Shabbir Yousuf Ali – 039081070

CMKE 136 – Capstone Project

syali@ryerson.ca

<https://github.com/shabbiryousufali/CKME136>

Abstract

Predict the Adult Income data using USA Census Data  
http://mlr.cs.umass.edu/ml/datasets/Adult

Adult income data

Final Project Report

# Introduction

Predict whether income exceeds $50K/year based on census data. Also known as "Census Income" dataset. Extraction was done by Barry Becker from the 1994 Census database. Adult income helps to determine the growth of the country/region. In this study I will try to understand which factor(s) is/are more important in improving individual income. I will do the classification, regression and prediction to determine whether the income of adult is greater than 50K or not. I will explore the key attributes / factors that could affect the adult income (like higher education, age, area etc.)

# Research Questions

1. How many adults have income less than 50K and greater than 50K
2. Individuals with higher education vs individuals with basis education
   1. Higher education means high income?
3. What factors/attributes should be considered to predict the income

# Literature Review

The adult dataset was compiled by Barry Becker, who extracted the data from the US 1994 Census database. This dataset was given to Ron Kohavi who used it to study the effectiveness of a new machine learning algorithm he’s proposed called the NBTree. According to Kohavi, the NBTree algorithm leverages the “surprising [accuracy]” of Naïve-Bayes and the scalability of Decision Trees. After the completion of Kohavi’s paper in 1996, the dataset was donated and now hosted by the Machine Learning Group at UC Irvine

<http://robotics.stanford.edu/~ronnyk/nbtree-talk.pdf>  
The paper by Ron Kohavi talks about a modified version of ID3 Decision Tree. The new algorithm is called NBTree, which induces a hybrid of decision-tree classifiers and Naïve-Bayes classifiers. The NBTree nodes contain univariate splits as regular decision-tree, but the leaves contain Naïve-Bayesian classifiers. Kohavi explains that the NBTree is a hybrid of naive bayes and a decision tree and is most suitable for scenarios where many attributes are significant in predicting the label, but they aren’t all necessarily

conditionally independent

<https://mpra.ub.uni-muenchen.de/83406/1/MPRA_paper_83406.pdf>  
S. M. Bekena, "Using decision tree classifier to predict income levels," MPRA Archive, 2017

In this study Random Forest Classifier machine learning algorithm is applied to predict income levels of individuals based on attributes including education, marital status, gender, occupation, country and others. Income levels are defined as a binary variable 0 for income <=50K/year and 1 for higher levels

<https://link.springer.com/content/pdf/10.1023/B:MACH.0000011804.08528.7d.pdf>   
The paper by Jinyan Li introduces a new algorithm doesn’t use distance as measurement but use frequency of an instance’s subsets and the frequency-changing rate of the subsets among training classes to perform both knowledge discovery and classification tasks.

<https://link.springer.com/content/pdf/10.1007/978-0-387-35592-4_12.pdf>   
Dennis P. Groth and Edward L. Robertson. An Entropy-based Approach to Visualizing Database Structure. VDB. 2002. The work by Dennis P. Groth talks about the use of entropy for visualizing database structure. Visualizing entropy of a relation provides a global perspective on the distribution of values and helps to identify areas within the relation where interesting relationships may be discovered.

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.387.5068&rep=rep1&type=pdf>   
A. Lazar, "Income prediction via support vector machine," in 2004 International Conference on Machine Learning and Applications, 2004. Proceedings., Louisville, Kentucky, USA, 2004. In this paper, the effects of data reduction on the classification results of the SVM algorithm are presented. The training time and the performance of a SVM classifier where compared for six different subsets of the adult data set. Despite the vertically reduced datasets, good classification accuracy was obtained in faster time. The conclusion is very important especially for datasets with many variables that can be reduced by using the PCA method

After reading the above reviews, I have decided to use logistic regression, random forest and naïve bayes for my project. It is further supported by the fact that the project of predicting income from census data is a binary classification problem, as the target variable having binary output and data has mixed numerical and nominal attributes

I will use Logistic regression to predict income based on the p values. Random forest can be used for better performance.

# Dataset

URL - [http://mlr.cs.umass.edu/ml/machine-learning-databases/adult/](http://mlr.cs.umass.edu/ml/machine-learning-databases/adult/adult.names)

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0))

* The dataset has 14 attributes and 48842 number of instances.

