|  |
| --- |
| If real wages aren't rising, how is household income going up? |
| DS6372 Project 2 Logistic Regression Adult Income  April 10, 2022 |
| |  | | --- | | Eric Laigaie, Rayon Morris & Douglas Yip | |

**1. Introduction**

This project will focus on logistic regression where we will be analyzing the response (pay\_reponse) that indicates if an individual is making either greater or less than $50,000.

The following report will contain a detailed analysis and conclusions of the following;

* Initial (Exploratory Data Analysis) EDA
* Building a Logistic Regression Model to predict the binary pay\_response
* Comparing and compiling different regression models, where at least one contains complex variables and at least one that is non-parametric.
* Conclusion and determination of our best model that can predict a binary outcome of in an individual make greater or less than 50,000.

**2. Data description**

For this project, we downloaded training and testing sets from a online census data source. Our data exploration will mainly take place in the training set, and the testing set undergo the same transformations that the training set does.

The training data set contains 32,561 records with 16 different attributes (Table 2.1). Further changes of the data set will be addressed in our exploratory data analysis. Below is a summary of the original file.

***Table 2.1. R output of the car data set that contains the 16 different variables.***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable Name | Data Type | Levels | Min | Max | Mean | Median |
| Id | Int | 32,561 | 1 | 32,561 | 16,281 | 16,281 |
| Age | Int | 73 | 17 | 90 | 38.58 | 37 |
| Workclass | Factor | 9 | -- | -- | -- | -- |
| fnlwgt | Int | 21648 |  |  |  |  |
| Education | Factor | 16 | -- | -- | -- | -- |
| Education\_num | Factor | 16 | -- | -- | -- | -- |
| Martial\_status | Factor | 7 | -- | -- | -- | -- |
| Occupation | Factor | 15 | -- | -- | -- | -- |
| Relationship | Factor | 6 | -- | -- | -- | -- |
| Race | Factor | 5 | -- | -- | -- | -- |
| Sex | Factor | 72 | -- | -- | -- | -- |
| Capital\_gain | Int | 119 | 0 | 99999 | 1082 | 0 |
| Capital\_loss | Int | 92 | 0 | 3770 | 87.9 | 0 |
| Hours\_per\_week | Int | 94 | 1 | 99 | 40.39 | 40 |
| Native\_country | Factor | 42 | -- | -- | -- | -- |
| Pay\_Response | Factor | 2 | -- | -- | -- | -- |

Source: Appendix 2.1 and 2.2**3. Exploratory Data Analysis**

**Missing Values**

Identifying missing values was necessary to obtain accurate summary statistics. We first evaluated whether there were any NA variables or blanks within both data sets. Based on our results, there were no NA. However, based on the summary output, we noticed that workclass, occupation and native\_country contained “?” values. After checking for “?” counts in each of these variables, we found that these ‘missing’ values were contained to less than 2000 records (Appendix 3.1). Given that we have a train dataset of 32,561 and that there was no logical method to impute the data, in our analysis, *we removed all ‘?’ rows from both train and test data to complete this study.*

**Unbalanced datasets**

As a result of this study being a logistic regression, we checked the response value to see if we had a balanced dataset for both training/test data sets. Based on our count (Appendix 3.3), we identified approximately 25% of the results showing a pay\_response greater than $50,000. This would suggest that we are dealing with an unbalanced data set. As such, *we will identify the optimal cut off to maximize accuracy in our logistic regression model and prediction*.

It is also important to note that there is no difference in severity between a false positive and false negative in this case. Therefore, overall accuracy may provide a better picture of model performance than specificity or sensitivity alone.

**Continuous variables collinearity check**

The correlation grid (Appendix 3.4) for the continuous variables provides no evidence that any of the variables are correlated. *No action was taken to the continuous variables as a result of the correlation grid*.

**Effects of continuous variables on pay response**

The cluster heat map (Appendix 3.5) to response was evaluated and we see an effect of capital gain and capital loss to the response variables. All other variables were difficult to determine to see if there was any separation of the response. *No action was taken to the continuous variable as a result of the heat map.*

**Categorical Variable Exploration**

Using a bar chart of pay\_response proportions by factor level, we examined each categorical variable to determine if level consolidation could take place. Below you will find a short breakdown of each variable we examined along with the actions we took after observation.

