ADULT INCOME

Does it depend only on your qualifications?

An analysis using various ML tools

2STAT4.5 – PROJECT DISSERTATION

By

Subhadeep Majumder

Disha Roy

Ananya Halder

Siddhertha Podder

Acknowledgement

This project has been great opportunity to gain lots of experience in Statistical Analysis, followed by the knowledge of how to actually observe. I got to learn a great deal about Logistic Regression as well as other Machine Learning Models.

I would like to express my special thanks of gratitude to our project supervisor Dr. Kiranmoy Chatterjee (Assistant Professor, Dept. of Statistics, Bidhannagar College, Salt Lake) who gave the golden opportunity to do this wonderful project. During the course of this project he taught me that- " You don't always have to be the Decision-Maker. As a statistician, you would serve your purpose even if you observe meticulously and report the facts without bias. " I would always be indebted to him for the valuable lessons he instilled in me during the course of this project as well as inside the walls of the classroom.

Also a special thanks to the Faculty of Department of Statistics, University of Kalyani for giving me enough Statistical knowledge to envision this project and bring it to reality.

Contents

SECTION 1:

INTRODUCTION

INTRODUCTION

Research says experienced well-being does not increase above incomes of $75,000/y(From an research article by Matthew A. Killingsworth).This finding has been the focus of substantial attention from researchers and the general public, yet is based on a dataset with a measure of experienced well-being that may or may not be indicative of actual emotional experience (retrospective, dichotomous reports). A study conducted over one million real-time reports of experienced well-being from a large US sample show evidence that experienced well-being rises linearly with log income, with an equally steep slope above $80,000 as below it. This suggests that higher incomes may still have potential to improve people's day-to-day well-being, rather than having already reached a plateau for many people in wealthy countries. So this shows us that predicting the income class of a person also lets us get an idea about their Wellbeing. In recent times, due to rising global inflation, a better income class definitely ensures much more wellbeing.

Traditionally, we are conditioned to the idea that increased educational Qualifications, Native Country and even Gender(we know this phenomenon as Gender Gap) can affect the income of a person.

As Harry A. Patrinos and George Psacharopoulos say in a World Bank Article-"Classical economists knew it. But it was only in the latter half of the 20th century that the link between education and earnings was established in theory and practice." In the present internet connected world, income shouldn't have depended on native country ideally. But it does. Since even the presence of internet and internet penetration is unequal for countries. Hence, an average person of a poor and developing nation would be limited to the earnings that his/her country can provide, which would not be at par with that of a developed nation.

Also, even in recent times, there is the issue of Gender Gap. According to Global Gender Gap Report 2021 - "On the other hand, overall income disparities are still only part-way towards being bridged and there is a persistent lack of women in leadership positions, with women representing just 27% of all manager positions." In other words, there is a persistent lack of women in High-paying jobs. In some cases, women even getting less pay for the same designated work.

Now, we knew about these factors affecting income quite intuitively. We can also think age and the workclass(private or government job),working hours per week and the particular occupation of the individual affecting the income class, though we might ask by how much or what degree?Also, we would find out if less obvious information like Marital status and relationship status, race and also amount of income from investment sources other than salary wages affect the income class of a subject.

OBJECTIVE

Our Objective in this sudy are :

i) To find the factors significantly determining the Chances of a person earning Higher Income.
ii) To find a model that best classifies a sample of individual according to their Income levels based on these factors.

SECTION 2:

OBSERVING THE DATASET

OBSERVING THE DATASET

We have a dataset containing various demographic variables as well as the income class of the person. The income class is divided on the basis - Yearly income “below or equal to 50K USD” or “above 50K USD”.

There are 15 variables in total and 48,842 observations. The variables being age, workclass, fnlwgt (Sampling weight), education(Qualification), educational.num(Number of years of education), marital.status(married, unmarried, widowed, etc.), occupation(Type of occupation like farming, Craft, etc.), relationship (Status of relationship like Husband, wife, etc.), race(White, Black, Asian,etc.), gender(Male or Female), capital.gain(Gain from investments made), capital.loss(Loss from investments made), hours.per.week(Working hours per week), native.country(Country of origin), income(Income above 50K or not). Here, Income is our response variable.

For the purpose of the analysis, we split the data into two parts:

- Training dataset
- Testing dataset

The Dataset

age	workclass	fnlwgt	education	education	marital-st	occupatio	relationsh	race	gender	capital-ga	capital-lo	hours-per	native-co	income
25	Private	226802	11th	7	Never-ma	Machine-(	Own-chilc	Black	Male	0	0	40	United-St	<=50K
38	Private	89814	HS-grad	9	Married-c	Farming-f	Husband	White	Male	0	0	50	United-St	<=50K
28	Local-gov	336951	Assoc-acd	12	Married-c	Protective	Husband	White	Male	0	0	40	United-St	>50K
44	Private	160323	Some-coll	10	Married-c	Machine-(	Husband	Black	Male	7688	0	40	United-St	>50K
18	?	103497	Some-coll	10	Never-ma	?	Own-chilc	White	Female	0	0	30	United-St	<=50K
34	Private	198693	10th	6	Never-ma	Other-ser	Not-in-far	White	Male	0	0	30	United-St	<=50K
29	?	227026	HS-grad	9	Never-ma	?	Unmarriec	Black	Male	0	0	40	United-St	<=50K
63	Self-emp-	104626	Prof-scho	15	Married-c	Prof-speci	Husband	White	Male	3103	0	32	United-St	>50K
24	Private	369667	Some-coll	10	Never-ma	Other-ser	Unmarriec	White	Female	0	0	40	United-St	<=50K
55	Private	104996	7th-8th	4	Married-c	Craft-repa	Husband	White	Male	0	0	10	United-St	<=50K
65	Private	184454	HS-grad	9	Married-c	Machine-(	Husband	White	Male	6418	0	40	United-St	>50K
36	Federal-g	212465	Bachelors	13	Married-c	Adm-cleri	Husband	White	Male	0	0	40	United-St	<=50K
26	Private	82091	HS-grad	9	Never-ma	Adm-cleri	Not-in-far	White	Female	0	0	39	United-St	<=50K
58	?	299831	HS-grad	9	Married-c	?	Husband	White	Male	0	0	35	United-St	<=50K
48	Private	279724	HS-grad	9	Married-c	Machine-(	Husband	White	Male	3103	0	48	United-St	>50K
43	Private	346189	Masters	14	Married-c	Exec-man	Husband	White	Male	0	0	50	United-St	>50K
20	State-gov	444554	Some-coll	10	Never-ma	Other-ser	Own-chilc	White	Male	0	0	25	United-St	<=50K
43	Private	128354	HS-grad	9	Married-c	Adm-cleri	Wife	White	Female	0	0	30	United-St	<=50K
37	Private	60548	HS-grad	9	Widowed	Machine-(	Unmarriec	White	Female	0	0	20	United-St	<=50K

An Overview of the Data

The first thing we do is to observe the extent of missing values in the data. Once we can detect the columns with maximum number of missing values we can observe and decide things more accurately for our further analysis. For this we plot a graph showing the percentage of missing values in each column.

