Building classifier for Income Prediction

Introduction

The problem of income inequality is prevalent in almost every country in the world including the United States. The government and the international institutions have been trying to reduce the income inequality and bring the people out of poverty. Economic upliftment of the poor has been the centre of policy making for different governments since decades. In fact, Economic well-being is one of the main criteria that assesses the welfare of a nation. This study aims to look at the problem of income equality in the United States and identifies the characteristics that are necessary to increase the income levels of an individual. This study applies the classification techniques of machine learning to determine the factors that affect the income. The goal is to build to a classifier to predict whether a person would earn more than $50k in a year or not based upon the certain characteristics. The data for the study has been taken from the UCI website.

Literature Review

Different studies have been undertaken by the researchers in the past to predict the income levels using machine learning techniques. Haojun Zhu [1] applied logistic regression as the modelling tool and various machine learning techniques such as random forest, classification, neural network and supporting vector machine to predict the income levels. Deepajothi et. al. [2] replicated the Bayesian networks, lazy classifier, decision tree induction to predict the income levels and produced a comparative analysis of the different predictive models. Chockalingam et. al. [3] also used various machine learning algorithms such as logistic regression, naïve bayes, k-nearest neighbor, decision trees, gradient boosting and other algorithms in predicting the income.

While on the other hand, there were researchers that applied advanced modeling techniques on the data. Topiwalla [4] used comples algorithms such as XGBOOST on the dataset and presented different ways to scale up the accuracy of the models by using techniques like Logistic Stack on XGBOOST. Similarly, Lemon et. al. [5] identified the important parameters in the dataset that could possibly help in optimizing complexity of machine learning models in the classification process.

Dataset

The dataset has been obtained from the UCI machine learning repository and is known as the Adult Data Set/ Census income dataset and is available in the training and the test set directly from the source. There are in total 32,561 records in the data and it consists of 14 attributes, out of which, 8 attributes are categorical and 6 attributes are continuous in nature. The details of the attributes are given in the table below:

Variable Name	Description	Type	Labels
Age	Age of the individual	Continuous	Numeric
Workclass	Class of Work	Categorical	?,Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwgt	Final Weight Determined by Census Org	Continuous	Numeric
Education	Education of the individual	Ordered Factor	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Education-num	Number of years of education	Continuous	Numeric
Marital-status	Marital status of the individual	Categorical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Occupation of the individual	Categorical	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces

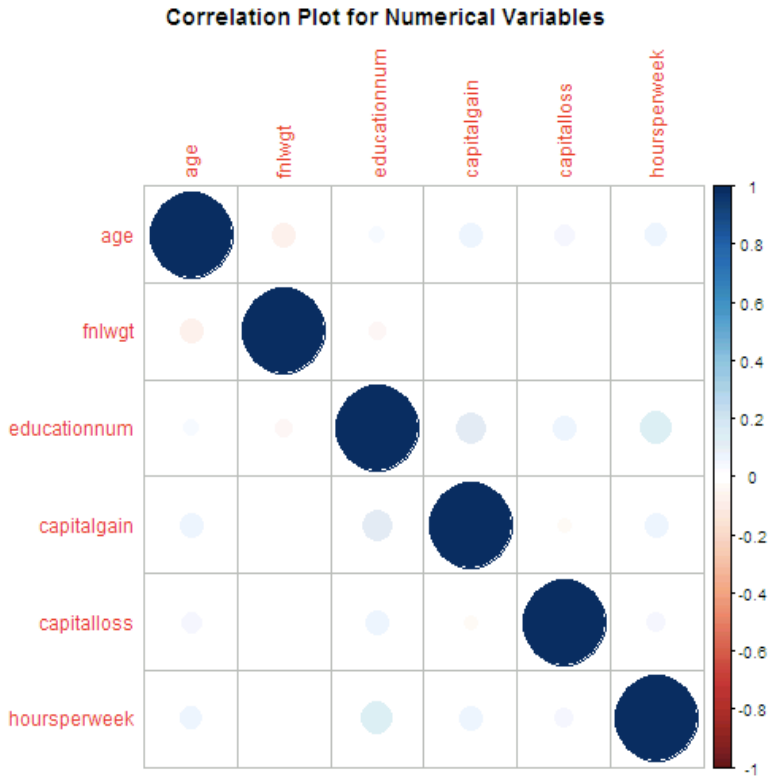
Relationship	Present relationship	Categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	Race of the individual	Categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Sex of the individual	Categorical	Female, Male
Capital-gain	Capital gain made by the individual	Continuous	Numeric
Capital-loss	Capital loss made by the individual	Continuous	Numeric
Hours-per-week	Average number of hours spent by the individual on work	Continuous	Numeric
Native-country	Average number of hours spent by the individual on work	Categorical	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

