Income Evaluation

Phillip Conrady

Sammer Khalaf

Vincent O’Dea

Layth Zubaidi

Fowler College of Business, San Diego State University

BA/MIS 749-01 Business Analytics

Spring 2022

Dr. Xialu Liu

Assignment Due Date: 05.11.2022

# 

# 

# INTRODUCTION

Our dataset analyzes an individual’s annual income results from various factors. Intuitively, it is influenced by the individual’s education level, age, gender, occupation, etc. This data comes from the U.S Census Bureau and is compiled from the period of 1994. With the dataset, we have applied machine learning algorithms in R Studio to conduct an analysis of the variables that are highly correlated to wealth. We have also implemented a training set to compare against a test set to evaluate the effectiveness of our models. The accuracy of our model was tested with cross-validation technique by splitting the data. Ultimately, we calculated prediction error by using model performance metrics.

# Data description

The data source comes from online, downloaded from Kaggle.com (<https://www.kaggle.com/datasets/wenruliu/adult-income-dataset>).

The data contains 15 variables and the income is divided into classes which are <=50K and >50K. We were able to obtain this information from Kaggle and had to clean up the data to proceed with the analysis. The population size of our dataset is about 48,842, and there are missing values. About 7% (3,419) have missing values, however, we will clean up the data by removing all entries with missing data. We also looked at our Income variable and would switch all values of <=50K and >50K into 0 and 1 respectively to represent those data points. It is interesting to examine which variables influence the income level. All variables are displayed in figure 1.

Analyzing census data is critical for developing accurate assessments of economic well-being for the Nation as a whole as well as for different racial, ethnic, and gender populations. Income statistics enable us to know about the distribution of income for a given population. The prominent inequality of wealth and income is a huge concern, especially in the United States. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. The principle of universal moral equality ensures sustainable development and improves the economic stability of a nation. We are curious about what variables have the biggest influence on an increase or decrease in income. It is important to analyze the data to compare the income of the people who belong to different sectors, countries, and occupations. The performance can be evaluated gender-wise and age wise. It primarily aims at learning the various factors that can help our evaluation process of what determines a salary less or greater than 50,000.

The information provided in the income evaluation dataset can be used to give better services, improve the quality of life, and identify the problems and solutions of a population. The fields of Data Analysis have not only exploited us for knowledge and discovery but also to explore certain hidden patterns and concepts which can lead to the prediction of future events. The problem of income inequality has been of great concern in recent years. Making the poor better off does not seem to be the sole criteria to be in the quest for eradicating this issue. People of the United States believe that the advent of economic inequality is unacceptable and demands a fair share of wealth in society. This model aims to conduct a comprehensive analysis to highlight the key factors that are necessary for improving an individual's income. Such an analysis helps to set focus on the important areas which can significantly improve the income levels of individuals.

Description of all Data Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Qualitative/Quantitative** | **Description** | **Levels for qualitative variables** |
| Age | Quantitative | The age of an individual. | Integer greater than 0. |
| Workclass | Qualitative | This is defined as a general term to represent an individual's employment status. | Private, Self­emp­not­inc, Self­emp­inc, Federal­gov, Local­gov, State­gov, Without­pay, Never­worked |
| Fnlwgt | Quantitative | This stands for the final weight. Essentially the number of people the census believes it represents. | Integer greater than 0. |
| Education | Qualitative | This is defined as an individual’s educational level. | ​​Bachelors, Some­college, 11th, HS­grad, Prof­school, Assoc­acdm, Assoc­voc, 9th, 7th­8th, 12th, Masters, 1st­4th, 10th, Doctorate, 5th­6th, Preschool. |
| Educational number | Quantitative | Highest level of education achieved in numerical format. | Integer greater than 0. From |
| Marital-status | Qualitative | This is defined as the marital status of an individual being separated as marriage of a civilian and armed force individual as seen between labels with Married-civ and Married-AF; there are 7 levels in this section. | Married­civ­spouse, Divorced, Never­married, Separated, Widowed, Married­spouse­absent, Married­AF­spouse |
| Occupation | Qualitative | This is defined as the occupation/career of an individual | Tech­support, Craft­repair, Other­service, Sales, Exec­managerial, Prof­specialty, Handlers­cleaners, Machine­op­inspct, Adm­clerical, Farming­fishing, Transport­moving, Priv­house­serv, Protective­serv, Armed­Forces. |
| Relationship | Qualitative | This is defined as the relation of an individual to another | Wife, Own­child, Husband, Not­in­family, Other­relative, Unmarried. |
| Race | Qualitative | This is the description of the race of an individual. | White, Asian­Pac­Islander, Amer­Indian­Eskimo, Other, Black. |
| Gender | Qualitative | This is the biological sex of an individual. | Male and female |
| Capital gain | Quantitative | Capital gains for an individual. | Integer greater than 0. |
| Capital loss | Quantitative | Capital loss for an individual. | Integer greater than 0. |
| Hours per week | Quantitative | The hours an individual has reported to work per week. | Integer greater than 0. |
| Native country | Qualitative | This is defined as the country of origin of the individual. There are 41 levels to this variables. | United­States, Cambodia, England, Puerto­Rico, Canada, Germany, Outlying­US(Guam­USVI­etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican­Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El­Salvador, Trinadad&Tobago, Peru, Hong, Holand­Netherlands. |
| Income | Qualitative | This variable is defined as whether or not an individual makes more than $50,000 annually. | There are 2 levels to this variable: <= 50K and > 50K |

