Machine Learning Project

Goktug Cengiz

Henry Qiu

June 2019

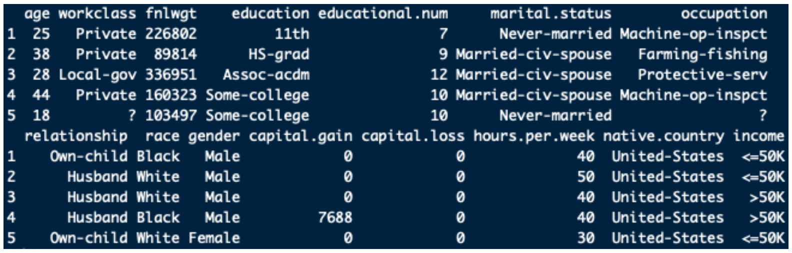
Contents

1 Introduction

The goal of this project is on estimating whether an individual makes more than 50.000 USD a year. The data set[1] used in this project was provided by the UCI Machine Learning Repository. One of the most significant question we wanted to answer with this project was to find how much of an impact does education play in determining an individual's annual income with using machine learning techniques and algorithms. In addition, we wanted to see what other factors play a role in determining income level and by how much.

2 Data Exploration Process

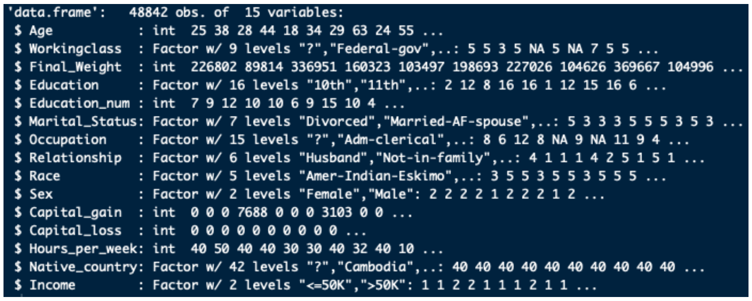
After reading the data in the .csv format we downloaded, we examined the data in the first 5 rows thanks to the head function. As you can see below, there are 2 types of variables, categorical and continuous. In addition, there are missing values (represented by as ?) and NA values will be assigned to them in the recoding section.



```
  age workclass fnlwgt    education educational.num     marital.status        occupation
1  25   Private 226802         11th               7      Never-married Machine-op-inspct
2  38   Private  89814      HS-grad               9 Married-civ-spouse   Farming-fishing
3  28 Local-gov 336951   Assoc-acdm              12 Married-civ-spouse   Protective-serv
4  44   Private 160323 Some-college              10 Married-civ-spouse Machine-op-inspct
5  18         ? 103497 Some-college              10      Never-married                 ?
  relationship  race gender capital.gain capital.loss hours.per.week native.country income
1    Own-child Black   Male            0            0             40  United-States  <=50K
2      Husband White   Male            0            0             50  United-States  <=50K
3      Husband White   Male            0            0             40  United-States   >50K
4      Husband Black   Male         7688            0             40  United-States   >50K
5    Own-child White Female            0            0             30  United-States  <=50K
```

Figure 1: Head of Data Set

According to Figure 2, it can be considered as the data set consist of 48842 observations and 15 variables. Moreover, we can see what the data types of data set are.



```
'data.frame':	48842 obs. of  15 variables:
 $ Age           : int  25 38 28 44 18 34 29 63 24 55 ...
 $ Workingclass  : Factor w/ 9 levels "?","Federal-gov",..: 5 5 3 5 NA 5 NA 7 5 5 ...
 $ Final_Weight  : int  226802 89814 336951 160323 103497 198693 227026 104626 369667 104996 ...
 $ Education     : Factor w/ 16 levels "10th","11th",..: 2 12 8 16 16 1 12 15 16 6 ...
 $ Education_num : int  7 9 12 10 10 6 9 15 10 4 ...
 $ Marital_Status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 3 3 5 5 5 3 5 3 ...
 $ Occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 8 6 12 8 NA 9 NA 11 9 4 ...
 $ Relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 4 1 1 1 4 2 5 1 5 1 ...
 $ Race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 3 5 5 3 5 5 3 5 5 5 ...
 $ Sex           : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 2 2 2 1 2 ...
 $ Capital_gain  : int  0 0 0 7688 0 0 0 3103 0 0 ...
 $ Capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Hours_per_week: int  40 50 40 40 30 30 40 32 40 10 ...
 $ Native_country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 40 40 40 40 40 40 ...
 $ Income        : Factor w/ 2 levels "<=50K",">50K": 1 1 2 2 1 1 1 2 1 1 ...
```

Figure 2: Data Types of Data Set

2.1 Data Pre-processing

Missing and outlier values cause many setbacks in our model. That's why we'll get rid of them.