***Education (Appendix 3.6-3.7)***

This variable initially had 16 levels, but the bar chart in appendix 3.6 indicates that many of these education levels display similar pay\_response proportions. Therefore, *we grouped education based on the groups below, reducing this variable to seven levels.*

* Preschool
* Grade School (grade 1-12 of original data)
* HS Grads (HS Grads + Some college of original data)
* Assocs (Assoc-voc + Assoc-acdm of original data)
* Bachelors
* Masters
* Docs/Profs (Prof-school + Doctorate of original data)

***Workclass (Appendix 3.8-3.9)***

The pay\_response for the government classified workers was similar in both data sets and *we grouped government (Local, State and Federal) workclass to reduce the levels from 7 to 5 level*s.

***Occupation (Appendix 3.10)***

A bar chart of pay\_response was performed on the occupation variable as we identified 15 levels in the data. While we could understand grouping education because the difference between 10th and 11th grade is negligible, occupations are much more different. Therefore, *we left the variable* as is.

***Marital\_Status (Appendix 3.11-3.12)***

After observing the initial distributions, we noticed the pay\_response proportions for the married and formerly married responses were similar in both data sets. Therefore, *we grouped and created “married” and “single was married” marital statuses to reduce the levels from 7 to 4.*

***Native\_Country (Appendix 3.13-3.14)***

This variable had many levels that did not seem to follow an exact pattern. In an effort to find some sort of logical grouping, we *grouped countries into the continents and reduced the levels from 42 to 7 levels.* One important distinction is the ‘South’ level. Since this country could be South Korea or South Africa, we decided to leave it as its own level.

**Redundancy in education and education \_num**

A box plot graph was made between the variable education and education\_num. As expected, we observed that these metrics were perfect matches and were therefore redundant. (Appendix 3.15) *Education is best viewed through factors (there isn't a numerical relationship in education levels), so we kept education in our final dataset.*

**Capital Gain and Loss**

After observing the variables capital\_gain and capital\_loss, we noticed that one value was always zero. Therefore, we decided to *consolidate these using the formula capital\_gain – capital\_loss.* This resulted in a new column, capital\_net.

**Mosaic plots to check multicollinearity for categorical variable**

Mosaic plots were made to check for multicollinearity for categorical variables. Four of the graphs (appendix 3.16) workclass & education; marital status & education; race & education; and race & marital status exhibit little to no evidence of correlation. However, relationship & marital status (appendix 3.17) should strong evidence of correlation. As a result, *the relationship column was removed in favor of marital status*.

**Removal of categorical ID and fnlgwt variables**

Two variables were removed from dataset prior to any modeling, as the variables provided do not support the prediction of income based on our research and knowledge. ‘ID’ is an arbitrary identifier column and ‘fnlwgt’ is used by the US Census to signify how many people are represented by that record. *We removed the two variables from the data*.

**4. Objective 1**

Build a logistic regression model to complete the analysis; 1) hypothesis test to see if we have any significant variables that could predict an individual having income greater or less than $50,000 2) Determine the model and variables used for the analysis, which include the interpretation of each variable and confidence intervals for each parameter of the model.

**Model Selection Methodology**

Logistic regression was selected for our model given that the response is a binary variable of <$50K and =>$50K. After our EDA, a total of 9 variables were used in stepwise logistic regression model. The following explanatory variables exists in the model; age, workclass, education, marital\_status, occupation, race, sex, hours\_per\_week, native\_country, and capital\_net. Additionally, a check for influential points was completed with the cooksd() function. We found that no points qualified as an influential outlier, so our full dataset was fine to put into the model.

The resulting model provided a large number of predictors with varying levels of statistical significance. With more than 20 predictors, interpretability was in danger. A majority of the least-significant predictors came from the occupation variable, so we decided to run the stepwise process again without it. After this new model was created, we noticed that many insignificant predictors were weeded out without sacrificing a practically significant amount of accuracy (Table 4.1).

**Table 4.1 Interpretation of Parameters for Final Simple Model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **# of betas at P=<0.01** | **Accuracy** | **Specificity** | **Sensitivity** |
| ***All predictor*** | ***20*** | ***83.9%*** | ***92.5%*** | ***57.6%*** |
| ***Occupation Removed*** | ***9*** | ***83.1%*** | ***92.9%*** | ***53.2%*** |

*Source: Appendix (4.1 and 4.2)*

**Test for Fit**

Given the training sample for our train is greater than +30,000 observation, we did not perform the Hosmer-Lemeshow tests to determine goodness of fit as it is not robust for large datasets. The combination of the training accuracy and the cross validation with the test set and the plot for the ROC curve (Appendix 4.3), we determine that our model is a good fit. In addition, a sensitivity of the cutoff was reviewed (Appendix 4.4) and we determine that the default 0.5 was a suitable cutoff to use to maximize our model’s accuracy.

