Poverty Level Prediction for Households

Chiranth J - PES1201701438
C Diya - PES1201700246
Namrata R - PES1201700921

Problem Statement

Prediction of poverty level on households based on socio-economic factors and classify them into 4 distinct categories.

Not Vulnerable

Mildly Vulnerable

Vulnerable

Extreme Poverty

Our Dataset

- Data collected from Costa Rican households
- Records a family’s observable household attributes.
- Proxy Means Test (PMT) is used for base income qualification.
- The training data contains 143 columns and 9557 rows
- Each row represents a member of a household and their corresponding attributes

Data Preprocessing

Data cleaning, Transformation and Reduction

Inconsistency in the dataset

All members of a household to be classified by the same label

```
#checking consistency of data, if all members of a given household belong to same target
#split(train, with(train, interaction(idhogar)), drop = TRUE)

#to handle this inconsistency, all other members will have same target value as the head
new_train<-train %>%
group_by(idhogar)%>%
arrange(desc(parentesco1),.by_group = TRUE)%>%
mutate(Target=Target[1])
```

Handling non numeric attributes

	dependency <fctr>	SQBdependency <dbl>
1	no	0.0000000
2	8	64.0000000
3	8	64.0000000
4	yes	1.0000000
5	yes	1.0000000

- Replace all “yes” with 1 and “no” with 0
- Uniquely identifying attributes: id and idhogar remain unchanged

Handling missing data

	names <fctr>	count <dbl>
rez_esc	rez_esc	7928
v18q1	v18q1	7342
v2a1	v2a1	6860
meaneduc	meaneduc	5
SQBmeaned	SQBmeaned	5

5 rows

rez_esc(Years behind in school)

- Setting “0” value for age groups other than 7 to 18
- Within the age group, replaced by the mean

V18q1 (number of tablets owned by a family)

- The missing data points take value 0

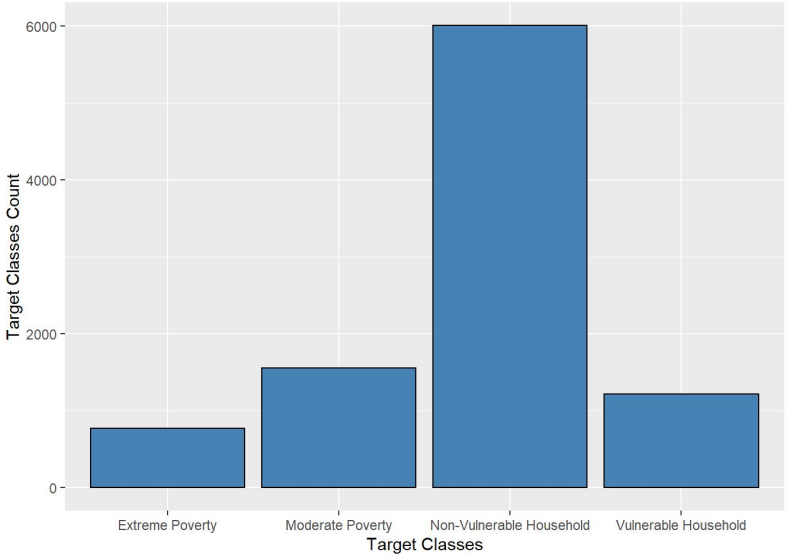
v2a1(Monthly rent payment)

- Replaced by 0 if head is the owner,else the mean rent

meaneduc(average years of education)

- Grouping the members based on household, replacing by mean
- SQBmeaned replaced with squared value of meaneduc