2.1.1 Missing Value Detection

We can interpret Figure 3 as the variables which are called as "Occupation", "WorkingClass", "Native_Country" have missing values.

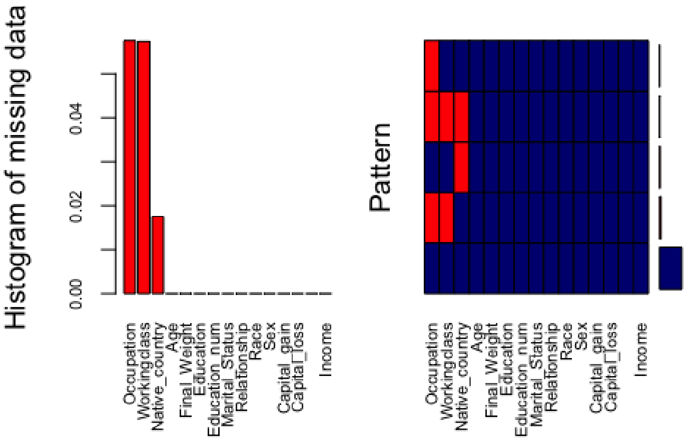


Figure 3: Missing Value Visualization

2.1.2 Missing Value Imputation

Missing values were imputed thanks to missForest packege in R. 'missForest' is used to impute missing values particularly in the case of mixed-type data. It can be used to impute continuous and/or categorical data including complex interactions and nonlinear relations. It yields an out-of-bag (OOB) imputation error estimate. Moreover, it can be run parallel to save computation time. Consequently, we have both of continuous and categorical variables to be imputed that's why we use it.

2.1.3 Outlier Detection

In outlier detection, outliers were detected by Box plot. If a value exceeds the box plot boundary line, it is considered as an outlier. In this case, all of continuous variables were accepted as outlier values.

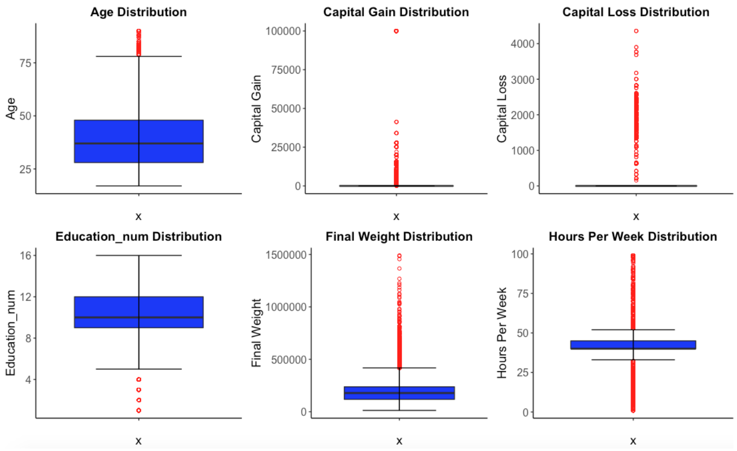


Figure 4: Outlier Detection by Box Plot

2.1.4 Outlier Treatment

NA values were assigned to the detected outliers and treated as if they were missing. Again, using Random Forest, missing values were estimated. Thus, outlier values have been treated.

2.2 Visualization

In this section, our cleaned data set will be visualized by means of pie charts, bar plots, box plots, histograms and they will be interpreted.

