**Predicting Potential Earning**

**Stats 6372 – Project 2**

Adeel Qureshi, Sadik Aman, Gavin Rick

07/31/2021

**Introduction**

The purpose of this project is to understand the earning potential of people based on multiple factors. The data was provided from the 1994 Census dataset. We will be doing statistical analysis and machine learning approach to see if we can predict if a person will make more than $50k or less than $50k. The models that we will be used will include logistic regression(basic), logistic regression(complex), LDA, random forest and random forest. We will see based on our finding if statistical or machine learning perform better.

**Data Description:**

Data for this study was provided by Barry Becker from the 1994 Census database. The dataset can be found at the UCI Machine Learning Repository. The data was filtered to take out anyone less than 16 years of age. The data consists of 15 variables and 32,561 observations. The data contains both continuous and categorical variables in the dataset. A few examples of variables used were Age of the people in the census which covered from 17 to 90 years old, Workclass provides the breakdown of if they work in private industry, self employed and etc, Education which provides high school, bachelors and other categorical information, Education\_num is another variable that provides the number associated with education level and Sex which provides Female or Male option. There were variables that are not self-explanatory like fnlwgt, education\_num and relationship. (See Appendix “Original Data Variables”).

**Exploratory Data Analysis (EDA):**

**Preprocessing Data:**

The first phase of our analysis was data preprocessing which included two crucial steps. Firstly, we explored the data to determine the variables that were best suited for our analysis. Secondly, we performed variable transformation. The details of each steps are highlighted below.

**Simplifying the Categorical Variables:**

While exploring the dataset we wanted to reduce the levels in the factors that were in the data. The three categorical variables that we simplified were:

***Workclass variable***

|  |  |
| --- | --- |
| Original Factor level | Transformed Level |
| Without – pay, Never – worked | Jobless |
| State-gov, local-gov | Govt |
| Self-emp-inc, self-emp-not-inc | Self employed |

***Marital\_status***

|  |  |
| --- | --- |
| Original Factor level | Transformed Level |
| Married-AF-spouse, civ-spouse, spouse-absent | Married |
| Divorced, Separated, Widowed | Not-Married |

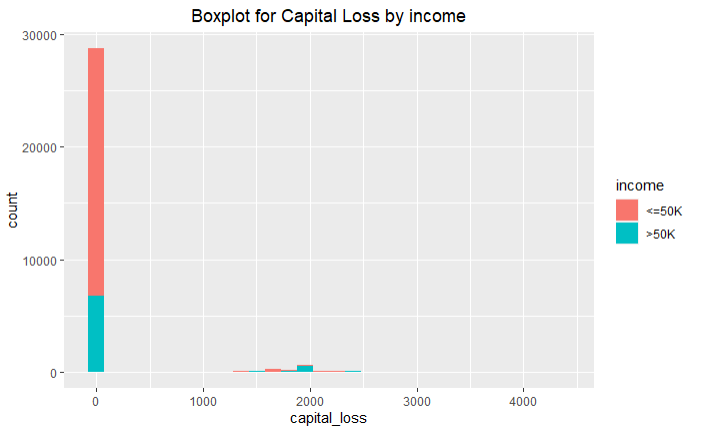
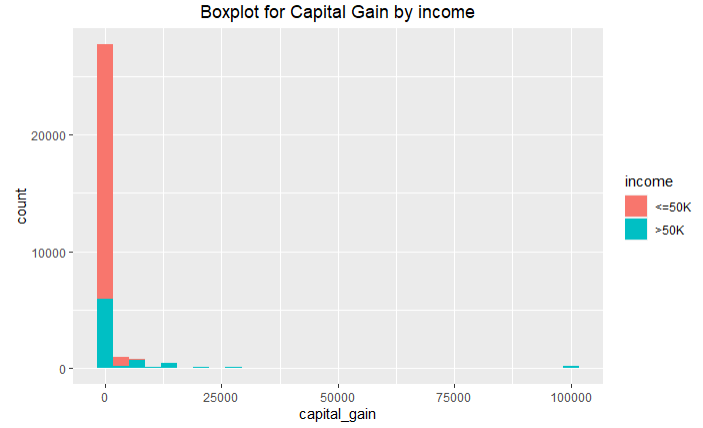
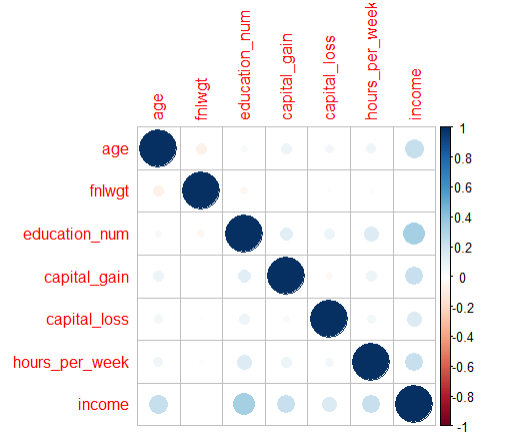
***Native\_country***

|  |  |
| --- | --- |
| Original Factor level | Transformed Level |
| Countries | continent |