Figure 1: Data Description

Summary Report of Quantitative Variables

|  | Mean | Median | Max | Min |
| --- | --- | --- | --- | --- |
| Age | 38.55 | 37 | 90.00 | 17.00 |
| Capital-Gain | 1101 | 0 | 99999 | 0 |
| Capital-Loss | 88.59 | 0 | 4356.00 | 0 |
| fnlwgt | 189735 | 178316 | 1490400 | 13492 |
| Hours per work | 40.94 | 40 | 99 | 1.00 |
| Educational Num | 10.12 | 10 | 16 | 1.00 |

Figure 2: Quantitative Data Summary

**Histogram of Quantitative Variables.**

The histograms presented provide further insight into the data set.

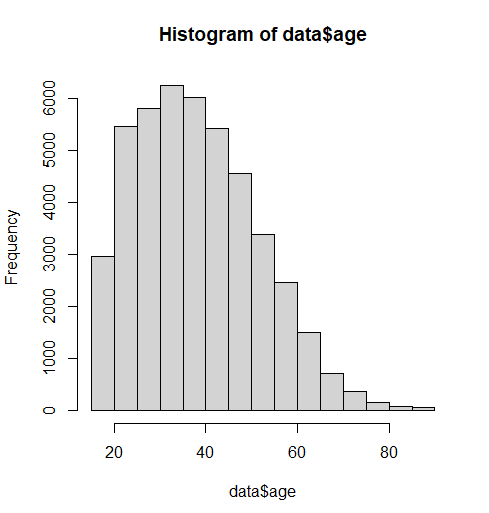
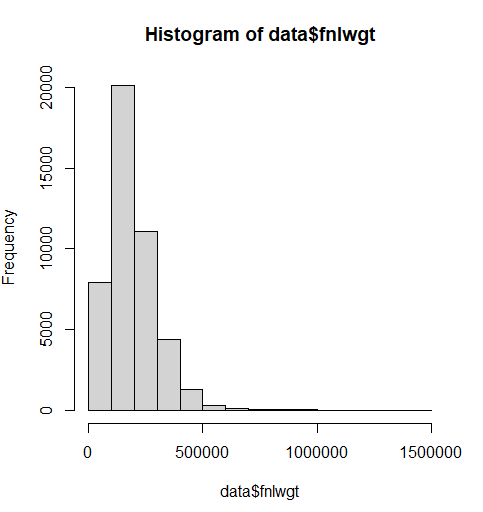
 

Figure 3: Age Figure 4: fnlwgt (Final Weight)

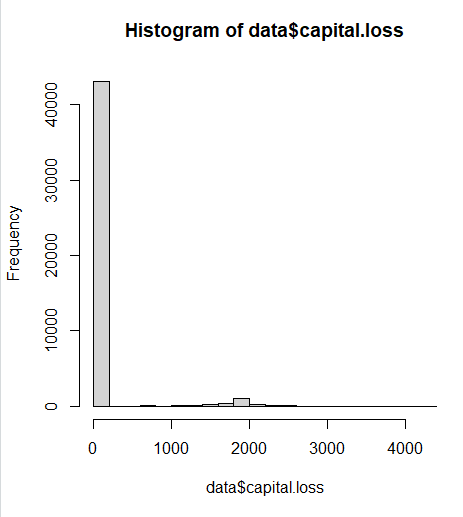
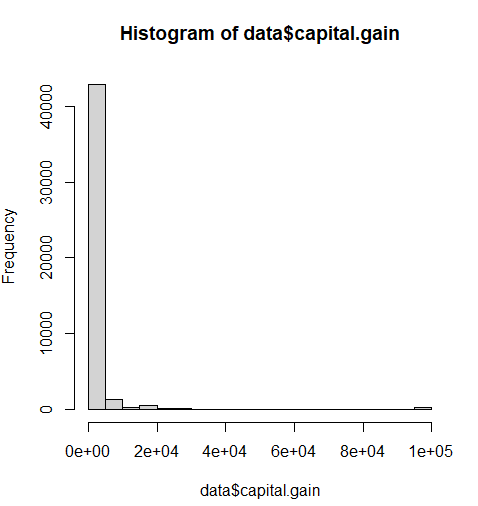


