**6372 Project 2**

Ryan Herrin,

Garrett Shankel,

Chad Kwong

**Introduction**

Society has always been curious about individual worth. In today’s economy, most measure individual worth through individual income. We chose to explore the 1994 US Gov Census data for possible factors that may help in predicting whether an individual made more or less than $50,000 at the time. Our expectations were to conduct an Exploratory Data Analysis on the data, construct a simple logistic regression model, and explore more complex predictive models for predicting the monetary status of an individual.

**Data Description**

The data used in this analysis was obtained from the UC Irvine Machine Learning repository, which initially obtained the data from the 1994 Government Census Database made publicly available. There were 15 total attributes containing 14 predictor factors and one response for whether the individual made above 50K in salary or not. Out of the 14 predictors, 8 are categorical variables with discrete levels and 6 are continuous. A full description of each variable can be found in **Table 1** within the **appendix**.

The dataset obtained contained 6465 total missing values spread across 3 distinct variables: Work Class, Occupation, and Native Country. Each missing value represents an instance where the field was not recorded for an individual. We extracted the rows containing missing values into a separate data frame and we discovered no correlation between the data. This leads us to assume that missing values resulted from individual responders choosing to not report that information on the census. With this in mind, we chose to omit the rows containing missing values.

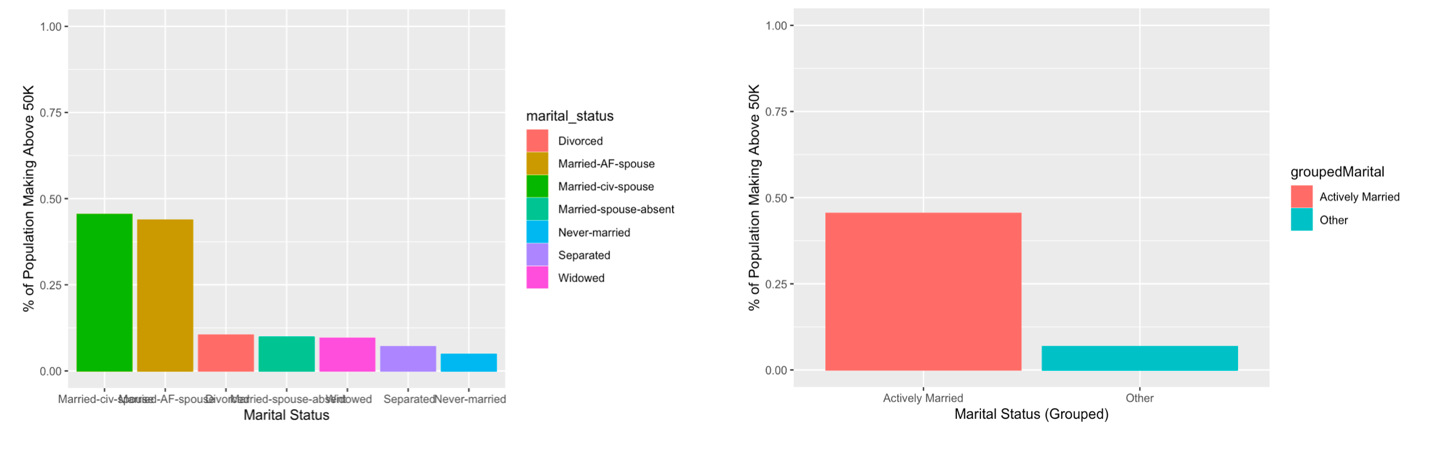
While we controversially chose to omit data from some of our models, we recognized that there were options for approaching these missing values. We constructed a secondary dataset that converted all NA values to “Unavailable” that will be used in tandem with our omitted dataset for direct comparisons between approaches. Furthermore, we theorized that a clustering method may be used to extrapolate and estimate these missing fields, however we did not have the available time to conduct this approach. If a further investigation was performed to minimize missing values, we would highly recommend taking this approach.