Figure 1 : Total Number of missing values in each independent variable

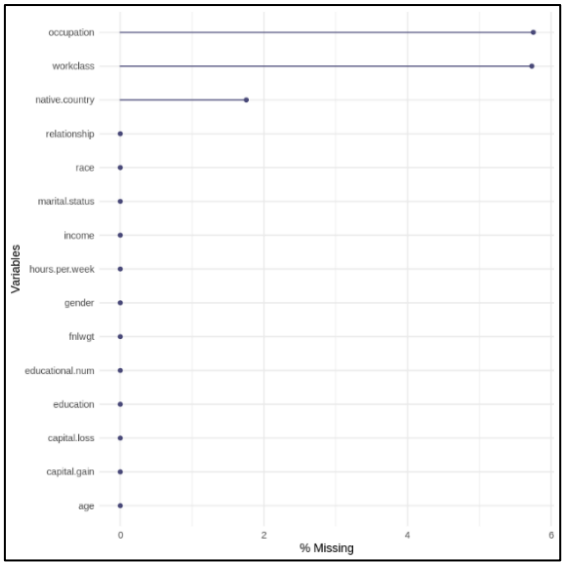
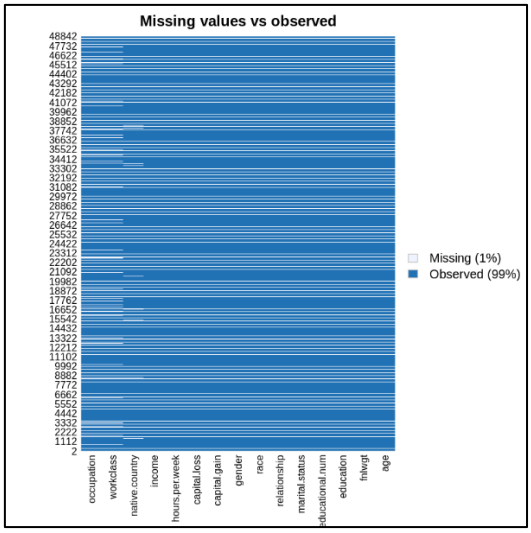


Figure 2: Missmap showing the location of the missing values



Interpretation: From the plot Figure 1 it is clear that Occupation, Workclass, and Native Country have a small percentage of missing values. The Missmap in Figure 2 shows that the missing values are few and far between. And constitute only 1% of the total observations.

Now, though imputation is a very lucrative option in order to not lose any of the sample data, but it comes with its own problems i.e. mean imputation does not preserve the relationships among variables. Also, mean imputation Leads to an underestimate of Standard Errors.

As just a small percentage of data is missing, we remove the corresponding rows for the purpose of our analysis.

Henceforth, we also replace "<=50k" by "0" and ">50k" by 1 in the Income column.

SECTION 3:

ANALYSING THE DATA

ANALYSING THE DATA

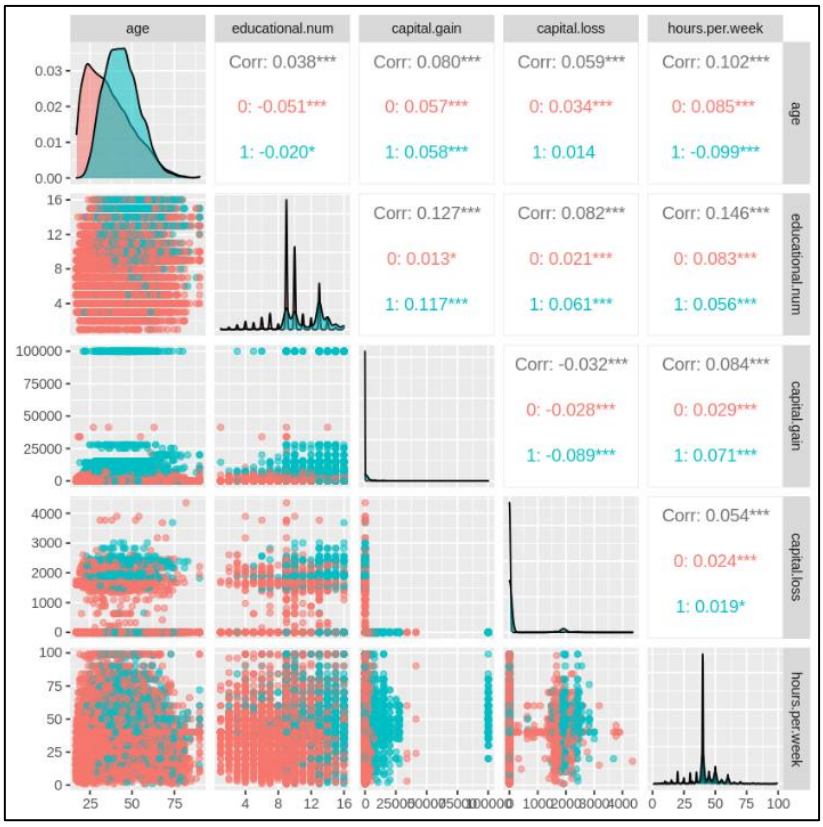
Part I: Studying the data based on different plots and charts and trying to understand relationships between different variables

We'll remove fnlwgt- "sampling weight" as it has no relation with income. Also we remove education as it is the categorical equivalent of educational.num.

a) Scatter Plot:

For our dataset we will look into the linear relationship between continuous independent variables. In this data set we have five continuous variable age, educational.num, capital.gain, capital.loss, hours.per.week. We will verify whether there is any linear relationship between these continuous variables. Income (0 or 1) have been shown in graphs by different colours of plot. 0 is represented by pink and 1 by blue.

Figure 3:Scatter plot between all the continuous variables



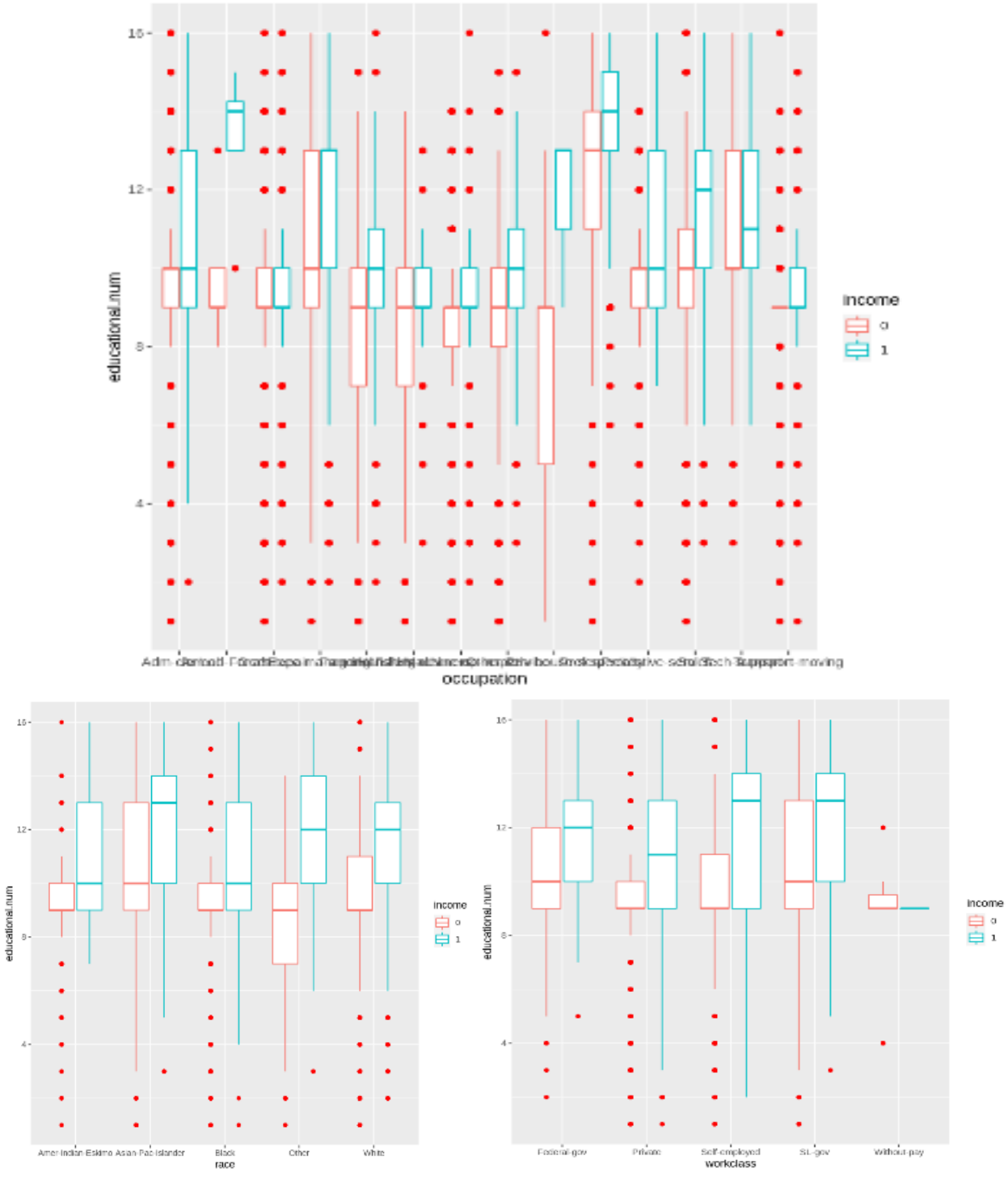
Interpretation: The correlations as shown in the upper triangle are all close to 0. The 0 and 1 values are all mixed together in the plots. Also the plots do not exhibit any linear relationship between the pairs of continuous variables.
So we can say there might not be any correlation between the various pairs of variables with respect to the Income status.

