OBA 455-555 Final Report
Maddie Frost, Ryan Maki, Ari Nguyen

Introduction

The data analysis done on the salary prediction data set enabled us to conclude the most important factors to annual income. From the raw data with more than 30,000 rows, we manipulated it down to around 20,000. This required numerous attempts of grouping and coding variables into a binary format. A KNN classification model, a Logistic regression model, and a Classification tree model were run on the data to develop a prediction. The models were compared against each other with accuracy measures and confusion matrices. Each model offered unique insights, ultimately the classification tree proved to be the most accurate in predicting if an individual makes over 50,000 US dollars per year.

Problem Description

Does an individual make equal to or more than $50,000 a year, yes or no? What components affect this outcome? Our problem is to determine what patterns will result in getting a salary above $50,000/year or equal to and below depending on variables such as education, experience, location, hours worked, etc.

Business Context

As we are entering the workforce, it is crucial to know the importance of certain factors that contribute to our employee composition. We are hoping to see which variables most affect one's salary and what they can do to make themselves more valuable employees.

Data Exploration

The data were chosen from Kaggle, extraction was done by Barry Becker from the 1994 Census database, and donated on 01-05-1996. The initial variables include; age, work class, final weight, education, education number, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours worked per week, native country, and salary (the target/outcome variable).

For our report, we have chosen to opt-out of "final weight", "workclass", "education-number", "occupation", "capital gain", and "capital loss" variables. As the "final weight" variable indicates the number of people the census believes to fall in the same background category, we believe that it is irrelevant to our analysis and have chosen to not use it in our model buildings. We also eliminate capital gain, and capital loss variables from our models due to the vast amount of bad data. Moreover, due to the nature of our problem description, we are trying to predict whether or not an individual would make a salary of over $50,000 a year based on their personal background before their current occupation status. After vetting our data set, we have decided to perform prediction on the response variable of more than or equal to $50,000

by the selected variables of “age”, “work class”, “education”, “marital status”, “relationship”, “race”, “sex”, “hours worked per week”, “native country” and “salary”.

To gain better insights into the variables of our choosing, we first performed our exploratory data analysis and data cleaning using Excel. The desirable results from this process mainly focus on testing our hypothesis, learning about the distribution of data concerning our target outcome (salary <= 50K or > 50K), and identifying possible restrictions that could affect the accuracy of our models.

Variable Descriptions

Sex[1]

The sex variable looks at only males and females. There is a disproportionate amount of males to females within the dataset, because of this there could be bias because they are pooled from different amounts, from the dataset only 32.43% are female. Comparing women to women, most of them make less than or equal to $50,000, the same as the males.

Race[2]

There are three main issues with the race variable in our data set. First off, the data is highly made up of individuals who classify as “White”. This may cause our data to be skewed in one way or another and not be representative of all of the United States. Secondly, some races such as American-Indian/Eskimo and Asian/Pacific-Islander are grouped which could cause the data to be skewed. Lastly, Hispanic isn’t even listed as an option for race and is either grouped in somewhere else with another race or entirely left out altogether.

Age[3]

The age variable in our data set is on a continuous scale and is set as yearly. The trend presented in this variable indicates that people who are in the range of 35 to 53 years old would have the highest probability of earning over $50,000. The main issue with this variable has a higher amount of people from 17 to around 45 than in other age ranges. Although we expect this to represent the actual variation in the data decently since this is a reasonable estimation of average working ages, this is something to keep in mind.

Marital status and Relationship[4]

For “marital status” and “relationship”, we have decided to look at these two variables together due to their connection to each other. The “marital status” variable describes the current relationship status of the people in our collected data in whether or not they are or have been married. The “relationship” variable describes the role of each individual in a family setting. We can see the recurring trend of males having a higher

[1] Graph and table for Sex variable
[2] Graph and table for Race variable
[3] Graphs for Age variable
[4] Graph for Marital Status and Relationship

probability of making more than $50,000 a year with the “husband” role in a “married civilian spouse” shown in the table having the highest probability of making more than $50,000. However, the data being collected lacks adequate information on married couples with a military spouse (“ married AF spouse”) to make an unbiased comparison with civilian married couples. One interesting trend we observed from the table was that people who are currently not married or have an absent spouse are less likely to make more than $50,000 a year.