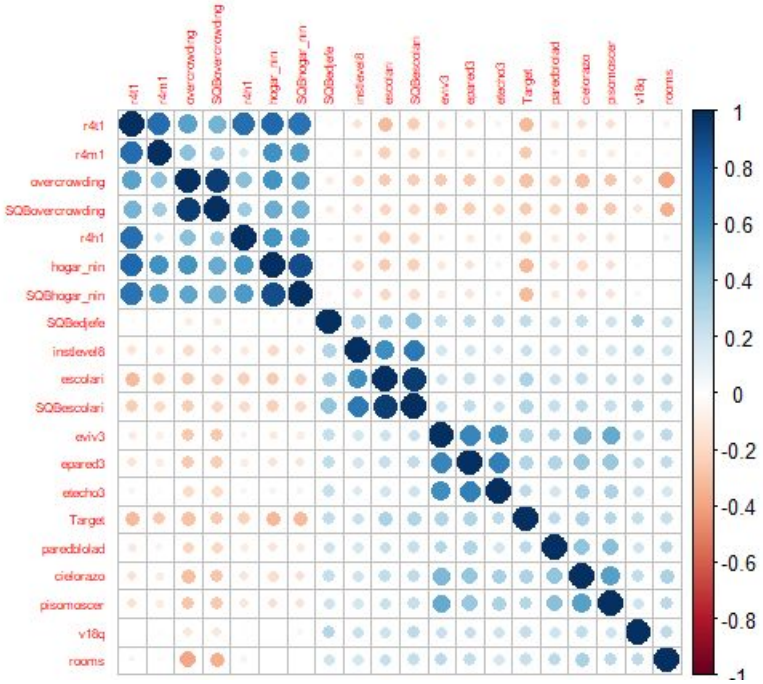
Upsampling



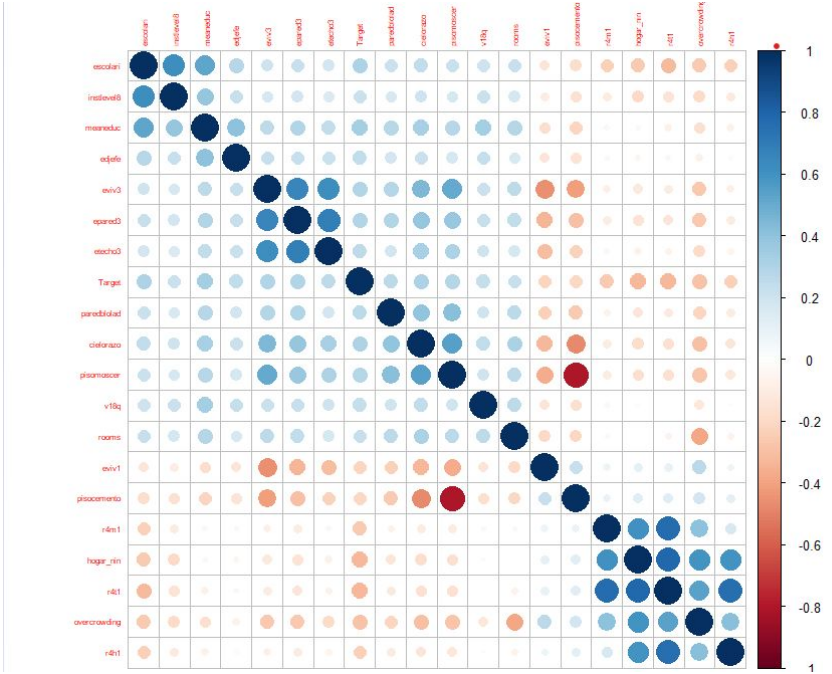
Skew in values for category 4 households

Upsampling of low frequency classes to make class size difference the same.