The overall summary and some descriptive statistics of the dataset are as given below:

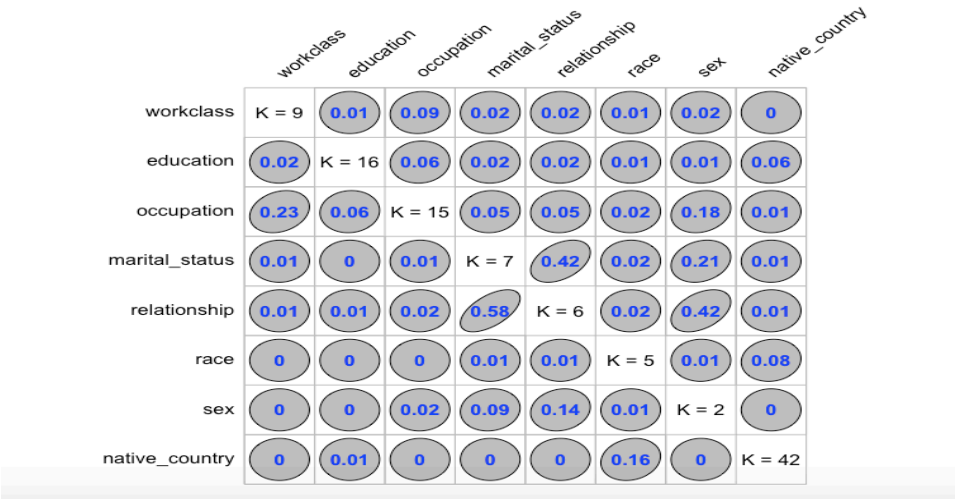
```
'data.frame':   32561 obs. of  15 variables:
 $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
 $ workclass     : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
 $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
 $ education     : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
 $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
 $ marital_status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
 $ occupation    : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
 $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
 $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
 $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
 $ capital_gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
 $ capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
 $ native_country: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
 $ income        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

