TAG Survey Report: Predicting Income

Vatsal Bajaj

September 21, 2020

Abstract

This report highlights my attempts at predicting income for households in the TAG survey (2018). We attempt to predict income quartiles, deciles and percentiles based on household assets, demographic characteristics and household education levels based on the IHDS survey (2011-2012). We are not able to predict income at any level for a sophisticated accuracy score. For example, in order to predict income deciles, the best training accuracy we get is 93.25%. However, the best 5-fold cross validation accuracy we get is 22.1%. Despite such disappointing results, I believe that this prediction problem is possible to solve. I believe this based on the variation of household asset ownership across different income groups.

1 Introduction

During the initial analysis of the Teenage Girl Survey (TAG), we noticed that the survey was missing Income Data on households. Without Income Data, it is difficult to conduct fruitful econometric analysis. Thus, this report highlights my attempts at predicting Income based on the Data we have in TAG. In order to do this, we use the Indian Household Development Survey II (IHDS II) conducted in 2012 to predict Income of households.

In order to predict household Income Quartiles, Deciles and Percentiles, we attempted various classification strategies to classify households into 4, 10 and 100 bins respectively based on their Income. All of our current analysis is based around the IHDS II Survey which contains Household Income Data.

2 Description of Data

The IHDS II survey consists on 42,152 observations. The TAG survey, however, only looks at households with teenage girls. The TAG survey consists of 61,672 observations. Thus, we choose a sampling frame of households in TAG which have one or more teenage girls. Our fixed sampling frame consists of 10,124 observations. The graph below shows how the sampling frame drastically reduces as we look at households with teenage girls. Here, we have filtered the data frame according to 'nteenf' column.

Note that this is not the exact same sampling frame as TAG since IHDS classifies teenagers to be in the age range of 15-20 while TAG classifies teenagers to be in the age range 13-20. Thus, further research into this topic should look at a more accurate sampling frame than the

one used during this study. In this study, 'IHDS II' will refer to the entire data set consisting of 42,152 observations and 'restricted sampling frame' will refer to the sampling frame consisting of 10,124 observations that correspond to households with a teenage girl. Since the results are not too different between the sampling frame and the entire data set, we use the entire data set for most of our analysis except the random forest regression at the end.

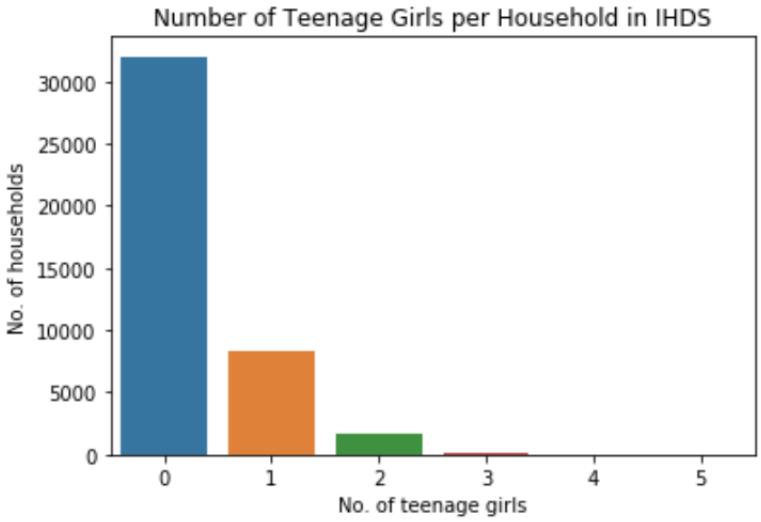


Figure 1: Number of Teenage Girls per Household in IHDS II

3 Feature Selection

Note that we will only be looking at features which are available in both TAG and IHDS II since our end goal is to predict income within the TAG data set. Thus, we look at three separate categories of features.

3.1 Household Assets

Our first set of features consists of household assets that are available in both IHDS and and TAG data sets. The following assets are available in both TAG and IHDS. The ownership and missing column only refer to the proportions in IHDS.

Table 1: Households Assets Description Table

Asset	Ownership	Missing	Comments
Electricity	0.874	0.004	N/A
Pressure Cooker	0.544	0.001	N/A
Chair/Table	0.778	0.001	TAG has separate chair and table columns
Electric Fan	0.754	0.001	N/A
Black and White TV	0.051	0.001	N/A
Color TV	0.615	0.001	N/A
Sewing Machine	0.261	0.001	N/A
Mobile Telephone	0.805	0.001	N/A
Landline Telephone	0.082	0.001	N/A
Computer/Laptop	0.07	0.001	N/A
Refrigerator	0.279	0.001	Fridge == Refrigerator is assumed
AC	0.178	0.001	Air Cooler == AC is assumed
Washing Machine	0.099	0.001	N/A
Watch/Clock	0.859	0.001	N/A
Bicycle	0.542	0.001	N/A
Motorcycle/Scooter	0.287	0.001	N/A
Car	0.05	0.001	N/A
Thresher	0.031	0.402	N/A

We notice from Table 1 that there are a lot of missing values for the household asset thresher. We also imagine that Thresher is not going to be a very important variable in predicting Income, so we remove it from our list of features and keep the rest.