**Exploring Continuous Variables:**

Before we begin our analysis and build the statistical and machine learning models. We need to analyze how the continuous variables are correlated to each other. How is the data spread out by doing multiple tests, first will be correlation between the continuous variables. Then we will be viewing the histograms of capital gain and capital loss to understand the data spread. What the shape of the data is in.

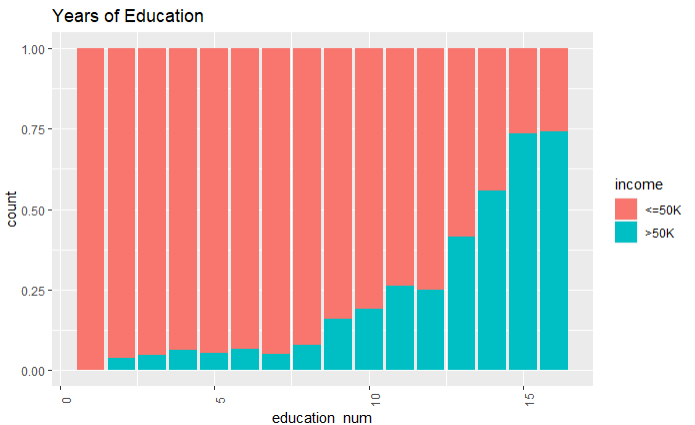
*We can see a weak relationship between income and fnlwgt. Adding another reason of dropping the variable.*



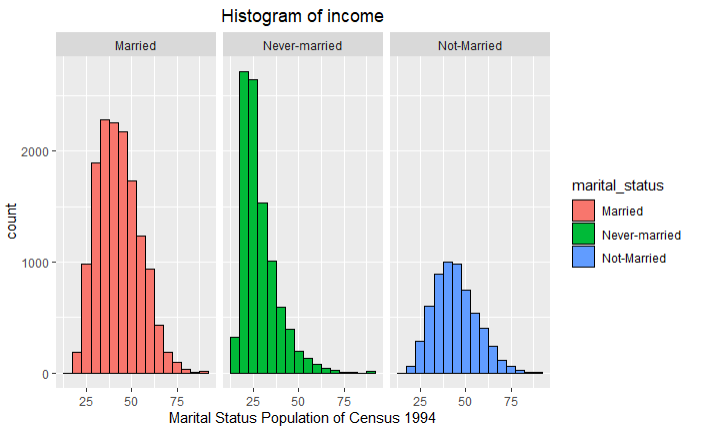
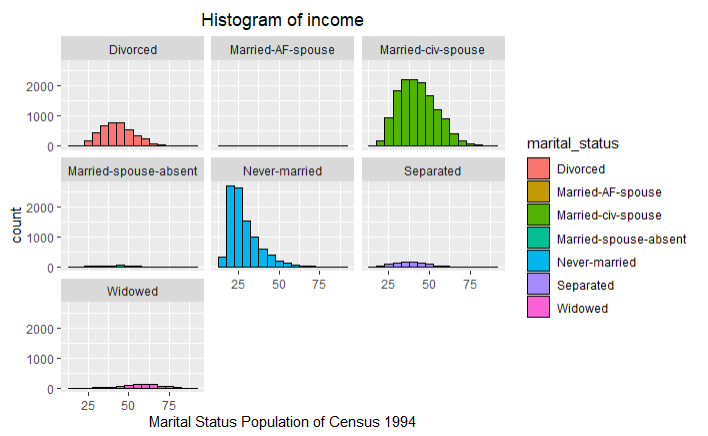
We can see the data for capital gain and capital loss is predominately 0. Which makes the variables as less useful. We can see the capital gain and capital loss variables can be combined but do not seem to add much value due to majority of the data in the variables being 0.

