Data Mining Project

Snehal Chemburkar
Indiana University Bloomington
snehchem@umail.iu.edu

Ruchi Gupta Neema
Indiana University Bloomington
rneema@umail.iu.edu

ABSTRACT
The aim of this project is to predict whether the income of an individual is greater than or less than 50K based on Adult dataset[1] by applying classification techniques. The dataset provides demographic information about individuals as predictors of income variable. We perform a comparative study of the performance of Naive Bayes [3], Decision Tree [2], and Random Forest [4] classification techniques for this dataset.

General Terms
Data Mining, Adult Data set, R

Keywords
Decision trees, Naive Bayes, Random Forest

1. INTRODUCTION
Our goal is to apply classification techniques to accurately predict the income level based on the demographic attributes of an individual. Our basic understanding is that age, occupation, and education play a major role in deciding an individual's income. The dataset used for this project is available at UCI Machine Learning repository as the Adult dataset and is extracted from the 1994 census database. The dataset consists of 15 attributes including income level which we will elaborate in the following section. In this project, we employ Naive Bayes, Decision Tree, and Random Forest classification models to accurately predict the income level. A comparison among these techniques can help us understand which model better suits this data set.

2. DATASET DESCRIPTION
The dataset consists of training data and testing data. The training data consists of lot of attributes of each person and whether they earned (>$50k or <=$50k) per year. For each person, the dataset contains total of 14 attributes which includes age, gender, education level, marital status as well as their job titles. The dataset has 32,561 records in training data and 16,281 records in testing data. The prediction task associated with this data set is to predict whether or not a person makes more than $50K a year using census data. The dataset contains total six continuous and eight nominal attributes which are summarized in table 1.

Attribute	Type
Age	continuous
Workclass	categorical
fnlwgt	continuous
education	ordered factor
education num	continuous
marital status	categorical
occupation	categorical
relationship	categorical
race	categorical
sex	categorical
capital gain	continuous
capital loss	continuous
hours per week	continuous
native country	categorical
incomelevel	(>$50k or <=$50k)

Table 1: Types of attributes in dataset

3. DATA PREPROCESSING
The data contains missing values which need to be handled before proceeding to the classification stage. The dataset contains a few attributes with string values which need to be transformed to a format suitable and understandable to R. These preprocessing steps are covered in detail here.

3.1 Deal with missing values
In this dataset the missing values are denoted by "?" instead of the generic "NA" or "NAN" format. The attributes with missing values are workclass, occupation, and native country. All the three attributes are categorical and the missing observations belong to the majority class i.e '<=50K'. Since the majority class has a probability of 76.07%, we can safely discard the observations with missing values from the data set as this will not affect the class probabilities as such.

3.2 Deal with categorical variable
As you can see from table 1, eight of the attributes are categorical variables with string values. We cannot string data to a classifier directly. Using the as.factor() function in R, this type of categorical variables are assigned numerical levels based on the categories. We applied this transformation on the string attributes with distinct levels to enable processing using R.

4. DATA VISUALIZATION
Before going into the details of the analysis, let us visualize the data and develop some understanding that can be later used to apply algorithms in an efficient way.

Figure 1 represents box plot of age distribution for different income level. It shows that the age variable has a wide range and variability and mean are quite different for both income levels. Also it shows that person having age group 45-50 have greater probability of earning >$50K. Therefore age can be a used as a feature variable for machine learning techniques.