3.2 Demographics

We use the following features for Demographics: State, Urban/Rural, Religion, Caste.

There were an insignificant amount of missing values for these variables.

3.3 Education

We use the following features for Education: Highest Male Education, Highest Female Education, Highest Adult Education.

Again, there were an insignificant amount of missing values. We use years of education as a continuous variable. Figure 9 in the Appendix shows a heat map showing the Pearson correlation between the education features.

4 Response Variable: Income

The following graph demonstrates the distribution of Mean Yearly Income for every Percentile.

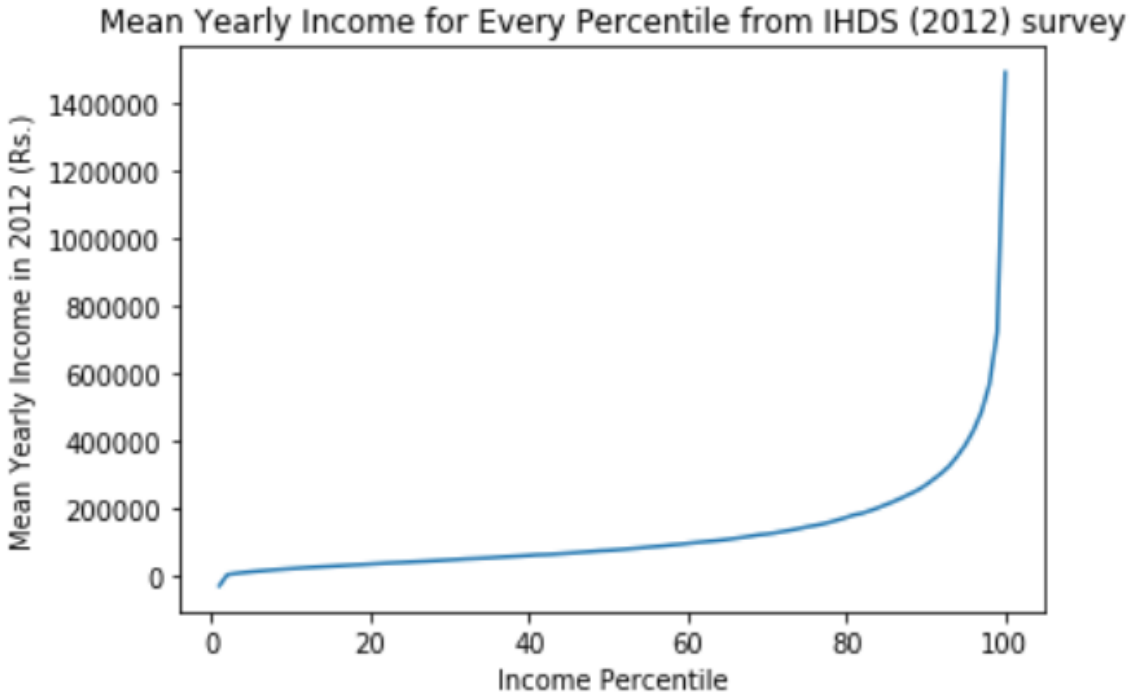


Figure 2: Mean Income for every Percentile

In order to conduct inference, I create three response variables which classifies the households into income percentiles, deciles and quartiles.

5 Inference

I divided my data set by creating a training set with 75% of the data and a test set consisting of 25% of the data. In the end, my training set for the IHDS II data set consisted of 28,505 observations.

6 Feature Engineering

I created two new features:

- Number of Household Assets Owned
- Wealth Score

The calculation for Number of Household Assets Owned was trivial.
In order to calculate a Wealth Score, I completed the following steps.

1. I grouped all the households into 10 deciles.

2. For every single decile, I calculated the proportion of people who owned a specific household asset.

3. Next, I calculated the standard deviation of these proportions for every single household asset across different deciles. This number is called the weight of a specific asset. The higher the deviation of ownership of a specific asset between deciles, the higher the weight.

4. Now, that I had a weight for every asset, I calculated a wealth score by multiplying the weight of a specific asset by a binary indicator for whether a household owned that asset. Then, I summed these numbers across all assets to calculate one wealth score for every household.