Education[5]

Most people who make more than $50,000 a year are high school graduates or have some college experience. The majority of people who do not make over $50,000 a year fall in the categories that are less than a high school graduate. There are very few people who make less than $50,000 and have higher educational status compared to those who make more. Even though there are people with less than a high school level education who make more than $50,000, the ratio is not significant compared to those with an education level higher than high school.

HPW[6]

By grouping the data we found that the overall probability of below five hours and over seventy hours was less than 1%; therefore, to better visualize the data, we set the cap at below five hours per week and above seventy hours per week. The data shows a trend of people who work over 35 hours to over 55 hours per week have a higher chance of getting a salary of over $50,000 than other hours range. However, the graph shows the trend of people who make less than or equal to $50,000 mirroring the trend of people with more than $50,000 a year. Consequently, looking at the “HPW” variable through the exploratory method alone, we do not find any substantial results on how they will affect the people's ability to make more than or less than or equal to $50,000 a year.

Approach (KNN Classification)

This approach focused on determining the best value of *K* to make a proper prediction and to determine how well the model operates in correctly predicting if an individual will make more than $50,000.

Data Analysis

The KNN Classification analysis steps include first cleaning up the data to only select the input variables that are significant and have no missing values. After this, the data was coded into binary format to prepare it for the classification model. The KNN observations were identified in the data set that is similar to the new observation we are trying to predict the outcome for. The predominant class is assigned to the new observation based on the class of the nearest neighbors. The data analysis steps

[5] Graph for Education variable

[6] Graph for Hours Worked per Week variable

include partitioning the data (27,786) with a seed of 30, and 70% (19,450) attributed to training data and 30% (8,336) to validation data. The train data with numeric and dummy variables are standardized and run against the validation data, then standardized again. The optimal k value was 14, with a result of 0.55278. The output is tracked and the KNN analysis then predicts the outcome variable for the new observation.

Results & Inferences[7]

The model did not prove to be very accurate. The confusion matrix reported that the model predicted false positives more frequently than true positives, meaning it classified an individual as making over $50,000 when they did not. The matrix also reflected that the model was able to correctly report a true negative compared to a false negative, meaning the model correctly predicted that an individual would not make more than $50,000 correctly more times than not. With an accuracy of 52.159%. The sensitivity rate, which indicates the model's ability to detect the important class correctly, was 0.649 along with the specificity rate, which indicates the model's ability to rule out the non-important class. The overall accuracy of the KNN classification model was 52.2%, which is not very high and does not measure the importance of the classes very well.

Key Insights

1. **From this data analysis, we were able to determine that being married versus being single will predict that you will make more than $50,000.**

1. **From this data analysis, we were able to determine that having children lessens the likelihood that you make more than $50,000.**

1. **From this data analysis, we were able to determine that higher education and making more than $50,000 are correlated.**

Approach (Logistic Regression)

For the approach of the logistic regression, our group wanted to answer the questions, "What is the best variable to predict salary?", and "How can employees set themselves apart from other candidates?"

Data Analysis

Before we ran the logistic regression, we had to ensure the data was ready to be used. We first took all "?"s out of the data to where the value was just a blank space that would not affect our analysis. we also had to take the "-"s out of the names of the columns because the glm function wouldn't accept them. Next, we changed the native country column to be set into whether you were from the United States, England, Germany, Italy, or any other country (due to the results from the previous first logistic

[7] KNN Classification Confusion Matrix and Statistics

regression). Lastly, I partitioned the data into a 70% train and a 30% validation split for the confusion matrix later on. This makes the train data 22,793 and the validation data 9768.