**Exploratory Data Analysis**

As the first part of our EDA, we chose to investigate each individual relationship between the predictor variables and the response variable, which we will refer to as **Money**. For the classification variables, we investigated relationships through direct comparisons of the populations of each group that make above 50K Salary in separate bar charts.

In our investigation into these individual relationships, we found that many of the classification variables had more than 3 factor levels, which would lead to a high quantity of individual level interpretations within later models. To ease the interpretation of these variables used in our later models, we chose to consolidate most of these variables where available. By comparing the percentages of the populations of each group that made above 50K a year, we were able to logically group certain levels together within the following categorical variables: Education, Race, Marital\_Status, and Relationship.

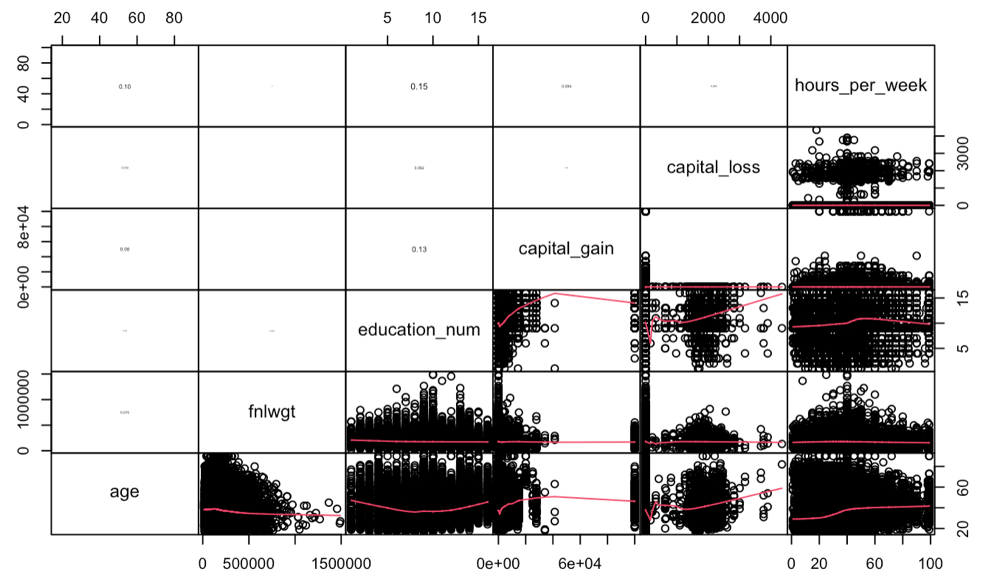
For each of the 4 listed categorical variables we found evidence for the following; Education could be partitioned into 3 groups, and Race, Marital\_Status, and relationship could be partitioned into 2 groups. We acknowledged that using ANOVA could help us determine exact differences between groups but ultimately used intuition to partition each group into different levels. Furthermore, we used intuition to divide the Hours\_per\_week variable into 3 factor levels: >40, =40, and <40. **Figure 1** below is an example of how Marital\_Status was divided, and figures for each insight can be found in **Figures 2-5** the **appendix**.



**Figure 1:** The plot on the left shows the marital status ungrouped while the plot on the right shows the marital status grouped

The last variable we observed was Native Country. This variable has 41 levels to interpret, so as an attempt to ease the interpretation of our later models, we chose to disregard this variable. An analysis could have been performed to determine which countries can be grouped together to reduce the number of levels, which would have required us to use contrasts through an ANOVA test to determine which groups are more similar. Additionally, we also could have logically created another column derived from Native Country that would convey whether the individual was native to the US or not. However, due to time constraints, we chose to simply leave this variable out of our model. A view of the histogram produced can be found in **Figure 6** in the **Appendix**.