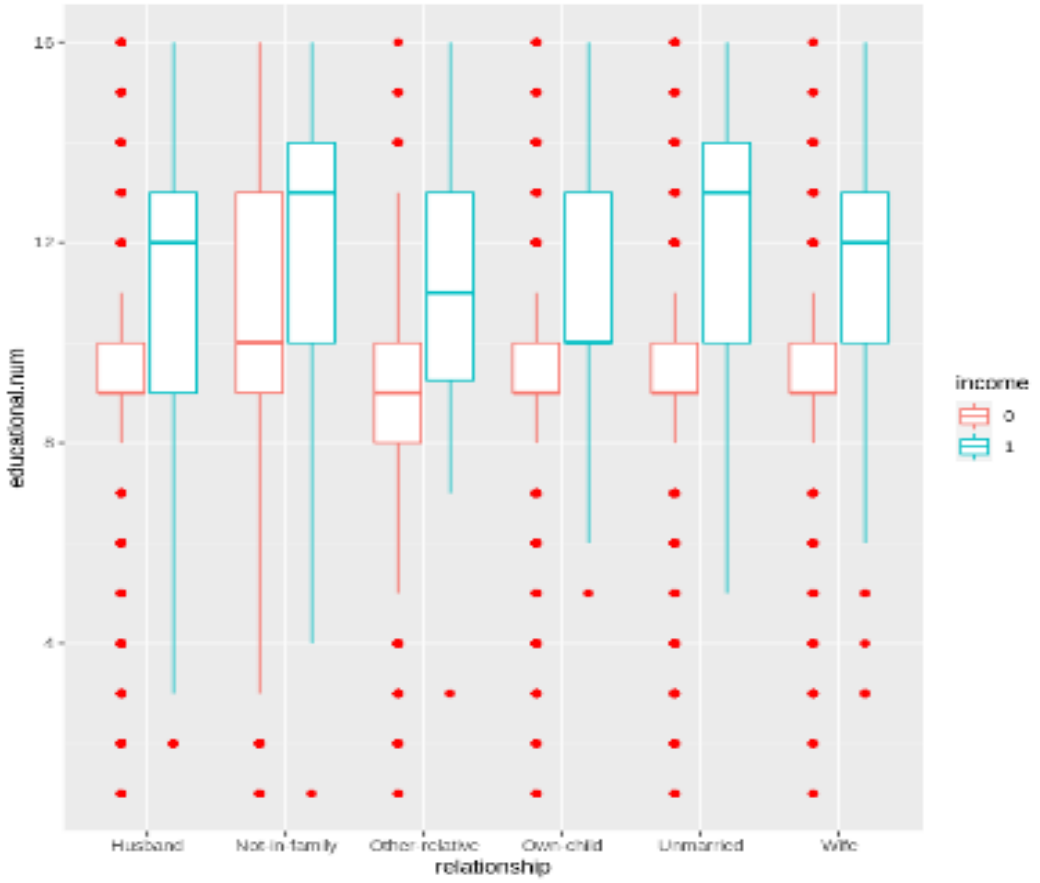
b) Box Plot

A boxplot is a standardized way of displaying the distribution of data based on a five-number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are.

For our data, we used this visualization to look for the outliers present in the data.

Figure 4: Boxplots for the categorical variables w.r.t Number of years of education




educational.num
relationship
Husband
Not in family
Other relative
Own child
Unmarried
Wife
income
0
1

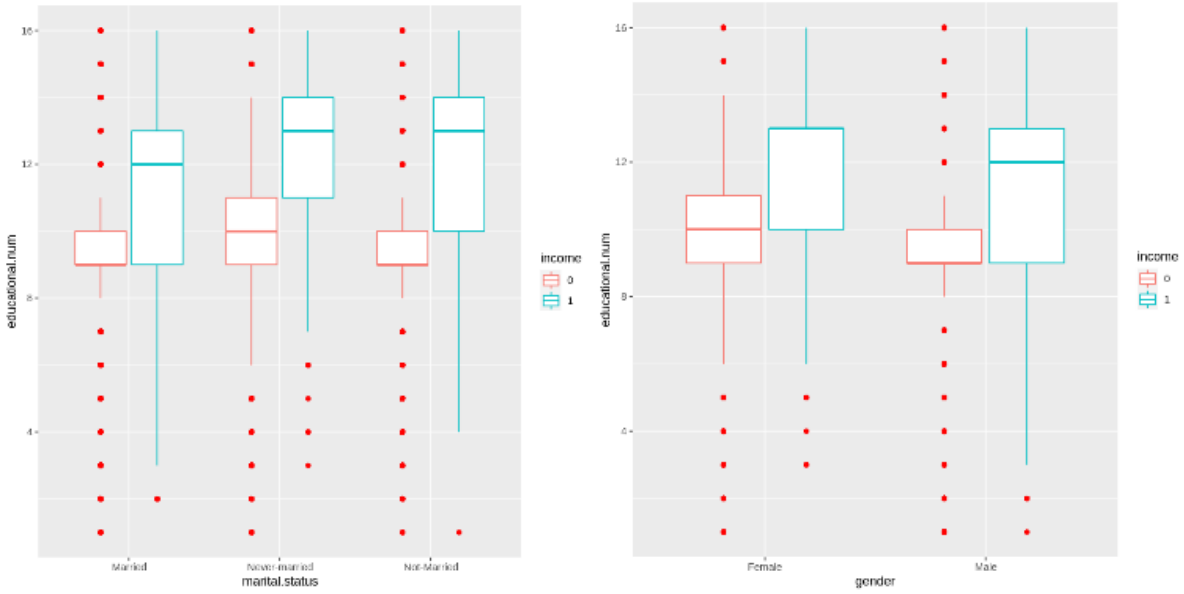
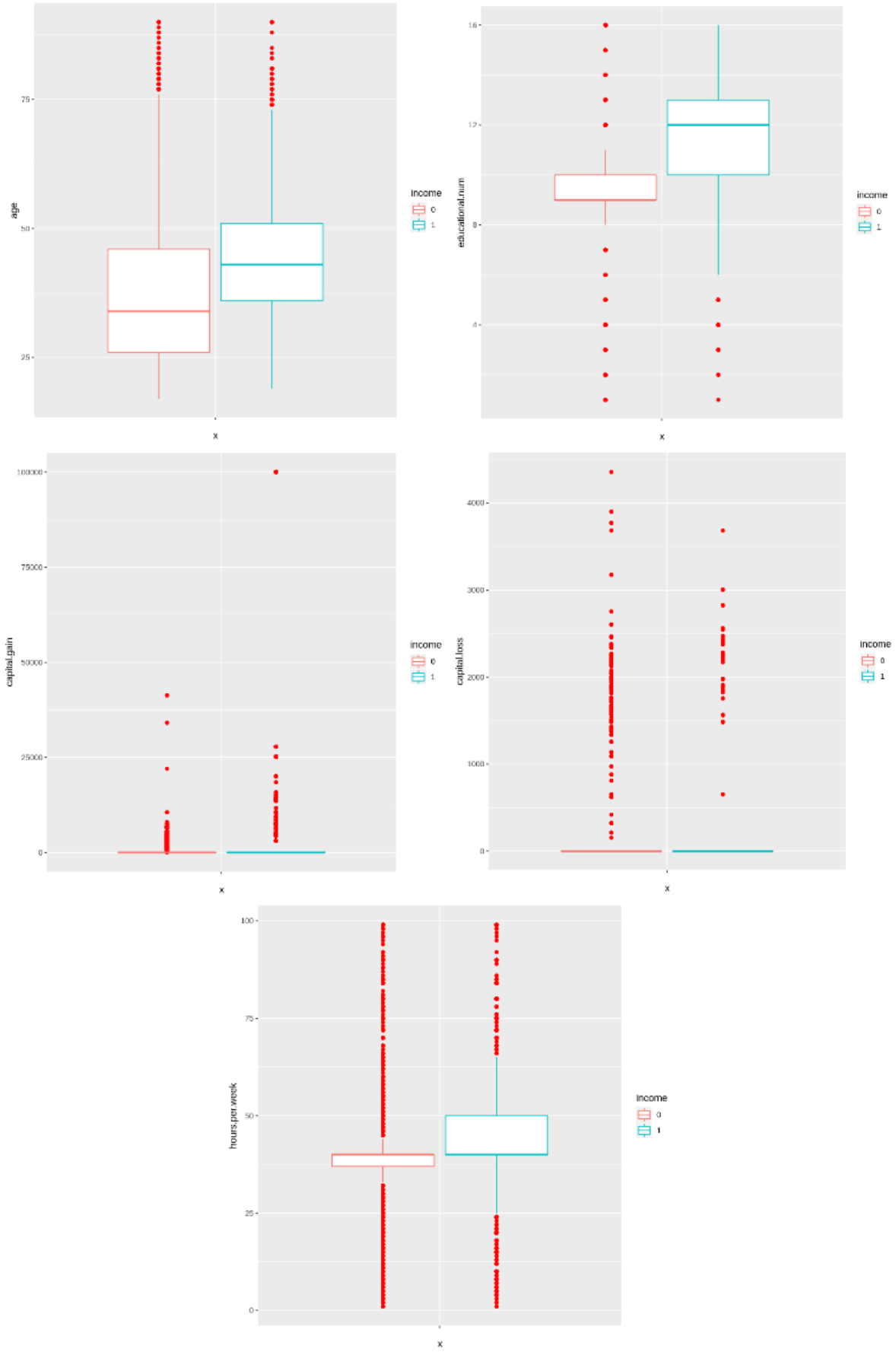