**Final Model**

|  |
| --- |
| **Final Simple Model** |
|  |

**Table 4.1 Interpretation of Parameters for Final Simple Model**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Odds Ratio** | | |  |
| **Variable (P<0.01)** | **Parameter** | **Value** | **Odds** | **CI 2.5%** | **CI**  **97.5%** | **Interpretation with CI** |
| Intercept |  | -24.43 | 0.00 | 0.00 | 0.00 | Odds of a person pay >50k are even when all is zero and after accounting for all categorical variables. |
| Age |  | 0.0268 | 1.027 | 1.024 | 1.03 | The odds of a person having >50k is 1.027 times higher than a person 1 year younger holding all other variables fixed. 95% CI (1.024,1.031). |
| WorkClass Self-emp-inc |  | 0.3395 | 1.404 | 1.175 | 1.679 | The odds of Self-employed with income workers having >50k is 1.404 times more than government worker holding all other variables fixed. 95% CI (1.17,1.68) |
| WorkClass Self-emp-not-inc |  | -0.621 | 0.537 | 0.467 | 0.618 | The odds of Self-employed with no income workers having >50k is 1.89 (1/0.5237) times less than a government worker holding all other variables fixed. 95% CI (1.61,2.14) |
| Marital Status Married |  | 2.136 | 8.461 | 7.533 | 9.505 | The odds of Married individual having >50k is 8.46 times more than a previously married individual holding all other variables fixed. 95% CI (7.53,9.505) |
| Marital Status  Never Married |  | -0.531 | 0.587 | 0.506 | 0.682 | The odds of Never married individual having >50k is 1.7 times less than a previously married individual holding all other variables fixed. 95% CI (1.97,1.47) |
| Race White |  | 0.0642 | 1.900 | 1.240 | 2.912 | The odds of White individual having >50k is 2.0 times more than Indian American individual holding all other variables fixed. 95% CI (1.24,2.912) |
| Sex Male |  | 0.0135 | 1.144 | 1.039 | 1.261 | The odds of males having >50k is 1.1 times more than females holding all other variables fixed. 95% CI (1.039,1.261) |
| Hours per week |  | 0.03 | 1.030 | 1.027 | 1.033 | The odds of a person having >50k is 1.030 times higher than a person working an additional hour holding all other variables fixed. (95% CI (1.027,1.033) |
| Capital Net |  | 0.0003 | 1.000 | 1.000 | 1.000 | The odds of a person having >50k is even to a person making $1 younger holding all other variables fixed. (95% CI (1.000,1.0000)) |

**Conclusion**

Unfortunately, this model, while interpretable, does not provide a high level of accuracy (90%+). Although, some important relationships can be pulled out of this model. We found that white males are much more likely to classify as >50K as opposed to other racial and gender demographics. Additionally, marriage plays a large part in this classification model. Therefore, a new question is raised: in this data, is the pay\_response variable representing the income of one person, or their entire household?

**5. Objective 2**

**Problem Statement**

In objective 1, our goal was to create a simple, highly interpretable model. For this objective, we will be using that model as a baseline and use various methods in an attempt to improve model performance. First, we well keep the variables produced in the stepwise model but add complexity through interaction terms. Then, we will use LDA and QDA to see if performs better that our 2 models. Lastly, these models will be contrasted with the stepwise model to compare performance and determine the most optimal solution.

**Complex Model**

Our method for creating this model was to run through many prospective models with varying terms and viewing how these interactions turned out in the model summary. If we found certain pairings of variables were often insignificant, we moved away from them and explored new interactions. To better understand any roadblocks for this model, we used the cooksd() function to provide us with any outlying and highly influential points. Similar to objective 1, we did not find any points returned here.

After reaching a final model, we investigated every cutoff value between 10 and 90 percent with an interval of .1%. Our code found that a classification cutoff of 48% provided the maximum accuracy of 82.71%. In Appendices 5.1-5.4, you can find this model’s summary, confusion matrix, accuracy vs. cutoff chart, and roc curve.