```
     age                  workclass        fnlwgt              education     education_num                  marital_status
 Min.   :17.00   Private         :22696   Min.   :  12285   HS-grad     :10501   Min.   : 1.00   Divorced             : 4443
 1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827   Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse    :   23
 Median :37.00   Local-gov       : 2093   Median : 178356   Bachelors   : 5355   Median :10.00   Married-civ-spouse   :14976
 Mean   :38.58   ?               : 1836   Mean   : 189778   Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
 3rd Qu.:48.00   State-gov       : 1298   3rd Qu.: 237051   Assoc-voc   : 1382   3rd Qu.:12.00   Never-married        :10683
 Max.   :90.00   Self-emp-inc    : 1116   Max.   :1484705   11th        : 1175   Max.   :16.00   Separated            : 1025
                 (Other)         :  981                     (Other)     : 5134                   Widowed              :  993
           occupation           relationship                    race           sex         capital_gain    capital_loss
  Prof-specialty :4140    Husband       :13193    Amer-Indian-Eskimo:  311    Female:10771   Min.   :    0   Min.   :   0.0
  Craft-repair   :4099    Not-in-family : 8305    Asian-Pac-Islander: 1039    Male  :21790   1st Qu.:    0   1st Qu.:   0.0
  Exec-managerial:4066    Other-relative:  981    Black             : 3124                   Median :    0   Median :   0.0
  Adm-clerical   :3770    Own-child     : 5068    Other             :  271                   Mean   : 1078   Mean   :  87.3
  Sales          :3650    Unmarried     : 3446    White             :27816                   3rd Qu.:    0   3rd Qu.:   0.0