Figure 5: Capital Gain Figure 6: Capital Loss

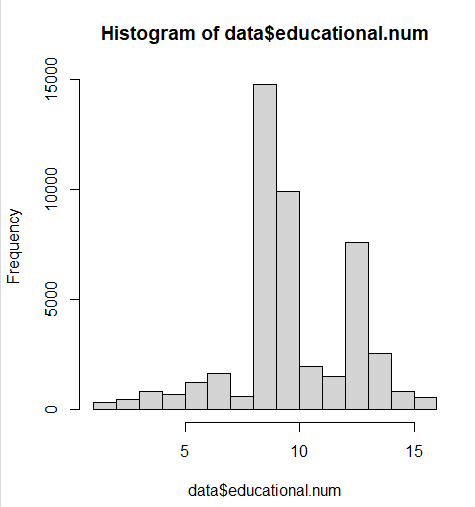
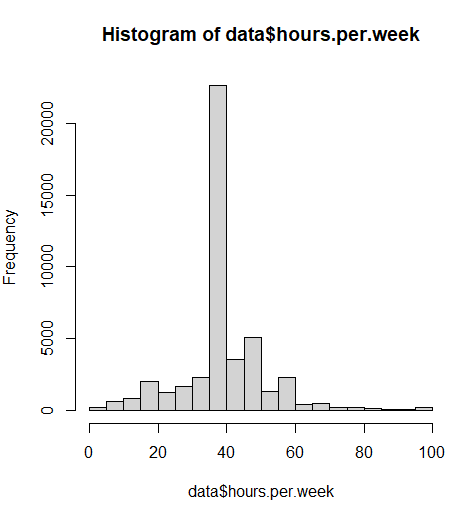


Figure 7: Hours Per Week Figure 8: Educational Num

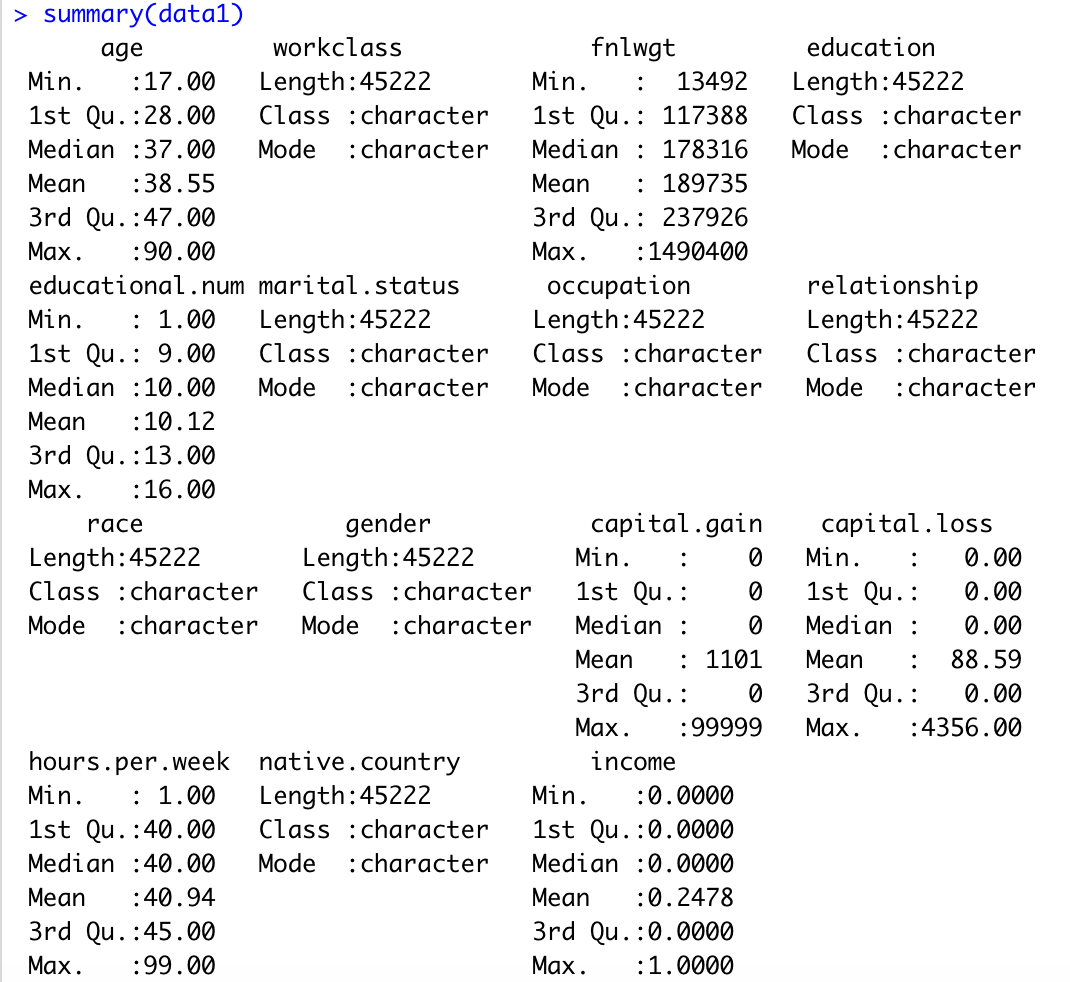
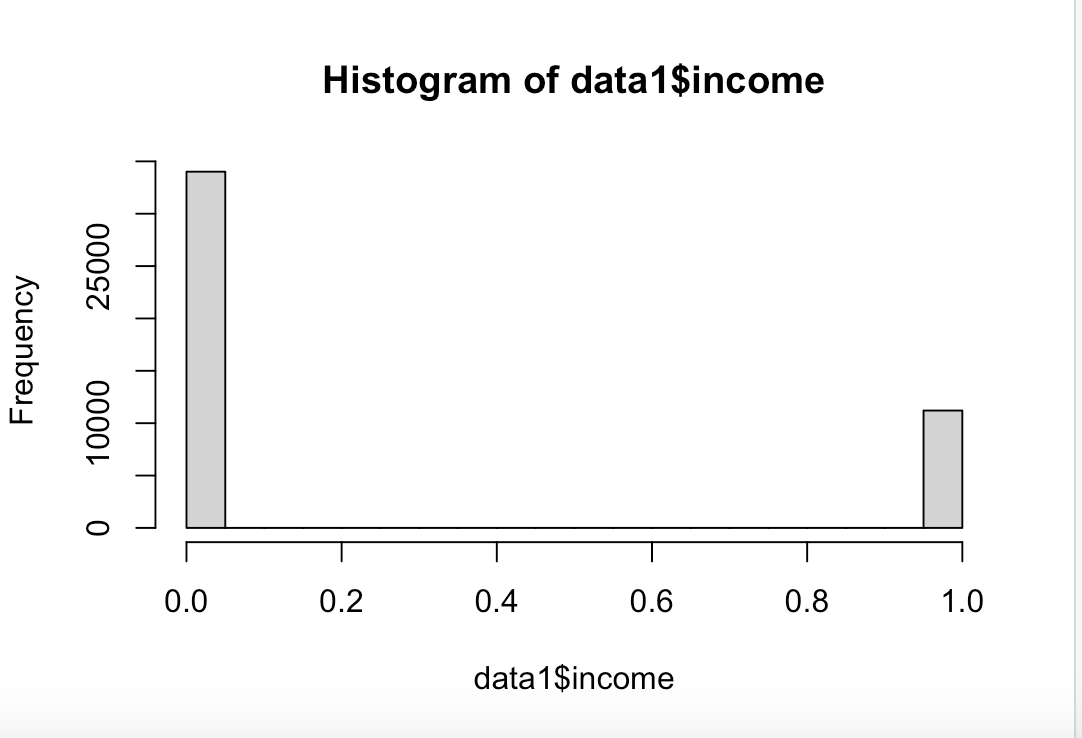