**LDA/QDA**

To give our LDA a chance to compete against the other models, we will start with the original data set to see how well LDA performs. The assumptions for LDA are Normality, and equal covariance matrices. If the equal covariance is not valid, we will need to utilize QDA. For our initial LDA model, we see that it fails the normality assumption and the equal covariance. Before proceeding with QDA, we log transformed the explanatory variables of the LDA model. Here we saw better separation of the data. However, the equal covariance assumption was still violated. For that reason, we ran a QDA model (See Appendix 6.1 -6.8). Our conclusion is that the LDA, though it failed its assumptions, gave a better accuracy that the QDA model. This is because the LDA model is robust to the equal covariance assumption.

**Random Forest**

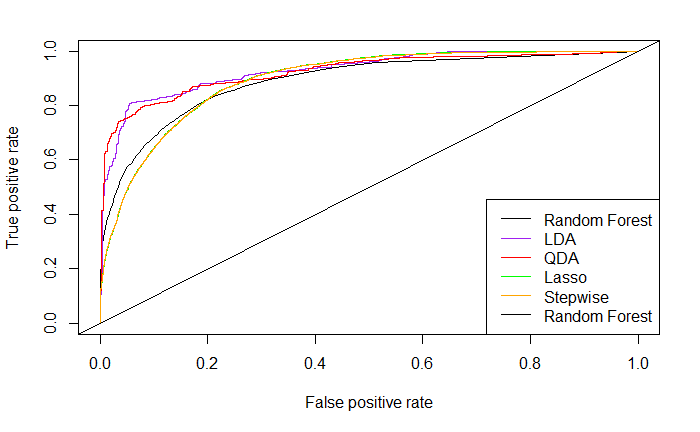
The random forest is a classification algorithm consisting of many decisions trees. The model builds decision trees on different samples and takes their majority vote for classification and average in case of regression. A MTRY model was run to identify the optimal MTRY to use for the Random Forest Model. We identified 9 to optimize accuracy of our model (Appendix 6.9). Utilizing Random Forest model and the optimal mtyr of 9, the achieved predication accuracy for the test set was 84.06%. (Appendix 6.10)

**Model Results**

Table 5.1 Summary of statics for the models predicting test and validation data.

|  |  |  |  |
| --- | --- | --- | --- |
| Predictive Models Test Statistics | Accuracy | Sensitivity | Specificity |
| Objective 1 Model | 83.12% | 92.88% | 53.16% |
| Complex Model | 82.71% | 90.87% | 57.65% |
| LDA | 83.87% | 74.15% | 89.74% |
| QDA | 81.87% | 89.87% | 68.64% |
| Random Forest | 84.06% | 92.29% | 62.67% |