Exploratory Data Analysis

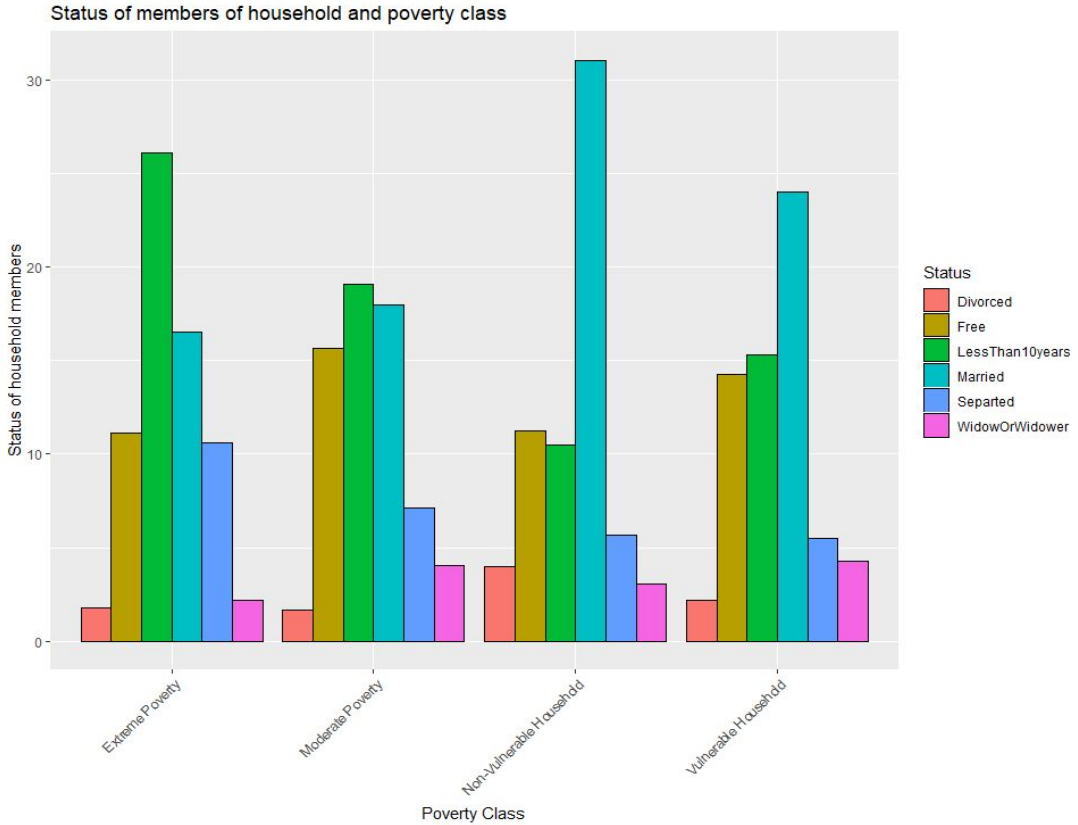


Before cleaning and processing



After cleaning and preprocessing

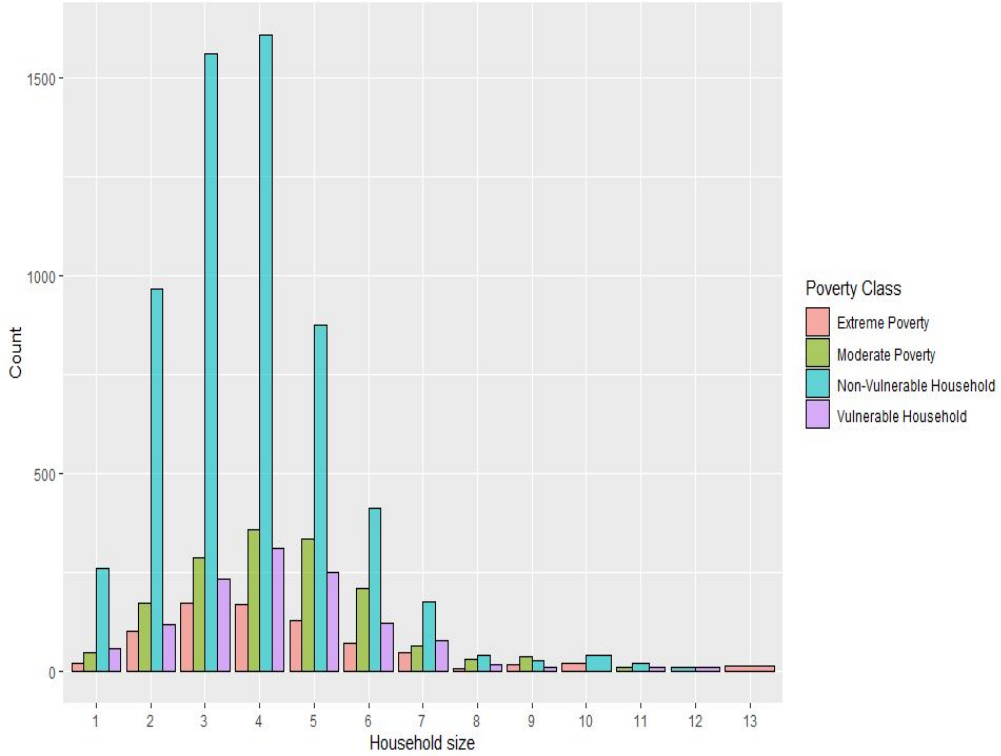
Exploratory Data Analysis



Inference:
The graph shows that the number of children(less than 19) does have a significant impact on the poverty level of each household.