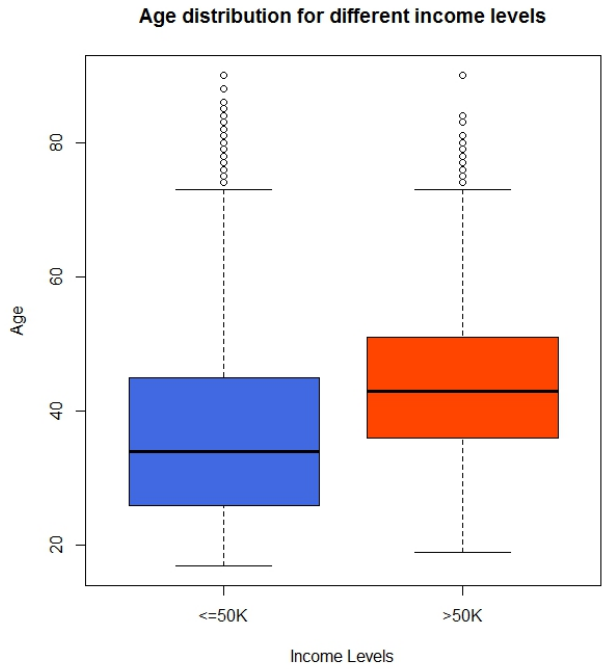

Figure 1: Age distribution for different income levels

Figure 2 represents distribution of education number for both income level. Most of the individuals in the dataset have at most a high school education while only a small portion have a doctorate. It shows that person having higher education have more chances of earning >$50K. Therefore education num can be a good predictor for income level.

Figure 3 represents distribution of hours per week for both income levels. It shows that person who spends more than 40 hours at work per week have greater chance of earning >$50K. It also shows that the percentage of individuals making over $50,000K drastically decreases for less than 40 hours per week. Therefore hours per week can be a good predictor for income level.

Figure 4 represents distribution of fnlwgt for both income levels which shows equal distribution for both income levels. Therefore, the feature variable fnlwgt can be removed from our analysis.

Figure 5 represents distribution of income with sex which shows that there is very less probability of earning >$50K for a person if it is female, while for male the chances for earning >$50KK is more. Therefore sex can be a very good predictor for different classification techniques.