*8 are categorical and 6 are numerical*

* 48842 instances, mix of continuous and discrete (train=32561, test=16281)
* 45222 if instances with unknown values are removed (train=30162, test=15060)

# Data Description

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Data Type | Description | Distinct Items |
| Age | Integer | Age of a person | 17 is Minimum  90 is Maximum |
| Workclass | Categorical | Work Class | 9 Distinct and Known Categories Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked  Unknown Category  ? = 1836 |
| Fnlwgt | Integer | Final Weight |  |
| education | Categorical | Highest education | 16 Distinct Categories  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 | Integer | # of years of education | 1 is Minimum  16 is Maximum |
| marital-status | Categorical | Marital Status | 7 Distinct Categories  Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| occupation | Categorical | Person Occupation | 15 Distinct and Known Categories  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  Unknown Category  ? = 1843 records |
| relationship | Categorical | Role in a family | 5 Distinct Categories  Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| race | Categorical |  | 5 Distinct Categories  White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| sex | Categorical |  | 2 Distinct Categories  Male, Female |
| capital-gain | Integer | Gain during a year |  |
| capital-loss | Integer | Losses during a year |  |
| hours-per-week | Integer | work hours in a week |  |
| native-country | Categorical | Native Country | 42 Distinct Categories  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  Unknown Category  ? = 583 records |

# Statistics

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SEQ | Name | Min | Max | Mean | Median | SD | 1st QU | 3rd QU |
| 1 | Age | 17.00 | 90.00 | 38.58 | 37.00 | 13.64 | 28.00 | 48.00 |
| 2 | Education Num | 1.0 | 16.0 | 10.58 | 10.00 | 2.57 | 9.00 | 12.00 |
| 3 | Capital Gain | 0.0 | 999 | 1078 | 0.0 | 7385.29 | 0.0 | 0.0 |
| 4 | Capital Loss | 0.0 | 4656.0 | 87.3 | 0.0 | 402.96 | 0.0 | 0.0 |
| 5 | Hours per week | 1.0 | 99.0 | 40.44 | 40.00 | 12.35 | 40.00 | 45.00 |

# Approach

## Step 1 – Data Preparation

The following data preparation tasks are conducted to make the data suitable for running the machine learning model

1. Download and load the data
2. Add headers to the data
3. Clean the data by removing the unknown values
4. Standard calculation can be performed on the integer attributes to get their statistical information (like mean, median etc.)
5. Categorical attributes (work class, occupation and native country) have missing information, we will remove the missing information to get more accurate result
6. Check if the data is normal or not and what attributes are the outliers

Result  
Train data will contain 70% of the original data (i.e. 24421 records)  
Test data will contain 30% of the original data (i.e. 8439 records)

## Step 2 – Data Analysis

We will do the complete analysis between attributes. Find the attribute relationship and dependencies; strong correlation vs weak correlation

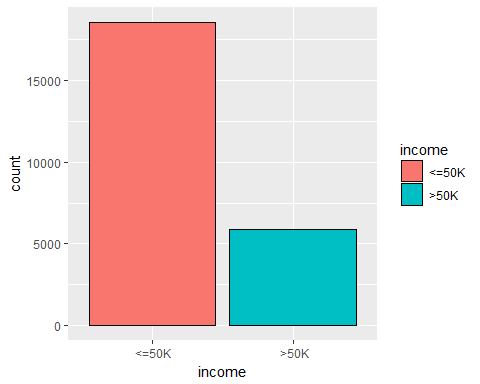
1. Does Higher education mean high income?

Study shows higher education means high income. As from the article published by Markus Hofmann in 2011 ([View Link](http://www.dataminingmasters.com/uploads/studentProjects/Earning_potential_report.pdf)) , on page 11, figure 4

*When the education number attribute was plotted for the class labels it was found that the lower values tend to predominate in the 50K class which may indicate some predictive capability*

1. Does age have any impact on the income?
2. Study by Sisay Menji Bekena 30 July 2017, clear shows that age has an impact on the income. The research on page 8 ([View Link](http://robotics.stanford.edu/~ronnyk/nbtree-talk.pdf)) shows the positive correlation with income. This show that marital status, capital gain, education, age and work hours (employment) determine much of the difference between low and high income levels
3. Which occupation pays the highest income?