educational.num
marital.status
Married
Never-married
Not-Married
income
0
1
gender
Female
Male

Figure 5: Boxplots for continuous variables



Interpretation:

It is observed from Figure 4 i.e. the Boxplots for the categorical variables w.r.t Number of years of education that across all sectors of Workclass, the instances of higher income class (>50k) consistently show more number of years of education than those in lower income class. A similar trend is observed for all factors of Marital Status, Relationships, Race and Gender. So there is a distinct relationship between the number of years of education and their Income Class.

From Figure 5 i.e. Boxplots for continuous variables it is observed that for Age, higher income class is more distributed towards higher ages. For Number of years of education, higher Income class is more distributed towards higher number of years as we observed earlier. Also, higher income class is more distributed towards higher Working hours per week.

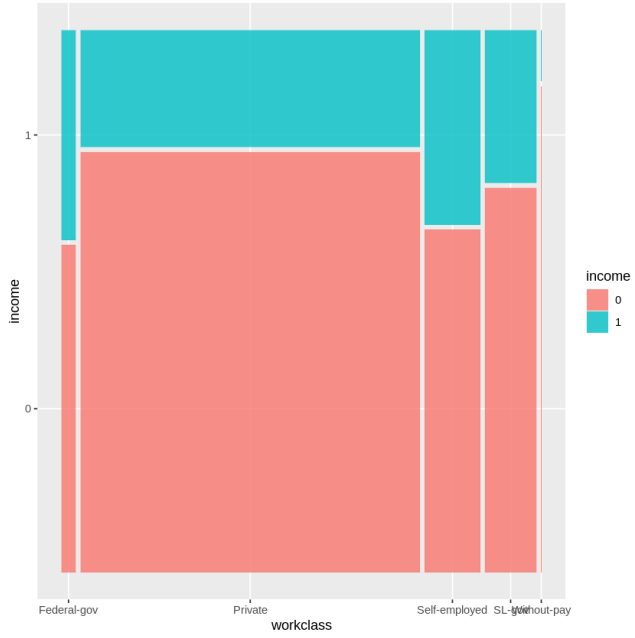
No pattern can be observed for Capital Gain or Loss as most of their values are zero and hence the box is limited to zero. The non-zero values are all shown as outliers. This might be due to the fact that most people do not venture into income form investments.

c) Mosaic Plot

A mosaic plot (also known as a Marimekko diagram) is a graphical method for visualizing data from two or more qualitative variables. It is the multidimensional extension of spine plots, which graphically display the same information for only one variable. It gives an overview of the data and makes it possible to recognize relationships between different variables.From these diagrams and a subsequent Test of Association between The Categorical Variables and Income Class, we have tried to understand which factor has more effect in determination of Income Class.

Null Hypothesis: There is no association between two variables vs,

Alternative Hypothesis: There is association between two variables

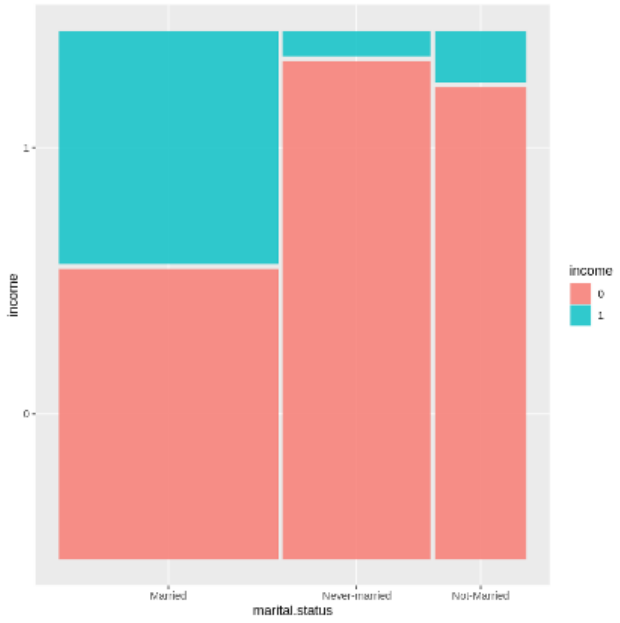


Test for independence of all factors:

Chisq = 736.2, df = 4,

p-value << 0.001.

Hence there is significant association between Workclass and Income Class (p values less than 0.05).

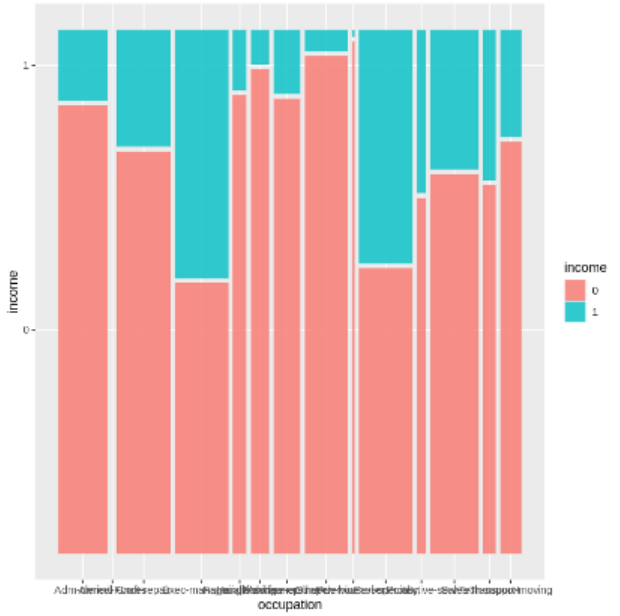


Test for independence of all factors:

Chisq = 8736, df = 2,

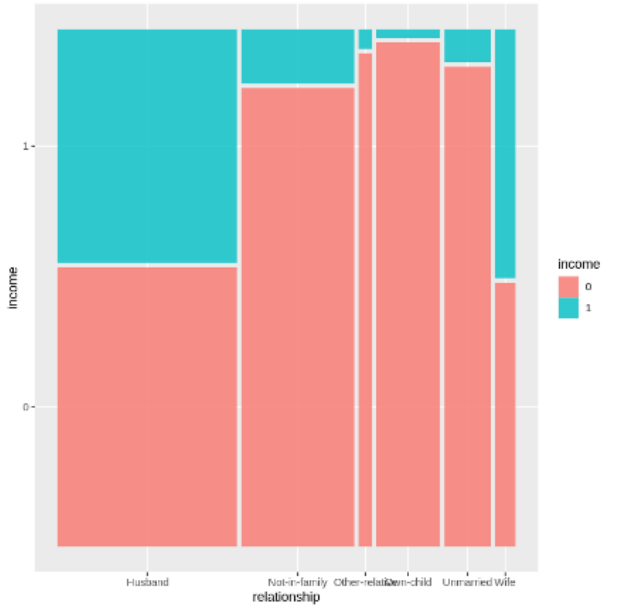
p-value << 0.001

Hence there is significant association between Marital Status and Income Class (p values less than 0.05).



Test for independence of all factors: Chisq = 5415, df = 13, p-value << 0.001

Hence there is significant association between Occupation and Income Class (p values less than 0.05).

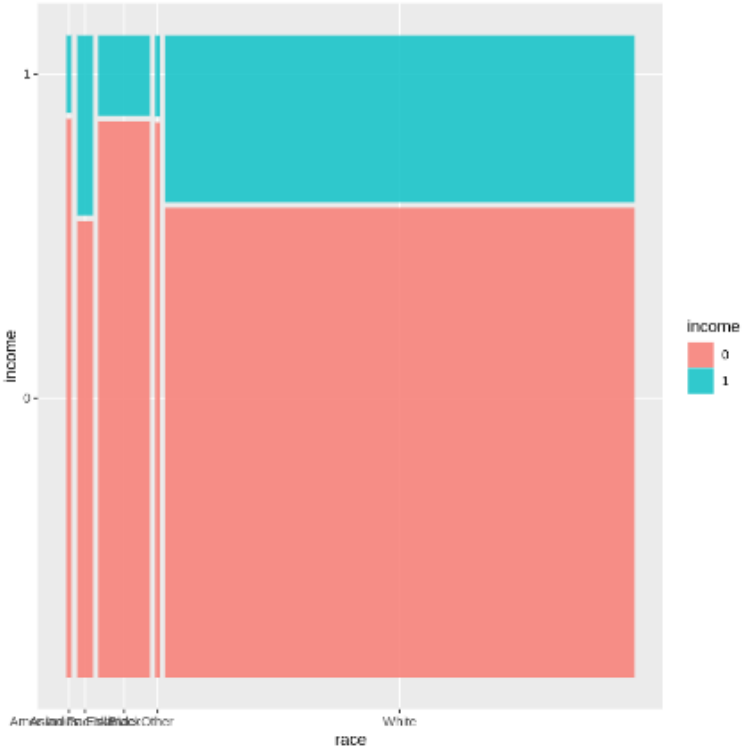


Test for independence of all factors:

Chisq = 9357, df = 5,

p-value << 0.001

Hence there is significant association between Relationship and Income Class (p values less than 0.05).

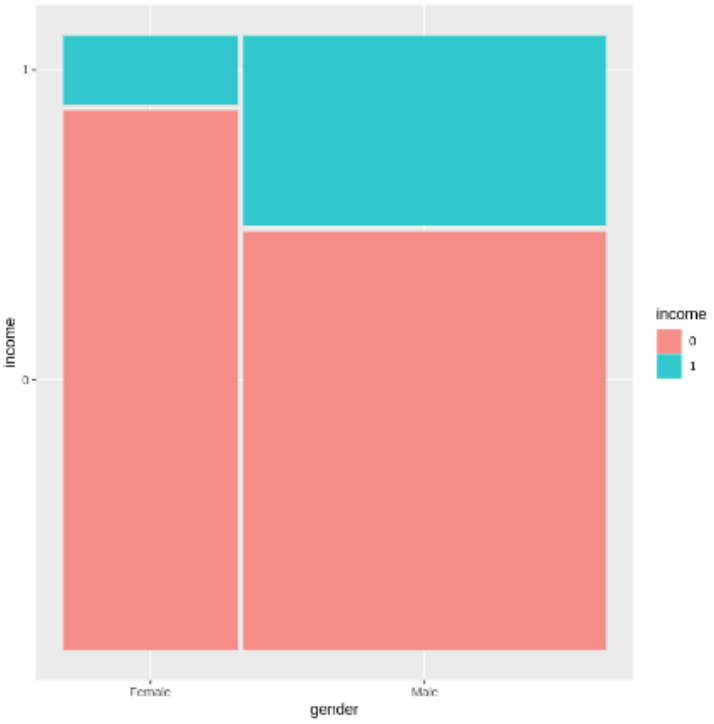


Test for independence of all factors:

Chisq = 452.3, df = 4,

p-value <<0.001

Hence there is significant association between Race and Income Class (p values less than 0.05).

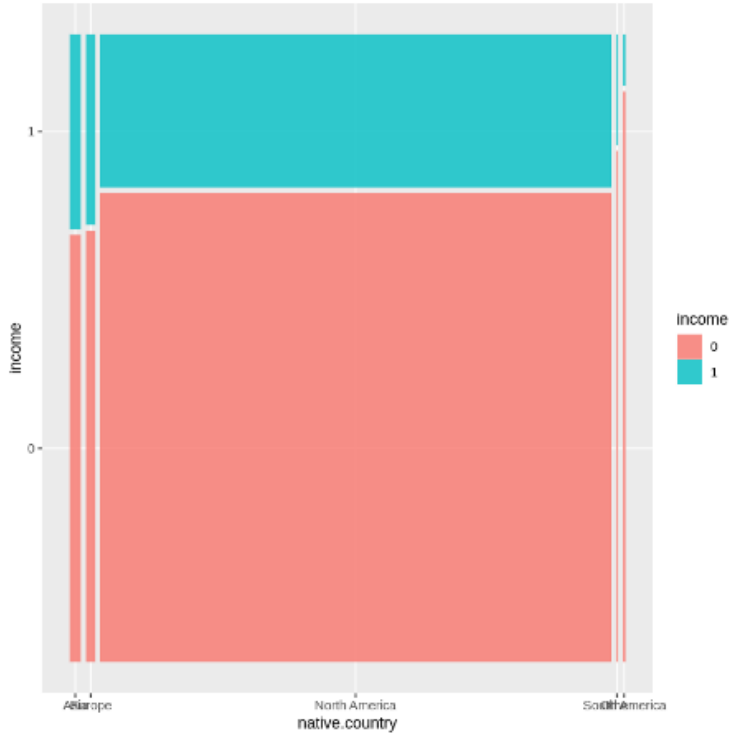


Test for independence of all factors:

Chisq = 2105.2, df = 1,

p-value <<0.001

Hence there is significant association between Gender and Income Class (p values less than 0.05).



Test for independence of all factors:

Chisq = 62.8, df = 4,

p-value << 0.001

Hence there is significant association between Native Continent and Income Class (p values less than 0.05).

d) VIF

	GVIF	Df	GVIF^(1/(2*Df))
age	1.180627	1	1.086567
workclass	1.508943	4	1.052771
educational.num	1.484732	1	1.218496
marital.status	17.671404	2	2.050302
occupation	2.346289	13	1.033345
relationship	42.483790	5	1.454864
race	2.567967	4	1.125120
gender	2.842858	1	1.686078
capital.gain	1.027210	1	1.013514
capital.loss	1.012449	1	1.006205
hours.per.week	1.120522	1	1.058547
native.country	2.498144	4	1.121249

The VIF value acts as a good indicator for multicollinearity. The VIF of a predictor is a measure for how easily it is predicted from a linear regression using the other predictors. Taking the square root of the VIF tells you how much larger the standard error of the estimated coefficient is respect to the case when that predictor is independent of the other predictors.VIF values more than 5 are considered as a sign of presence of multicollinearity. Here, we see that Relationship and marital.status have VIF's more than 5.