The following graph illustrates the importance of the wealth score. For every single Income Decile, I have calculated the mean wealth score and plotted it on the following bar graph.

Figure 3: Mean Wealth Score for every Income Decile

7 Models

7.1 Decision Trees

In order to analyze the importance of different features. I created different decision trees where I was predicting income deciles based on my features.

The decision trees are located in the appendix. I ran decision trees with different groups of features as listed: all features, only household assets, only demography parameters and only education parameters. Additionally for every group of features, I either restricted the tree by depth or pruned it using a parameter (ccp_alpha) described in the appendix. Based on the decision trees, I have listed down the top three tree divisions for every type of decision tree that I created. Division 1 refers to the division at the first node. Division 2 refers to the division if the first node led to a True decision. Division 3 refers to the division if the first node led to a False decision.

One could think of these three to be the three most important features for every category.

Table 2: Important Features based on Decision Tree Results

Category	Depth/Prune	Division 1	Division 2	Division 3
All	Depth	Wealth Score $\leq$ 1.225	Wealth Score $\leq$ 0.652	num_assets $\leq$ 12.5
All	Prune	Wealth Score $\leq$ 1.225	Wealth Score $\leq$ 0.652	num_assets $\leq$ 12.5
Assets	Depth	Refrigerator = 0	Color TV = 0	Computer/Laptop = 0
Assets	Prune	Refrigerator = 0	Color TV = 0	Computer/Laptop = 0
Demography	Depth	urban = 0	Forward Caste 2 = 0	Muslim 2 = 0
Demography	Prune	urban = 0	Muslim 2 = 0	Forward Caste 2 = 0
Education	Depth	Highest Adult $\leq$ 11.5	Highest Adult $\leq$ 7.5	Highest Male$\leq$14.5
Education	Depth	Highest Adult$\leq$11.5	Highest Adult$\leq$7.5	Highest Male$\leq$14.5

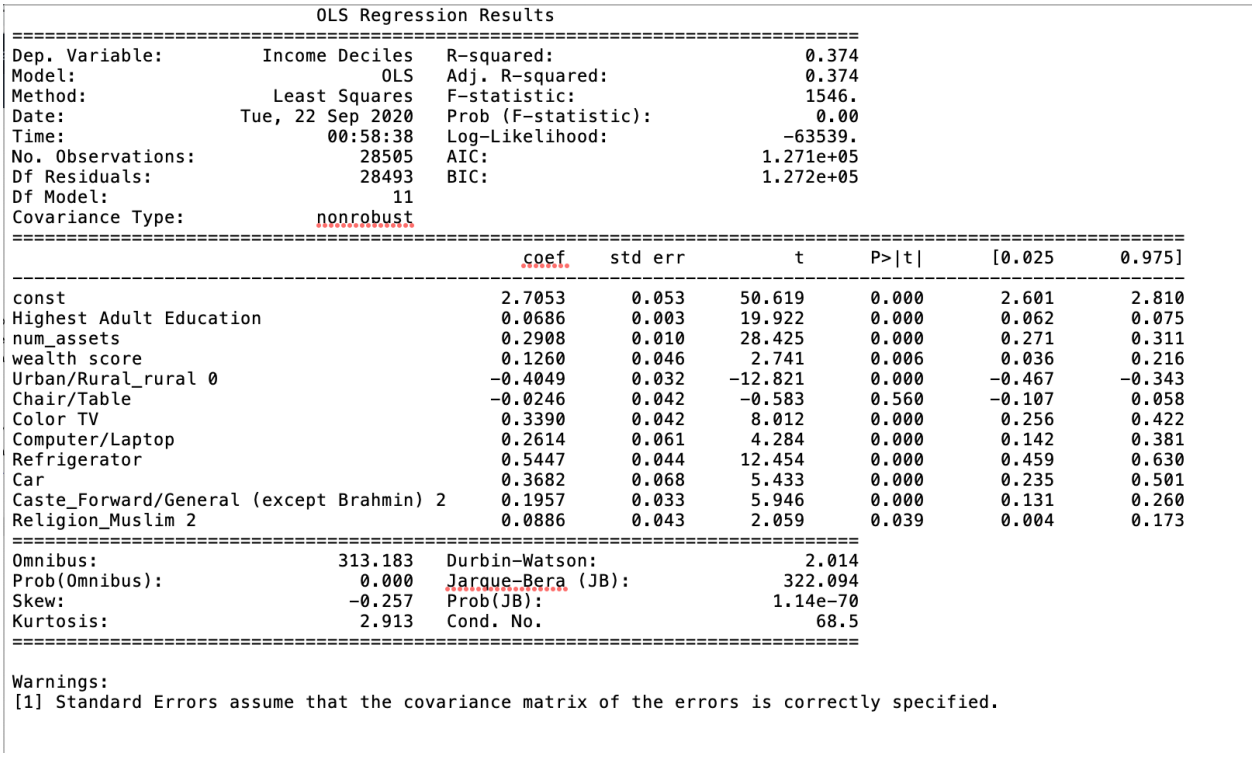
7.2 General Linear Regressions

First, we attempt to run a linear regression in order to predict Income Deciles. First, I chose to run an econometrics style OLS and judge the regression table. My response variable was income deciles from 1 - 10 which I treated as a continuous variable. I used the statsmodels library to run the following ols regression.

