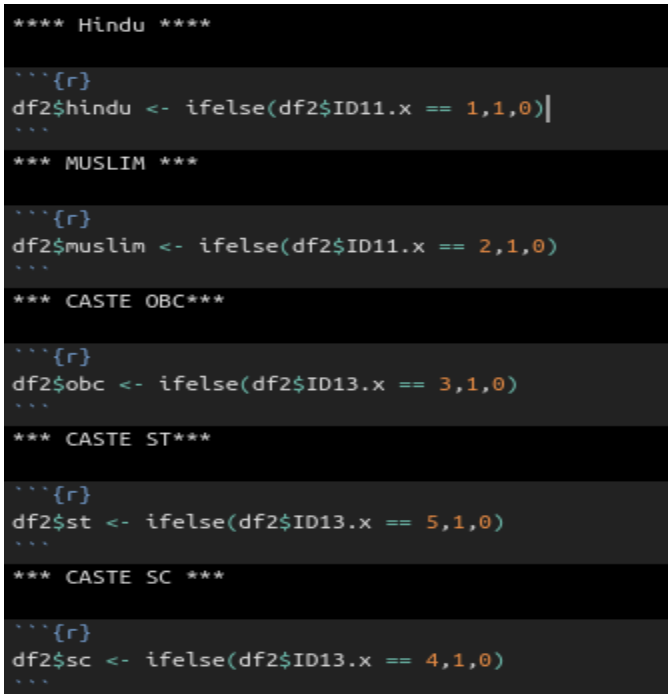
- After studying the dataset we create new variables that are needed for understanding the regression analysis. For example -

**** VARIABLE FOR FEMALE ****

```r
df2$female <- ifelse(df2$RO3 == 2,1,0)
```

Here we checked that RO3 was a variable that represents female category but the categorical variable took value 2 when the gender was female . But while doing regression we prefer the categorical variable to have values as 0 or 1 . So the code above does exactly this .

Another example is the categorical variable ID11 representing the caste took values 1 for ‘hindu’ , 2 for ‘muslim’ , 3 4 and 5 for ‘SC’ ‘ST’ and ‘OBC’ respectively . So we created 5 new variables each representing the above 5 categories where each column takes a value of 0 or 1 .



In similar fashion we created more new variables that were helpful in our regression analysis.

- The following for loop iterates over each numeric column and performs the following steps:

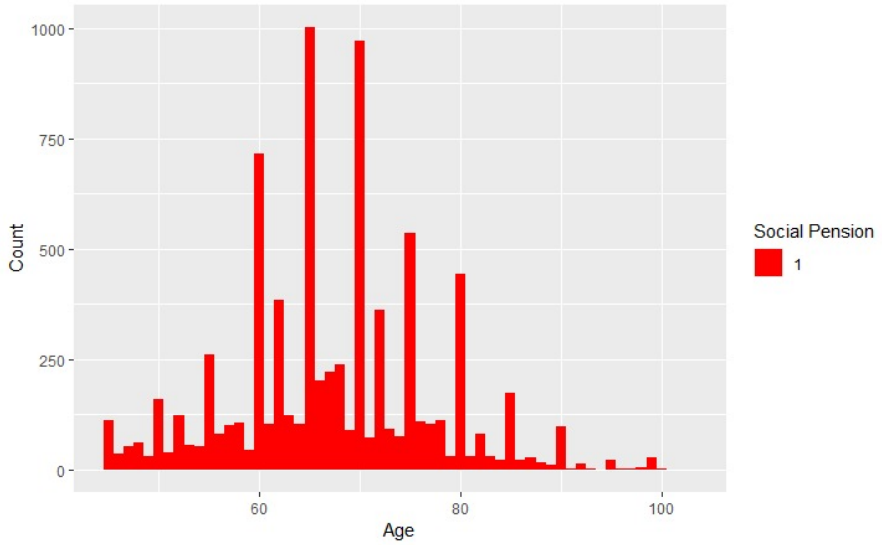
    - Creates a boxplot for the column using boxplot(my_data[[col]], horizontal = TRUE, main = col)

    - Identifies potential outliers using the interquartile range (IQR) method. This is done by calling boxplot.stats(my_data[[col]]), which returns the lower and upper whisker limits of the boxplot.Increments the row_count counter for each row that contains a potential outlier.