Part II : Model Fitting

As we see that the dependent variable of our data is Binary Variable which has only Two levels. (Income Class ->0 and 1). For this, general linear model will not be applicable here as the predicted value will cluster into two classes which is not proper. We need a model such that it will help us to classify the Income Class of a given entry if other informations are provided. For this we will apply Logistic Regression model. To apply this logistic regression model on our data, we will use R software and its package. But before that we should know what is logistic regression and why it will help on our data in detail.

a) Logistic Regression

In the linear regression model $\boldsymbol{X}\beta + \varepsilon$, there are two types of variables – explanatory variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_k$ and study variable $\boldsymbol{y}$. These variables can be measured on a continuous scale as well as like an indicator variable. When the explanatory variables are qualitative, then their values are expressed as indicator variables, and then dummy variable models are used. When the study variable is a qualitative variable, then its values can be expressed using an indicator variable taking only two possible values 0 and 1. In such a case, the logistic regression is used. For example, y can denote the values like success or failure, yes or no, like or dislike, which can be denoted by two values 0 and 1.

The log odds:

$$\frac{p}{1-p} = b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$$

By simple algebraic manipulation (and dividing numerator and denominator by $\boldsymbol{b}^{\boldsymbol{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}$), the probability that Y = 1 is

$$p = \frac{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k} + 1} = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$
$$= S_b(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

Where S_b is the sigmoid function with base b . The above formula shows that once β_i are fixed, we can easily compute either the log-odds that Y = 1 for a given observation, or the probability that Y = 1 for a given observation. The main use-case of a logistic model is to be given an observation $(x_1, x_2, \ldots, x_k)$, and estimate the probability p that Y = 1 . In most applications, the base b of the logarithm is usually taken to be e. However, in some cases it can be easier to communicate results by working in base 2, or base 10.

First we start off by building a model with all the categorical variables:

```
Null deviance: 35469 on 31661 degrees of freedom
Residual deviance: 24508 on 31628 degrees of freedom
AIC: 24576
```

Then we add the Continuous variables to the model and see whether the AIC and Residual Deviance decreases.

```
Null deviance: 35469 on 31661 degrees of freedom
Residual deviance: 20531 on 31623 degrees of freedom
AIC: 20609
```

Both the AIC and the Residual Deviance decreases substantially. Hence, the inclusion of Continuous variables improves the model. We call this model Full-Model as it uses all the 12 variables.

Let's check the Accuracy, Specificity, Sensitivity, and Precision of this model for future reference.

	FALSE	TRUE
0	9417	787
1	1352	2004

Accuracy = 0.84225, Sensitivity = 0.71802
Specificity = 0.87445, Precision = 0.59713

Choosing the optimal model

A model which uses all the contributing predictor variables available to it might seem like a good idea. But the more features we have the more likely our model will suffer from overfitting.

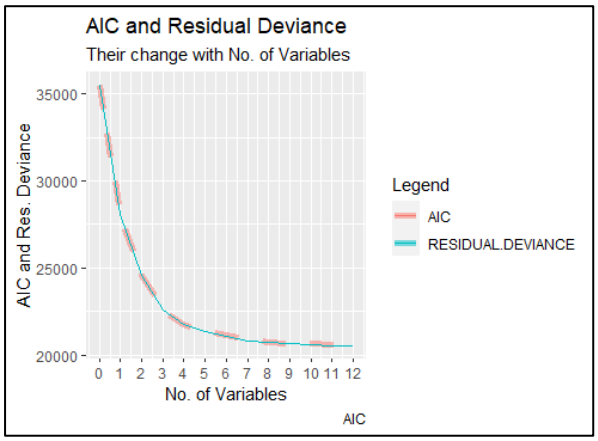
This is the reason we drop variables that we can from a model to make it optimal. We can do that by various method. Here we use the following on the Logistic Model:

i) Stepwise Selection using AIC
ii) Information Value of the Predictors

i) Stepwise Selection using AIC

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. AIC is founded on information theory. When a statistical model is used to represent the process that generated the data, the representation will almost never be exact; so, some information will be lost by using the model to represent the process. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model. In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model.

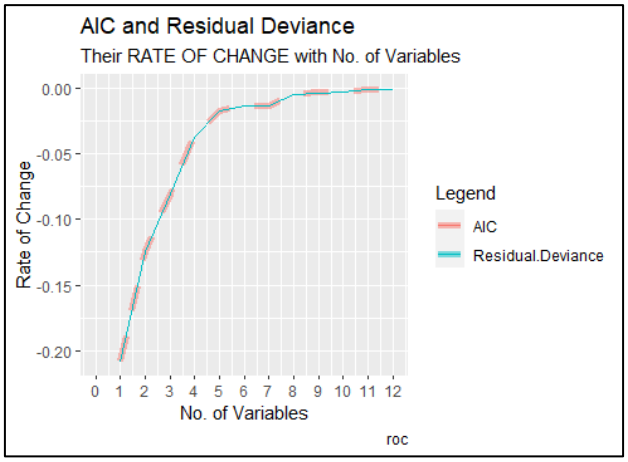
In time of analysis, we started with a Logistic Regression model having all independent variable.
We calculated AIC value of the model. Now our goal is to reduce the AIC as much as possible while also selecting the most optimal number of features.
Now while performing the Backward elimination of the variables based on AIC, the process stops at the Full-Model (AIC: 20609) itself indicating that elimination of none of the variables from the Full-Model reduces the AIC.
So, we try the Forward Selection. Even in this the Least AIC is achieved when the model selects all the variables i.e. we arrive at the Full-model. So in order to optimize, we look more closely at the entire Forward Selection process and note the AIC values after inducting a new variable into the model. Then we plot a graph of the AIC and Residual Deviance values varying against No. of variables in the model.



As we can see the plot exhibits a decay which is quite expected a Forward Selection selects the variable first which causes the most reduction in AIC. The residual deviance and AIC graphs coincide at most places.

Now in order to observe if there is an optimum number of variables after which the AIC does'nt change much. To observe this, we plot the Rate of Change of AIC and Residual Deviance varying against No. of variables.

From this graph we can see that the Rate of Change of bot AIC and Residual Deviance remains constant after 9 variables in the model.
So the optimum number of variables in the model is 9, after which the AIC value is not reduced much.



So, we observe the sequence in which the variables enter the Model and stop the sequence after the 9th variable i.e. workclass. This eliminates 3 variables i.e. marital.status, native.country and race.

Let's look at the AIC and the accuracy, sensitivity, specificity and Precision.

```
Null deviance: 35469 on 31661 degrees of freedom

Residual deviance: 20626 on 31633 degrees of freedom

AIC: 20684
```

```
   FALSE  TRUE
0   9427   777
1   1353  2003
```

Accuracy = 0.84292,	Sensitivity = 0.72050
Specificity = 0.87448,	Precision = 0.59684

ii) Information Value of the Predictors

Information Value analysis is a data exploration technique that helps determine which columns in a data set have predictive power or influence on the value of a specified dependent variable.

From the p-values of the Full- model we have seen that all the continuous variables are significant for the prediction. So, we find the Information Value for the Categorical predictors.

	VARS	IV	STRENGTH
	<chr>	<dbl>	<chr>
4	relationship	1.486080501	Highly Predictive
2	marital.status	1.279001637	Highly Predictive
3	occupation	0.753309772	Highly Predictive
7	gender	0.296342929	Highly Predictive
1	workclass	0.080637302	Somewhat Predictive
5	race	0.066064849	Somewhat Predictive
6	native.country	0.008575208	Not Predictive

It shows that native.country might not have predictive power or influence on the Income class.