Figure 9: Income (0 -> <=50K, 1 -> >50K) Figure 10: Summary (R Script) of the Dataset.

# 

# 

# 

# 

# 

# 

# 

# 

# – For qualitative variables: how many levels? The frequency of different levels.

# **Workclass:**

* Federal-gov = 1406
* Local-gov = 3100
* Private =33307
* Self-emp-inc= 1646
* Self-emp-not-inc = 3796
* State-gov= 1946
* Without-pay =21

**Education:**

* Preschool: 72
* 1st-4th: 222
* 5th-6th: 449
* 7th-8th: 823
* 9th: 676
* 10th: 1223
* 11th: 1619
* 12th: 577
* HS-grad: 14783
* Assoc-acdm: 1507
* Assoc-voc : 1959
* Some-college: 9899
* Bachelors: 7570
* Prof-school: 785
* Masters: 2514
* Doctorate : 544

**Marital-status:**

* Divorced: 6297
* Married-AF- Spouse: 32
* Married-civ-spouse: 21055
* Married-spouse-absent: 522
* Never-married: 14598
* Separated: 1411
* Widowed: 1277

**Occupation:**

* Adm-clerical: 5540
* Armed-Forces: 14
* Craft-repair: 6020
* Exec-managerial: 5984
* Farming-fishing: 1480
* Handlers-cleaners: 2046
* Machine-op-inspct: 2970
* Other-service:4804
* Priv-house-serv:232
* Prof-specialty: 6008
* Protective-serv:976
* Sales:5408
* Tech-support: 1420
* Transport-moving: 2316

**Relationship:**

* Husband: 18,666
* Not-in-family: 11,702
* Other-relative: 1349
* Own-child: 6626
* Unmarried: 4788
* Wife: 2091

Race:

* Amer-Indian-Eskimo : 435
* Asian-Pac-Islander: 1303
* Black: 4228
* Other: 353
* White: 38903

**Gender:**

* Male: 32650
* Female: 16192

**Native Country:**

* Cambodia: 26
* Canada: 163
* China: 113
* Columbia: 82
* Cuba: 133
* Dominican - Republic: 97
* Ecuador: 43
* El-Salvador: 147
* England: 119
* France: 36
* Germany: 193
* Greece: 49
* Guatemala: 86
* Haiti: 69
* Holand-Netherlands: 1
* Honduras: 19
* Hong Kong: 28
* Hungary: 18
* India: 147
* Iran: 56
* Ireland: 36
* Italy: 100
* Jamaica: 103
* Japan: 89
* Laos: 21
* Mexico: 903
* Nicaragua: 48
* Outlying-US(Guam-USVI-etc): 22
* Peru: 45
* Philippines: 283
* Poland: 81
* Portugal: 62
* Puerto-Rico: 175
* Scotland: 20
* South: 101
* Taiwan: 55
* Thailand: 29
* Trinadad&Tobago: 26
* United States: 41292
* Vietnam: 83
* Yugoslavia: 23

**Income:**

* <= 50k: 34,014
* >50k: 11,208

# 

# Data methods

# Since the data set at hand is a classification problem, the methods presented in class are applied. Each of these models has its advantages and disadvantages, which are presented below.