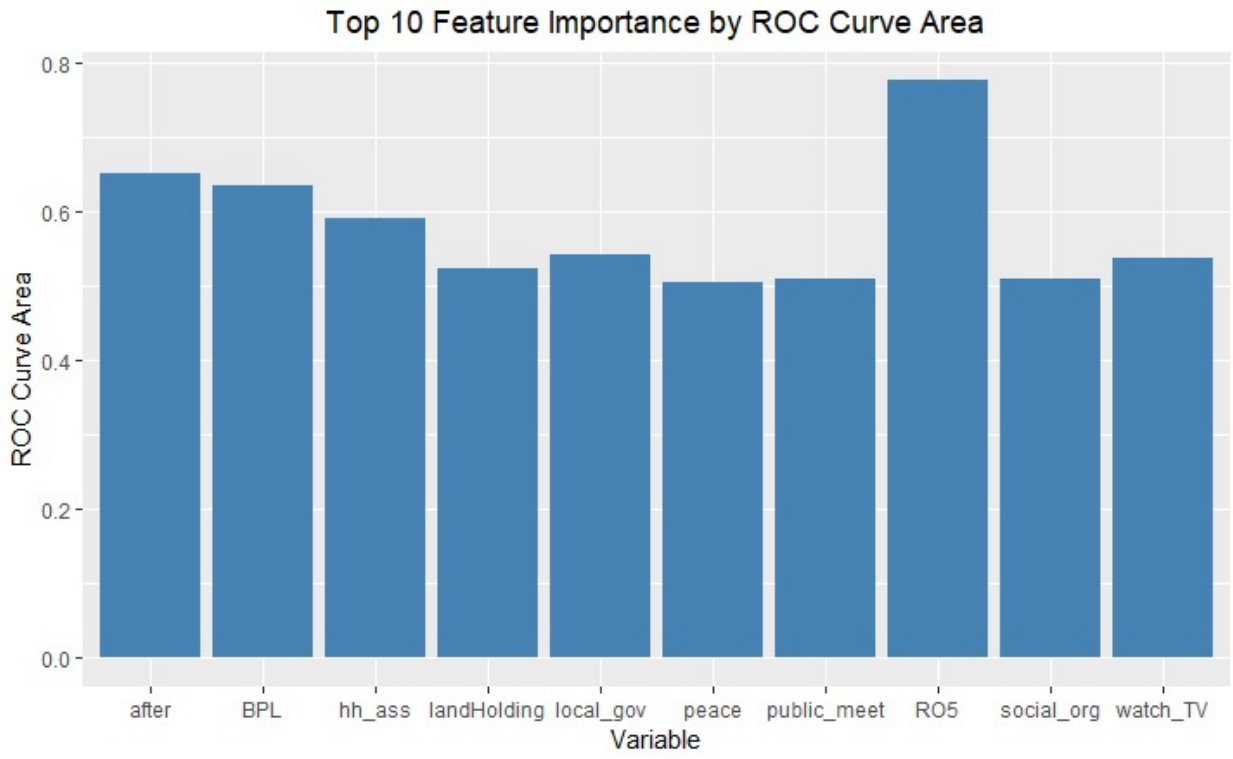
- Finally, the code identifies rows that appear in more than 5 numeric columns using outlier_rows <- which(row_count > 5) and removes these rows from the dataset using my_data <- my_data[-outlier_rows,].In summary, the code detects potential outliers in numeric columns using boxplots and the IQR method and removes rows that contain many outliers.

```r
# Identify numeric columns
num_cols <- sapply(my_data, is.numeric)

# Initialize counter for each row
row_count <- rep(0, nrow(my_data))

# Check for outliers in numeric columns
for (col in names(my_data)[num_cols]) {
  # Create boxplot
  invisible(boxplot(my_data[[col]], horizontal = TRUE, main = col))

  # Identify potential outliers
  stats <- boxplot.stats(my_data[[col]])
  col_outliers <- which(my_data[[col]] < stats$stats[1] - 1.5 * IQR(my_data[[col]]) | my_data[[col]] > stats$stats[5] + 1.5 * IQR(my_data[[col]]))

  # Increment counter for each outlier row
  row_count[col_outliers] <- row_count[col_outliers] + 1
}

# Identify rows that appear in more than 100 columns
outlier_rows <- which(row_count > 25)

# Remove outliers
my_data <- my_data[-outlier_rows, ]


my_data
```