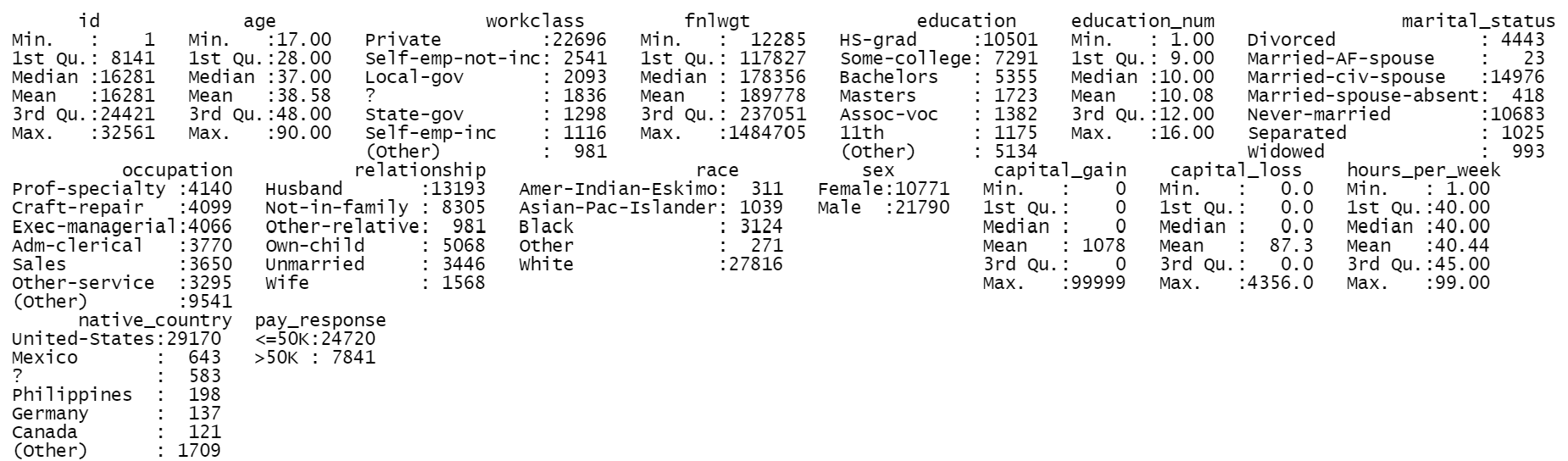
**6. Final Summary**

Through both objectives, we built a menu of models to be evaluated both alone and together. Although we attempted to build high-accuracy models, our results led us to conclude that this dataset leaves a lot to be desired (Figure 5.1). With only three numeric features, the regression models we built have struggled to achieve a strong accuracy. Our marginally best model, Random Forest model, achieved a 84.06% test accuracy. However, this research and model development was beneficial because it can be used to point future modelers towards success. A few suggestions for next steps include deeper exploration into related datasets that could bolster the census data and provide more continuous variables to use as predictors. Additionally, more complicated models such as k-nearest neighbors and neural networks could be used in an effort to drive up complexity and, hopefully, accuracy.

**Figure 5.1 Summary of ROC plots**

**7. Appendix**

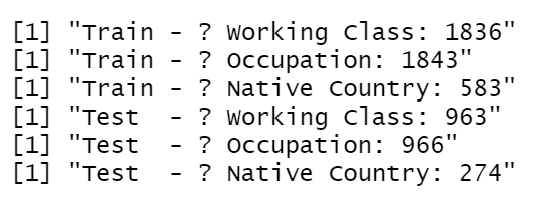
**Appendix 2.1 – Initial Summary Statistics**

****

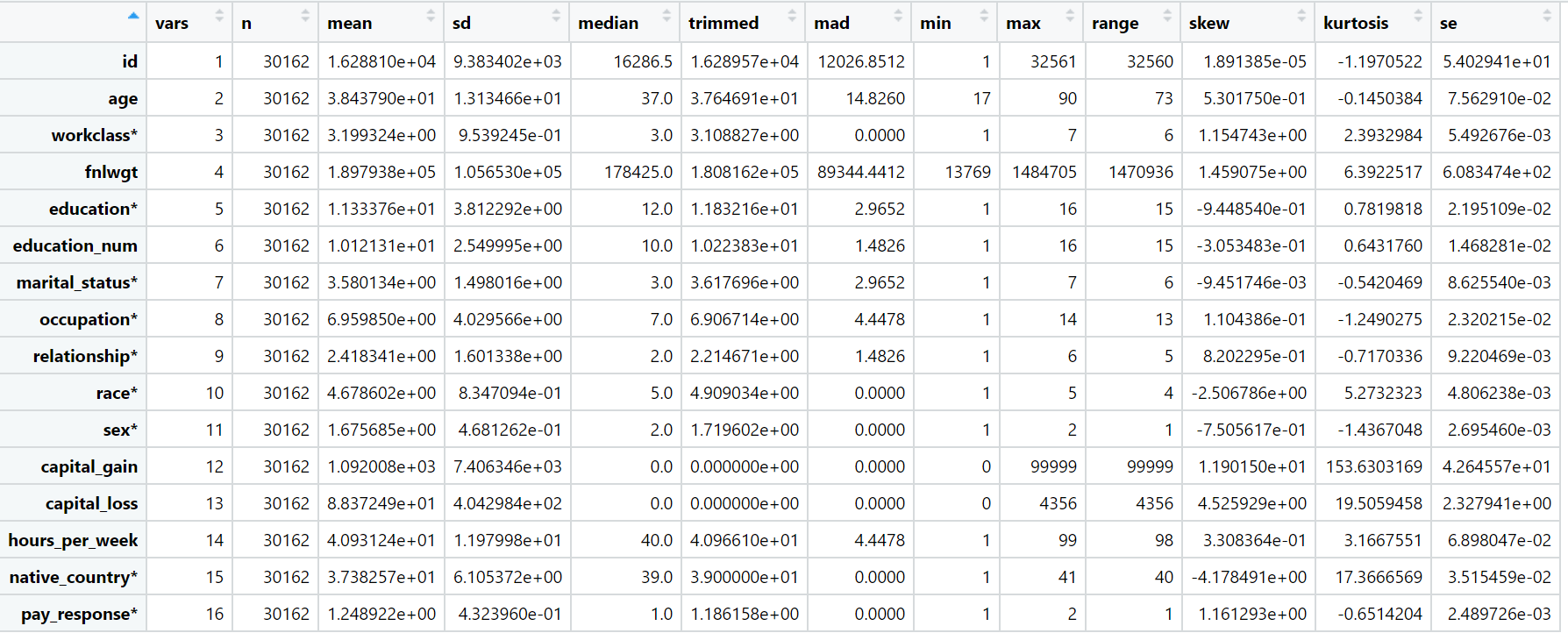
**Appendix 2.2 – Number of levels by Variable**