## Income distribution

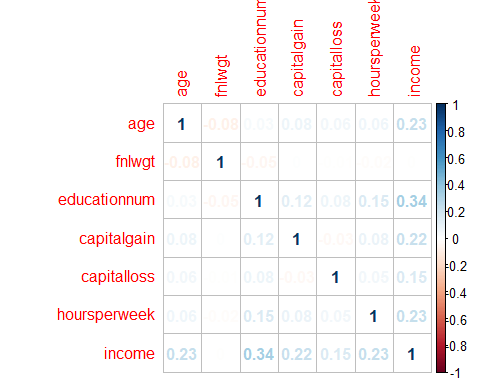


Above diagram shows the distribution of the income attribute.

Income <= 50K is approximately 76% where income > 50K is approximately 24%

## Correlation

* 1. Between Numerical Attributes



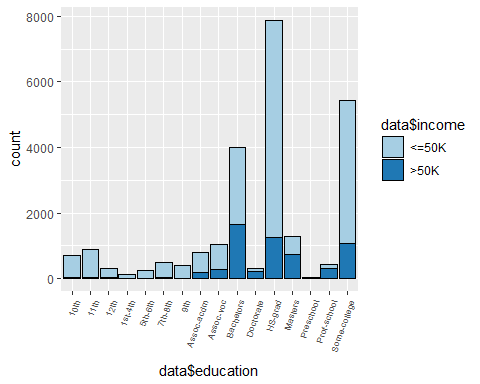
The correlation matrix created between the numerical attributes only.

Correlations shows that numeric attributes are related but they are not strongly correlated. Education has the highest correlation 0.33 with income followed by the Capital gain 0.22, age 0.24 and hours worked 0.23. The variables are positively correlated with each other

* 1. Between Categorical and Numerical (i.e. income)

**Correlation between Education and Income**

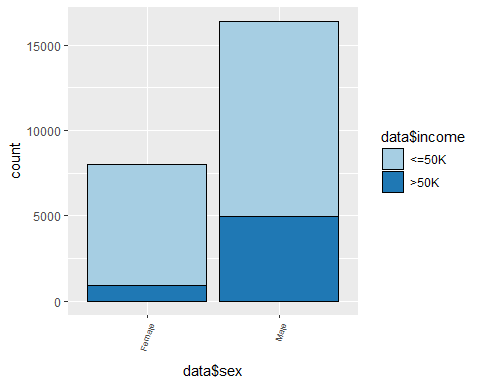
ggplot(data, aes(x=data$education,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



Result shows adults with higher education has earning > 50K. Adults with bachelor’s degree have maximum number of earnings > 50K, followed by doctorate and masters. Adults with lower education level have maximum portion of income <= 50K

**Correlation between Sex and Income**

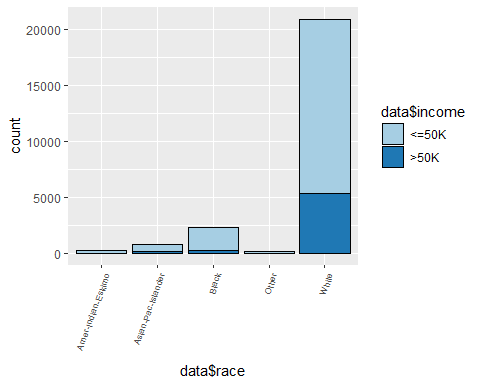
ggplot(data, aes(x=data$sex,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



Result shows the ratio of male earning income > 50K is more than female

**Correlation between Race and Income**

ggplot(data, aes(x=data$race,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")

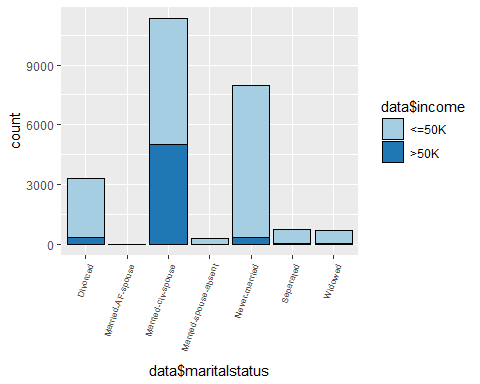
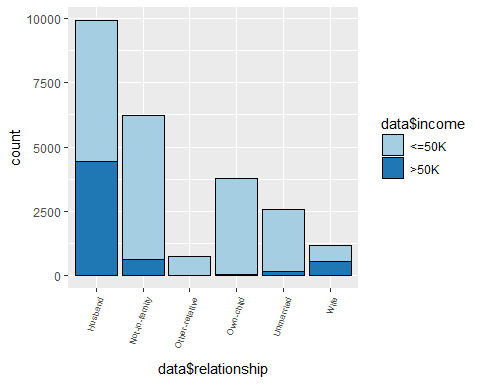