Moving on to the continuous variables, we constructed a correlation matrix of all the continuous variables within the dataset. From this correlation matrix, we found no evidence of collinearity between the variables. The produced correlation matrix can be viewed in **Figure** **7** both below and within the **Appendix**. Once we had explored all of the relationships between the variables, we moved on to the construction of our Logistical Regression models.



**Figure 7:** Correlation matrix of continuous variables. The red lines displayed on the lower quadrants are lines of best fit while the upper quadrants are used for displaying the Correlation constant in text size relative to the magnitude.

**Objective 1**

The next step after our Exploratory Data Analysis was to investigate the assumptions for Logistic Regression. We can assume that the observations are independent as they are individual responses recorded for a single year. If the dataset spanned across multiple years, then we would need to address this violation of independence. The next assumption for us to check is whether our response variable, money, is binary or not. To avoid violations of this assumption we have verified that the entire dataset’s money column is binary with the following responses: **<=50K** and **>50K**. The last assumption for us to check is for collinearity between the independent variables. This assumption will strictly apply to continuous variables, as categorical variables cannot be colinear since there is no degree of linearity between categorical responses. For this assumption we analyzed the relationship between our continuous variables: age, fnlwgt, education\_num, capital\_gain, capital\_loss, and hours\_per\_week. Through an analysis of a correlation matrix of all the continuous variables, we found no concern for collinearity (**Figure** **7** in the **appendix**).

After verifying the assumptions, we then moved on to splitting the data into testing and training sets. While the data was initially split when it was obtained, we chose to perform our own testing and training splits. We recombined the data together and then performed an 85/15 training test split on the data. We used a predetermined seed for replication purposes that can be found within our code file. In addition to this split we also created separate training sets containing the response variable and a matrix of predictors. Further information can be found within our provided R Markdown file.

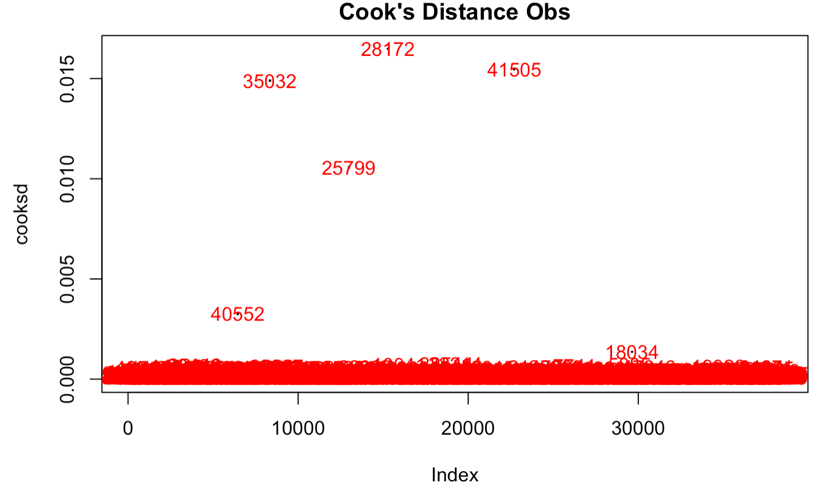
With our dataset split into testing and training sets, we were ready to begin constructing a model. Feature selection methods were used for the initial construction. We had many options to choose from for the feature selection including using LASSO, Forwards, Backwards, Stepwise, Ridge, etc. For this analysis, we decided to proceed with LASSO and Backwards, and stepwise as our main selection methods to compare.

Chart, line chart

Description automatically generated When constructing a LASSO model, we chose to use cross validation to obtain the optimal penalty value and an overall CV error rate. Using the GLMNet package within R, we were able to construct a model with an overall CV error rate of .1516 as shown below in **Figure 8**. From this model, we found the optimal penalty value of 1.675e-3 and applied this to create a model using LASSO feature selection. Within our LASSO model, we observed the coefficients for each level to determine which ones to remove from our final model. In reference to **Figure 9** within the **Appendix**, we examined the weight of each coefficient along with their p-value. Furthermore we compared the LASSO results with our Backwards and Stepwise selection models and found that all 3 models contained similar reductions. Based off this information combined with our intuition we began to construct our final model for this objective.