  Other-service  :3295    Wife          : 1568                                               Max.   :99999   Max.   :4356.0
 (Other)         :9541
 hours_per_week        native_country     income
 Min.   : 1.00   United-States:29170    <=50K:24720
 1st Qu.:40.00   Mexico       :  643    >50K : 7841
 Median :40.00   ?            :  583
 Mean   :40.44   Philippines  :  198
 3rd Qu.:45.00   Germany      :  137
 Max.   :99.00   Canada       :  121
                 (Other)      : 1709
```

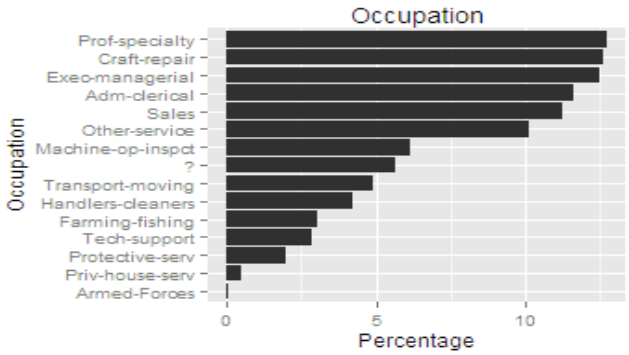
Exploratory data analysis:



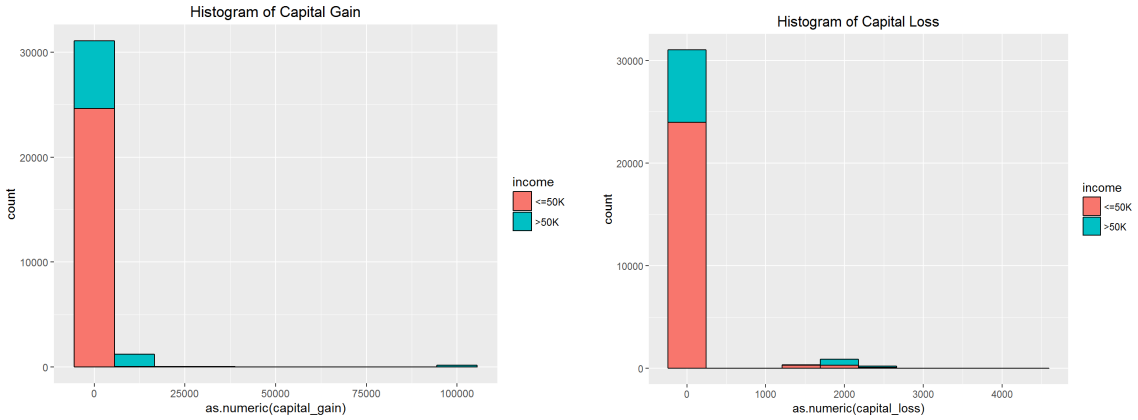
From the correlation matrix, it is clear that the numerical variables do not have a high strength of correlation.



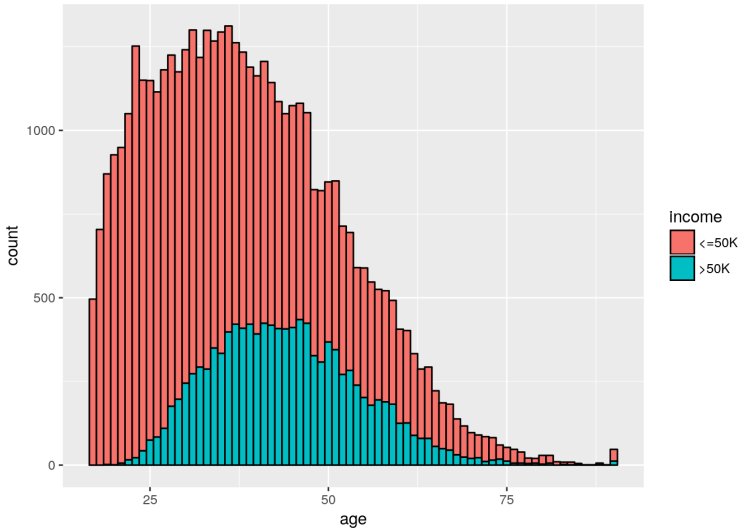
Correlation chart for the categorical Variables



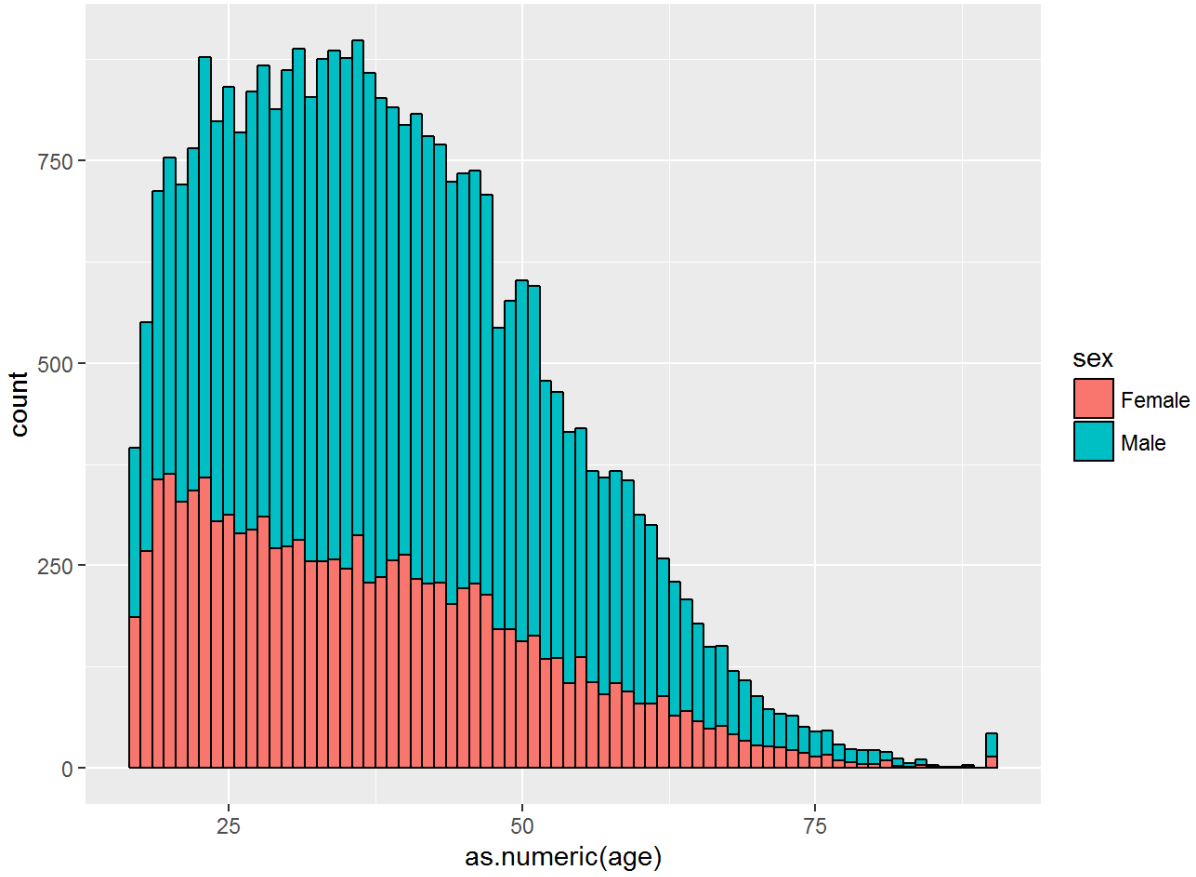
The above figure shows the distribution of occupation, in terms of percentage.



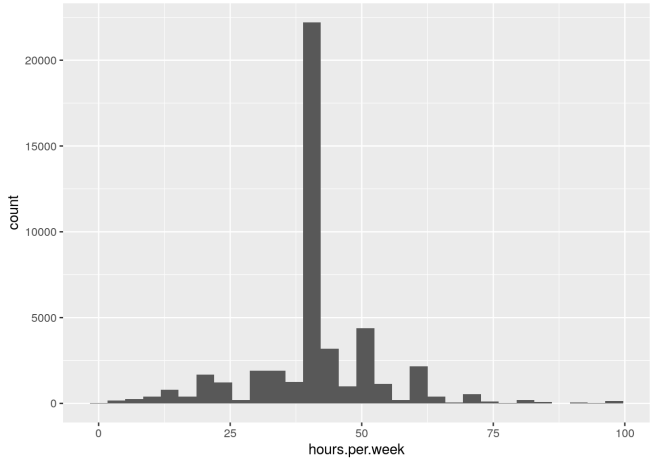
The histograms for capital gain and capital loss are shown above. As most of these records are equal to zero, therefore, they are dropped for the purpose of analysis. Similarly, the native country attribute is also skewed as the most entries are from the United States.



The above figure shows the distribution of income with respect to the age. People between the ages of 25-50 have the highest share of people who earn more than 50K.

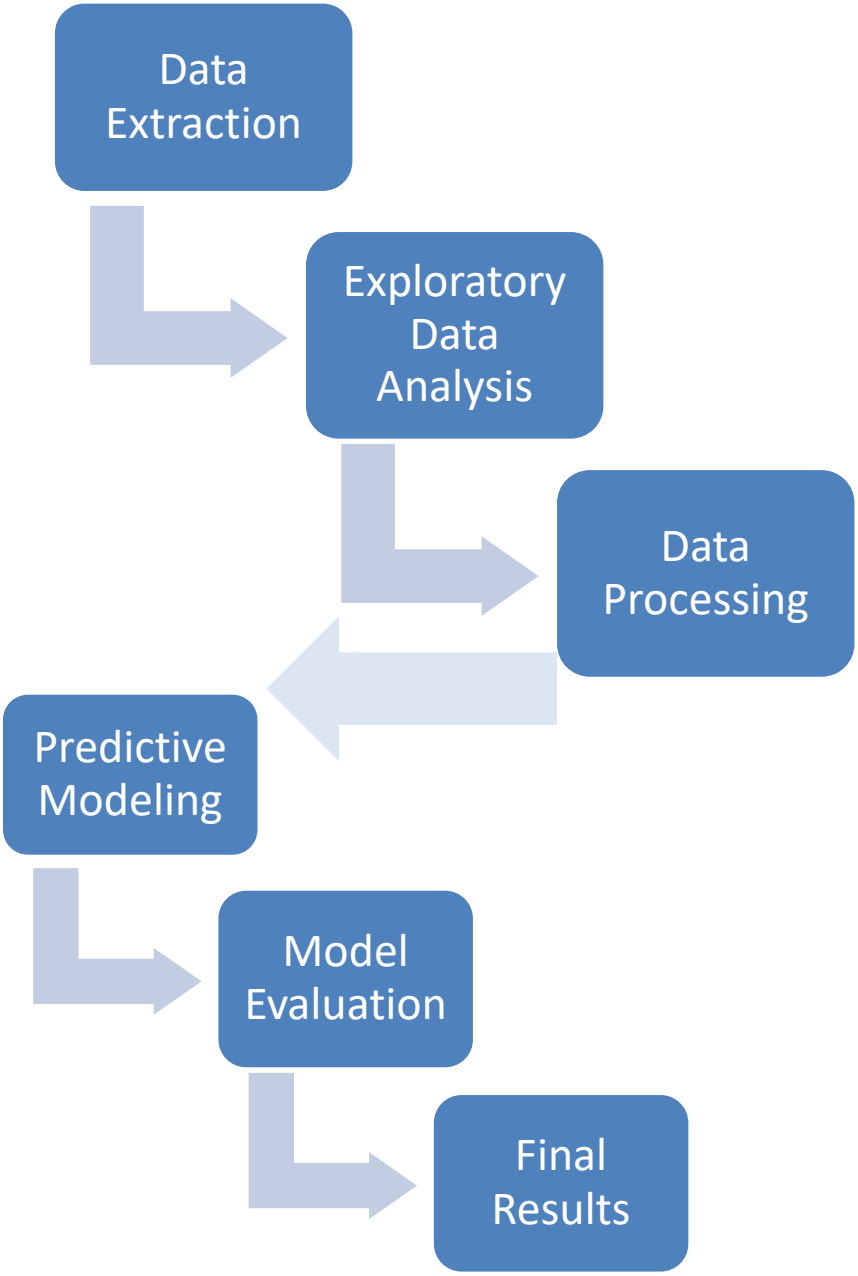


The above figure shows the count of age with respect to the gender. Males are highly represented in the data.



The above chart shows the distribution of hours per week. People working around 40 hours per week has the highest share.

Approach



Step 1: Data Extraction