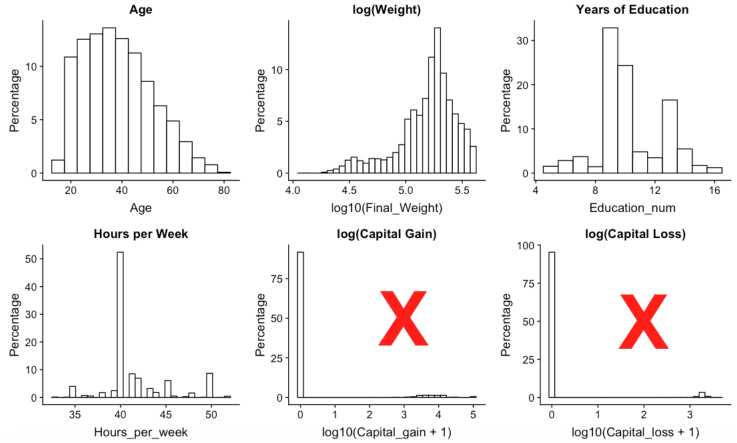


Figure 5: Data Visualization by Histogram for Continuous Variables

According to Figure 6, Age, Final_Weight, Education_num, and Hours_per_week have broad distributions, therefore will be considered for part of building model. However, Capital Gain and Capital Loss have very narrow distributions (more than 90

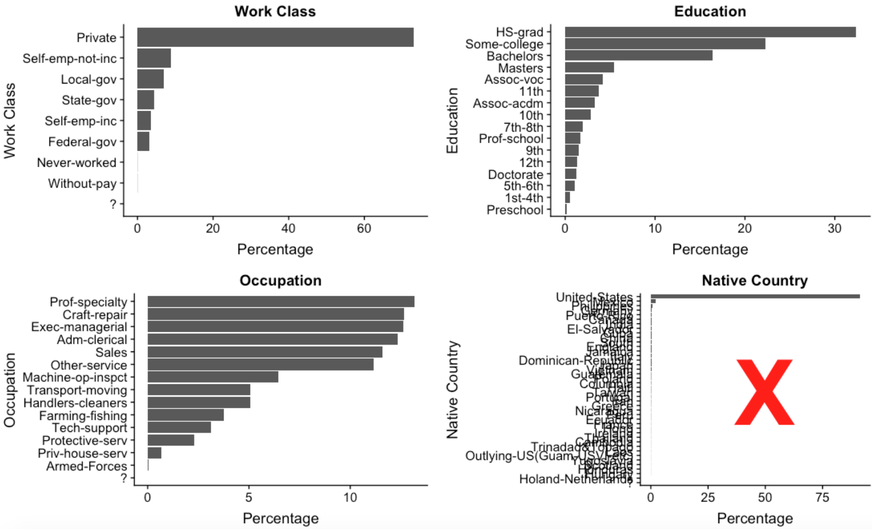


Figure 6: Data Visualization by Histogram for Continuous Variables 2

According to Figure 7, Native_country has a very narrow distribution (%90 of population coming from the United States). Therefore, it will be excluded from the model.

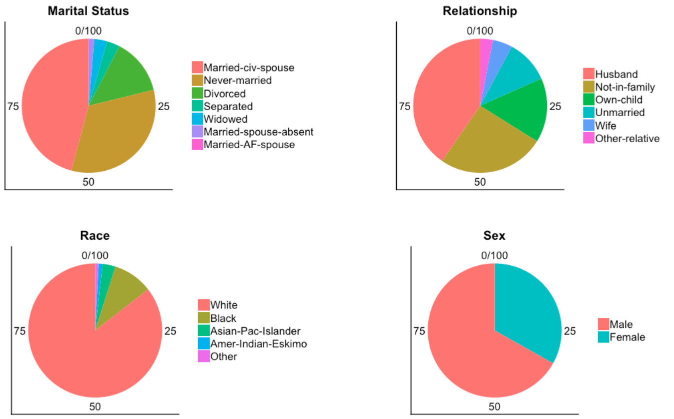


Figure 7: Data Visualization by Pie Chart for Categorical Variables

In the figure above, we can say that all these variables have reasonable spread of distribution, so they will be considered in part of building model. In the following figure 8, the correlation between the continuous values is shown in graphical form. We can say that the numerical variables are nearly uncorrelated.