**Figure 8:** Plot of the misclassification error versus log(lamda). The minimum value described can be found within the provided R Markdown file.

Our final simple logistic regression model contained 6 variables in total: age, sex, educlevel, groupedHours, occupation, and groupedRelationship. The model produced an AIC of 28601 along with an overall accuracy of 81.69% (deduced from Confusion Matrix), sensitivity score of .9226, and specificity score of .7057. Additionally, a measure of Cook’s Distance on the residuals was also performed for this model. Any hints of outliers were all well below the critical Cook’s Distance with a max value of .016 as shown below in **Figure 10**.



**Figure 10:** The plot on the left displays the calculated cook’s distance for each individual point. We can see that the highest value is just above .015.

Additionally, we obtained Wald confidence intervals for each level within the model. The following is a list of interpretations of the parameters with their confidence intervals as shown in the R output provided in **Figures 11** and **12** within the **Appendix**:

* For **age**, holding all other variables constant, an increase of one year in age increases the odds of having an income over 50k by a factor of e0.029  = 1.029. (95% CI (0.0269, 0.0319))
* For **sex**, holding all other variables constant, the odds of having an income greater than 50k for males is e0.169  = 1.184 (95% CI 0.0833, 0.254) times that of having an income >50k for females.
* For **education**, holding all other variables constant, the odds of having an income >50k for HS-Bachelors is e 1.32  = 3.74 (95% CI 1.184, 1.454) times that of having an income >50k for <HS. The odds of having an income >50k for Post-Bachelors is e 2.34  = 10.38 (95% CI 2.173, 2.506) times that of having an income >50k for <HS
* For **hours worked**, holding all other variables constant, the odds of having an income >50k for 40 hours worked is e 0.783  = 2.188 (95% CI 0.684, 0.882) times that of having an income >50k for <40 hours worked. The odds of having an income >50k for >40 hours worked is e 1.309  = 3.702 (95% CI 1.208, 1.411) times that of having an income >50k for <40 hours worked.
* For **occupation**, holding all other variables constant, the odds of having an income >50k for:
  + Individuals working in the **Armed forces** is e 0.451  = 1.57 (95% CI –1.46, 2.092) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working in **Craft-repair** is e –0.281  = 0.755 (95% CI –0.406, -0.155) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working in **Exec-Management** is e 0.853  = 2.347 (95% CI 0.734, 0.973) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working in **Farming-fishing** is e –1.286  = 0.276 (95% CI –1.503, -1.074) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working as **Handlers-cleaners** is e –0.925  = 0.396 (95% CI –1.152, -0.704) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working as **Machine-op-inspect** is e –0.709  = .492 (95% CI –0.87, -0.54) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working in **Other-services** is e –1.08  = 0.34 (95% CI –1.27, -0.89) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working in **Priv-house-serv** is e –1.304  = 0.271 (95% CI –2.72, -0.31) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working in **Prof-specialty** is e 0.817  = 2.264 (95% CI 0.69, 0.94) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working in **Protective-serv** is e 0.284  = 1.328 (95% CI 0.09, 0.47) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working in **Sales** is e 0.223  = 1.25 (95% CI 0.10, 0.35) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working in **Tech-support** is e 0.601  = 1.824 (95% CI 0.42, 0.78) times that of having an income >50k for individuals working in Adm-clerical.
  + Individuals working in **Transport-moving** is e –0.476  = 0.621 (95% CI –0.64, -0.32) times that of having an income >50k for individuals working in Adm-clerical.
* For **relationships**, holding all other variables constant, the odds of having an income >50k for those that are unmarried is e –2.32  = 0.1 (95% CI –2.32, -2.17) times that of having an income >50k for those that are married.