Using the decision trees to restrict features, I ended up using the following 10 features: Highest Adult Education, num_assets, wealth score, Urban/Rural_rural 0, Chair/Table, Color TV, Computer/Laptop, Refrigerator, Car, Caste_Forward/General (except Brahmin) 2, Religion_Muslim 2. They are addressed from x_0 to x_{10} in the regression equation below. α represents the constant term. β_0 to β_10 refers to the regression coefficients.

$$Y = \alpha + \sum_{i=0}^{10} \beta_i x_i + u$$

Here is the regression table showcasing the results.



OLS Regression Results

Dep. Variable:	Income Deciles	R-squared:	0.374
Model:	OLS	Adj. R-squared:	0.374
Method:	Least Squares	F-statistic:	1546.
Date:	Tue, 22 Sep 2020	Prob (F-statistic):	0.00
Time:	00:58:38	Log-Likelihood:	-63539.
No. Observations:	28505	AIC:	1.271e+05
Df Residuals:	28493	BIC:	1.272e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.7053	0.053	50.619	0.000	2.601	2.810
Highest Adult Education	0.0686	0.003	19.922	0.000	0.062	0.075
num_assets	0.2908	0.010	28.425	0.000	0.271	0.311
wealth score	0.1260	0.046	2.741	0.006	0.036	0.216
Urban/Rural_rural 0	-0.4049	0.032	-12.821	0.000	-0.467	-0.343
Chair/Table	-0.0246	0.042	-0.583	0.560	-0.107	0.058
Color TV	0.3390	0.042	8.012	0.000	0.256	0.422
Computer/Laptop	0.2614	0.061	4.284	0.000	0.142	0.381
Refrigerator	0.5447	0.044	12.454	0.000	0.459	0.630
Car	0.3682	0.068	5.433	0.000	0.235	0.501
Caste_Forward/General (except Brahmin) 2	0.1957	0.033	5.946	0.000	0.131	0.260
Religion_Muslim 2	0.0886	0.043	2.059	0.039	0.004	0.173

Omnibus:	313.183	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	322.094
Skew:	-0.257	Prob(JB):	1.14e-70
Kurtosis:	2.913	Cond. No.	68.5

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 4: Linear Regression Table

In order to judge the effectiveness of the linear regression. I converted my predictions to numbers 1 - 10 by rounding each prediction. Then, I calculated an accuracy score against the real income deciles for this household. The accuracy score is the proportion of results that I got exactly right.

I also ran a linear regression with all the features instead of the restricted set using sklearn. Additionally I calculated 5-fold cross validation results for the sklearn regression.

In addition, I ran a logistic regression model on sklearn with all the features.

My results from the above two regressions are as follows:

Table 3: Linear Regression Results

Model	Features	Accuracy	CV Accuracy	r^2	CV r^2	Mean Squared Error
statsmodel OLS	Restricted	15.9 %	N/A	-0.655	N/A	5.137
sklearn Linear Regression	All	15.9 %	16.6 %	0.378	0.300	4.812
sklearn Logistic Regression	All	24%	22.6%	0.108	0.102	N/A

Clearly, the linear regression is not the best method for this problem. However, I imagine the OLS regression table gives some insight into our features. If one wanted to improve the OLS and try it, that would require box-cox transformations, analysis trying to identify a linear relationship and a better response variable i.e. predicting income percentiles since they can be treated as a

continuous variables. I doubt, however, that OLS will yield favorable results for this specific problem.

The Logistic Regression shows some promise. I believe that further work into the logistic regression might yield better results. For reference and future work, please follow and implement this paper. This paper by June Y. T. Po, Jocelyn E. Finlay, Mark B. Brewster, David Canning deals with our problem of estimating household income from ownership of physical assets.

7.3 Random Forest

I used a randomized Grid Search in order to calculate the best possible Random Forest Accuracy Scores with the current features. I did a grid search over the max depth of a tree, the pruning parameter ccp_alpha as described before, the number of features used as determined by sklearn selectKBest function which selects the best features to maximize accuracy and the number of estimators i.e. the number of trees one wants to build before taking the maximum voting or averages of predictions. I used sklearn in order to run this randomized search.