Our team then ran an initial logistic regression of all of the variables and found that for the most part, workclass, education-number, occupation, capital gain, and capital loss either weren't significant or relevant to our model. We took these variables out of the model so we could better predict whether or not an individual makes over $50K per year and therefore get a higher accuracy rating.

Results & Inferences [8][9]

After running the logistic regression, we found the most significant variables to be the level of education you have, your marital status/relationship, hours per week worked, and whether your native country was the United States or not (as shown below). One thing to keep in mind when looking at the results of the logistic regression is that the "positive class" of the model is 0, meaning that it is predicting whether the individual will make less than $50K.

1. For every 5-year increase in age, an individual will have a higher salary at the odds of 1.138 (2.718^(0.025908*5)) with all other variables being held constant.
2. An individual who is male will have a higher salary at the odds of 2.21 (2.718^0.792934) relative to a being a female holding all other variables constant.
3. An individual who has a doctorate degree will have a higher salary at the odds of 40.08 (2.718^3.690868) relative to a sophomore in high school.
4. An individual who is a married civilian will have a higher salary at the odds of 6.013 (2.718^1.794) with all other variables being held constant, compared to an individual who is divorced.
5. An individual who is a child will have a lower salary at the odds of 3.513 (2.718^1.2567) compared to a husband with all other variables being held constant.
6. An individual who is white will have a higher salary at the odds of 1.51 (2.718^0.414) compared to American-Indian-Eskimo with all other variables being held constant.
7. For every increase of 10 hours worked per week, holding other variables constant, increases the odds that the individual makes over $50K by a factor of 1.38 (2.718^0.032511).

[8] Logistic Regression Results
[9] Confusion Matrix for Logistic Regression

8. An individual whose native country is not the United States, Italy, Germany, or England will have a lower salary at the odds of 2.849 (2.718^1.047) compared to someone who is from England with all other variables being held constant.

After we ran the logistic regression, we set the data up to return the accuracy measures of our model by predicting the salary of the individual and running a confusion matrix. As shown below, the model correctly classified 7846 of the salaries (1160+6686) and incorrectly classified 1763 of the salaries (1195+598). The accuracy of the model is 0.8165 so it isn't incredible at predicting whether or not someone makes below $50K but it isn't bad enough to ignore the results of the logistic regression altogether. The p-value of the logistic regression model is pretty much 0 (<2.2e-16) which means that the observed values are unlikely to have occurred by random chance. The sensitivity, or the ability to detect the class members correctly, was 0.4926 so still not that great for this purpose. The specificity, or the ability to rule out non-important class members, is 0.9217. The sensitivity for the logistic regression was the lowest of all of the models while the specificity was the highest of all of the models.

Key Insights

1. From looking at the data from the logistic regression, it is easy to see that the higher your education level is, the higher you are expected to be making in salary per year. This is pretty much exactly what we were expecting to see from running this so this is not at all surprising to us.
2. The next key insight is that the data is suggesting that compared to the husband, the wife is supposed to have a higher salary while holding all other variables constant. This is a weird indication in the data that we were not able to explain because this contradicts the findings higher up in the logistic regression saying males have a higher salary on average holding all other variables constant. We were expecting the husband to make more money than the wife in most cases because of both the wage gap and the fact that women usually take a longer parental leave. Although we are not explaining why this is, more research should be done to see if this is true and why.
3. Anyone who is from a country that isn't the United States, Italy, Germany, or England will have a lower salary compared to someone who is from one of those countries. Reversely, this means that anyone who is a U.S. Citizen or an immigrant from Italy, Germany, or England will have a higher salary on average compared to an immigrant from any other country.
4. The factor of race, compared to being an American-Indian-Eskimo, is only slightly significant if the individual is either Asian-Pacific-Islander or White. Although we cannot be sure that this plays a factor in income, the logistic regression suggests that there might be some correlation in the data.