**Exploring Categorical Variables:**

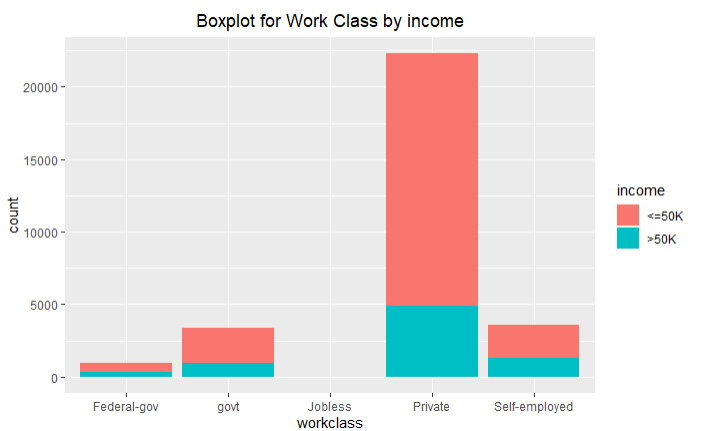
We can see below the that correlation between as education level increases so does the potential to make above $50k on average.



We can see pre-transformed marital status data, has normal distribution and right skewness. But there were other classification that were not adding much weight to the analysis. The simplified version is cleaner and provides a better relationship in the data.



Analyzing the work class variable we can see the highest of amount of people working that make above $50k are employed by private companies and second highest is self-employed. Both of them make sense since practically gov jobs are always on the lower end of the spectrum.



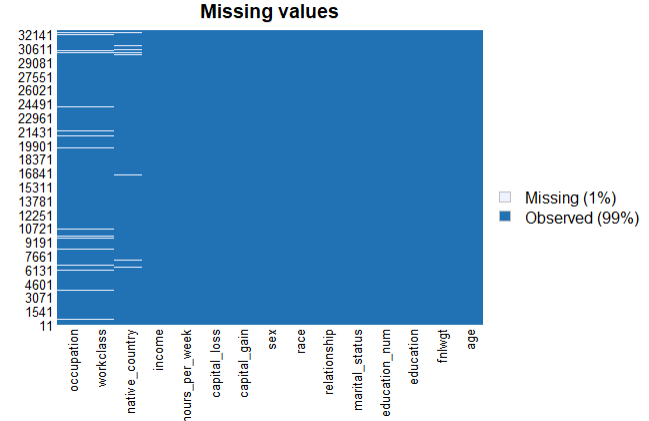
**Variable Selection:**

After through exploratory data analysis we decided to exclude the variables ‘education’ from our model. This is due to the fact that ‘education’ seems highly correlated with another variables ‘education-num’. We also proceeded with dropping the variable fnlwgt since it did not add any value and provide information. The relationship was dropped due to not providing any relevant information. The rest of the variables were kept in the dataset for our statistical models and machine learning models.

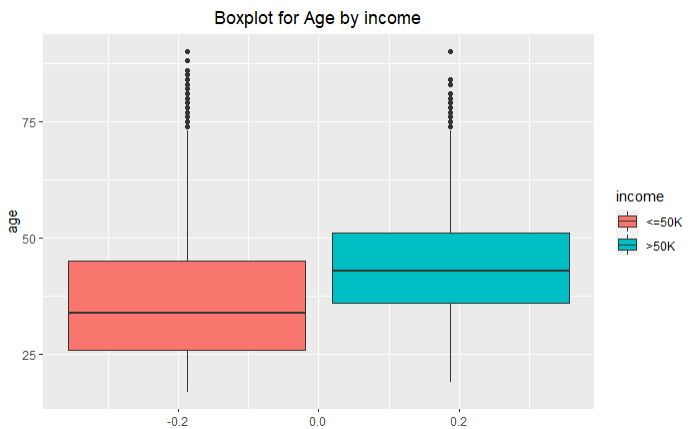
**Handling Missing Values:**

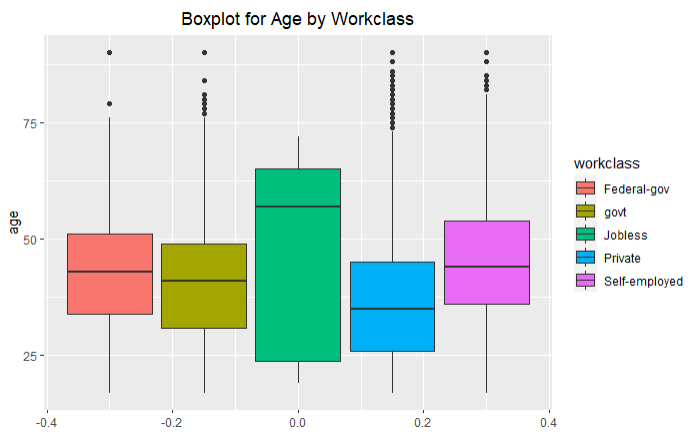
Considering income as our response value, we would begin with looking for the NA or missing values in the dataset. First we had to convert the “?” missing values to NA’s for us to see how many observations were missing and what variable. Based on our findings, we saw there are total 3,679 observations that were missing from two columns Occupation and Workclass. We looked at couple of options is either imputing or dropping the NA’s for a clean dataset that we could build upon. (See Appendix “EDA missing values”)