So, we make another model with all the features except native.country.

```
Null deviance: 35469 on 31661 degrees of freedom
Residual deviance: 20557 on 31627 degrees of freedom
AIC: 20627
```

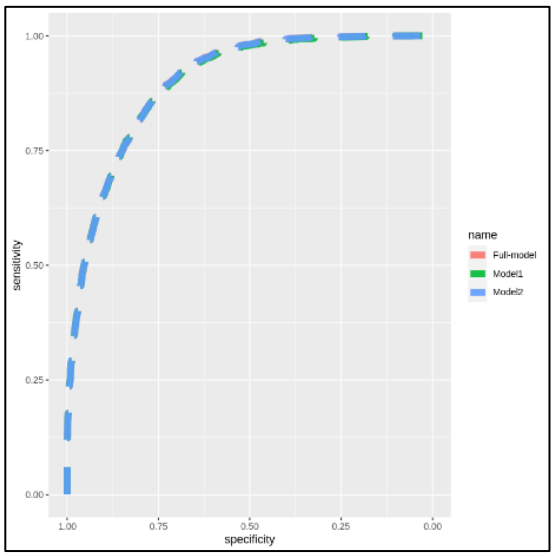
```
   FALSE  TRUE
0   9420   784
1   1357  1999
```

Accuracy = 0.84210,	Sensitivity = 0.71828
Specificity = 0.87408,	Precision = 0.59564

Comparing all the Logistic models:

Let's compare the two Logistic Models and the Full-model based on their performance and the number of variables in the model.

FULL MODEL:	MODEL 1:	MODEL 2:
Number of Variables = 12	Number of Variables = 9	Number of Variables = 11
AIC = 20609	AIC = 20684	AIC = 20627
Residual deviance: 20531	Residual deviance: 20626	Residual deviance: 20557
Accuracy = 0.84225	Accuracy = 0.84292	Accuracy = 0.84210
Sensitivity = 0.71802	Sensitivity = 0.72050	Sensitivity = 0.71828
Specificity = 0.87445	Specificity = 0.87448	Specificity = 0.87408
Precision = 0.59713	Precision = 0.59684	Precision = 0.59564
AUC = 0.89841	AUC = 0.89697	AUC = 0.89814



Interpretation:

The three ROC curves coincide and only differentiate slightly towards the end.

This also reflects in the AUC values of these curves which are very close to each other.

We notice here that the AIC values and Residual Deviance of Model1 and Model2 are higher than that of Full-model. But, Model 1 uses the least number of variables i.e 9. It also has the highest Accuracy, Specficity and Sensitivity. Furthermore, there is only a difference in the 3rd decimal place of AUC for the 3 models.

Looking at this observations, we select Model 1 to be the most optimum model in terms of Model Performance. Though it has a high AIC, but we have already seen that the Rate of Change of AIC after 9 variables in the model stagnates to a minimum.

Diagonostic checking of the optimal model and assessing its Goodness of fit:

I. Checking for Overdispersion

When we apply logistic regression in data, then we have to check that whether the observed variance is larger than the expected from the logistic model. If the dispersion is higher than expected, then overdispersion exists. It is a situation where the residual deviance of the model is large relative to the residual degrees of freedom. If Overdispersion exists, then it indicates that the model does not fit the data well. The explanatory variables may not describe the model. If there exists overdispersion, one potential solution Beta-Binomial family and Quasi-Likelihood method. One thumb rule to detect overdispersion exists or not is

$$\frac{Residual\ Deviance}{d.f_{\text{-}residual}} > 1.5$$

If the ratio value is greater than 1.5, then overdispersion exists, otherwise it is not.

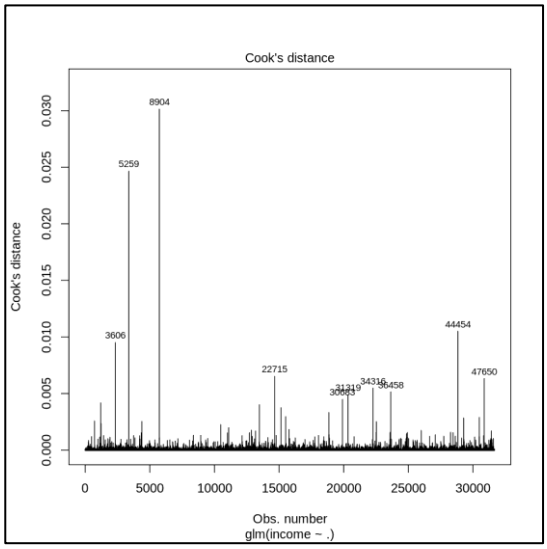
For Model1, Dispersion = 1.0059 < 1.5
Hence, Overdispersion doesn't exist.

II. Checking if Model1 fulfils the Assumptions of Logistic Regression

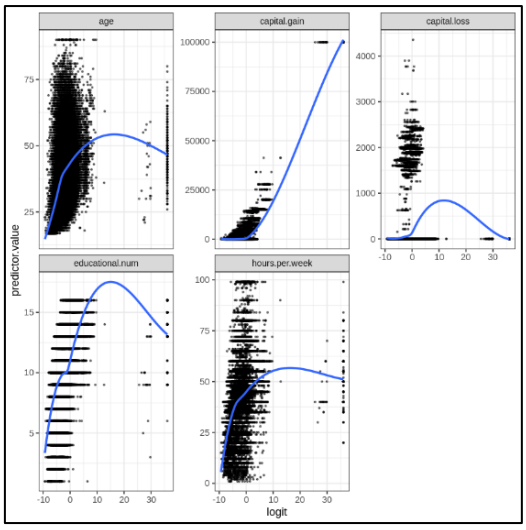
a) The response variable Income Class is binary(0 & 1).
b) There is no Multicollinearity since we eliminate marital.status and the VIF values for the variables in the model are less than 5.

	GVIF	Df	GVIF^(1/(2*Df))
relationship	3.120097	5	1.120513
educational.num	1.473693	1	1.213958
capital.gain	1.026506	1	1.013166
occupation	2.293388	13	1.032439
capital.loss	1.011418	1	1.005693
hours.per.week	1.116483	1	1.056637
age	1.124927	1	1.060626
gender	2.806481	1	1.675256
workclass	1.489168	4	1.051037

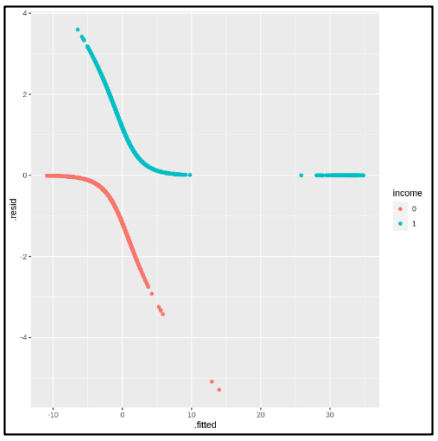
c) From Cook's distance plot we can see there are outliers in the data. But we keep them and we know it is mostly due to the Capital.loss and Capital.gain columns. But we have already explained the behaviour due to the presence of a lot of zero values in these columns.



d) Logistic regression assumes a linear relationship between Explanatory variables and the logit of the Response Variable. Here for the continuous variables, age, capital.loss, hours.per.week show non-linearity. But we keep them, as we saw at the beginning that the model with the Continuous variables and Categorical variables has significantly lower AIC than the ones without the Continuous variables.