Approach (Classification Tree) [10]

For the Classification Tree, our goal is to identify the most dominant decision variables and to divide the variable space up to be as homogeneous or pure as possible based on the splitting rule.

Data Analysis

The classification tree was built based on the cleaned data set where all the bad data was eliminated or converted into blank spaces to maximize the accuracy of the model. We set the seed at 30 and partition the data into training and validation at the ratio of 70-30 respectively. During our analysis, we failed to achieve the goal of exhausting all the possible splits due to a large amount of data and our classification tree would be prone to overplotting. To avoid overplotting and minimize possible errors in the model, the chosen optimal N-split was 13 to prune the tree where the relative errors start to stabilize. We would then plot the tree accordingly to the optimal cp with association to our chosen N-split and run an evaluation on our model.

However, after the initial classification tree was plotted, the tree showed some challenges in identifying the important components that contribute to the target outcome due to the wide array of components in each variable. To fix this problem, our team went back into our model building and grouped nominal data in the “education”, “work class” and “ marital status” variables.

Results & Inferences [11]

According to the results indicated from our classification tree, we were able to identify our dominant decision variables which are “relationship”, “age” and “hpw”. The model accuracy measure indicates that the model has an overall accuracy rate of 82.57% with a sensitivity of 93.09% and a specificity rate of 50.23% relative to the positive class. The “positive class” - which is the more important outcome that the model decided we should predict - are the people who are making less than or equal to $50,000 a year (<= 50K). The results from our classification tree indicate that the model has a high ability to accurately identify people from the important class at 0.93 and an adequate ability to rule out people from the non-important class correctly at 0.50.

Key Insights

1. From observing our classification tree and its identified dominant component, we can see that our primary decision node consists of “relationship” in which the role of that individual in a family setting would be “Own-child”, “Not-Married”, “Other-Relative”, “Unmarried”. The data shows that people who have these characteristics in the “relationship” variable are more likely to make less than

[10] Classification Tree
[11] Confusion Matrix for Classification Tree

$50,000 a year and vice versa. This could be an indication that people who are currently married have a higher probability of making more than $50,000 a year.

2. Our second most important decision variable is “education” where people who only have a high school degree, an associate degree and have never finished high school are less likely to make more than $50,000. In contrast, people who have a higher education level than what we have mentioned above and work more than 35 hours per week would have a higher chance of making more than $50,000 a year. Even though we did not get many valuable insights to determine how “hpw” would affect an individual’s ability to make more than $50,000 a year in our exploratory process, our decision tree analysis identified how “hpw” can interact with other variables like “education” to make the prediction.
3. The third most important decision factor identified by our classification tree is “age”. The model predicts that people who are younger than 34 years old and have a high school degree or an associate degree would be more likely to make less than or equal to $50,000 a year. We believe that this prediction can come from seniorities where people who are older with more experience could get paid better than those who do not.

Main Takeaway

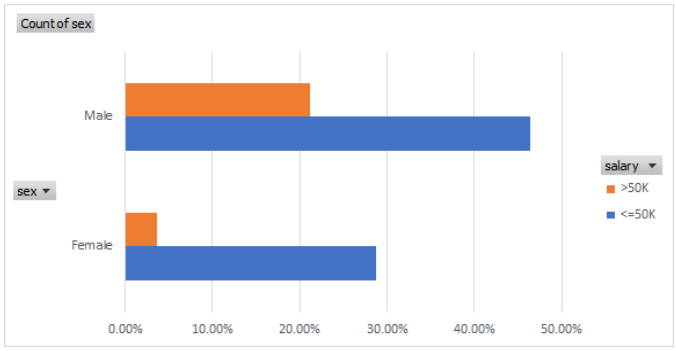
We decided that the classification tree was the best model for predicting an individual’s pay. The most important variables regarding higher pay include education, relationship, age, and hours per week worked. By using these findings in a corporate setting, managers can better compensate their employees. These findings can also help inform employees whether or not they are being adequately compensated for their work.