Analysis -



This is a plot of age against number of people who get pension

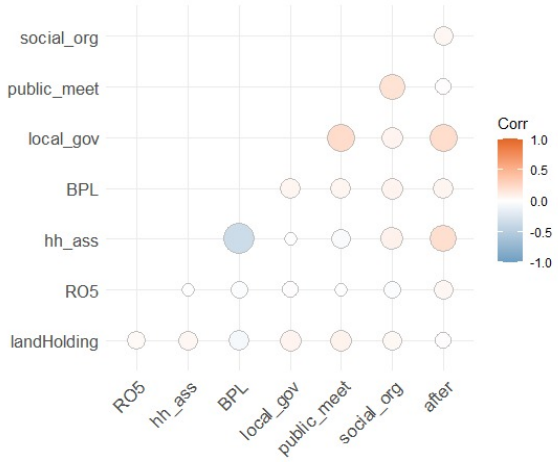


Out of the 20 important features in the research paper we have listed above the 10 most important features among them , based on ROC curve area.

- The code below was implemented to plot the correlation between the important features.

```r
library(ggcorrplot)
library(caret)
# create a correlation plot using ggplot2
cor_mat <- cor(final_df[, c("landHolding", "RO5", "hh_ass", "BPL", "local_gov", "public_meet", "social_org", "after")])
ggcorrplot(cor_mat, type = "lower", method = "circle", colors = c("#6D9EC1", "white", "#E46726"))

# check for high correlations (absolute values > 0.7)
high_cor <- findCorrelation(cor_mat, cutoff = 0.7, verbose = TRUE)

# print the names of the highly correlated variables
colnames(cor_mat)[high_cor]
```

And the plot obtained can be seen below -



Level of significance of regression coefficients -

```r
```{r}
tester<-  subset(final_df, select = c(RO5,BPL,hh_ass,landHolding,local_gov,public_meet,social_org,watch_TV,peace,after))
# dff <- subset(dff, select = -c(v000, caseid,bw4,bw5,bw6,bw7,bw8,bw9,bw10,bw11,bw12,bw13,bw15,v008a,b18,s412a))
# Calculate feature importance using ROC curve area as score
roc_imp <- filterVarImp(x =tester[,1:ncol(tester)], y = final_df$socialPen,
                        nonpara = TRUE, func = roc.area)

# Sort the score in decreasing order
roc_imp <- data.frame(cbind(variable = rownames(roc_imp), score = roc_imp[,1]))
roc_imp$score <- as.double(roc_imp$score)
roc_imp[order(roc_imp$score,decreasing = TRUE),]
```
```

Description: df [10 × 2]

	variable <chr>	score <dbl>
1	RO5	0.7762338
10	after	0.6511967
2	BPL	0.6346512
3	hh_ass	0.5915470
5	local_gov	0.5427705
8	watch_TV	0.5367605
4	landHolding	0.5230921
6	public_meet	0.5094032
7	social_org	0.5088844
9	peace	0.5055206