I got the following results.

Table 4: Accuracy Table

Simple	Granularity	Num Features	Max Depth	CCP_alpha	N_estimators	CV Accuracy
Full	Percentiles	25	5	0	160	2.91%
Full	Deciles	23	10	0	110	22.1%
Full	Quartiles	23	10	0	10	46.4%
Teenage Girl HH	Percentiles	25	5	0	110	2.8%
Teenage Girl HH	Deciles	18	5	0	60	20.5%
Teenage Girl HH	Quartiles	24	10	0	160	45.2%

Note that I did not use the wealth_score or num_assets features when creating this accuracy table. I find that using these two additional features does not change the results by much.

Also note that we are looking at 5-fold Cross Validation Accuracy here. When we look at training accuracy, we get an accuracy ranging from 90% - 100% for predicting Income deciles, quartiles or percentiles. However, since we will be predicting income for households, CV accuracy should be our metric of interest rather than training accuracy.

8 Conclusions

We have not been able to successfully predict with our current model and features. However, there is a lot of promise when we look at the Logistic Regression and the Random Forest. Our results till now have not been up to the mark, but I believe there is definitely a solution to this problem. For anyone who decides to pursue this problem, here are some routes you can take:

- Use a different data set other than IHDS.
- Use different models such as Neural Nets, MLP Classifier or Bayes Nets.

- Use Auto ML.
- We can use price data, and create a fixed effects logistic regression model as described in the following research paper.
- Re-implement all the models in this paper using an sklearn grid search.
- Utilize Box-Cox Transformations and other relevant techniques to make this problem suitable for a linear regression.
- Utilize less features to get rid of over-fitting. Create correlation heat map to see why the classifications are not accurate.
- TAG Data set has a column which divides household into 4 income categories. Find out how that is implemented and use that specific feature in IHDS.

I am very optimistic that this is a doable prediction problem. Even though my results have not been up to the mark, I argue that this is possible to accomplish because there is a clear distinction between different income deciles when it comes to the proportion of people owning a household asset in a specific income decile. The following table shows great variation of ownership proportion across different income deciles.

Y_dec	Electricity	Pressure Cooker	Chair/Table	Electric Fan
1	0.743377	0.290563	0.586507	0.533526
2	0.743799	0.257682	0.574232	0.532395
3	0.786566	0.331528	0.687613	0.631997
4	0.845451	0.416579	0.716544	0.685283
5	0.885714	0.488153	0.777003	0.762021
6	0.904387	0.540439	0.809116	0.795751
7	0.938942	0.642469	0.866011	0.864993
8	0.956102	0.720652	0.910875	0.899568
9	0.972548	0.839638	0.947439	0.931704
10	0.987508	0.930638	0.976003	0.965155

Y_dec	Black and White TV	Color TV	Sewing Machine	Mobile Telephone
1	0.053394	0.352235	0.117964	0.609272
2	0.053314	0.310255	0.104776	0.614957
3	0.056338	0.406645	0.136873	0.696641
4	0.056551	0.495961	0.175623	0.768177
5	0.053310	0.596167	0.211847	0.831010
6	0.061343	0.642563	0.258053	0.864291
7	0.048168	0.757802	0.299525	0.918589
8	0.046558	0.823412	0.341869	0.947123
9	0.040174	0.885169	0.423502	0.961500
10	0.044050	0.950690	0.530901	0.981920

Y_dec	Landline Telephone	Computer/Laptop	Refrigerator	AC
1	0.038079	0.019040	0.097682	0.059603
2	0.017031	0.009256	0.057756	0.046649
3	0.022030	0.008667	0.076923	0.071145
4	0.026695	0.009835	0.116614	0.095188
5	0.041812	0.019512	0.164460	0.112892
6	0.040781	0.023646	0.201851	0.145990
7	0.069539	0.047829	0.309023	0.194030
8	0.094114	0.066179	0.394746	0.238444
9	0.152662	0.143957	0.573485	0.340141
10	0.294543	0.348455	0.784681	0.497699

Y_dec	Washing Machine	Watch/Clock	Bicycle	Motorcycle/Scooter	Car
1	0.031043	0.695778	0.492550	0.155215	0.014901
2	0.011847	0.718993	0.542392	0.095150	0.007775
3	0.015890	0.783315	0.581076	0.125677	0.007223
4	0.022831	0.842290	0.568318	0.135230	0.010186
5	0.034146	0.862718	0.564111	0.188153	0.012195
6	0.040096	0.912954	0.581905	0.237491	0.019877
7	0.079715	0.932836	0.574966	0.336839	0.033582
8	0.109079	0.962421	0.582641	0.418690	0.039242
9	0.207901	0.976230	0.570807	0.554737	0.086374
10	0.428008	0.987179	0.565746	0.740302	0.264300

Figure 5: Proportion of Household Asset Ownership per Income Decile

The table above shows the proportion of household asset ownership per income decile. For

example, 1.5% of households in the lowest income decile own a car while 26.4% of households in the highest income decile own a car.