****

**Appendix 3.1 R output of number of unknown variables coded “?”**

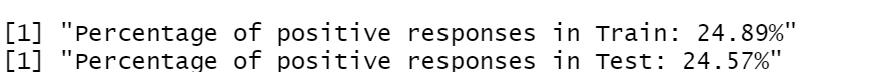
****

**Appendix 3.2 Summary of data set after removal of “?”**

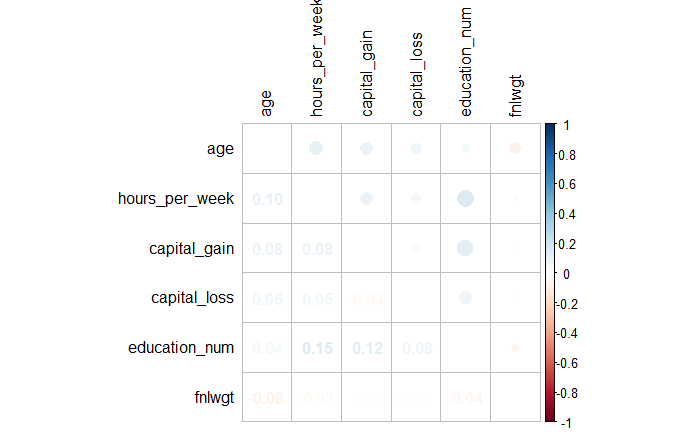
****

**Appendix 3.3 R output of Unbalanced response for both the train and test data sets**

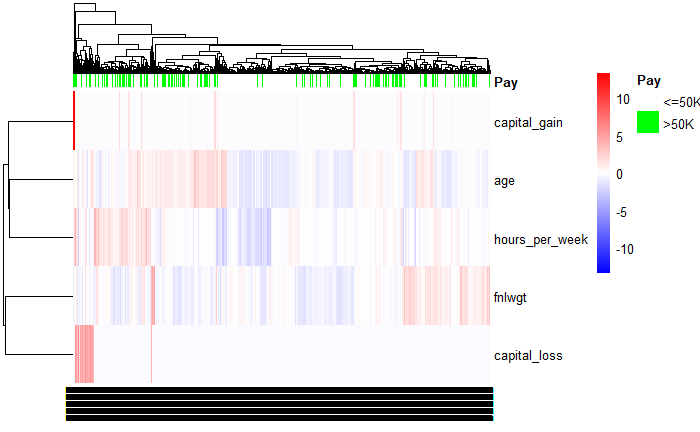
****



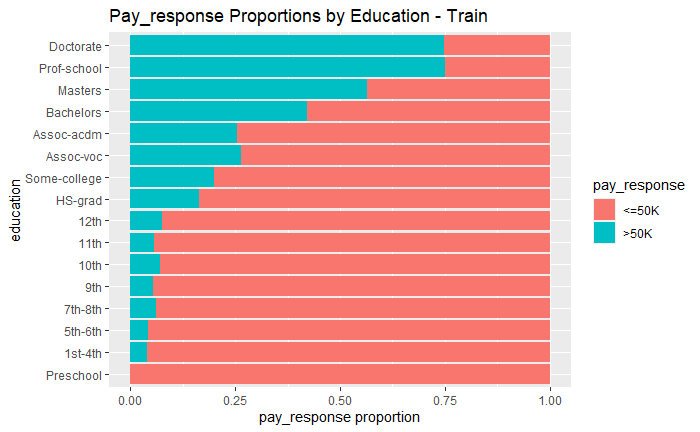
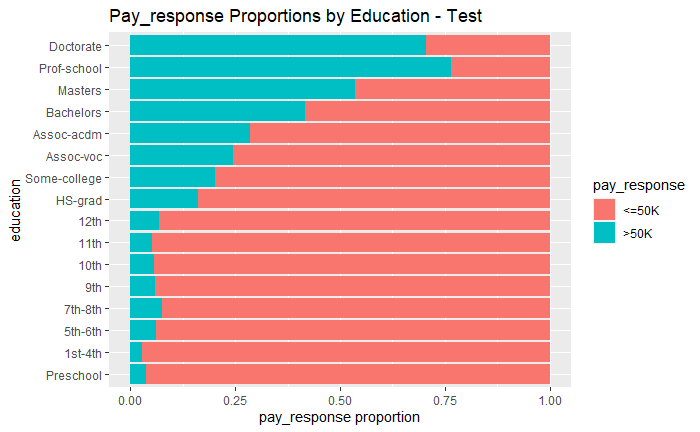
**Appendix 3.4 Correlation plot of continous variables.**