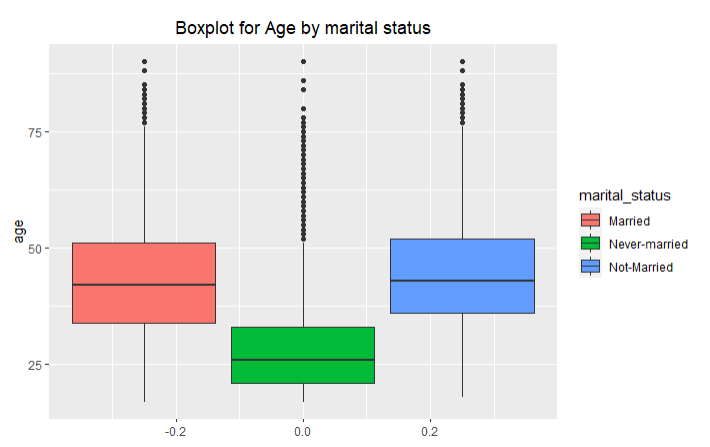
Missing values in the dataset:



To analyze the distribution of data, box plots of age versus income were created. Based on the overall spread, we can see that there is a as age increase so does the income level. There are quite a few outliers that we can see in the set that are the age of 70. Further examination shows that outliers are on the private sector in the workclass categorical variable. If see by marital status the “never-married” has majority of the outliers in the dataset.







**Objective 1:**

*Problem and Approach*

Many factors affect the persons income in United States, and most often the different factors can have different affects depending on which workclass and education level is being studied. This study did not a single out a native country but looked at the continent level to see how the race can affect the income level. The data collected was analyzed and some questions formed, such as; how much impact education level has on potential of income. Does native country affect how much someone is making and there might additional factors that dataset might not cover like how well off the participants were and how well their parents. Since that can affect a person’s potential by having resources available.

*Model Selection:*

The approach for the model selection process centered on the questions mentioned above. We looked at multiple ways on how we could build the logistic regression model, whether we wanted to manually pick and choose the variables and transformed the data. But we ended up building the simple logistic model by using all the variables.

*Assumptions:*

*Multicollinearity*:

Based on the variables that we selected the pairs plot below shows, that none of the continuous variables or categorical variables have high correlation with each other. Even when we transformed the data it did not help with the correlation.

Table

Description automatically generatedDiagram

Description automatically generated

*Outlier:*

The outliers from the dataset were removed to meet the no outlier’s assumption from the data.

Chart

Description automatically generated with medium confidenceChart

Description automatically generated with medium confidence

*Checking Assumptions:*

Chart, line chart

Description automatically generated

*Residuals:*

Chart, scatter chart

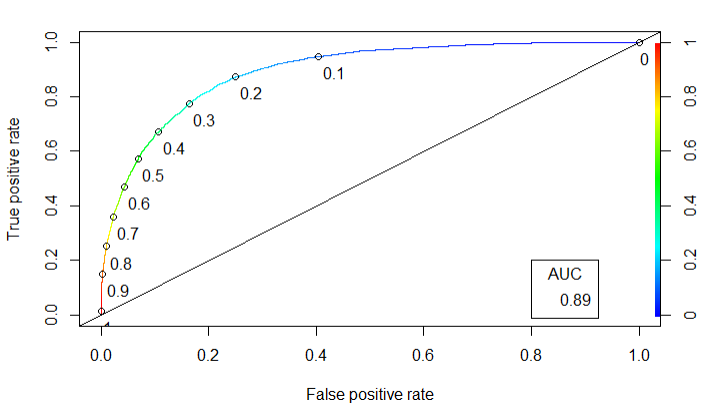
Description automatically generated

*Parameters Interpretations:*

|  |  |
| --- | --- |
| Accuracy | 84.7% |
| Sensitivity | 87.8% |
| Specificity | 73.1% |
| AIC | 7209 |
| AUC | 89.1% |
| Confidence Interval | (0.38, 0.56) |

Text

Description automatically generated**ROC Curve**



In order to achieve model that met all of the assumptions of a logistic regression model, we had to ensure there were no outliers and variables had no multicollinearity. We used the backwards technique in building the model. With that in mind for the variables were not transformed and we kept the object 1 model simple. The model was able to achieve decent accuracy of 84.7% and we relied on sensitivity as our measure with the performance of the model. We can see that AIC score after data cleaning was at 7209 and AUC at 89%. We are quite confident that this model will be able to perform well.