The extreme poverty class has the highest number of children as compared to the other others.

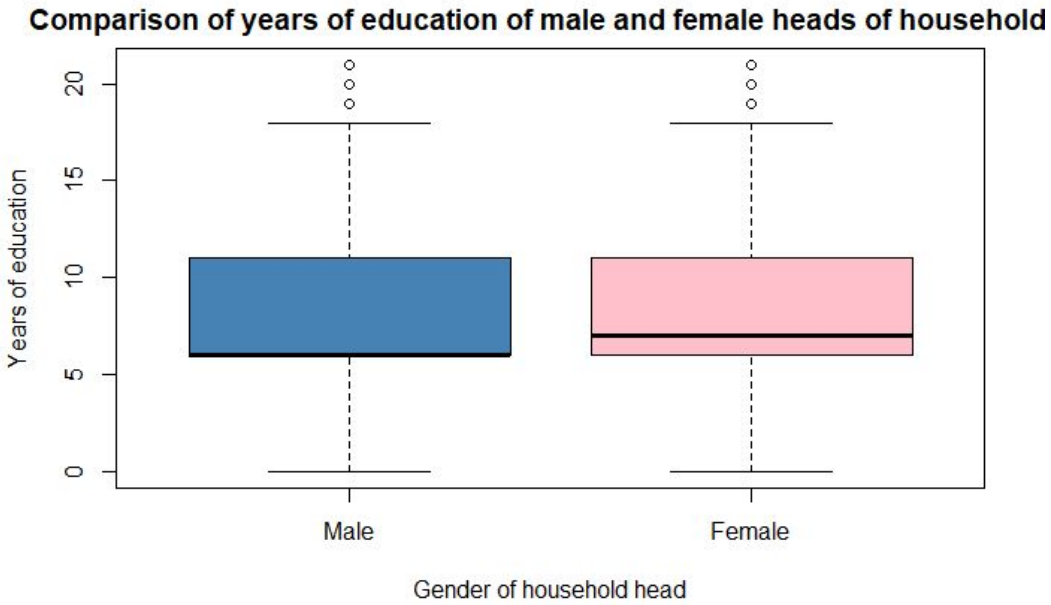
Exploratory Data Analysis



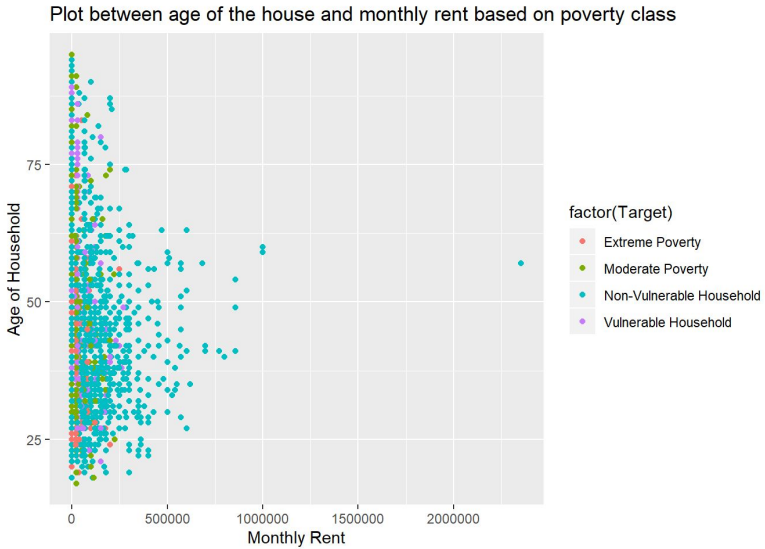
Inference:
This graph shows that a household size of three,four and five is common across the poverty classes.

There is evidence that Class 4(non -vulnerable) data as the count is significantly higher than others. Thus, data balancing methods would be required.