Result shows the highest earning adults are white followed by Black and Asia pacific

**Correlation between Marital Status and Income**

ggplot(data, aes(x=data$maritalstatus,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")

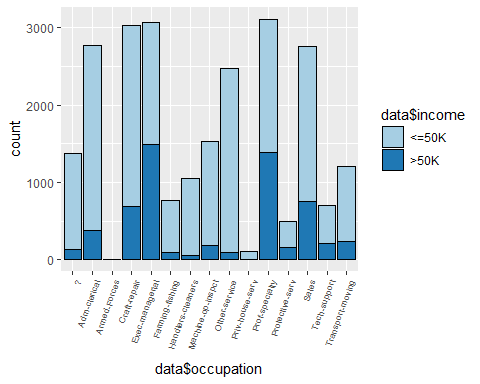
ggplot(data, aes(x=data$relationship,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")

Results in both the graphs show that Male and married people are earning more than 50K, as compared to female and unmarried people

**Correlation between Occupation and Income**

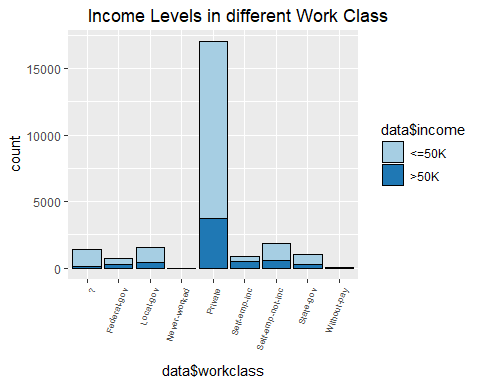
ggplot(data, aes(x=data$occupation,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



Result shows adults with higher position like Manager, Professor are earning > 50K

**Correlation between Work class and Income**

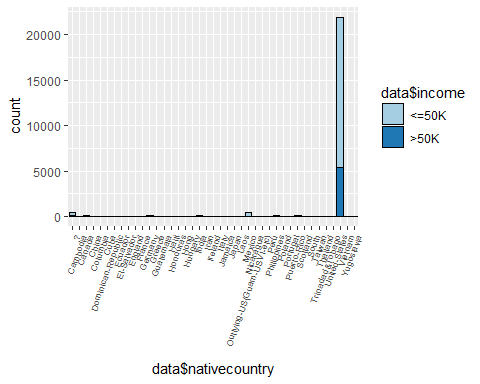
ggplot(data, aes(x=data$workclass,fill=data$income)) + geom\_bar(position = "stack", color = "black") + ggtitle(' Income Levels in different Work Class')+ theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



Result shows adults in private sector have maximum number of earning of > 50K

**Correlation between Native Country and Income**

ggplot(data, aes(x=data$nativecountry,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



Result shows majority of the adults belongs to the United States

## Step 3 – Test and Train Data

Different classification algorithm will be used to test the train and test data; algorithms like Logistic Regression, Random Forest and Naïve Bayes. Similar approached was used in the paper by Ron Kohavi talks about a modified version of ID3 Decision Tree ([View Link](http://robotics.stanford.edu/~ronnyk/nbtree-talk.pdf))

1. Train data will contain 70% of the original data (i.e. 24421 records)
2. Test data will contain 30% of the original data (i.e. 8439 records)

## Step 4 – Model Selection

S. M. Bekena use the Random Forest Classifier machine learning algorithm to predict the income level ([View Link](https://mpra.ub.uni-muenchen.de/83406/1/MPRA_paper_83406.pdf)). He achieved 85% accuracy by using the key attributes like marital status, capital gain, education, age and hours per week

Model used for the train and test data are

* 1. Decision Tree
  2. Linear Regression
  3. Random Forest

Random Forest Classifier will be used because the outcome (target) variable is binary variable (income level >50K or not). Also, the random forest has better accuracy as compared to Naïve Bayes classifier

The model with the highest accuracy will be selected as the final model

## Decision Tree

Dectree<- rpart(income~ age+ workclass+ education+maritalstatus+ occupation+ sex +hoursperweek, data = traindata, method='class',cp =1e-3)