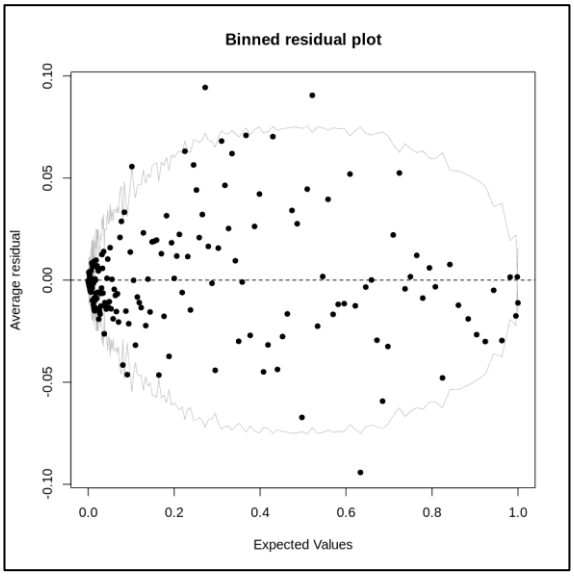


III. Assessing the model fit:



For this we plot the residual vs fitted plot. Now, we have to remember that our independent variable is categorical variable with two levels. On this kind of data, whether we can apply Negative Binomial or Poisson distribution or Logistic Regression. Now if our Residual vs Fitted Plot showed scattered points plotted over our plot then we should decide that we have to plot Negative Binomial or Poisson and if our plot shows points are clustered on our plot we should apply Logistic regression. Since we see kind of discrete pattern in our plot, our decision to fit Logistic Regression Model is correct.

Residual vs. Fitted plots are typically not very useful for Logistic Regression because of the discrete nature of residuals from these models.



Binned Residual plots as recommended by Gelman & Hill (2007) can be used to assess both the overall fit of regression models for binary outcomes and the inclusion of Continuous variables.
The grey lines represent $\pm 2SE$ bands, which we would expect to contain about 95% of the observation.
This model looks reasonable (though there is a slight bunching on the left side of the plot) in that majority of the fitted values seem to fall within the SE bands.

Thus, both the plots indicate a good fit of the Logistic regression Model1.

b) Random Forest

Random forest is a supervised learning algorithm. The “forest” it builds is an ensemble of decision trees, usually trained with the “bagging” method. random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Random forest is also a very handy algorithm because the default hyperparameters it uses often produce a good prediction result.

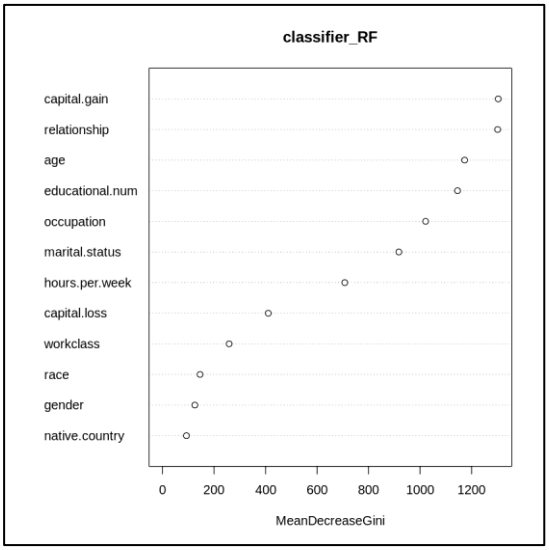
```
y_pred
     0    1
0 9556  648
1 1239 2117
```

Accuracy = 0.86084,	Sensitivity = 0.76564
Specificity = 0.88522,	Precision = 0.63081

Understanding the hyperparameters is pretty straightforward, and there's also not that many of them. One of the biggest problems in machine learning is overfitting, but most of the time this won't happen thanks to the random forest classifier. If there are enough trees in the forest, the classifier won't overfit the model.

Here, we fit the Random Forest model on all the 12 variables and measure the accuracy, sensitivity, specificity and Precision.

Choosing the optimal model for Random Forest Model:



The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The higher the value of mean decrease Gini score, the higher the importance of the variable in the model. From the graph we see that native.country, gender and race have the least Mean Decrease Gini score indicating less importance in the model.

Here for facilitating better comparison with the Logistic Model1 (9 variable), we also bring this one down to 9 variables. We call the new Random Forest model RF1.

a) Support Vector Machine(SVM)

Support Vector Machine (SVM) is a relatively simple Supervised Machine Learning Algorithm used for classification. It is more preferred for classification but is sometimes very useful for regression as well. Basically, SVM finds a hyper-plane that creates a boundary between the types of data. In 2-dimensional space, this hyper-plane is nothing but a line. In SVM, we plot each data item in the dataset in an N-dimensional space, where N is the number of features/attributes in the data. Next, we find the optimal hyperplane to separate the data. SVM works very well without any modifications for linearly separable data. Linearly Separable Data is any data that can be plotted in a graph and can be separated into classes using a straight line.

Here, we fit the data in a Linear SVM model and measure the accuracy, sensitivity, specificity and Precision.

```
Parameters:

SVM-Type: C-classification

SVM-Kernel: linear

cost: 1

Number of Support Vectors: 11154 ( 5587 5567 )
```

y_pred		
	0	1
0	9431	773
1	1360	1996

Accuracy = 0.84277, Sensitivity = 0.72083
Specificity = 0.87396, Precision = 0.59475

Choosing the comparable model for SVM:

For making the SVM model comparable in terms of no. of variables, we take the 9 variables used in the Logistic Model 1. We call the new SVM model SVM1.

Part III : Model Selection:

LOGISTIC MODEL 1:	RF 1:	SVM 1:
Number of Variables = 9	Number of Variables = 9	Number of Variables = 9
Accuracy = 0.84292	Accuracy = 0.85597	Accuracy = 0.84269
Sensitivity = 0.72050	Sensitivity = 0.75611	Sensitivity = 0.72083
Specificity = 0.87448	Specificity = 0.88124	Specificity = 0.87396
Precision = 0.59684	Precision = 0.61710	Precision = 0.59475

For the same number of Variables, Random Forest model has the most Accuracy, Sensitivity, Specificity and Precision. The Best Logistic Model ranks 2nd in terms of Accuracy, while the linear SVM comes 3rd(indicating that it probably would have needed a different kernel other than Linear).

SECTION 4

CONCLUSION

CONCLUSION

Whether a person has a higher income or not, it is generally associated with Number of years of education i.e. the qualifications of the individual, Number of Work-hours, Work-class (Private or Government) and type of Occupation. Socially, we have also attached Gender, Native Country and age as determining factors of higher income. Here in this project, our goal is to determine objectively the information that are important to decide whether an individual has a higher income or not. This project is carried out to select such information that effect the Income class and then should be kept in mind while making predictions about it.

During the analysis, we first prepared the data by removing missing values and redundant columns. This was followed by some exploratory data analysis to understand the structure of the data as well as the relationships between various variables through various plots. We noticed there isn't much interdependency between continuous variables. Then we selected the optimum Logistic Model with 9 variables for classifying Income class using Rate of change of AIC values and Accuracy measures.

In this project we have also used Random Forest and Linear Support Vector Machines to classify the income class. As expected, Random Forest provided the most Accuracy of classification, followed by the Logistic Regression Model. In terms of Accuracy, The Linear SVM model was close but least accurate among the 3 models. This could be due to the fact that we used a Linear kernel. If we would have used any other kernel, it might have given us better result but it would have further complicated the calculations.

Though Random Forest performs the best, these algorithms are like black boxes. It is difficult to understand how the individual predictors are being used in the models. By contrast,a logistic model, though not quite as strong predictively in this case, really shines when it comes to describing relationships among variables. In short, logistic regression will often (but not always) underperform the best ML algorithms in prediction but will outperform them in description.

An additional consideration is speed; the logistic function is very fast compared to random forest, and supports quick iterative learning about the structure of a dataset.

To conclude, we can say that the performances of the choices though in a hierarchy, are still very close i.e the accuracies of classification by 9 variabled Random Forest, Logistic and SVM model are 0.85597, 0.84292 and 0.84269 respectively. Also we see from the Logistic Regression model that there are other

variables other than Number of years of Education like Relationship status, Gain or Loss from Investments, type of Occupation, Working hours per week, age, gender and workclass that determines the Income Class of a person.

Bibliography

1. https://bookdown.org/jefftemplewebb/IS-6489/logistic-regression.html#assessing-logistic-model-performance
2. https://www.statology.org/assumptions-of-logistic-regression
3. http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r
4. Practical Guide to Logistic Regression by Joseph M. Hilbe
5. Statistical Inference – G. Casella & R.L. Berger