Figure 6 represents the income levels for different types of

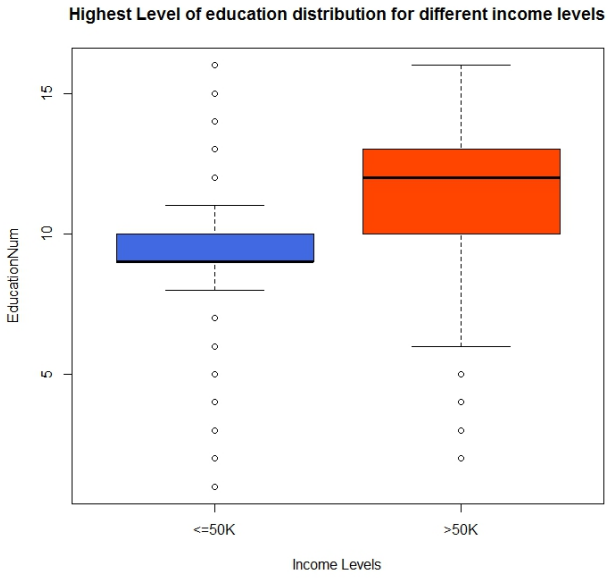

Figure 2: Age distribution for different income levels

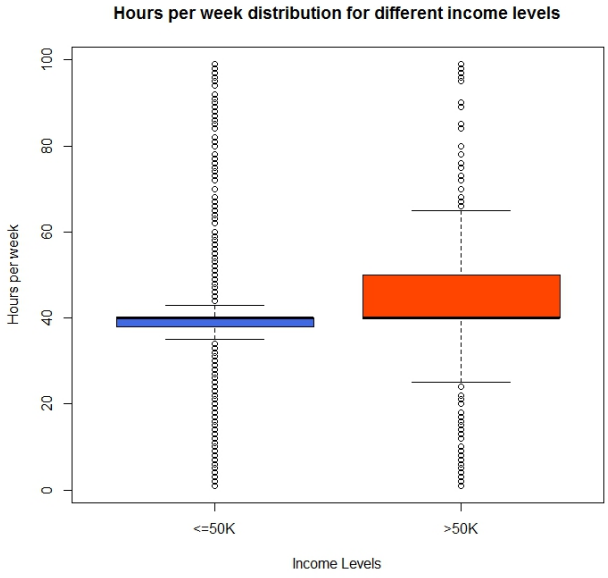

Figure 3: Hours per week distribution for different income levels

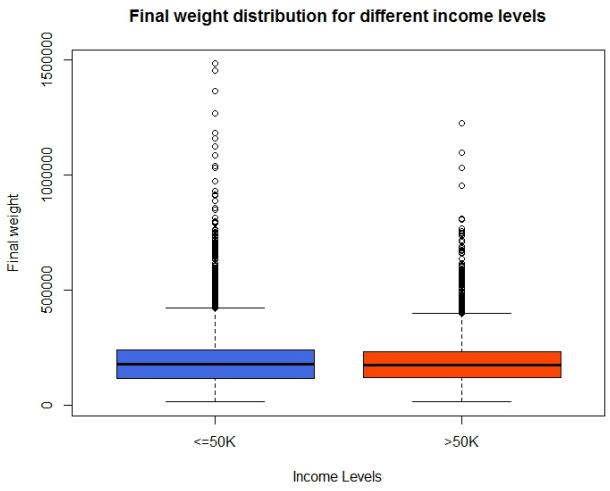

Figure 4: Final Weight distribution for different income levels

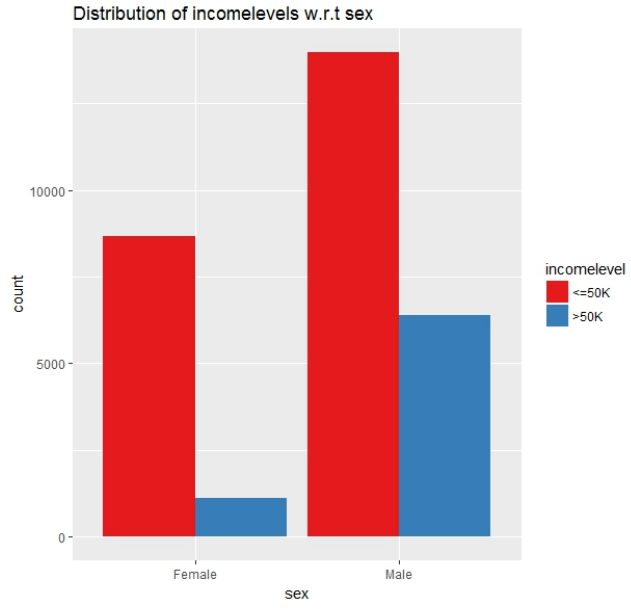

Figure 5: Sex vs Income

race. It can be seen from the plot that people from white race have more chances of earning >$50K. Therefore it can be used as a feature variable for our classification techniques.

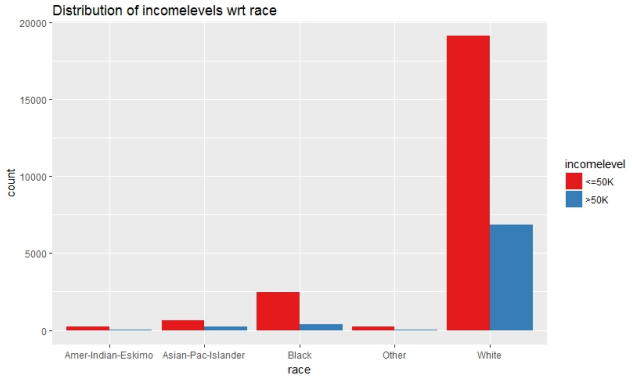

Figure 6: Race vs Income

Figure 7 reflects the variation in income level with respect to workclass attribute. Workclass defines if the person is employed at a private institution, self-employed, government employed and so on. It can be seen from the plot majority of the individuals who lie in the income level >$50K belong to private institutions. Thus we can say private institutions are more likely to pay >$50K as compared to rest of the categories.

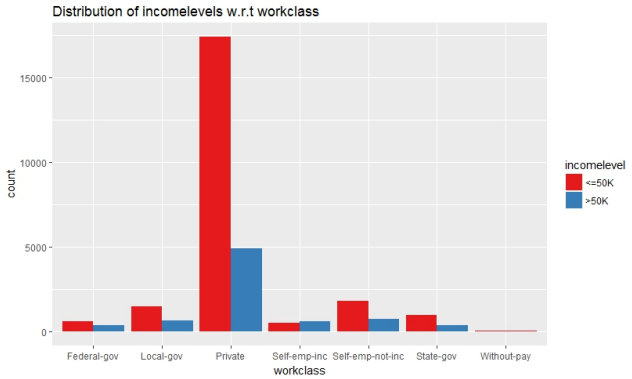

Figure 7: Workclass vs Income

Figure 8 reflects the variation in income level with respect to relationship attribute. Relationship defines if the person is married, divorced, single and so on. It can be seen that majority of the individuals who lie in the income level >$50K belong to married-civ-spouse category.

