**Data Mining and Data Cleaning:**

The dataset I will be using for this assignment is the “Adult” dataset hosted on UCI’s Machine Learning Repository. It contains 15 variables. The dependent variable that in all cases we will be trying to predict is whether or not an “individual” has an income greater than $50,000 a year.

The variables in dataset is as follow:

* age – The age of the individual
* type\_employer – The type of employer the individual has. Whether they are government, military, private, an d so on.
* fnlwgt – The \# of people the census takers believe that observation represents. We will be ignoring this variable
* education – The highest level of education achieved for that individual
* education\_num – Highest level of education in numerical form
* marital – Marital status of the individual
* occupation – The occupation of the individual
* relationship – A bit more difficult to explain. Contains family relationship values like husband, father, and so on, but only contains one per observation. I’m not sure what this is supposed to represent
* race – descriptions of the individuals race. Black, White, Eskimo, and so on
* sex – Biological Sex
* capital\_gain – Capital gains recorded
* capital\_loss – Capital Losses recorded
* hr\_per\_week – Hours worked per week
* country – Country of origin for person
* income – Boolean Variable. Whether or not the person makes more than \$50,000 per annum income.

Firstly I remove to variables: education\_num, fnlwgt. The reason for this is they clutter the analysis. education\_num is simply a copy of the data in education and the fnlwgt is observation represents weight which we don’t want to use in our prediction model. Though if we were running more advanced analysis we could weigh the observations by fnlwgt, we won’t be doing that right now.

When we upload the dataset the variables stored as factors. Therefore, in the next step, we I changed them as a character string.

I block the variables which has a relative frequency in some groups. The following lines explain it in detail:

On “type-employer “given “Never worked” and “Without-Pay” are both very small groups, I combined them to form a “Not Working” Category. This help to reduce the number of categories significantly.

On occupation, a simple way to block the categories would include blue collar versus white collar. I separate out service industry, and other occupations that I didn’t see fitting well with the other groups into their own group.

The variable education also blocked. For some methods this vastly simplifies the calculations, as well as to make the output more readable. I choose to block all the dropouts together.

 In all cases, blocking drastically simplifies the models we use. For categorical variables, on mathematically based methods, the actual calculation sees a dummy variable for each level within a categorical variable. So for example if you have 2 categorical variables, with five and three categories respectively, the calculation will see 8 variables added to the equation. For something like this dataset where the bulk of the variables are categorical, and each of them has several levels, it’s easy to see how the number of variables in the model equation will rise. Blocking reduces this problem.

**Exploratory Data Analysis and Feature Selection**

### **Histograms of numerical variables**

% of data with zero Capital Gain: 91.67   
% of data with zero Capital Loss: 95.33

**Feature Analysis:** Age, log(Weight), Years of Education, and Hours per Week have broad distributions, therefore will be considered for regression analysis. Capital Gain and Capital Loss, however, have very narrow distributions (more than 90% of data are clustered at zero) therefore they will be excluded from the model.

### **Bar plots and Pie charts**

All these variables have reasonable spread of distribution, therefore will be considered.

### **Correlation between numerical variables**

### The numerical variables are nearly uncorrelated.

### **Get the training and test datasets**

### It’s important to have a constant sample left out that the classifier is not trained on, to see how well the trained classifier will generalize to new data. Here, I use random sampling, approximately thirty percent of the data into a validation set that I will then use to check the accuracy of my model.

**Performance of Classifiers on Dataset**

I try two model for predict dependent variable:1-logistic regression,2- decision tree.

## Logistic Regression

### Generalized linear model (glm)

|  |
| --- |
| GLM coefficients (sorted in ascending order of p-value) |

Estimate Std. Error z value Pr(>|z|)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.34033083 0.131982279 -25.3089343 2.547004e-141

educationBachelors 1.18491059 0.062568963 18.9376736 5.580917e-80

occupationWhite-Collar 1.07066976 0.063905848 16.7538622 5.306643e-63

educationMasters 1.46916105 0.090711199 16.1960272 5.379189e-59

educationProf-School 2.31251361 0.156725510 14.7551832 2.849168e-49

age 0.02655452 0.001861273 14.2668637 3.520595e-46

hr\_per\_week 0.02629088 0.001855979 14.1655058 1.497845e-45

educationDoctorate 2.08278972 0.177665652 11.7230860 9.707029e-32

relationshipWife 1.28258606 0.116164077 11.0411591 2.418924e-28

occupationProfessional 0.75002805 0.074186339 10.1100560 4.985577e-24

educationDropout -0.90193671 0.093915545 -9.6036998 7.712513e-22

sexFemale -0.83588560 0.087378785 -9.5662305 1.108746e-21

occupationOther-Occupations 0.76358193 0.087304430 8.7461992 2.206594e-18

occupationSales 0.52601394 0.068148030 7.7186963 1.175256e-14

educationAssociates 0.55605271 0.076798653 7.2403967 4.473741e-13

occupationService -0.79606846 0.122329025 -6.5076008 7.636049e-11

maritalNever-Married -2.04218602 0.329799319 -6.1922081 5.932715e-10

educationHS-Graduate 0.31768294 0.058248192 5.4539536 4.926207e-08

maritalNot-Married -1.53163658 0.334507329 -4.5787833 4.676884e-06

relationshipOwn-child -1.29192708 0.337231419 -3.8309808 1.276335e-04

maritalWidowed -1.35575800 0.364896237 -3.7154617 2.028330e-04

relationshipOther-relative -0.85879412 0.302268385 -2.8411642 4.494916e-03

occupationAdmin 0.22796931 0.083113098 2.7428807 6.090280e-03

raceOther -0.94181979 0.357318195 -2.6358014 8.393885e-03

raceAmer-Indian -0.25469734 0.249290955 -1.0216870 3.069290e-01

raceBlack -0.07209827 0.085119078 -0.8470283 3.969793e-01

raceAsian -0.10214631 0.121501644 -0.8406990 4.005166e-01

relationshipUnmarried -0.24156682 0.348399137 -0.6933623 4.880822e-01

occupationMilitary 0.33429519 1.399499061 0.2388677 8.112081e-01

relationshipNot-in-family 0.07060607 0.332937512 0.2120700 8.320524e-01

### **Feature analysis**

* The top three most-relevant features include Education, work class, and Age, all of which are numeric variables.
* At the bottom of the table, instances of  Relationship, and Marital Status are observed. However these variables are seen all across the table, therefore cannot be eliminated from the model.
* From the remaining two categorical variables, Sex appears near the top of the table and is highly ranked, however Racecategories are ranked near the bottom of the table and therefore can be removed to avoid model over-fitting.

## Cross-Validation

**ROC Curve:**

An ROC curve demonstrates several things:

1. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
2. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

3. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system:

* .90-1 = excellent (A)
* .80-.90 = good (B)
* .70-.80 = fair (C)
* .60-.70 = poor (D)
* .50-.60 = fail (F)

According to above explanation, our model is a good and accurate model.

Area under the curve: 0.8855

The accuracy of our model to predict the dependent model is 83.5% that is shows high performance of our model.

## Decision Tree

The result of cross validation of decision tree is as follow:  
accuracy = 83.56 %  
missclassification = 16.44 %

The results of our models show that both logistic and decision tree model are appropriate model to predict dependent variable.