The first step consists of extracting the data from its source. In this study, the data was extracted in the form of training and test set data directly from the source of the data, that is, UCI machine learning repository.

Step 2: Exploratory Data Analysis

The Second step consists of doing the preliminary analysis on the data. This involves checking for correlation among the variables, analyzing the descriptive statistics. In this study, these steps were performed on the training data set and a number of different measures were used as correlation matrix

to assess the correlation among the numerical variables, histograms of different variables with respect to the income attribute, box plots of different attributes and summary statistics of a few variables.

Step 3: Data Processing

This steps consists of preparing the data for the purpose of analysis. This involves the treatment of NA's in the dataset and variable selection. In this study, there were direct records with NA's, however, there were certain records with '?' in the 'work class' and 'native-country' attribute. So, for the purpose of analysis and predictive modeling stage, these records, which around 1800 were dropped from the analysis. Similarly, variables such as 'fnlwt', 'education', 'relationship', 'capital gain', 'capital loss' and 'native country' were dropped based upon different feature selection methods and as such they had no impact on the income attribute. The entire dataset was cleaned before the application of any predictive model.

Step 4: Predictive Modeling

This step consists of applying the classification models to the dataset. The building of the model is done on the training data and it is applied for analysis on the test dataset. This study applies logistic regression, naïve bayes and random forest model to predict the income attribute.

Step 5: Model Evaluation