5. CLASSIFICATION TECHNIQUES

We apply some of the well known classification algorithms to predict the income level for the adult dataset.

5.1 Naive Bayes Classifier

The first classifier we apply on dataset is Naive Bayes classifier. It assumes independence of the predictor variables and Gaussian distribution (given the target class) of metric predictors [3].An advantage of Naive Bayes is that it only

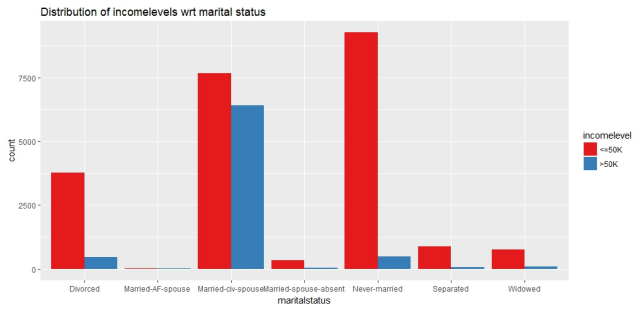

Figure 8: Relationship vs Income

requires a small number of training data to estimate the parameters necessary for classification. Using this classifier on our dataset we got an accuracy of 81.2% on test data and 82.7% on training data.

We have used all the available feature variables as predictors for the model. Also since we assumed that each of these feature variables are independent from one another, which is not true in reality because a person which works in an IT industry will have different levels of income depending on his education level. Therefore, the assumption we took is wrong and we can further improve the accuracy using different classifier such as decision tree and random forest.

5.2 Decision trees

Decision Trees is a supervised learning method that intelligently classify the given dataset. It predicts value by taking one variable at a time and try to draw boundaries between sections of data. Each drawn boundary acts as a 'decision' in the classification of different data.

We build a decision tree using all the feature variables available in the dataset. The resulting model uses only the three most important variables from the dataset namely capital-gain, education and relationship. The model tree is shown in figure 9 Using this model we achieved 83.89% accuracy on test data.

5.3 Random Forest

The last classification technique that we apply to the adult dataset is Random Forest (RF). Random forest is a machine learning algorithm that overcomes some disadvantages of decision tree. Decision tree is a simple technique, but may or may not be able to classify the data correctly. However, a deep tree can be made that works well on training data but has a problem of over fitting. Apart from this disadvantage of decision tree, it is also constructed using greedy algorithm and thus can be sub-optimal. To address these shortcomings, RF algorithm came into picture.

We apply this technique on the adult dataset using the same regressors as that of decision tree and obtain an accuracy of 82.27%. Random forest is supposed to give a better answer than decision tree, however for us it didn't perform well because the decision tree approach itself produced a simple tree and gave good results. Also, there is some randomness in the data that is keeping the accuracy around the same for all the methods.

6. COMPARISON OF DIFFERENT TECHNIQUES

Table 2 shows the comparison of different algorithms. We can see that all the algorithms performed almost the same except when naive bayes is used as a classifier. Decision tree and random forest worked better than Naive Bayes here which shows that the data is more or less non linear.

Also in decision tree we found that three attributes that are education, capital gain and relationship are good enough to classify a sample. Adding other attributes does not improve the results drastically. This means that other parameters are either correlated with these three or they can not explain the dependent variable.

Also since the original dataset contains a distribution of 23.93% entries labeled with >50k and 76.07% entries labeled with <=50k, therefore we used random forest as a classification technique after decision tree. But it did not perform well on this dataset as there is some randomness in the data that is keeping the accuracy around the same for all the methods.

Algorithm	Training data	Testing data
Naive Bayes	82.77%	81.20%
Decision Tree	84.10%	83.89%
Random Forest	83.19%	82.27%

Table 2: Results

7. CONCLUSION

Truly, all different machine learning algorithms have their advantages and disadvantages, and are appropriate for a particular type of dataset. However, for adult dataset, we have seen that the non-linear algorithm works equally good. This means that the data is more or less non-linear. This also suggest that there is some randomness in the data which cannot be avoided. Thus we conclude that any non-linear algorithm can work as a good predictor for adult dataset.