To summarize this objective, we successfully constructed a custom Logistic regression model with reference to models created from feature selection methods. Additionally, we analyzed cooks distance measurements for each point in the dataset and found nothing noteworthy. Furthermore, we provided in depth interpretations of the resulting parameters of our final custom model along with their Wald’s confidence intervals.

**Objective 2**

Our final objective in our analysis of the 1994 US Government Census data, we chose to explore complex models for predicting whether an individual will make more or less than $50K a year. By focusing on predictions and retracting attention from interpretations of the model, we were able to put our entire effort into creating a model that would return the best accuracy statistics. As a result, we constructed two primary complex models to compare to our final model from the previous objective: one LDA model and one Random Forest model.

The first complex model we created was a Linear Discriminant Analysis (LDA) model. We theorized this would be an excellent approach to predicting Money using only the continuous variables. The results from constructing this model given only the continuous variables resulted in an overall accuracy of 78.87%. The LDA sensitivity comes in at 0.94 and the specificity at 0.70. Compared to the final model the LDA model performed slightly better in sensitivity and specificity but fell behind in terms of accuracy by ~2.8%. This could be attributed to LDA only utilizing the continuous variables while other models take advantage of continuous and categorical variables.

The next complex model we created was an improvement on our original logistic regression model. We retained the modified categories, added transformations and interactions. The transformation was added to the age category. From F**igure 13** we can see that the age category has some strong right skewness to it. To smooth out the skewness we performed a log transform on the age category. Next we added an interaction term between race and relationship **Figure 14**. This resulted in an accuracy of 83.3%, a sensitivity of 0.93, and a specificity of 0.52. The main advantage here was in the increase of AUC at a value of 0.881. We also constructed an ROC curve for this plot to explore the relationship between sensitivity and specificity within our model. From **Figure 15** displayed below, we can see a favorable curve that strays away from the diagonal comparison line. However, we acknowledge that this curve could be more favorable if it strayed further from this line. We speculate that this curve does not appear ideal because our model contains complex transformations.

Our final complex model used for predicting the money status was a random forest model. Given that the dataset used has a mix of continuous and categorical variables, we theorized that this algorithm would return predictions with the best statistics. When constructing our model, we found that the model performed much better when using the full dataset including all additional columns. When using this Full dataset within the model, our error rate decreased from 17% to 13%. The accuracy of our full model results in 86.26% while the sensitivity is 0.94 and the specificity is .80. When comparing this model to our final model you will find that the random forest model slightly makes improvements on all categories.

From our overall analysis, we have found evidence to suggest that when focusing predictions without interpretations, more complex models produce more favorable statistics when compared to our final model from **Objective 1**. Furthermore, we found a slight increase in accuracy and sensitivity when increasing the complexity of the model while the specificity slightly decreased. We also discovered that the random forest model performs the best overall with increases to all statistics all around.

**Conclusion**

From this analysis we have explored relations between the predictors and the response variable, and we have successfully constructed three models with increasing complexity to predict whether an individual made more than 50K. Furthermore, we were able to predict this response with accuracies as high as 86% and produce acceptable sensitivity and specificity scores. Interestingly enough, we found evidence to suggest that random forest models perform the best out of all available models. The data is almost 30 years old from the time of this paper (2022) however, so it is also worth noting that this data does not reflect today’s trends. In fact, after adjusting for inflation for the past 28 years, $50,000 from 1994 is worth about 99,969$. With this in mind, we cannot infer any correlation or causation in relation to today’s economy. However, we can use this analysis in tandem with any future studies to help analyze trends over time. Finally from our analysis, we recommend pursuing random forest models when predicting whether an individual makes above a certain amount of money.