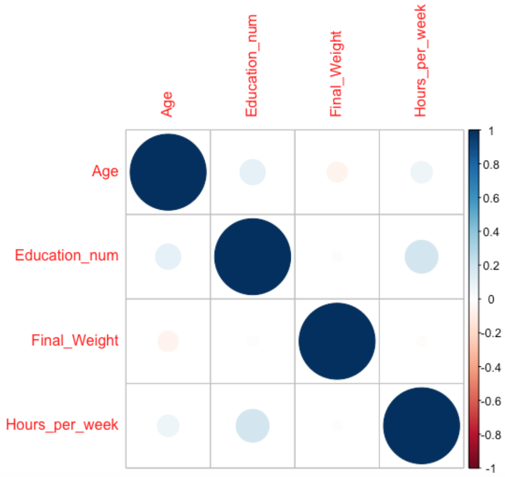


Figure 8: Correlation between Continuous Variables

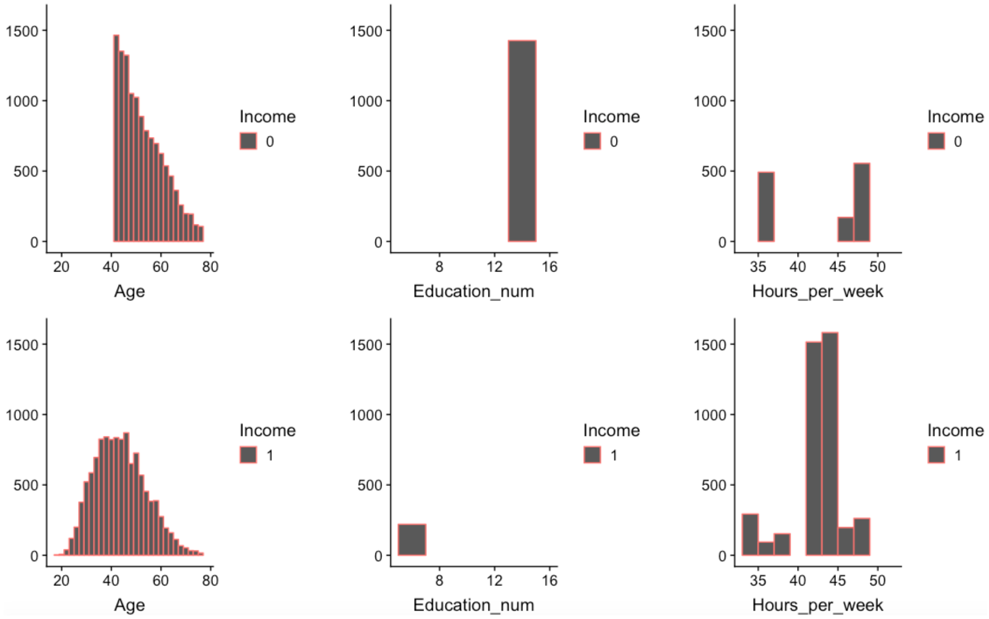


Figure 9: Income vs rest of continuous variables

According to figures (10-11-12-13), the variables workingclass, occupation, maritalstatus, relationship, education and sex all show good predictability of the income level variable.