8. REFERENCES

[1] Adult dataset. webpage.
[2] Decision tree. webpage.
[3] Naive bayes classifier. webpage.
[4] Random forest. webpage.

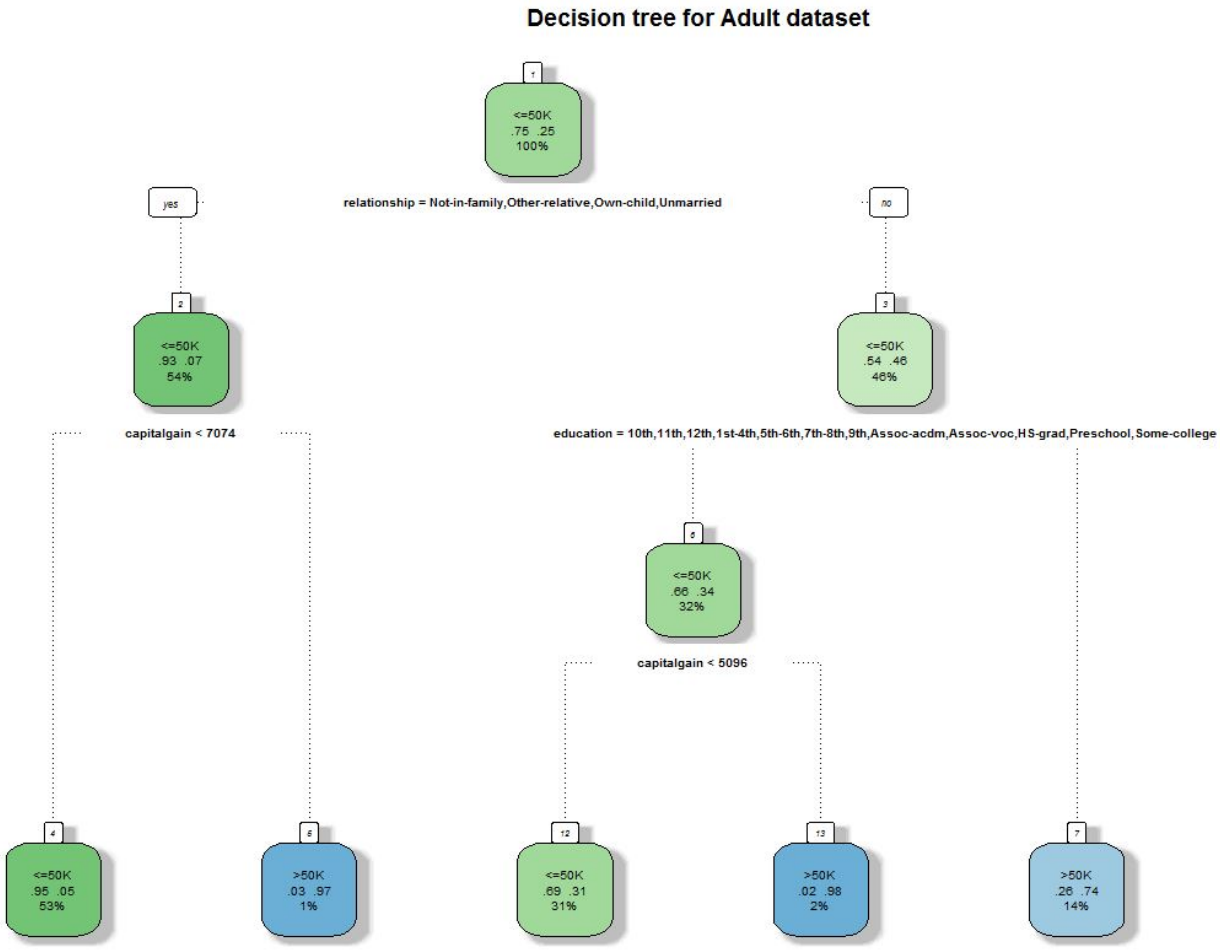

Figure 9: Decision Tree