Once different predictive models have been applied to the dataset, the effectiveness of each algorithm is evaluated using different characteristics. In this study, the models were evaluated on the accuracy, recall, ROC area and precision values.

Step 6: Final Results

This step consists of the drawing the final conclusion once all the models on their parameters have been evaluated. This section highlights the main findings of the study.

Attribute Selection:

The first stage in applying classification algorithms was choosing the right attributes for the predictive model. To do so, different methods were applied on the raw dataset to determine the best attributes:

1. Correlation subset evaluation:

This method correlates each attribute with the class attribute and ranks all the attributes in the order of their correlation with the class attribute. This technique was also applied 10 fold Cross Validation set and all the parameters were set as default. The outcome of this method is given below:

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0.335 +- 0.002     1.1 +- 0.3     6 education_num
 0.331 +- 0.001     1.9 +- 0.3     7 marital_status
 0.27  +- 0.001     3   +- 0       9 relationship
 0.234 +- 0.001     4   +- 0       2 age
 0.23  +- 0.001     5   +- 0      14 hours_per_week
 0.223 +- 0.001     6   +- 0      12 capital_gain
 0.216 +- 0.001     7   +- 0      11 sex
 0.151 +- 0.001     8   +- 0      13 capital_loss
 0.109 +- 0.001     9   +- 0       5 education
 0.097 +- 0        10   +- 0       8 occupation
 0.082 +- 0.002    11   +- 0      10 race
 0.071 +- 0.002    12   +- 0       3 workclass
 0.033 +- 0.002    13   +- 0      15 native_country
 0.009 +- 0.002    14   +- 0       4 fnlwgt
 0.005 +- 0.002    15   +- 0       1
```