Recommendations

From the classification tree model and cross-referencing the results with the odds from logistic regression, we recommend people further their education to have a higher chance of earning over $50,000 a year. Throughout all of our models, this has been the leading variable in earning over $50,000. Even though the data set is taken from 1994, we believe that it would still align with our current society and could serve as a shortcut to earning more throughout our careers. We also believe that, statistically, getting married and staying in a marriage can lead to a higher chance of earning over $50,000. Ultimately, with an adequate education background, working long hours can lead to getting a bigger paycheck.

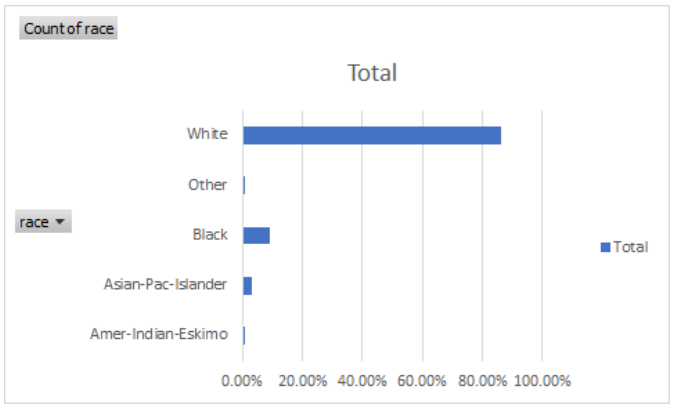
Appendix

1. Graph & Table for Sex Variable



Gender	<=50K	>50K	Grand Total
Female	28.74%	3.69%	32.43%
Male	46.37%	21.20%	67.57%
Grand Total	**75.11%**	**24.89%**	**100.00%**

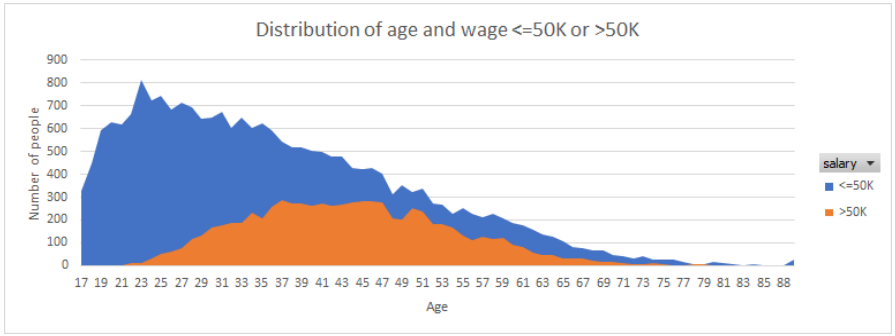
2. Graph & Table for Race Variable

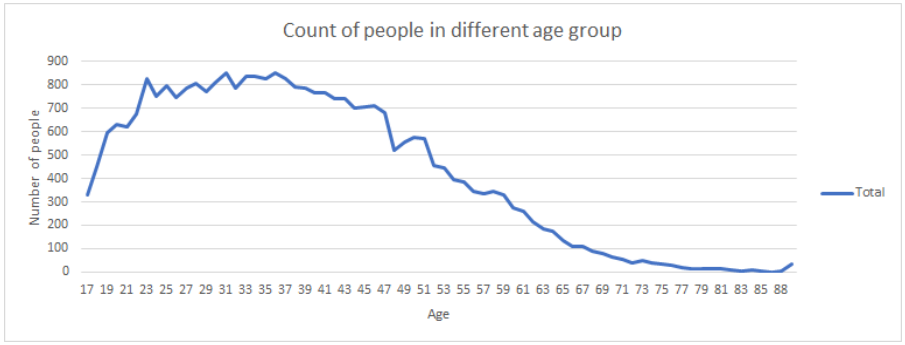


Race	Percentage
Amer-Indian-Eskimo	0.95%
Asian-Pac-Islander	2.97%
Black	9.34%
Other	0.77%