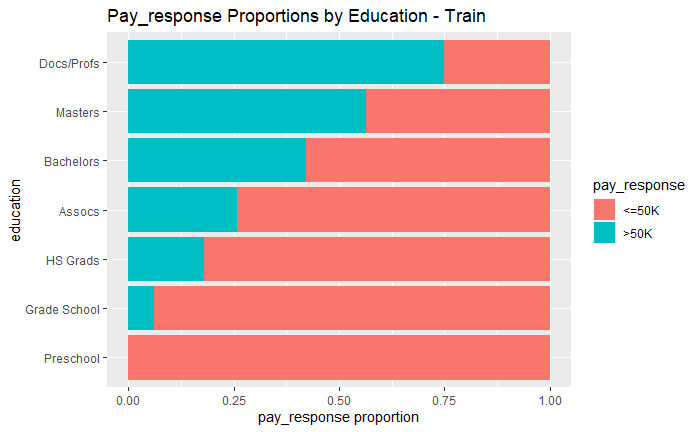
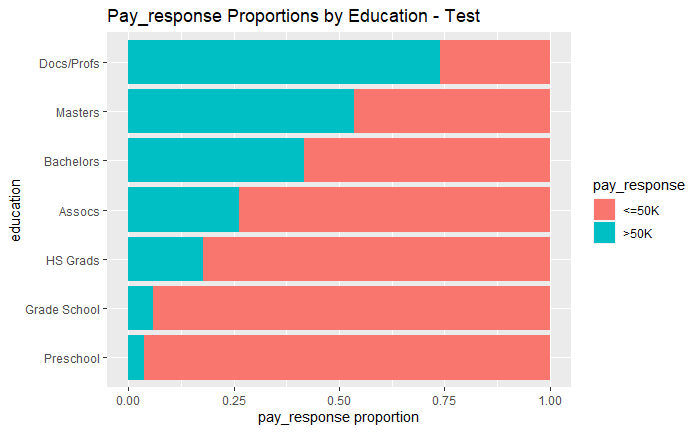
**Appendix 3.5 Cluster Heat map to determine the impact to pay response**



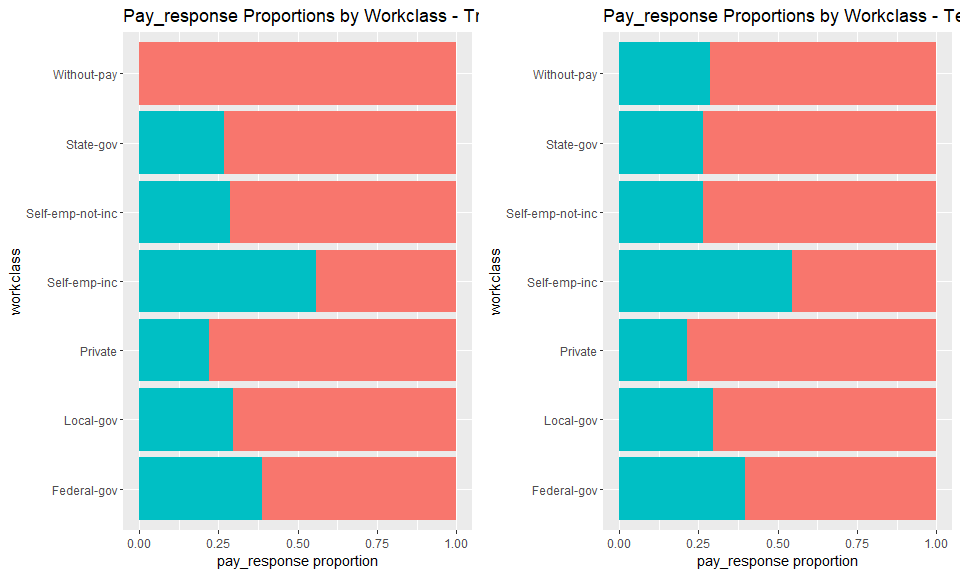
**Appendix 3.6 Pay response for Education pre transformation (Training and Test Sets)**