As we can observe, there is great variation of ownership for specific household assets based on the income decile. Thus, I believe this is a solvable prediction problem. Yet, the techniques used in this paper have not been successful in reaching a sophisticated level of CV accuracy. I hope that my results guide whoever decides to try their hand at this problem.

9 Appendix

The following represents a heat map showing the pearson correlation between the education features.

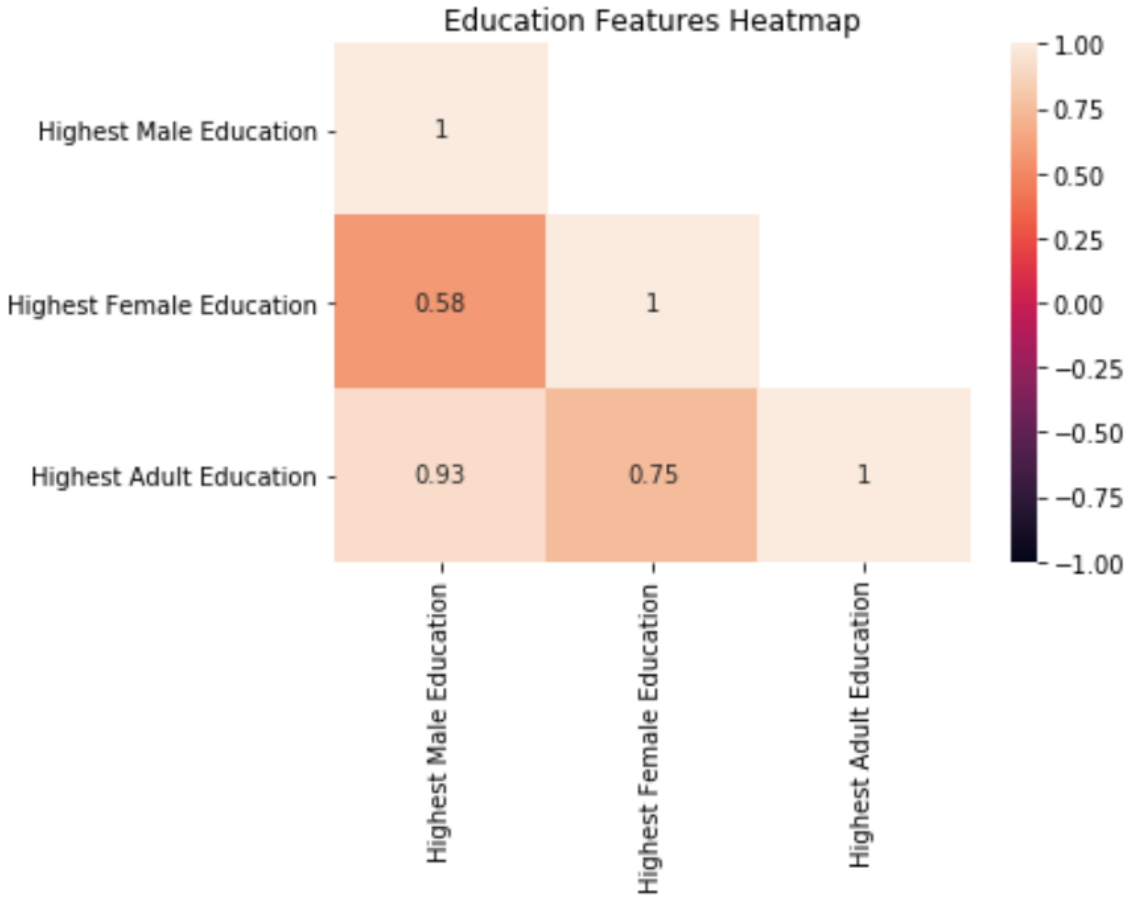


Figure 6: Heat map

The following represents the many decision trees that I developed.

The first element in every node of the decision tree shows the feature on which the node was split. The second element saying 'gini' shows the gini impurity. Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly

labeled. The 'sample' shows the proportion of data on which that node split up. The values array show the class probabilities. It shows the proportion of the sample that belong to the 10 classes. The class parameter signifies which class the node belongs to.