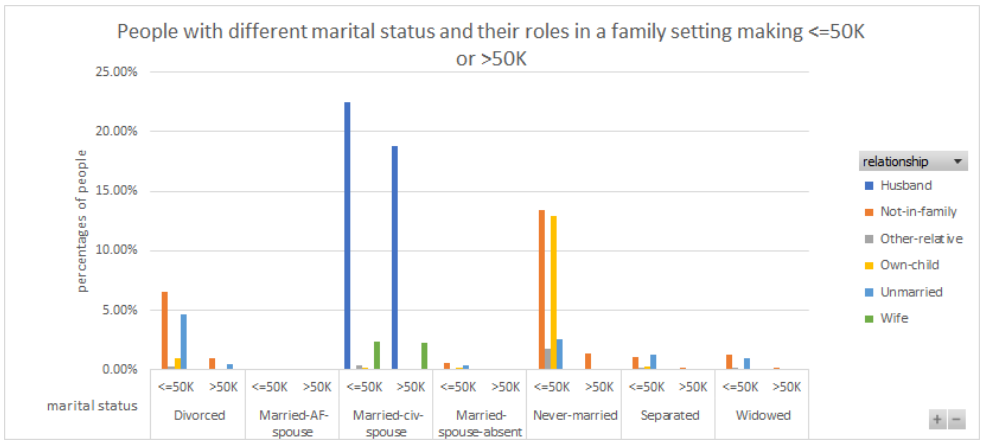
White	85.98%
Grand Total	**100.00%**

3. Graphs for Age Variable

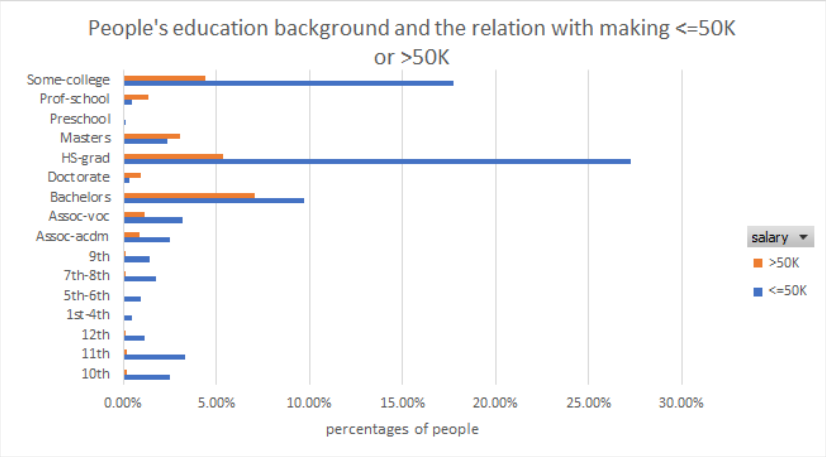




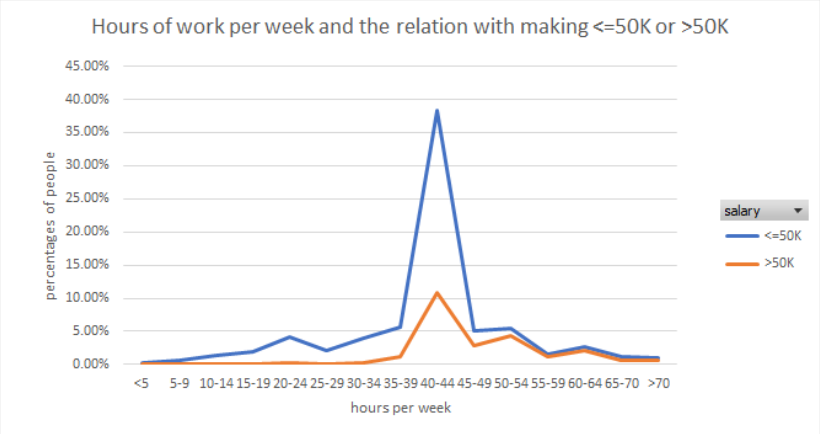
4. Graph for Marital Status and Relationship