1-10 of 10 rows

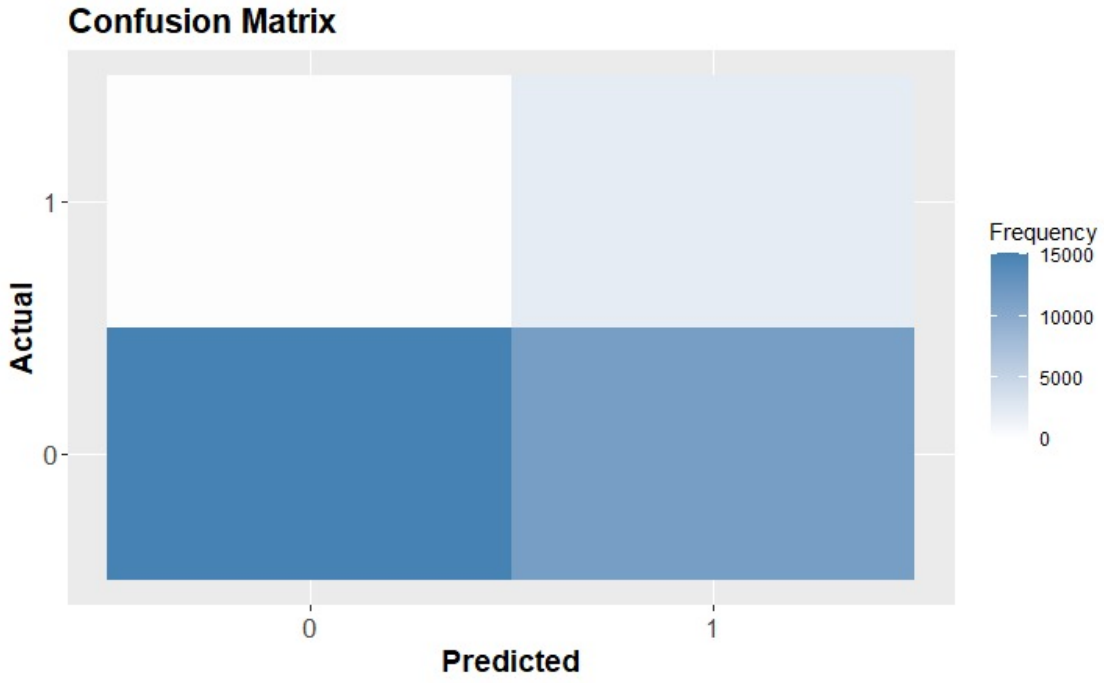
These were the values obtained from our regression model depicting level of significance.

RMSE and R^2 Values -

```r
# Calculate mean squared error
mse <- mean((predictions -  prediction_mean)^2)
cat("MSE:", mse, "\n")

# Calculate R-squared
ss_res <- sum((predictions - test$socialPen)^2)
ss_tot <- sum((test$socialPen - mean(test$socialPen))^2)
r_squared <- 1 - (ss_res / ss_tot)
cat("R-squared:", r_squared, "\n")
```

```
MSE: 0.009619183
R-squared: 0.1300831
```

These values obtained are very similar to those seen in the research paper .



From the above confusion matrix we can see that the predicted values which correspond to the actual values have a high frequency.

Descriptive Statistics for Variables -

	IHDS 2004-05				IHDS 2011-12				
	mean	**sd**	**min**	**max**	**mean**	**sd**	**min**	**max**	**Variable category**
Social pension	0.033696	0.180449	0	1	0.12621	0.332085	0	1	Dependent Variable
Age	57.35317	10.19083	45	116	58.3808	10.64868	0	99	Independent Variables
BPL card	0.336300	0.472448	0	1	0.39635	0.489144	0	1	
Household assets	12.97429	6.353467	0	30	15.7197	6.228551	0	30	
Land holding	2.433287	7.169449	0	360	2.53349	5.571165	-6	400	
Local government connection	0.108237	0.310683	0	1	0.28194	0.449946	0	1	
Public meeting	0.303314	0.459694	0	1	0.31136	0.463054	0	1	
Social organization	0.369087	0.482562	0	1	0.42218	0.493912	0	1	
Watching TV	0.440364	0.496436	0	1	0.54736	0.497756	0	1	Control Variables
Reading newspaper	0.207574	0.405574	0	1	0.20759	0.405582	0	1	
Literate	0.490821	0.499921	0	1	0.51702	0.499715	0	1	
Education	3.775156	4.688516	-1	15	4.1003	4.813647	0	16	
Highest adult education in household	7.807	5.059	0	15	8.703	5.095	0	16	
Working	0.990077	0.099117	0	1	0.51914	0.499638	0	1	
Permanent job in household	0.180980	0.385006	0	1	0.07313	0.260344	0	1	
Electrificatio n rate	0.200492	0.400373	0	1	0.89615	0.305070	0	1	
Village collaboration	0.582055	0.493226	0	1	0.72943	0.444260	0	1	