Now, I have created my decision trees by setting either the sklearn 'max_depth' parameter or the sklearn "ccp_alpha" parameter. The max depth parameter restricts the depth of the tree to 4 levels in my case. The ccp_alpha parameter is a pruning parameter which uses minimal cost complexity pruning. Minimal cost complexity pruning recursively finds the node with the "weakest link". The important thing to note is that higher values of ccp_alpha leads to more pruning.

First, I created a decision tree with all features and a depth restricted to 4 levels. I got the following tree:

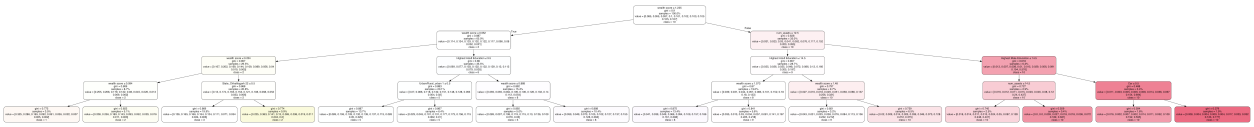


Figure 7: Decision Tree for All Features with Depth 4

Second, we have a decision tree with all features and the pruning parameter on sklearn set to 0.0005.

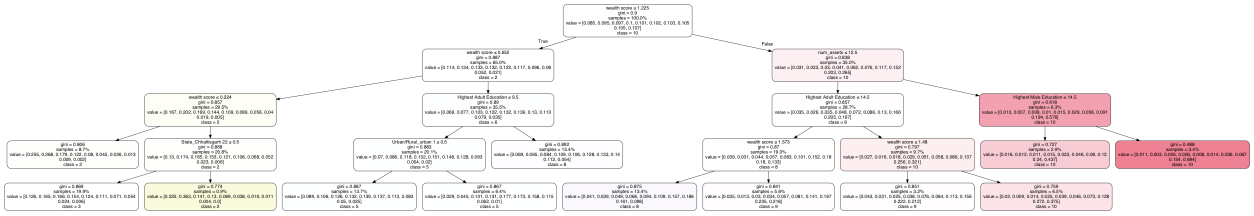


Figure 8: Decision Tree for All Features with Pruning Parameter 0.0005

Third, we have a decision tree with only household assets with a depth restricted to 4 levels.

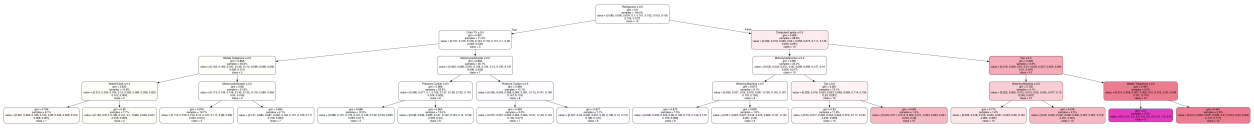


Figure 9: Decision Tree for Household Assets with Depth 4

Fourth, we have a decision tree with only household assets with the pruning parameter set to 0.0005.

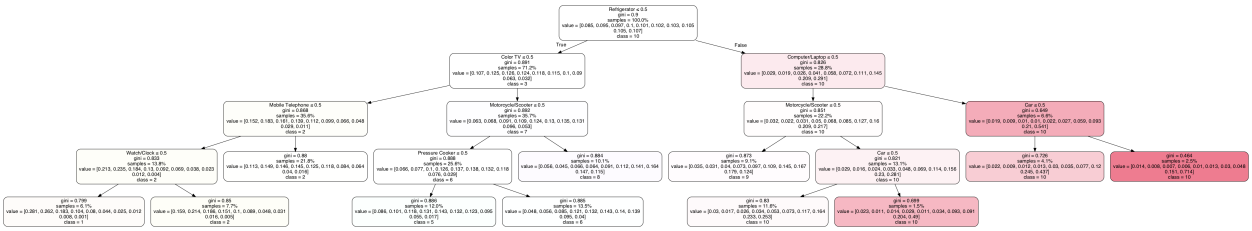


Figure 10: Decision Tree for Household Assets with Pruning Parameter 0.0005

Fifth, we have a decision tree with only demography parameters with a depth restriction of 4 levels.

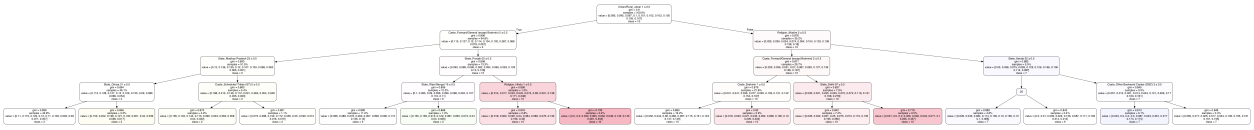


Figure 11: Decision Tree for Demography Parameters with Depth 4

Sixth, we have decision tree with only demography parameters with the pruning parameter set to 0.0005.