5. Graph for Education Variable



6. Graph for Hours Per Week Worked



7. KNN Classification Confusion Matrix and Statistics

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1080  489
         1 1023 5744

               Accuracy : 0.8186
                 95% CI : (0.8102, 0.8268)
    No Information Rate : 0.7477
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4751

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5136
            Specificity : 0.9215
         Pos Pred Value : 0.6883
         Neg Pred Value : 0.8488
             Prevalence : 0.2523
         Detection Rate : 0.1296
   Detection Prevalence : 0.1882
      Balanced Accuracy : 0.7175

       'Positive' Class : 0
```

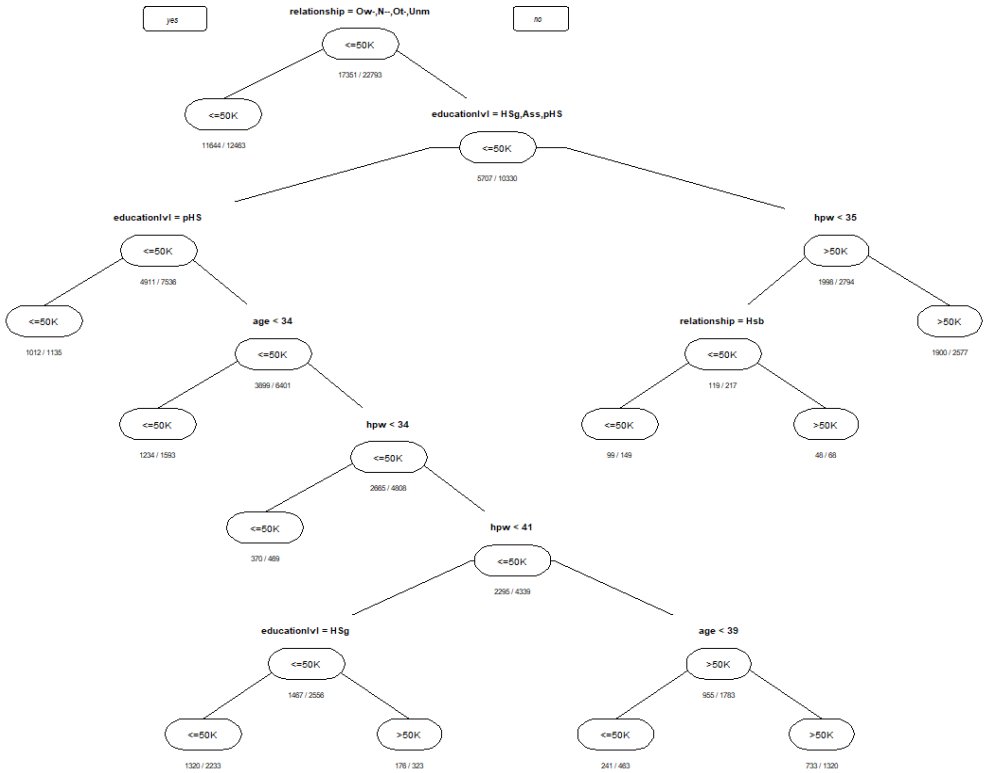
8. Logistic Regression Results

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6268   0.0412   0.2343   0.5962   2.7248

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       6.684204   0.571597  11.694  < 2e-16 ***
age                              -0.025908   0.001775 -14.594  < 2e-16 ***
Male                             -0.792934   0.085247  -9.302  < 2e-16 ***
education11th                     0.086830   0.240145   0.362 0.717672
education12th                    -0.224319   0.312719  -0.717 0.473177
education1st-4th                  0.753123   0.551410   1.366 0.171998
education5th-6th                  0.913920   0.430605   2.122 0.033804 *
education7th-8th                  1.032163   0.289468   3.566 0.000363 ***
education9th                      0.564939   0.301867   1.871 0.061278 .
educationAssoc-acdm              -1.633025   0.193512  -8.439  < 2e-16 ***
educationAssoc-voc               -1.519044   0.186833  -8.130 4.28e-16 ***
educationBachelors               -2.413020   0.171894 -14.038  < 2e-16 ***
educationDoctorate               -3.690868   0.234487 -15.740  < 2e-16 ***
educationHS-grad                 -0.817267   0.170029  -4.807 1.53e-06 ***
educationMasters                 -2.959368   0.182282 -16.235  < 2e-16 ***
educationPreschool               11.695366 134.853162   0.087 0.930889
educationProf-school             -3.507510   0.217890 -16.098  < 2e-16 ***
educationSome-college            -1.363711   0.171580  -7.948 1.90e-15 ***
marital.statusMarried-AF-spouse  -1.726140   0.611973  -2.821 0.004793 **
marital.statusMarried-civ-spouse -1.793633   0.324903  -5.521 3.38e-08 ***
marital.statusMarried-spouse-absent -0.063482   0.256327  -0.248 0.804397
marital.statusNever-married       0.384773   0.096072   4.005 6.20e-05 ***
marital.statusSeparated          -0.133175   0.170968  -0.779 0.436009