**Objective 2:**

*Problem and Approach*

In Part 1, a simple logistic regression using manually selecting the variables. Besides the three variable’s we dropped we found no additional justification in dropping any further. The assumptions were me building the simple model once we took out the outliers in the model.

Complex Model:

Our goal for object 2 was to build a complex logistic regression model using multiple techniques. Again we ended up using the backwards selection technique due to the lower AIC score and highest accuracy it produced. We did not transform any of the variables and choose to include interactions between variables to increase the complexity. We used only continuous variable to build the complex model to see if that would be more advantageous. We found the AIC score to increase and AUC to decrease.

*Parameters Interpretations:*

*Complex Logistic Regression*

|  |  |
| --- | --- |
| Accuracy | 81.3% |
| Sensitivity | 82.3% |
| Specificity | 72.0% |
| AIC | 8662 |
| AUC | 81.1% |
| Confidence Interval | (0.41, 0.76) |

*Linear Discriminate Analysis:*

Linear Discriminate Analysis, overall, resulted in lower accuracy over logistic regression. The most probable reason for this is because linear discriminate analysis assumes normally distributed predictors, while logistic regression does not. It achieved an accuracy of 84.04%, which came lower than the both the simple logistic regression and complex logistic regression.

Trees Model:

The decision trees model produced an accuracy of 85% and had the second highest sensitivity score among the models tested at 94.12%.

Text

Description automatically generated

Random Forest:

The random forest model produced the highest sensitivity score at 95.15% but accuracy was lower than the logistic model.

A picture containing text, receipt, screenshot

Description automatically generated

Chart, line chart

Description automatically generated

*Analysis Content:*

We can see based on all the tests done to complete with objective 1 and see if a complex model can outperform a simple model. We build an additional 4 models to test against the simple complex model. The random forest, linear discriminate analysis, decision tree and continuous complex logistic regression model. All the models were close to each other, I do not think any of them performed terribly. If we deployed out all 5 of the models, we believe they would do the job adequately.

**Conclusion:**

While the exact mechanisms behind income are nuanced and diverse, this analysis was able to use census data to predict individuals’ income based on multiple factors. As with many real-world analyses, the way in which socio-economic factors affect a person’s income are not quite linear. This is why, we deployed logistic regression model to help with predicting the income. We found that non-parametric approach (random forest) allowed us to uncover more nuanced relationships among persons income. Due to the confounding variables and using manual approach we found the non-parametric models like decision trees and random forest were the better ones. Statistical methods can work in predicting a person’s income, but there are hidden relationships that do not get covered. I think more data is needed to help with the future analysis since 1994 census data limited to certain segment of the population does not tell us overall story of the country. So we would need to look at additional years to see if the trend exists and if the model can hold up in the future. To conclude we would recommend using the random forest for predicting a person’s income due to the high sensitivity score of 95.15% which we used to pick out the top model.

**Appendix**

**Original Data Variables:**

***Adult Data frame***

|  |  |
| --- | --- |
| **Variable** | **Data Type** |
| Age | Int |
| Workclass | Chr |
| Fnlwgt | Int |
| Education | Chr |
| Education\_num | Int |
| Marital\_status | Chr |
| Occupation | Chr |
| Relationship | Chr |
| Race | Chr |
| Sex | Chr |
| Capital\_gain | Int |
| Capital\_loss | Int |
| Hours\_per\_week | Int |
| Native\_country | Chr |
| income | chr |

EDA for NA’s and missing values

|  |  |  |
| --- | --- | --- |
| Variable | No. Missing | Pct Missing |
| Occupation | 1843 | 5.66 |
| Workclass | 1836 | 5.63 |
| Age | 0 | 0 |
| Education\_num | 0 | 0 |
| Marital\_status | 0 | 0 |
| Race | 0 | 0 |
| sex | 0 | 0 |
| Capital\_gain | 0 | 0 |
| Capital\_loss | 0 | 0 |
| Hours\_per\_week | 0 | 0 |
| Native\_country | 0 | 0 |
| income | 0 | 0 |