**Appendix**

**Link to GitHub Containing All Related Files:**

<https://github.com/Abillelatus/MSDS-6372-Project-2>

**Table 1**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Type** | **Description** |
| **Money** | Categorical | Response on whether the individual reported making above 50k or equal to or below 50k a year |
| **Age** | Continuous | Individual’s Age |
| **Workclass** | Categorical | The work class of the individual’s occupation/job |
| **Fnlwgt** | Continuous | The weights on the CPS files are controlled to independent estimates of the civilian noninstitutional population of the US. |
| **Education** | Categorical | The highest level of education obtained by the individual |
| **Education\_num** | Continuous | Numerical representation of the highest level of education obtained by the individual |
| **Marital\_status** | Categorical | Marital status of the individual |
| **Occupation** | Categorical | Occupation/job of the individual |
| **Relationship** | Categorical | Denotes the type of relationship that the individual has reported being in |
| **Race** | Categorical | Race of the individual |
| **Sex** | Categorical | Sex of the individual |
| **Capital\_gain** | Continuous | Yearly profits earned from sales of personal assets |
| **Capital\_loss** | Continuous | Yearly losses from sales of personal assets |
| **Hours\_per\_week** | Continuous | Hours worked per week of the individual |
| **native\_country** | Categorical | The country that the individual was born in |

**Figure 2:** Plot on the left displays a bar chart of education groups that make above 50k before grouping while the plot on the right displays the effects of grouping the variable

**Chart, bar chart

Description automatically generated**

**Figure 3:** Plot on the left displays a bar chart of Racial groups that make above 50k before grouping while the plot on the right displays the effects of grouping the variable

**Chart, Teams

Description automatically generated with medium confidence**

**Figure 4:** Plot on the left displays a bar chart of Relationship Status groups that make above 50k before grouping while the plot on the right displays the effects of grouping the variable

**Chart, bar chart

Description automatically generated**

**Figure 5:** Plot of the Hours\_per\_week variable after grouping it into 3 distinct levels.

**Chart, bar chart

Description automatically generated**

**Figure 6:** Bar plot of the native country variable

**Chart

Description automatically generated**

**Figure 7:** Correlation matrix of continuous variables. The red lines displayed on the lower quadrants are lines of best fit while the upper quadrants are used for displaying the Correlation constant in text size relative to the magnitude.