rate									
Peaceful village rate	0.545605	0.497921	0	1	0.59159	0.491545	0	1	Control Variables
Head of household	0.512375	0.499852	0	1	0.51697	0.499716	0	1	
Widow	0.202224	0.401663	0	1	0.21299	0.409421	0	1	
Household size	6.384	3.252	1	38	5.964	2.944	1	33	
Number of adults	3.509379	1.610726	1	18	3.55701	1.587877	0	18	
Urban	0.669715	0.470320	0	1	0.35083	0.477233	0	1	
Other backward castes	0.347133	0.476063	0	1	0.4083	0.491523	0	1	Time-invariant Characteristics
Scheduled castes	0.179227	0.383546	0	1	0.18717	0.390054	0	1	
Female	0.488268	0.499867	0	1	0.50999	0.499905	0	1	
Hindu	0.856245	0.350844	0	1	0.8223	0.382259	0	1	
Muslim	0.101866	0.302476	0	1	0.10572	0.307476	0	1	
Scheduled tribes	0.07414149	0.2620039	0	1	0.07765	0.267623	0	1	
Asset poor	0.397567	0.489400	0	1	0.38697	0.487061	0	1	Asset poverty measure

Interpretation of the coefficients-

1) LPM with the baseline specification presented below.

	Linear probability model with individual fixed effects: 2004–05 to 2011–12		
variables	**coefficient**	**standard error**	**Level of significance (t-value)**
After	0.0885	0.0021	41.819
Age	0.0070	0.0001	73.402
BPL Card	0.0686	0.0022	30.577
Household Assets	-0.0034	0.0002	-19.597
Land holdings	-0.0010	0.0002	-6.618
Local government connection	0.0111	0.0026	4.218
Public meeting	0.0010	0.0023	0.433
Social organization	-0.0079	0.0021	-3.790
Observations	96536		
Number of id	48268		
Adjusted within R-squared	0.125		

From the table we can interpret the following -

- If the person has the pension scheme in 2011 then there are 8.85% more chances that the person gets pension.
- The coefficient of age tells that for every unit increase in age pension increases by 0.7%.
- Similarly if a person has a BPL card then that person increases his chance of getting pension by 6.8%
- Note that the effect of land ownings is negligible as is seen from the table.
- On the other hand Local government and public meetings have a positive influence on target variable whereas social organization

and household assets have a small negative influence on the chances of getting a pension.

- Lastly , local government connections increase the chances of getting pension by 1%

From here we can conclude that owning a BPL card is a big factor that influences the chances of getting a pension and hence it directly supports the existing literature which recommends a reform of the allocation of BPL cards and suggests alternative targeting approaches for social pensions such as the use of clear exclusion criteria that at least prevent clearly non-poor older people from accessing social benefits targeted at the poor and facilitate access to social pensions for the poor older people.

2) Including interaction with time dummy variable

Variables	Coefficients	Standard Error
BPL card	-0.032	0.01
After X BPL card	0.145478	0.012
Local government connection	0.003453	0.014
After X local government connection	0.025093	0.018
Public meeting	0.006939	0.011
After X public meeting	0.003765	0.014
Social organization	-0.028652	0.009
After X social organization	-0.003961	0.012

We can draw the following inferences from the above table -

- Prior to the reform, local government officials were requested to select individuals for the national social pension scheme based on the destitution criterion. Accordingly the results show that if an individual lives in a household that holds a BPL card in 2004–05, his or her likelihood of gaining access to social pensions is reduced by 3.0 percentage points

- In 2011–2012, for an individual who lives in a household that holds a BPL card, the likelihood of receiving a social pension increases by 14.5 percentage indicating that the centrally reformed eligibility criterion was implemented (at least to some extent) by the state governments in panchayats and Municipalities.