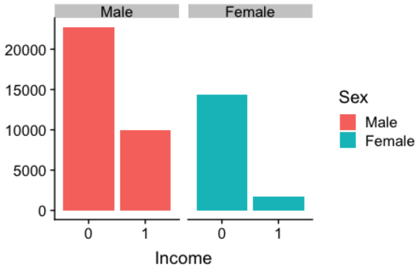


Figure 10: Sex vs Income

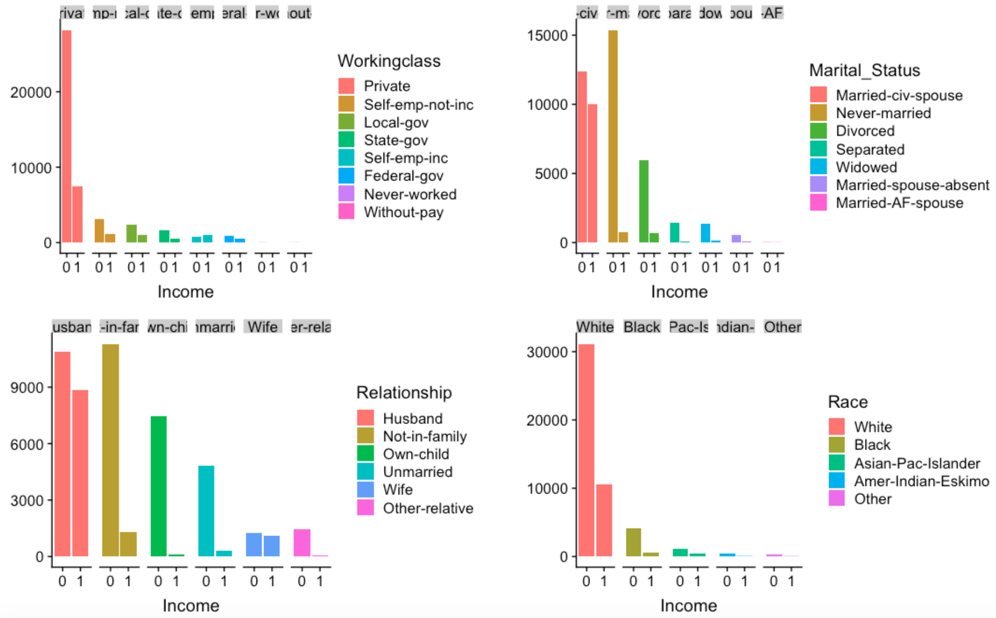


Figure 11: Income vs rest of categorical variables

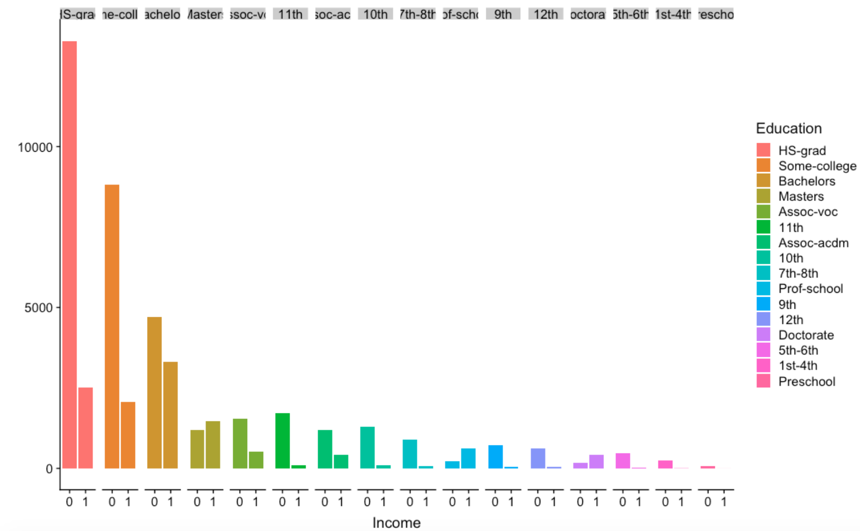


Figure 12: Education vs Income

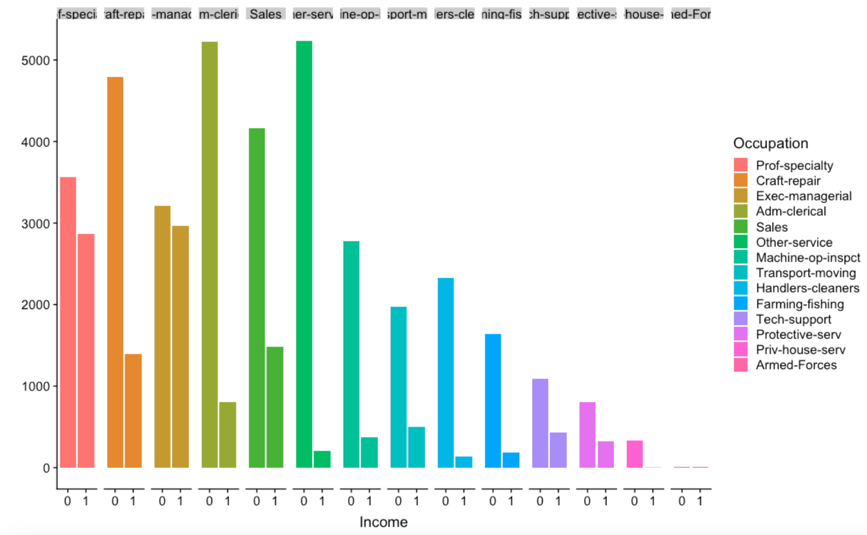


Figure 13: Occupation vs Income

2.3 Variable Recoding

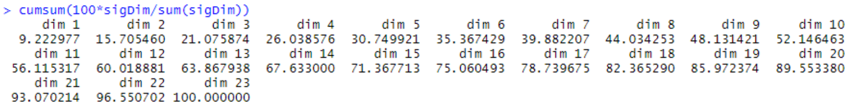
So as to make working with the data easier in the long run, we recoded Income, to be either 1 or 0 instead of ">50K" or "<=50K".