Exploratory Data Analysis

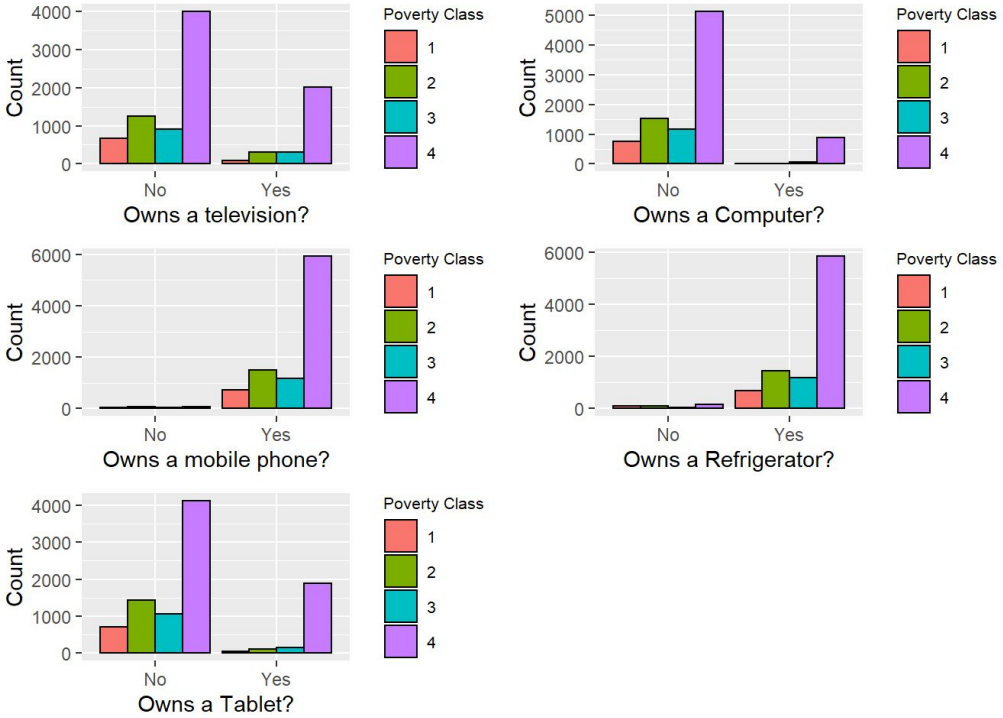


It can be inferred that the average years of education for female household head is significantly lesser than male household head.



This graph helps understand the distribution of the rent paid depending on the age per poverty class. Outliers can also be seen for the monthly rent.

Exploratory Data Analysis



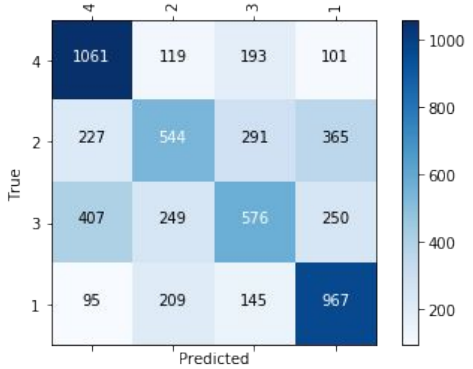
Visualizations pertaining to the households was also made.

It can be inferred that a mobile phone, refrigerator and television may be owned by households of all poverty classes.

However, owning a tablet or a computer in household other than Class non-vulnerable is not common.