- Direct connections with local government officials gained importance over time. In 2004–05 the relevant coefficient is insignificant but the estimation of the marginal effects for 2011–12 indicates that living in a household that has a connection to the local government is associated with a 2.5 percent higher likelihood of receiving social pensions.

- Lastly , we can see that public meeting has a small positive influence in the likelihood of receiving pension and on the other hand social organization has a small negative influence on the target variable which diminishes even further after the pension reforms of 2011.

3) Including interaction with time and assetPoor variable

Variabes	Coeffiicients	Standard Errors
BPL Card	-0.032	0.013
BPL card X after	0.035	0.016
BPL card X asset poor	0.0127056	0.021
BPL card X after X asset poor	0.1092667	0.027
Local government connection	0.006	0.015
Local government connection X after	0.0216704	0.019
Local government connection X asset poor	0.0065993	0.032
Local government connection X after X asset poor	0.0027878	0.04
Public meeting	-0.0031567	0.011
Public meeting X after	0.01183	0.015
Public meeting X asset poor	-0.011648	0.023
Public meeting X after X asset poor	-0.0080511	0.031
Social organization	-0.0327048	0.01
Social organization X after	-0.0082471	0.013
Social organization X asset poor	-0.02046	0.019
Social organization X after X asset poor	0.0013179	0.027

- The above two results potentially mask heterogeneity in the factors playing a role for older people from poor and non-poor households.

- To examine the heterogeneity between these two groups for access to social pension benefits before and after the reform, I include triple interaction terms of the time dummy, the variables of interest and the dummy for living in an asset poor household.

- As explained before, this approach is preferable to using a dummy variable for being poor based on comparing consumption expenditures to the Tendulkar poverty line since the latter are directly affected by the social pension income.

- After the reform, BPL card holding is relevant for individuals living in asset poor and asset non-poor households. For individuals living in asset poor households, BPL card holding is associated with a higher likelihood of receiving social pensions.

- Further, the effect of local government connections on social pension receipt seems to be primarily driven by individuals living in asset non-poor households who are also in general more likely to have better connections to the local government.

Logistic Regression - Model Suggestion

We tried various models on our dataset and we have reached the conclusion that a better model to be suggested for better and stronger model for prediction of "Targeting for Social Pension among Elderly people" would be :

Logistic Regression

Introduction :

Logistic regression is a popular statistical method used to model the relationship between a binary response variable and one or more predictor variables. It estimates the probability of the occurrence of the response variable based on the values of the predictors. The logit function is used to transform the probability values to a continuous range.

Model hyperparameters and training :

```r
# Randomly undersample the majority class
set.seed(123)
majority <- subset(training, socialPen == 0)
minority <- subset(training, socialPen == 1)
n_majority <- nrow(majority)
undersampled_majority <- majority[sample(n_majority, size = nrow(minority)), ]

# Combine the undersampled majority class and the minority class
balanced_training <- rbind(minority, undersampled_majority)

# Train the model on the balanced data
ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 1)

lambda_grid <- expand.grid(alpha = 1, lambda = seq(0.12, 0.15, by = 0.005))

model <- train(socialPen ~ .,
               data = balanced_training,
               method = "glmnet",
               family = "binomial",
               preProc = c("center", "scale"),
               trControl = ctrl,
               tuneGrid = lambda_grid)
```

Model specifications after training

```
glmnet

1122 samples
  75 predictor
   2 classes: '0', '1'

Pre-processing: centered (75), scaled (75)
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 1010, 1010, 1009, 1010, 1010, 1009, ...
Resampling results across tuning parameters:

  lambda  Accuracy   Kappa
  0.120   0.7674384  0.5348849
  0.125   0.7611884  0.5223644
  0.130   0.7611884  0.5223644
  0.135   0.7585256  0.5170530
  0.140   0.7576248  0.5152481
  0.145   0.7531685  0.5063283
  0.150   0.7513906  0.5027622

Tuning parameter 'alpha' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were alpha = 1 and lambda = 0.12.
```