2.4 Feature Selection

Some of variables excluded from the model such as Capital_gain, Capital_loss and Native_country. Moreover, Education and Education_num gives us same information. Education_num will be excluded from the model, because Education more easily interpretable.

2.5 Multiple Correspondent Analysis

For the exploration of the latent information in the data set we have also performed a Multiple Correspondent Analysis over the categorical variables. We have performed MCA instead of PCA because the most data we have are categorical type. So we though performing an analysis of the categorical variables will probably give more explanation of the data set. Thanks to the results of the MCA we will be able to reduce the dimensionality of the data and perform clustering only with the significant dimensions. It must be mentioned that we had to do the analysis on a selected training data set, that is a sample of 4000 observations, this is because our computer was not able to deal with the all the observations.



```
> cumsum(100*sigDim/sum(sigDim))
    dim 1      dim 2      dim 3      dim 4      dim 5      dim 6      dim 7      dim 8      dim 9     dim 10 
 9.222977  15.705460  21.075874  26.038576  30.749921  35.367429  39.882207  44.034253  48.131421  52.146463 
   dim 11     dim 12     dim 13     dim 14     dim 15     dim 16     dim 17     dim 18     dim 19     dim 20 
56.115317  60.018881  63.867938  67.633000  71.367713  75.060493  78.739675  82.365290  85.972374  89.553380 
   dim 21     dim 22     dim 23 
93.070214  96.550702 100.000000 
```

Figure 14: Accumulative percentage of contribution of the dimensions

With the eigenvalues obtained we first performed Kaiser rule filtering the dimensions with eigenvalues greater than the average, and then we choose the dimensions with more than 80% of accumulative contribution, as result we got 18 significant dimensions.

2.6 Clustering

The method used is hierarchical clustering with a bottom-up approach, as k-means are dependent on initial centroids, and this can be determined by our randomly chosen seed initially.

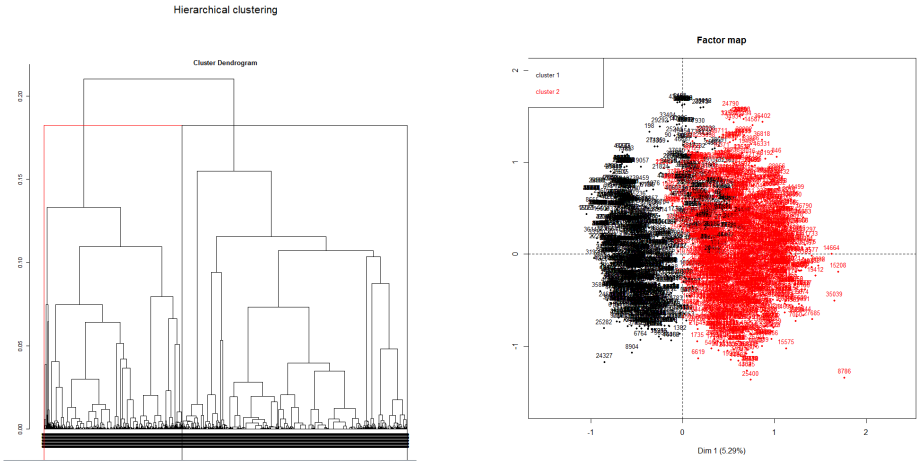


Figure 15: Clustering results.

From the results we can observe that big part of the observation is assigned to one cluster, it seems that the results are not quite good, better performance will be obtained with prediction models.

3 Modeling

As you can see in the figure below, we've detected an imbalance between Income values. To solve this problem, we received an equal number of observations from the two values in the data set. After that, we divided 4000 of the 5000 observations into training and 1000 of them as tests.