Before the data can be analyzed, the data must be adjusted. First, the income variable is converted to 0 (for income below 50k) and 1 (for income above 50k). In addition, all quantitative variables have to be converted into factors in order to be able to use them for the classification methods. Likewise, cross validation is performed by splitting the data set. The training set contains 70% of the data and the test data set contains 30% of the data. The cross validation is to estimate a first test error.

K Nearest Neighbors

Pros: These algorithms can be used for classification, ranking, regression (using neighbors average or weighted average), recommendations, missing value imputation etc.

Cons: It is a distance based-approach hence the model can be badly affected by outliers, it’s prone to overfitting.

Logistic Regression

Pros: Easy to separate response into 0 and 1 indicating <=50k and >50k.

Cons: It can overfit in high dimensional datasets and does not support non-linear relationship between the predictor and the outcome..

Decision Tree

Pros: Easy to understand and interpret, perfect for visual representation.

Cons: It is very sensitive . Small change in the data can affect prediction greatly (High variance).

For the decision tree methods, in addition to the unpruned approach, the prune, bagging and random forest approaches can be used to obtain better results.

Quadratic Discriminant Analysis

Pros: Classification is usually more accurate and tends to outperform KNN and LDA.

Cons: This model uses Gaussian assumption and complex matrix ops.

Linear Discriminant Analysis

Pros: It is a simple, fast and portable algorithm.

Cons: It requires normal distribution assumption on features/predictors.

Naive Bayes

Pros: NB classifier performs better compare with other models like logistic regression and you need less training data.

Cons: There is also the assumption of independence in predictors. In real life, it is almost impossible that we get predictors which are completely independent.

# VARIABLE HANDLING AND MANIPULATION

Due to the sheer size of the dataset running simple r-code was difficult due to the numerous variables. As a result we systematically tried to eliminate variables that were cluttering the data and causing handling difficulties.

Educational.num = We nulled this variable as the education variable already explains this.

Fnlwgt = Excluded as the final weight was calculated by the Census Bureau at the time and has become redundant.

Relationship = Details an individual's role in a family, we believe that marital status and gender sufficiently cover that demographic.

Native.country = Reluctantly this variable was earmarked for removal due to the high level of skewness within the variable. 96.92% of individuals were from the United States.

Workclass = This variable was restructured to allow for greater interpretability levels federal, state and local government were combined into a Government column. Additionally Self-emp-not-inc, and Self-emp-inc were fused to form Self-employed.

# COMPARISON OF DIFFERENT MODELS

Logistic Regression (GLM):

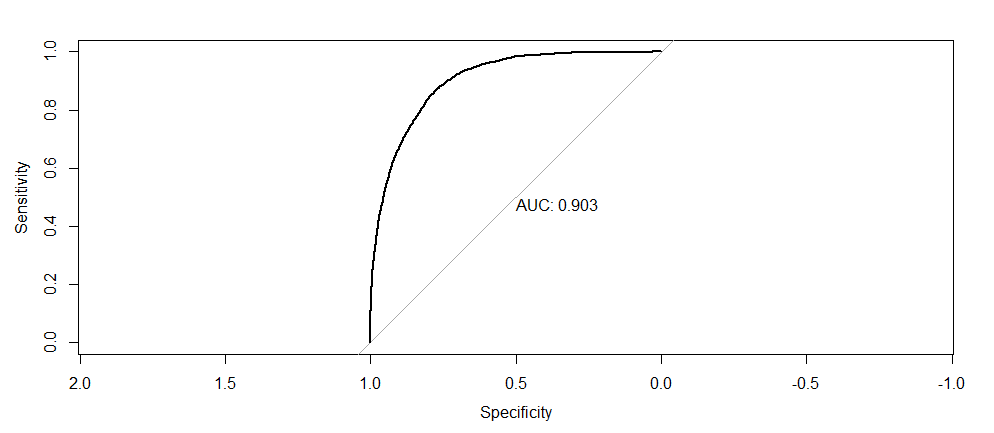
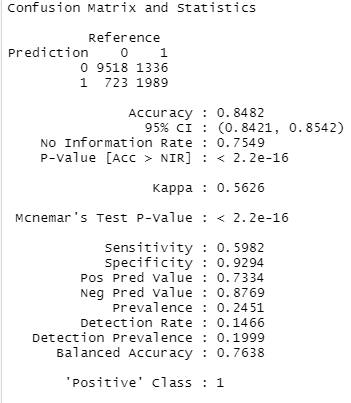


Figure 11: Confusion matrix -GLM Figure 12: Area Under the Curve - GLM

Linear Discriminant Analyst (LDA):