**Result using Traindata**

Dectree.Ptrain <- predict(Dectree,newdata= traindata, type = 'class')  
confusionMatrix(traindata$income,Dectree.Ptrain)

|  |  |
| --- | --- |
| Prediction <=50K >50K<=50K 17328 1212>50K 2576 3305 | Accuracy : 0.8449Sensitivity : 0.8706Specificity : 0.7317Pos Pred Value : 0.9346Neg Pred Value : 0.5620 Prevalence : 0.8150 |

**Result using Testdata**

Dectree.pred.prob <- predict(Dectree, newdata = testdata, type = 'prob')  
Dectree.pred <- predict(Dectree, newdata = testdata, type = 'class')  
confusionMatrix(testdata$income,Dectree.pred)

|  |  |
| --- | --- |
| Prediction <=50K >50K  <=50K 5700 479  >50K 888 1072 | Accuracy : 0.832  Sensitivity : 0.8652  Specificity : 0.6912  Pos Pred Value : 0.9225  Neg Pred Value : 0.5469  Prevalence : 0.8094 |

## Linear Regression

linReg <- glm(income ~ age+ workclass+ education+maritalstatus+ occupation+ sex +hoursperweek, data = traindata, family = binomial('logit'))

pred1 <=50K >50K  
<=50K 17168 2644  
>50K 1372 3237

## Random Forest

randforest <- randomForest(income ~ age+ workclass+ education+ maritalstatus+occupation+ sex+hoursperweek, data = traindata, ntree = 500)  
randforest.pred.prob <- predict(randforest, newdata = testdata, type = 'prob')  
randforest.pred <- predict(randforest, newdata = testdata, type = 'class')

|  |  |
| --- | --- |
| Prediction <=50K >50K  <=50K 5699 480  >50K 830 1130 | Accuracy : 0.839  Sensitivity : 0.8729  Specificity : 0.7019  Pos Pred Value : 0.9223  Neg Pred Value : 0.5765  Prevalence : 0.8022 |

## 

## Step 5 – Result and Recommendation

Result generated based on the previous steps

## Compare the Algorithm

**DECISION TREE**prtree <- prediction(Dectree.pred.prob[,2],testdata$income)  
perftree <- performance(prtree,measure="tpr",x.measure="fpr")  
DTFrametree <- data.frame(FP=perftree@x.values[[1]],TP=perftree@y.values[[1]])  
auctree <- performance(prtree, measure='auc')@y.values[[1]]  
auctree

Result = 0.8500693

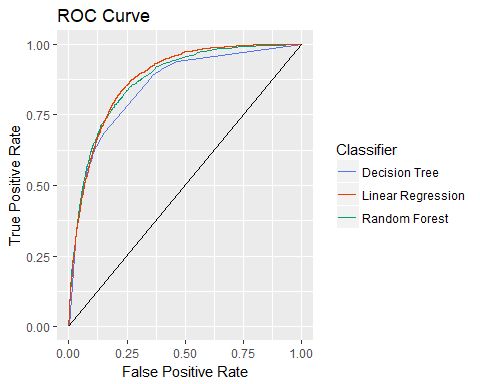
RANDOM FOREST  
prRForest <- prediction(randforest.pred.prob[,2],testdata$income)  
perfRForest <- performance(prRForest,measure="tpr",x.measure="fpr")  
DTFrameRForest <- data.frame(FP=perfRForest@x.values[[1]],TP=perfRForest@y.values[[1]])  
aucFtree <- performance(prRForest, measure='auc')@y.values[[1]]  
aucFtree

Result = 0.8733921

LINEAR REGRESION  
pr <- prediction(prob,testdata$income)  
perf <- performance(pr,measure="tpr", x.measure="fpr")  
DtFrameReg <- data.frame(FP=perf@x.values[[1]],TP=perf@y.values[[1]])  
aucRegresion <- performance(pr,measure='auc')@y.values[[1]]  
aucRegresion

Result = 0.879603

Now plot the graph using Area Under Curve



auc <- rbind(aucRegresion,auctree,aucFtree)  
rownames(auc) <- (c('Decision Tree', 'Random Forest', 'Linear Regression'))  
colnames(auc) <- 'ROC Curve Area'  
round(auc, 6)

ROC Curve Area  
Decision Tree 0.879603  
Random Forest 0.850069  
Linear Regression 0.873392

## Recommendation

Linear regression has the highest area under curve (AUC) value, then random forest and lowest is with the decision tree. So, we selected linear regression as our final model for predicting income of an individual as it gives the largest area under the curve