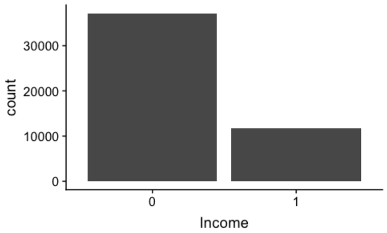
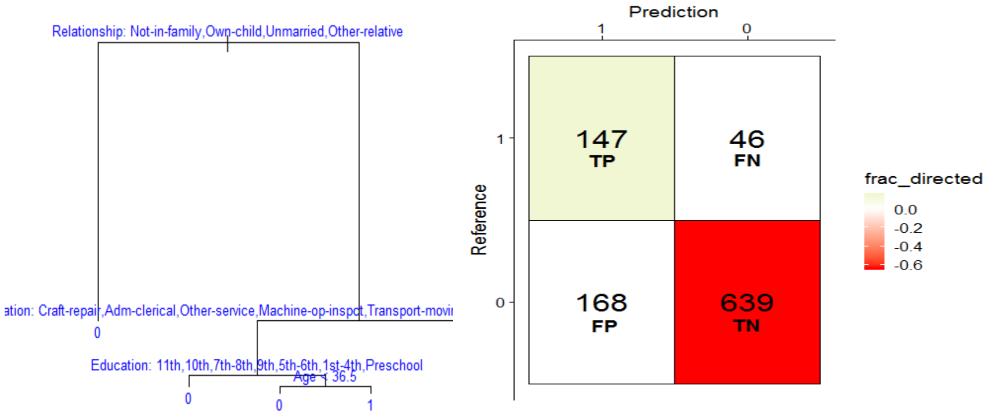


Figure 16

We used a couple of linear - nonlinear methods and performed **10-fold cross-validation** to choose the best threshold that minimizes the misclassification error rate in all models.

3.1 Decision Tree



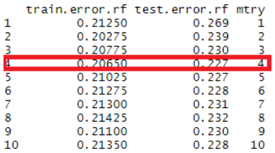
(a) Optimal Decision Tree (b) Confusion Matrix for Decision Trees

Figure 17: Decision Tree Results

We can see above the result of the decision tree after pruning, which is already the optimal tree. From the tree we can see the most discriminant variable is Relationship. We can observe other variables such as occupation, education and age also has relevant importance.

3.2 Random Forest

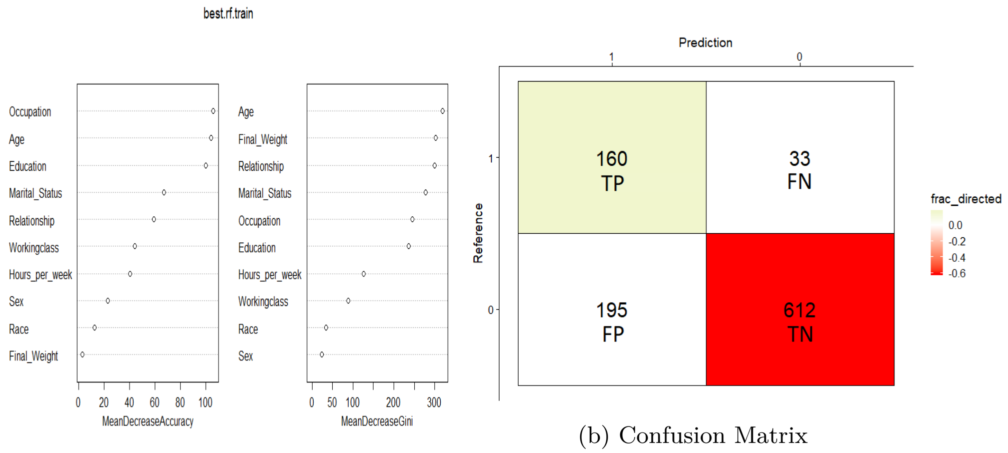
Using a smaller number predictors and a large number of trees is optimal while using Random Forests. So as to find the optimal number of predictors, we run a loop comparing different number of selected predictors and determine which gives the lowest misclassification error rate. According to that, the fourth one is the best. As a result, to create the Random Forest model we used 4 predictors and 2000 trees.



```
   train.error.rf test.error.rf mtry
1         0.21250         0.269    1
2         0.20275         0.239    2
3         0.20775         0.230    3
4         0.20650         0.227    4
5         0.21025         0.227    5
6         0.21275         0.228    6
7         0.21300         0.231    7
8         0.21425         0.232    8
9         0.21100         0.230    9
10        0.21350         0.228   10
```