The cut-off was set to 0.071. All the attributes equal to and above this value were selected for the purpose of the analysis, while the others were rejected.

2. Info Gain subset evaluation:

The info gain method provides the information gain of all the attributes and ranks them in the order of the highest to the lowest. All the parameters were set as default and applied on the 10 fold CV set. The output is given below:

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0.165 +- 0.001     1   +- 0       9 relationship
 0.157 +- 0.001     2   +- 0       7 marital_status
 0.115 +- 0.001     3   +- 0      12 capital_gain
 0.098 +- 0         4   +- 0       2 age
 0.094 +- 0.001     5.3 +- 0.46    5 education
 0.093 +- 0.001     6.3 +- 0.9     8 occupation
 0.093 +- 0.001     6.4 +- 0.49    6 education_num
 0.058 +- 0.001     8   +- 0      14 hours_per_week
 0.05  +- 0.001     9   +- 0      13 capital_loss
 0.037 +- 0        10   +- 0      11 sex
 0.022 +- 0.001    11   +- 0       3 workclass
 0.009 +- 0        12.2 +- 0.4    15 native_country
 0.008 +- 0        12.8 +- 0.4    10 race
 0     +- 0        14   +- 0       4 fnlwgt
 0     +- 0        15   +- 0       1
```

The cut-off was set to 0.022. All the variables above this threshold were selected while others were rejected. It is noteworthy to say that the variables such ‘capital gain’, ‘capital loss’ and were ranked higher than the threshold but still they were rejected as they had most values equal to zero.

The continuous variable ‘fnlwgt' represents final weight, which is the number of units in the target population that the responding unit represents. For the purpose of this analysis, this variable is discarded. Total number of years of education can be represented by the highest education level completed. Similarly, role in the family can be determined from gender and marital status. Thus, the 3 variables, namely, fnlwgt, education and relationship are dropped.

Overall, 8 attributes were retained while 6 were dropped for the prediction of the class attribute.

Selected attributes:

Attributes
age
workclass
education_num
marital_status
occupation
race
sex
hours_per_week
income

Predictive Modeling/Classification

This section looks into the different classification algorithms applied on the dataset. The aim of this section is to explain the rational involved in the process of attribute selection and choosing the best classification model.

Classification Algorithms:

The next step in the predictive model building is applying the different classification algorithms on the selected attributes dataset. In total, three different algorithms were applied, namely, the logistic regression, the Naïve Bayes and the Random Forest. All these algorithms were applied on the 10 fold Cross Validation set and not on the full training or the test set because using k fold Cross Validation

technique is simple to understand and it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times. The number of folds was set 10 as k=10 is widely used in the area of machine learning and it results in a model skill estimate with low bias a modest variance.

1. Naïve Bayes Algorithm:

The first algorithm applied to our dataset was the Naïve Bayes classification. The results of this classification are given below:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       26699               81.9969 %
Incorrectly Classified Instances      5862               18.0031 %
Kappa statistic                          0.52
Mean absolute error                      0.217
Root mean squared error                  0.3541
Relative absolute error                 59.3355 %
Root relative squared error             82.8258 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.869    0.336    0.891      0.869    0.880      0.521    0.876     0.957      <=50K
                 0.664    0.131    0.617      0.664    0.640      0.521    0.876     0.681      >50K
Weighted Avg.    0.820    0.286    0.825      0.820    0.822      0.521    0.876     0.891

=== Confusion Matrix ===

     a     b   <-- classified as
 21492  3228 |     a =  <=50K
  2634  5207 |     b =  >50K
```

From the above figure, we can conclude that:

Overall Accuracy: 81.9%

Overall Precision: 82.5%

Overall Recall: 82%

2. Random Forest Algorithm:

The second classification applied to the dataset was the Random forest. The output of this classification algorithm are given as follows:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       26560               81.57   %
Incorrectly Classified Instances      6001               18.43   %
Kappa statistic                          0.4797
Mean absolute error                      0.2167
Root mean squared error                  0.3619
Relative absolute error                 59.2734 %
Root relative squared error             84.6466 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.893    0.429    0.868      0.893    0.880      0.481    0.859     0.946      <=50K
                 0.571    0.107    0.629      0.571    0.599      0.481    0.859     0.641      >50K
Weighted Avg.    0.816    0.351    0.810      0.816    0.813      0.481    0.859     0.873

=== Confusion Matrix ===

     a     b   <-- classified as
 22079  2641 |     a =  <=50K
  3360  4481 |     b =  >50K
```

From the above figure, we can conclude that:

Overall Accuracy: 81.57%

Overall Precision: 81%

Overall Recall: 81.6%

3. Logistic Regression:

The last classification applied to the dataset was the Logistic Regression. The output of this classification algorithm are given as follows:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27130               83.3205 %
Incorrectly Classified Instances      5431               16.6795 %
Kappa statistic                          0.5088
Mean absolute error                      0.2276
Root mean squared error                  0.3385
Relative absolute error                 62.2415 %
Root relative squared error             79.1731 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.923    0.451    0.866      0.923    0.894      0.514    0.883     0.959      <=50K
                 0.549    0.077    0.694      0.549    0.613      0.514    0.883     0.690      >50K
Weighted Avg.    0.833    0.361    0.825      0.833    0.826      0.514    0.883     0.894

=== Confusion Matrix ===

     a     b   <-- classified as
 22823  1897 |     a =  <=50K
  3534  4307 |     b =  >50K
```