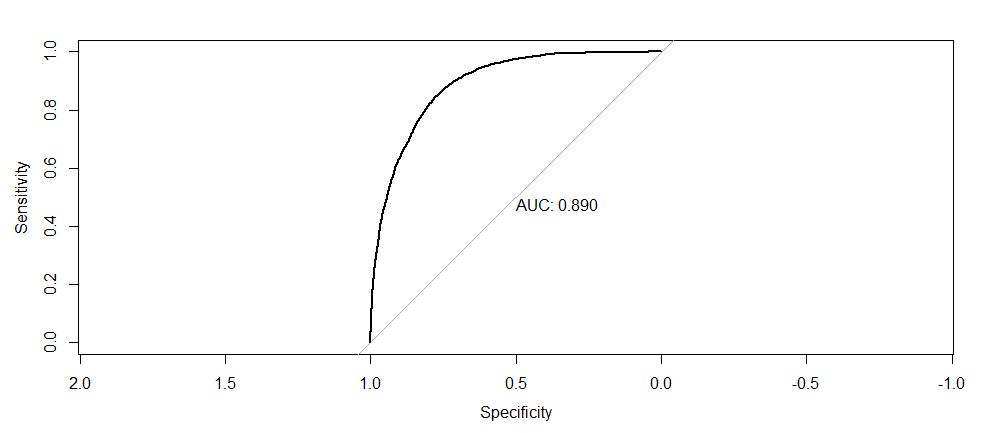
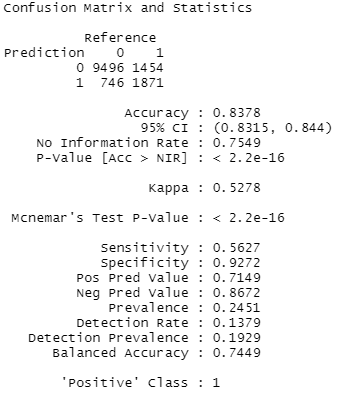


Figure 13: Confusion Matrix -LDA Figure 14: Area Under the Curve -LDA

Quadratic Discriminant Analysis (QDA):

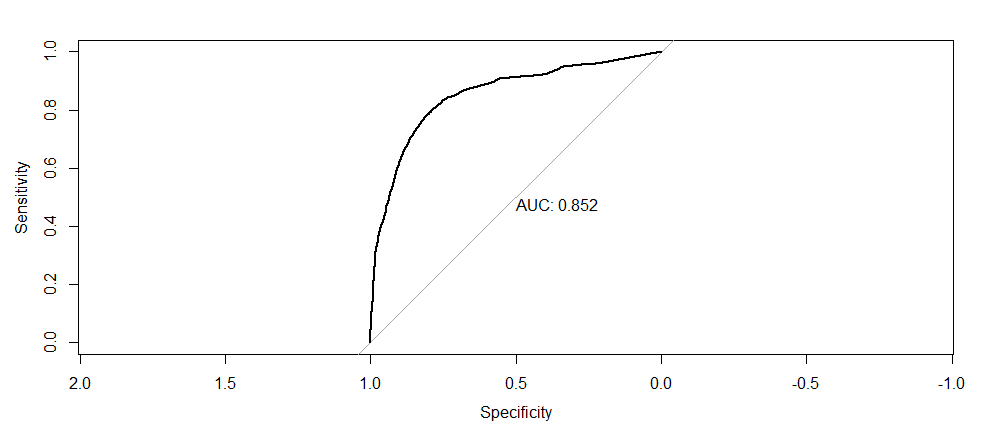
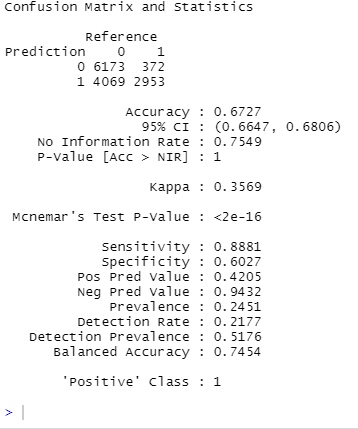


Figure 15: Confusion Matrix - QDA Figure 16: Area Under the Curve - QDA

Naive Bayes:

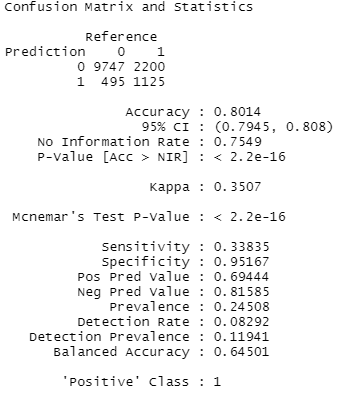


Figure 17: Confusion Matrix - Naive Bayes

Table. 1 Model Performance

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Misclassification Rate | AUC |
| Logistic Regression | 0.8482235 | 0.1517765 | 0.9033 |
| LDA | 0.8378418 | 0.1621582 | 0.8903 |
| QDA | 0.6726616 | 0.3273384 | 0.8524 |
| Naive Bayes | 0.8013562 | 0.1986438 | / |