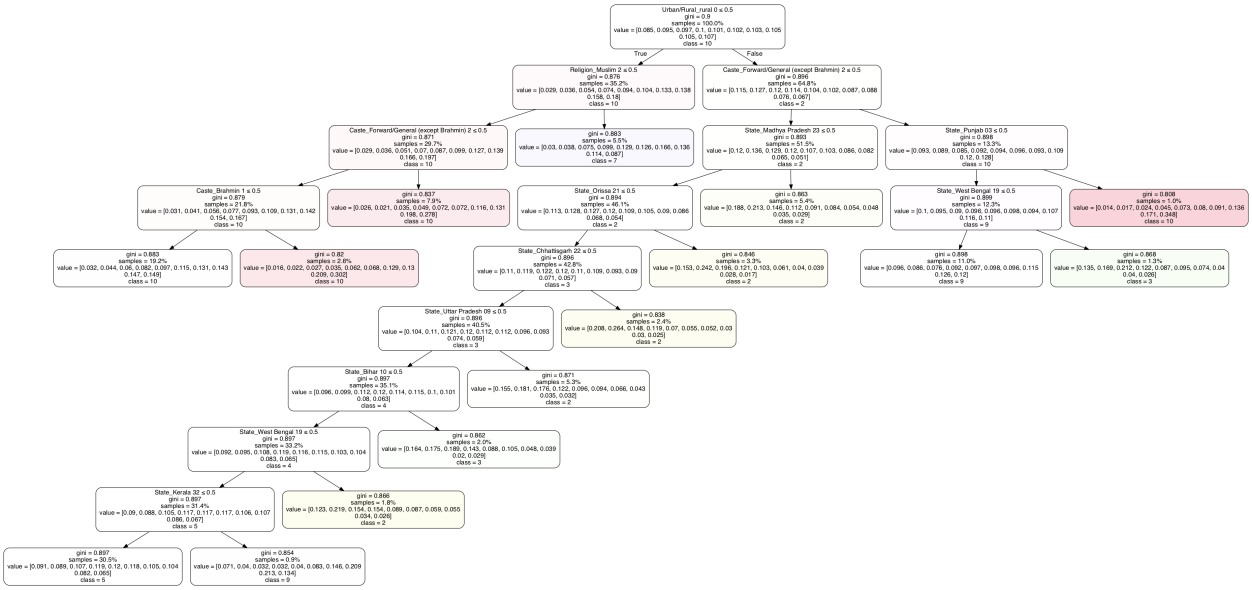


Figure 12: Decision Tree for Demography Parameters with Pruning Parameter 0.0005

Seventh, we have a decision tree with only education parameters with a depth restriction of 4 levels.

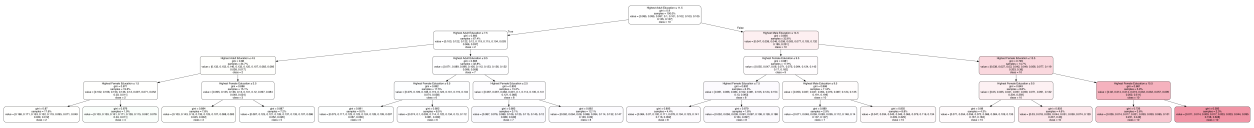


Figure 13: Decision Tree for Education Parameters with Depth 4

Finally, we have a decision tree with only education parameters with the pruning parameter set to 0.0005.

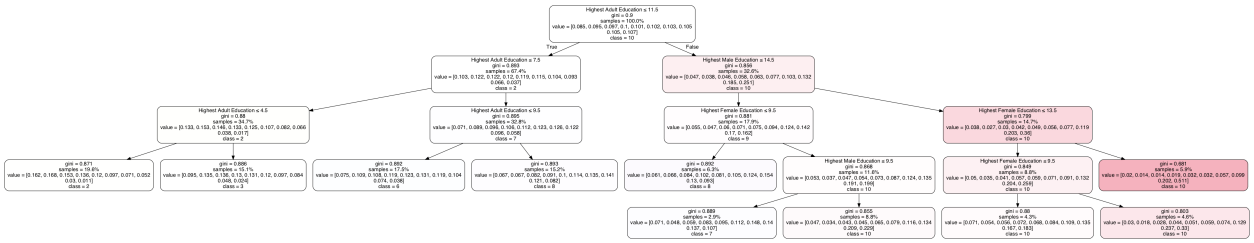


Figure 14: Decision Tree for Education Parameters with Pruning Parameter 0.0005