From the above figure, we can conclude that:

Overall Accuracy: 83.32%

Overall Precision: 82.5%

Overall Recall: 83.3%

Coefficients of different attributes under logistic Regression:

```
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -9.304867   0.278827 -33.371  < 2e-16 ***
## age                       0.028424   0.001632  17.413  < 2e-16 ***
## workclassOther/Unknown   -1.226062   0.812990  -1.508   0.1315
## workclassPrivate          0.065641   0.053401   1.229   0.2190
## workclassSelf-Employed   -0.143134   0.068836  -2.079   0.0376 *
## educatoin_num             0.316604   0.009367  33.799  < 2e-16 ***
## marital_statusMarried     2.009367   0.067075  29.957  < 2e-16 ***
## marital_statusSeparated  -0.118183   0.162774  -0.726   0.4678
## marital_statusSingle     -0.456110   0.082510  -5.528 3.24e-08 ***
## marital_statusWidowed    -0.049819   0.153751  -0.324   0.7459
## occupationOther/Unknown   0.890005   0.811907   1.096   0.2730
## occupationProfessional    0.758835   0.067266  11.281  < 2e-16 ***
## occupationSales           0.493151   0.063295   7.791 6.63e-15 ***
## occupationService         0.156311   0.066728   2.343   0.0192 *
## occupationWhite-Collar    0.775103   0.052391  14.795  < 2e-16 ***
## raceAsian-Pac-Islander    0.127056   0.248409   0.511   0.6090
## raceBlack                 0.307972   0.237403   1.297   0.1945
## raceOther                -0.466651   0.370702  -1.259   0.2081
## raceWhite                 0.489306   0.227223   2.153   0.0313 *
## sexMale                   0.387568   0.051550   7.518 5.55e-14 ***
## hours_per_week            0.030626   0.001629  18.802  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Model Evaluation

All the different algorithms were evaluated on accuracy, precision, recall and ROC area. The results are all given below:

Classifier	Accuracy	Precision	Recall	ROC area
Logistic Regression	83.32%	82.5%	83.3%	0.883
Random Forest	81.57%	81%	81.6%	0.859
Naïve Bayes	81.9%	82.5%	82%	0.876

Logistic Regression appears to be the best classifier with the best figures across the evaluation perimeters. It has the highest accuracy value of 83.32%, joint highest precision value of 82.5%, highest recall value of 83.3% and the maximum area under the ROC curve of 0.883. To evaluate ROC curve, the closer the value to 1, the better it is. So, logistic regression is the best classifier out of all the classifiers.

Results

Overall, logistic regression emerged to be the best classifier for predicting whether the person would earn more than 50K in a year or not. The top characteristics that emerged from the analysis were marital status and occupation. It is interesting to note that within occupation, a person must be working in white collar jobs or as a professional to be able to earn 50K or more. Within the stage of marital status, it comes out that a person who is married is likely to earn greater than 50k in a year.

References:

[1] Haojun Zhu: “Predicting Earning Potential using the Adult Dataset”, https://rstudio-pubsstatic.s3.amazonaws.com/235617 51e06fa6c43b47d1b6daca2523b2f9e4.html

[2] S.Deepajothi and Dr. S.Selvarajan: “A Comparative Study of Classification Techniques On Adult Data Set”, International Journal of Engineering Research Technology (IJERT), ISSN: 2278-0181 Vol. 1 Issue 8, October2012.

[3] Vidya Chockalingam, Sejal Shah and Ronit Shaw: “Income Classification using Adult Census Data”, https://cseweb.ucsd.edu/classes/wi17/cse258- a/reports/a120.pdf.

[4] Mohammed Topiwalla: “Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting”, University of SP Jain School of Global Management.

[5] Chet Lemon, Chris Zelazo and Kesav Mulakaluri: “Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques”, https://cseweb.ucsd.edu/ jmcauley/cse190/reports/sp15/048.pdf.