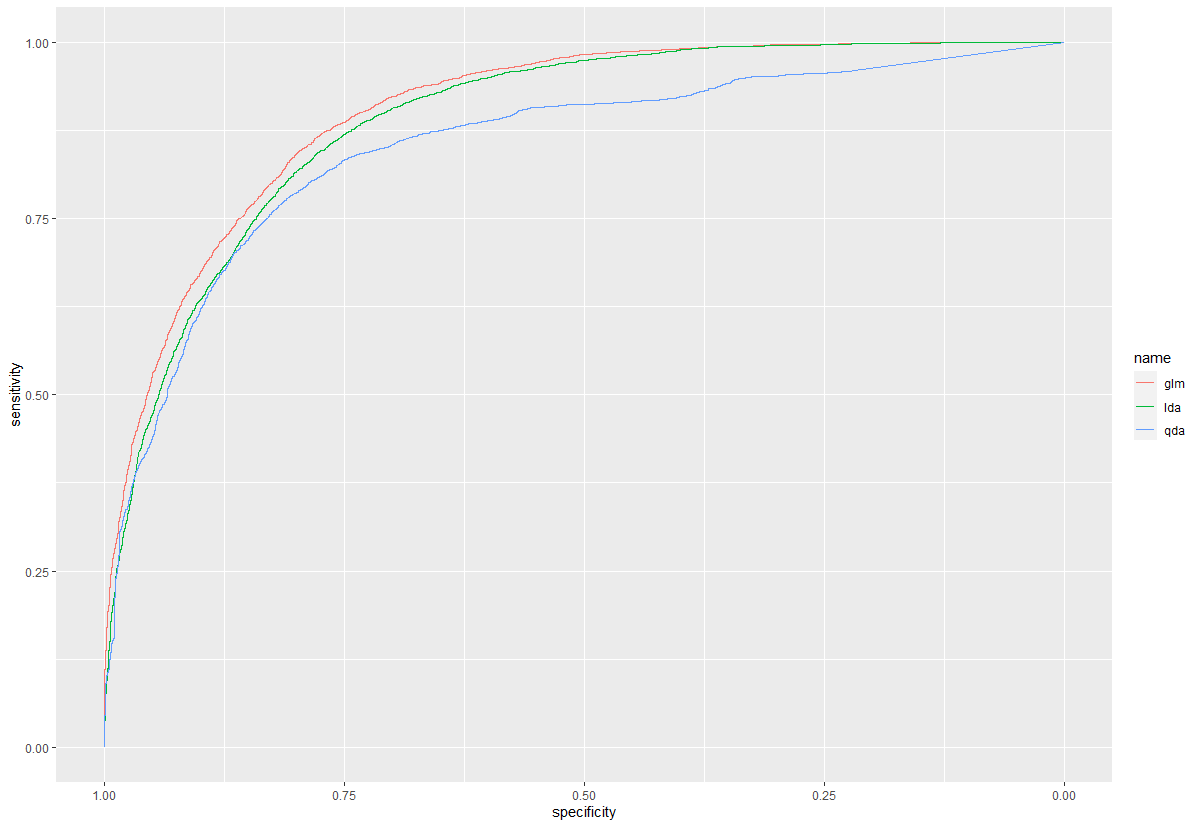


Figure 18: ROC Curves

Classification Tree:

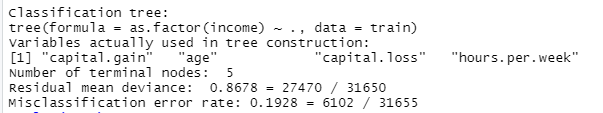


Figure 19: Classification Tree Summary

Decision Tree:

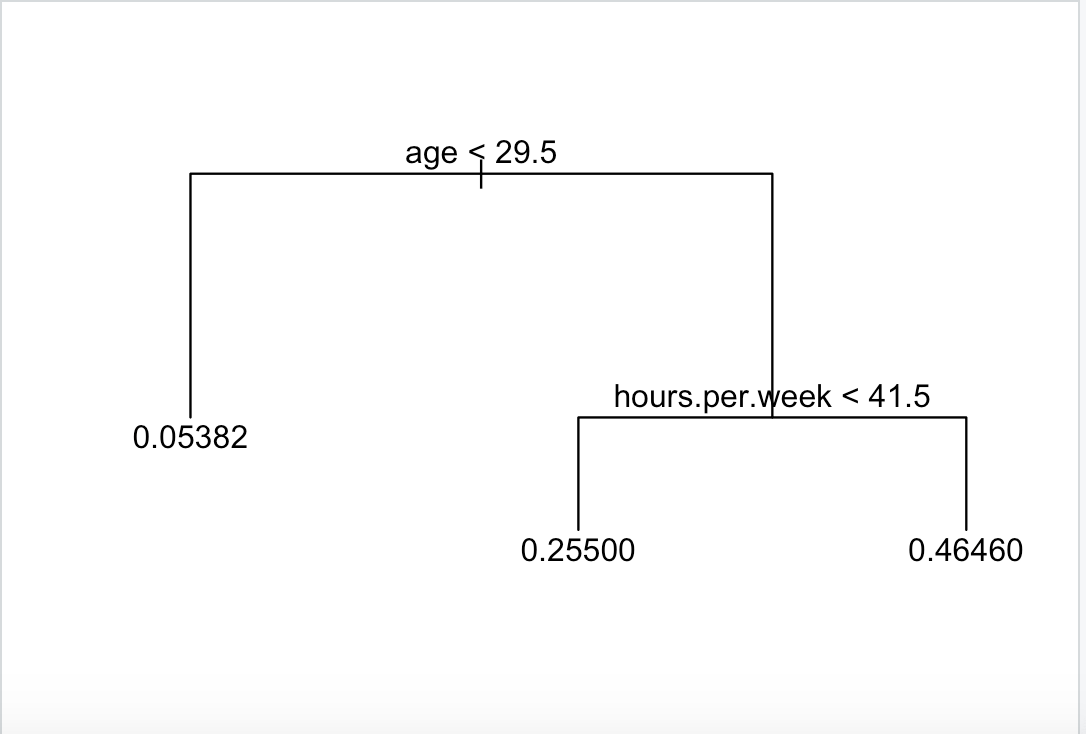


Figure 20: Decision Tree - Pruned

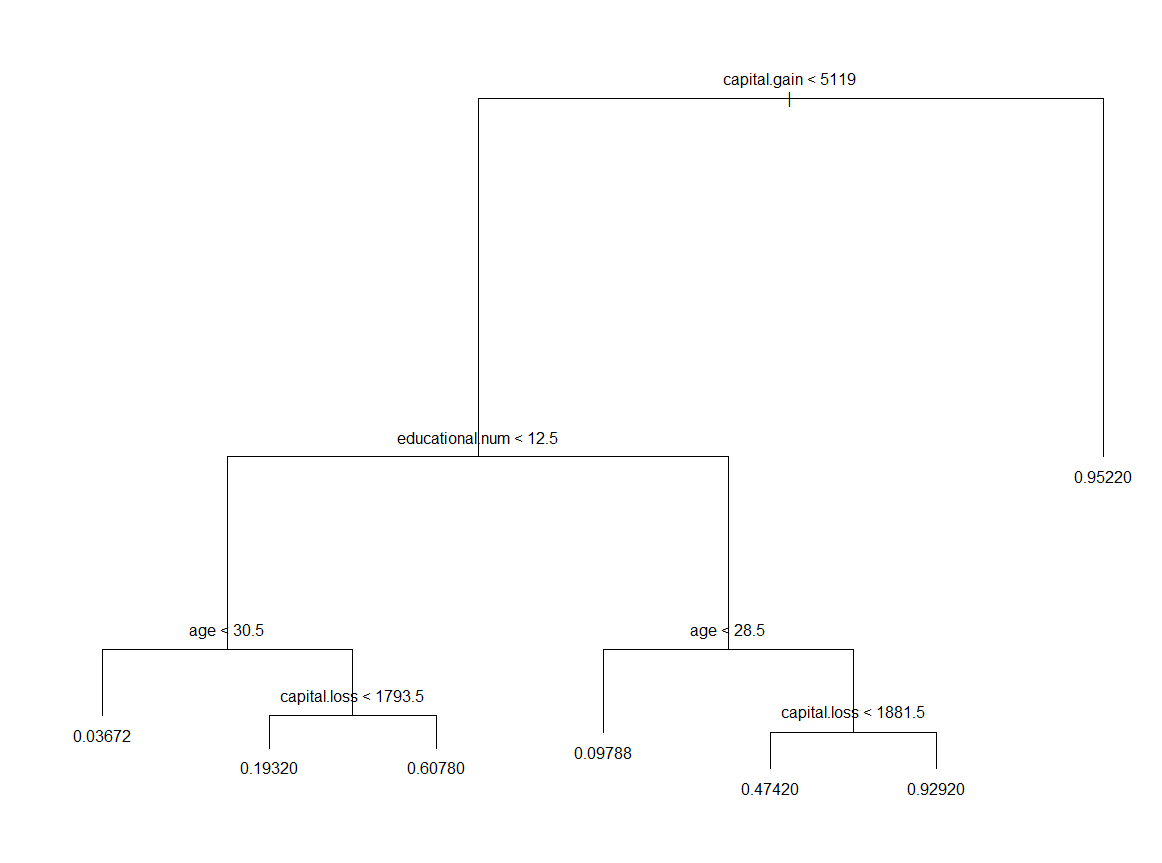


Figure 21: Decision Tree - Unpruned

# 

# 

# 

# 

# 

# 

# 

# Results

In conducting a ROC curve the Logistic Regression model had the highest AUC compared to other models. The test error rate for the Logistic Model was the lowest as well, which indicated it was the most accurate model. Occupation and Age had the highest influence on the binary result of >= 50K and <50K. Also, we were able to find that most people who tested >=50K had a High School education and some college education was next.

The decision tree model had interesting results, although the unpruned one was an overfit we were still able to conduct quality results from the pruned decision tree. We have indicated that age was a very important terminal node which was similar among the logistic regression test. With gaining such similarity we were able to conclude that age was a significant variable as well. To understand the dataset, we looked at the average age and discovered it was 39 years old.

# Conclusion

Classification supervised learning gave us the ability to create multiple trends within the data set to create recommendations of which variables produce significance. We initially decided to look at a handful of different analytical models for our data set, that being Logistic Regression, Quadratic Discriminant Analysis, Linear Discriminant Analysis, Naive Bayes, and Classification Tree. We found the Quadratic Discriminant Analysis to be the least accurate function for our datasets. Through cross validation we were able to assure that our models weren’t overfitting or underfitting the data set. Through our findings the models test error rate were compared to see which yielded highest results. Ultimately after comparing all these functions we found that Logistic Regression produces the most accurate results for our dataset.

Through the power of our machine learning algorithms the dataset was capable of being able to predict individuals who would make >= 50K annual income. We saw how certain variables such as occupation (like executive managerial positions) were a strong point associated with whether a person would make more or less than 50K annually. Age was in second for the highest variable associated with making more than 50k annually. From these results our group can suggest which careers are lucrative and the variables associated with these individuals. Our results can also give us a better understanding of which professions and associated variables were underperforming in the market.