A picture containing graphical user interface

Description automatically generated

**Figure 9:** R Output containing estimates from LASSO regression

A picture containing text, newspaper, receipt

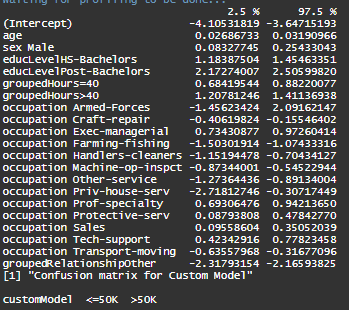
Description automatically generated

**Figure 11:** R Output containing estimates calculated for our final model

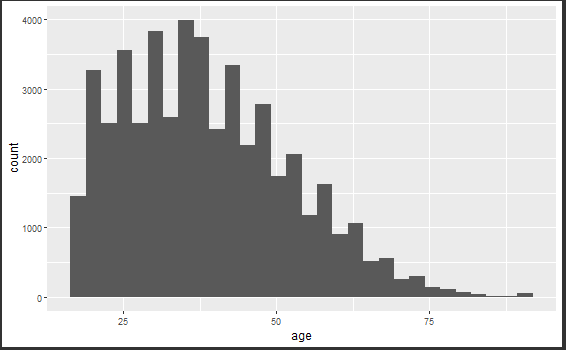
Text

Description automatically generated

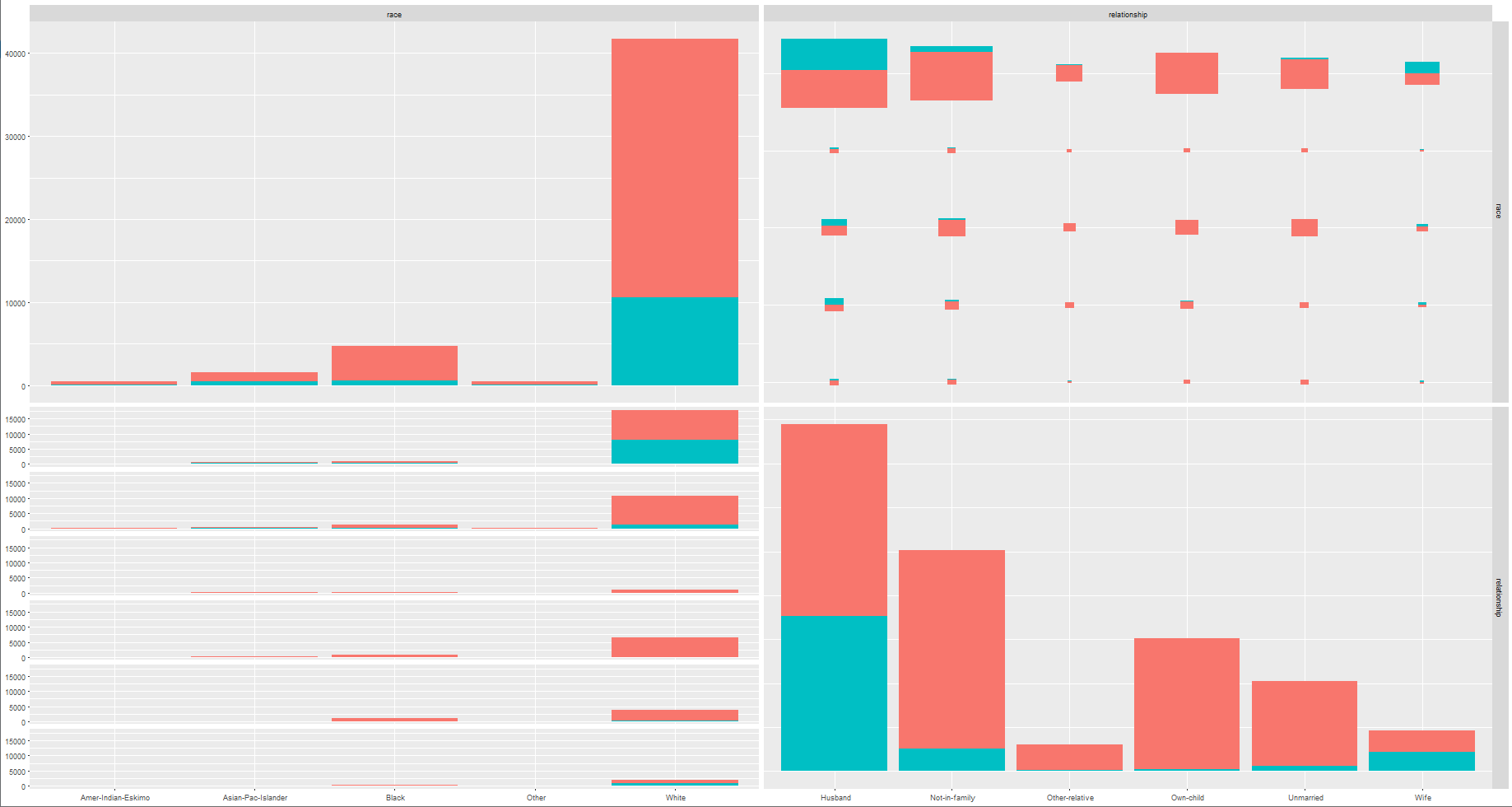
**Figure 12:** R Output containing Wald Confidence Intervals



**Figure 13:** Histogram of the age variable



**Figure 14:** R output containing the ggpairs for race vs relationship. Color coded for the money variable.



**Figure 15:** ROC curve for complex logistic regression model