Results :

```
```{r}
confusionMatrix(predictions, as.factor(testing$socialPen))
```

Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 2125   57
         1 530 183

               Accuracy : 0.7972
                 95% CI : (0.7821, 0.8117)
    No Information Rate : 0.9171
    P-Value [Acc > NIR] : 1

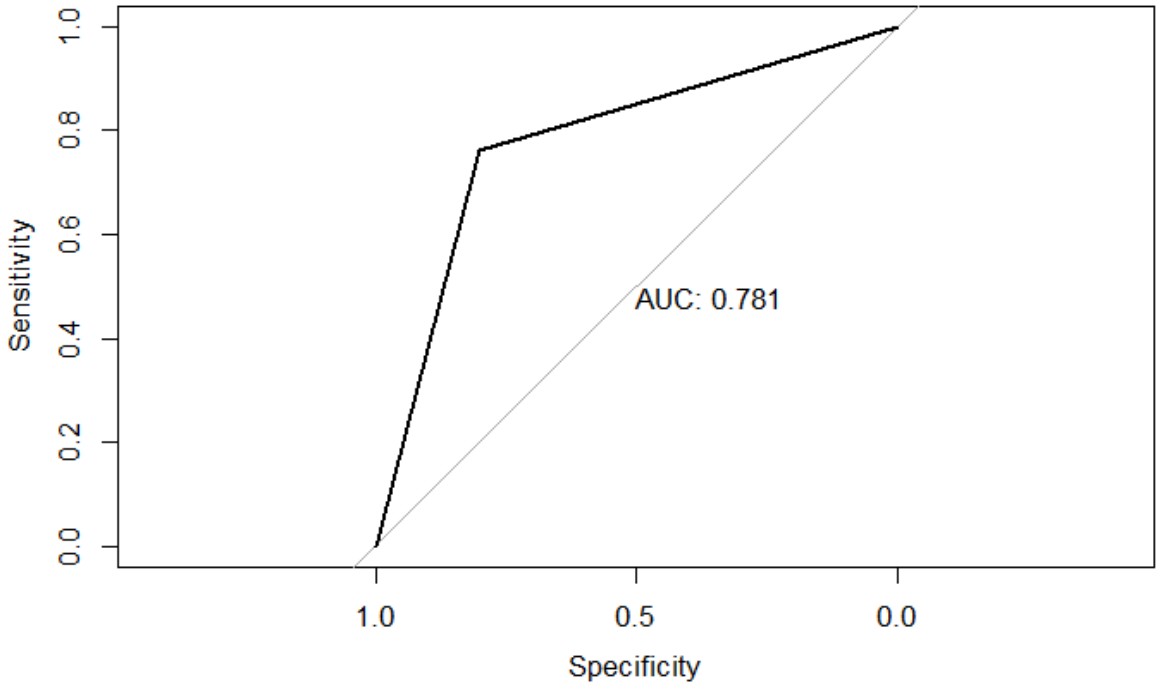
                  Kappa : 0.2968

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8004
            Specificity : 0.7625
         Pos Pred Value : 0.9739
         Neg Pred Value : 0.2567
             Prevalence : 0.9171
         Detection Rate : 0.7340
   Detection Prevalence : 0.7537
      Balanced Accuracy : 0.7814

       'Positive' Class : 0
```

AUC curve for this model :



As we can see , the accuracy has significantly increased to 79.72% from 59.68% of LinReg Model.
And the model specificity and sensitivity has also improved.

Explanation for better performance :

Logistic regression performs better than linear regression in predicting binary outcomes because it models the relationship between the predictor variables and the probability of the binary outcome, rather than the outcome itself. Linear regression assumes a linear relationship between the predictors and the outcome variable, which is not appropriate for binary outcomes. Logistic regression models the probability of the binary outcome using a logistic or sigmoid function,

which constrains the probability values between 0 and 1. This allows for a better fit to the data and improved predictive accuracy. Additionally, logistic regression outputs probabilities, which can be easily converted to class labels using a threshold value, providing more nuanced and interpretable results than linear regression.