marital.statusWidowed            -0.178519   0.162619  -1.098 0.272303
relationshipNot-in-family        -0.209511   0.321548  -0.652 0.514679
relationshipOther-relative        0.622635   0.290785   2.141 0.032256 *
relationshipOwn-child             1.256714   0.329826   3.810 0.000139 ***
relationshipUnmarried            -0.072446   0.336574  -0.215 0.829576
relationshipWife                 -1.243029   0.113003 -11.000  < 2e-16 ***
raceAsian-Pac-Islander           -0.475586   0.271437  -1.752 0.079757 .
raceBlack                        -0.181162   0.247758  -0.731 0.464654
raceOther                         0.106998   0.389884   0.274 0.783751
raceWhite                        -0.414091   0.235284  -1.760 0.078414 .
hpw                              -0.032511   0.001769 -18.377  < 2e-16 ***
native.country.newGermany         0.333902   0.435092   0.767 0.442826
native.country.newItaly           0.047162   0.486716   0.097 0.922808
native.country.newother           1.047177   0.350380   2.989 0.002802 **
native.country.newUnited-States   0.498897   0.336692   1.482 0.138404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

9. Confusion Matrix for Logistic Regression

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1160  568
         1 1195 6686

               Accuracy : 0.8165
                 95% CI : (0.8086, 0.8242)
    No Information Rate : 0.7549
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4552

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4926
            Specificity : 0.9217
         Pos Pred Value : 0.6713
         Neg Pred Value : 0.8484
             Prevalence : 0.2451
         Detection Rate : 0.1207
   Detection Prevalence : 0.1798
      Balanced Accuracy : 0.7071

       'Positive' Class : 0
```

10. Classification Tree



11. Confusion Matrix for Classification Tree

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  6860 1194
     >50K    509 1205

               Accuracy : 0.8257
                 95% CI : (0.818, 0.8331)
    No Information Rate : 0.7544
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4794

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9309
            Specificity : 0.5023
         Pos Pred Value : 0.8518
         Neg Pred Value : 0.7030
             Prevalence : 0.7544
         Detection Rate : 0.7023
   Detection Prevalence : 0.8245
      Balanced Accuracy : 0.7166

       'Positive' Class : <=50K
```

Sources

UCI Machine Learning Repository: Census Income Data Set, https://archive.ics.uci.edu/ml/datasets/Census+Income.