Modelling Behaviour

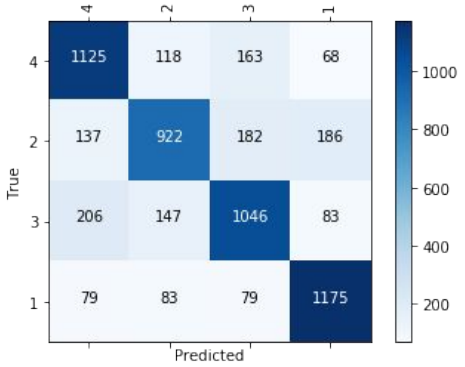
Ridge Regression



Confusion matrix for Ridge Classifier model

F1 Score	0.53
Precision	0.53
Recall	0.54

XGBoost Classifier

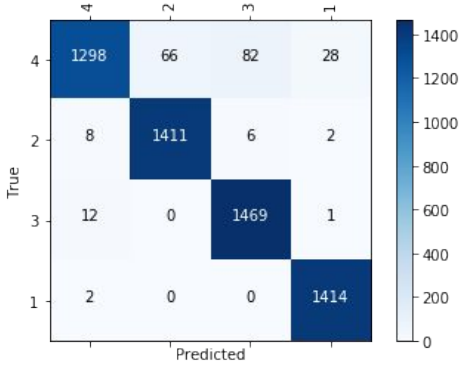


Confusion matrix for XGBoost Classifier model

F1 score	0.73
Precision	0.73
Recall	0.73

Modelling Behaviour

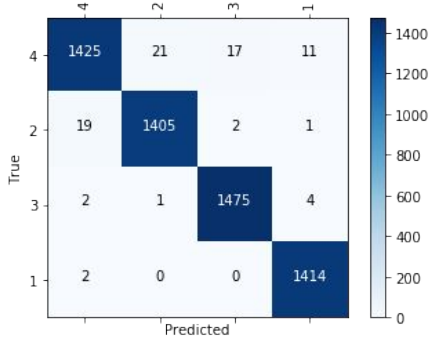
Decision Trees



Confusion matrix for Decision Tree model

F1 Score	0.96
Precision	0.96
Recall	0.96

Random Forest

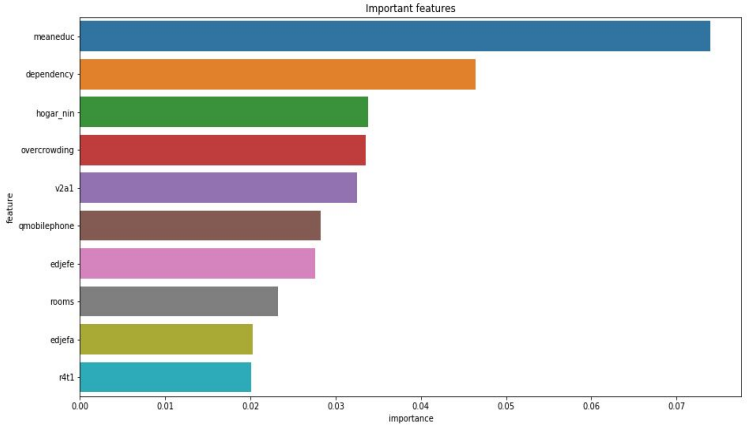


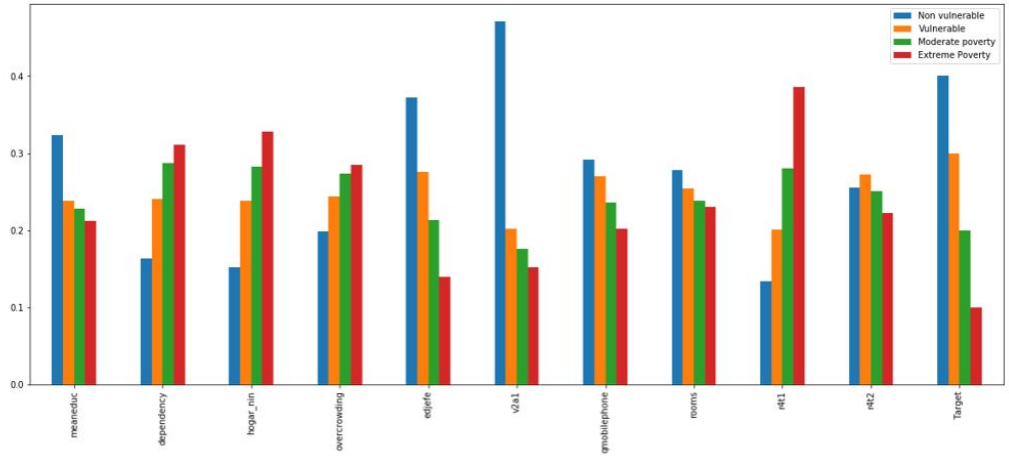
Confusion matrix for Random Forest Classifier model

F1 score	0.98
Precision	0.98
Recall	0.98

Analysis and Conclusions

Random forest gives us the best results with regard to our dataset and its schema





Plot shows top 10 attributes contributing to each poverty level

Plot shows the mean values of the top 10 attributes for a given poverty level