Figure 18: Train and Test Error for Random Forest

According to Figure 19, we can say that "Occupation", "Age" and "Education" made the most difference in determining Income. In addition, "Relationship", "Occupation", and "Marital_Status" should also be taken into due to their high MeanDecreaseGini values. Finally, we build the model and found that the accuracy rate was 0.772. We have visualized how the data is classified with the confusion matrix.



(b) Confusion Matrix

(a) Importance Variables

Figure 19: Random Forest Results

3.3 Support Vector Machine

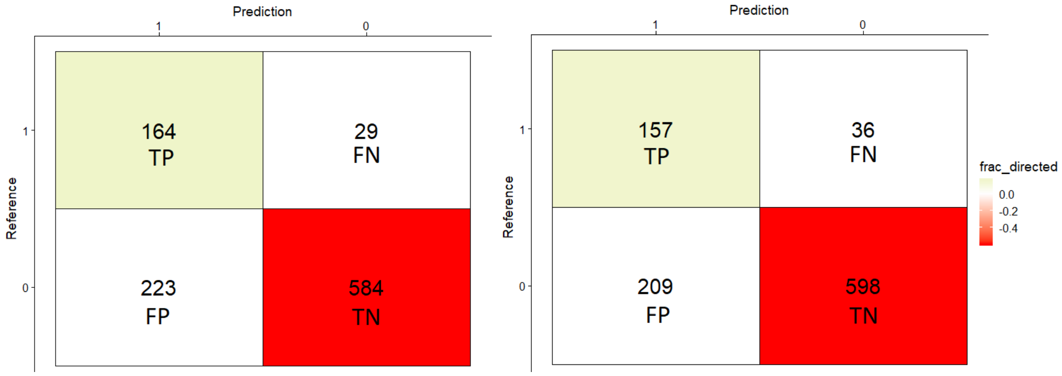


Figure 20: Linear SVM vs Non-linear SVM

These are the results obtained by performing linear SVM and none-linear SVM respectively, it seems that we obtain more true positives in linear approach, while in non linear we obtain more true negatives.

3.4 Naive Bayes

The Naive Bayes classifier is a simple and powerful method that can be used for binary and multiclass classification problems. Therefore, it was a natural choice to try. We have visualized how the data is classified with the confusion matrix.

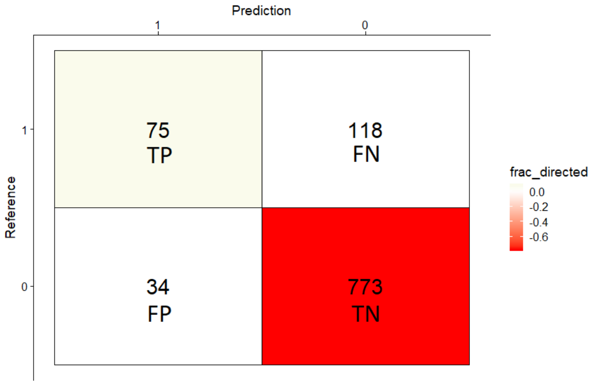


Figure 21: Confusion Matrix for Naive Bayes

With Naive bayes the number of true negatives has been increased hugely, however, with the number of True positives have decreased significantly.

4 Scientific and personal conclusions

Our best-running model with an accuracy of 0.848 was Naive Bayes. Then, decision trees are 0.78, random forest 0.77, nonlinear svm 0.755 and the lowest linear svm 0.748. When we look at these results, we can not get the desired results in except naive bayes, the reason why we can not get the reason for all of our models PPV ratio is lower than NPV. This means that our models are directed towards negative classification. Although we do model tuning, especially on the basis of svm did not have any effect. Nevertheless, this huge data set went through a good preprocessing process and an examination of the relationships of variables.

5 References

[1] https://archive.ics.uci.edu/ml/datasets/adult - Adult Data Set - ByRonny Kohavi and Barry Becker - 1996-05-01.1