**Appendix 3.7 Pay response for Education post transformation (Training and Test Sets)**

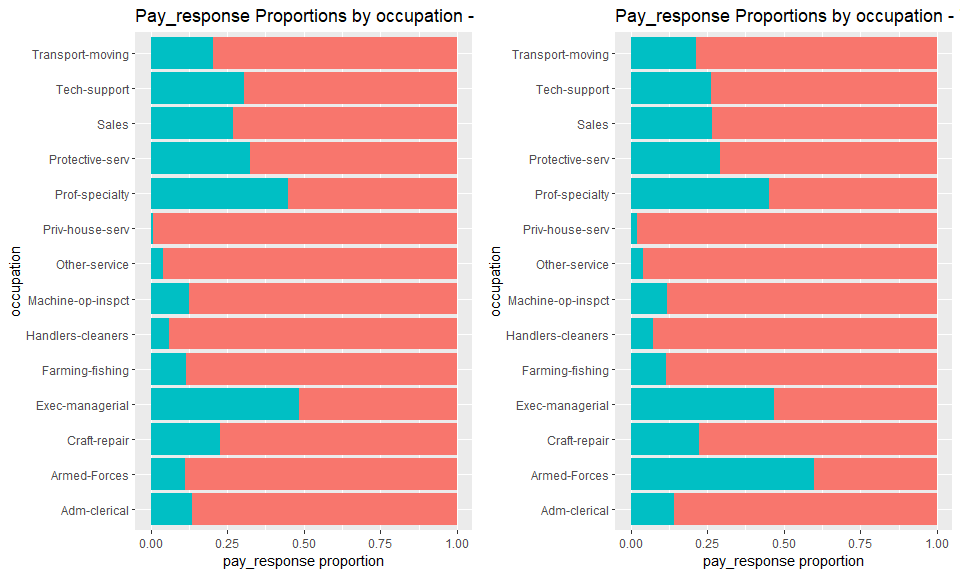
**Appendix 3.8 Pay response for Workclass pre transformation (Training and Test Sets)**



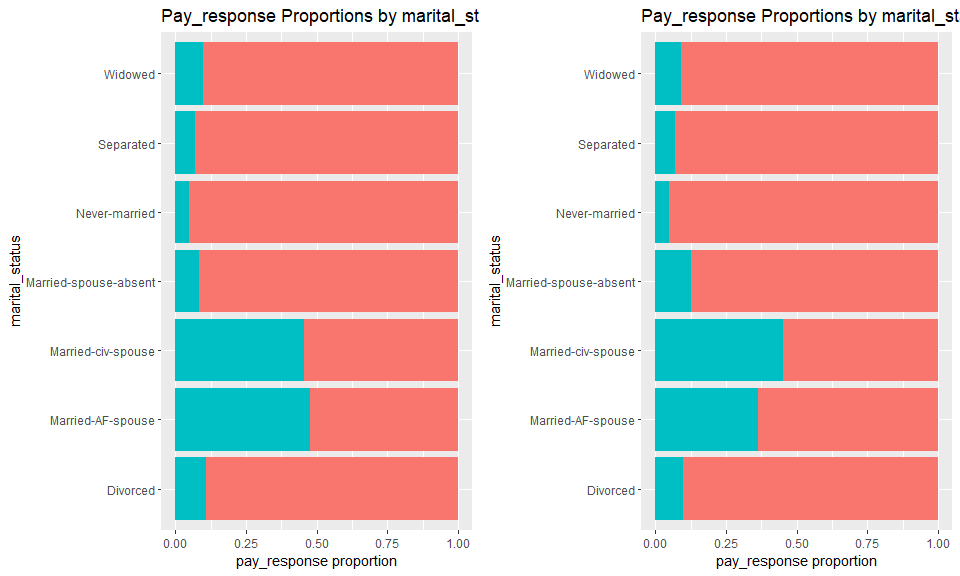
**Appendix 3.9 Pay response for Workclass post transformation (Training and Test Sets)**



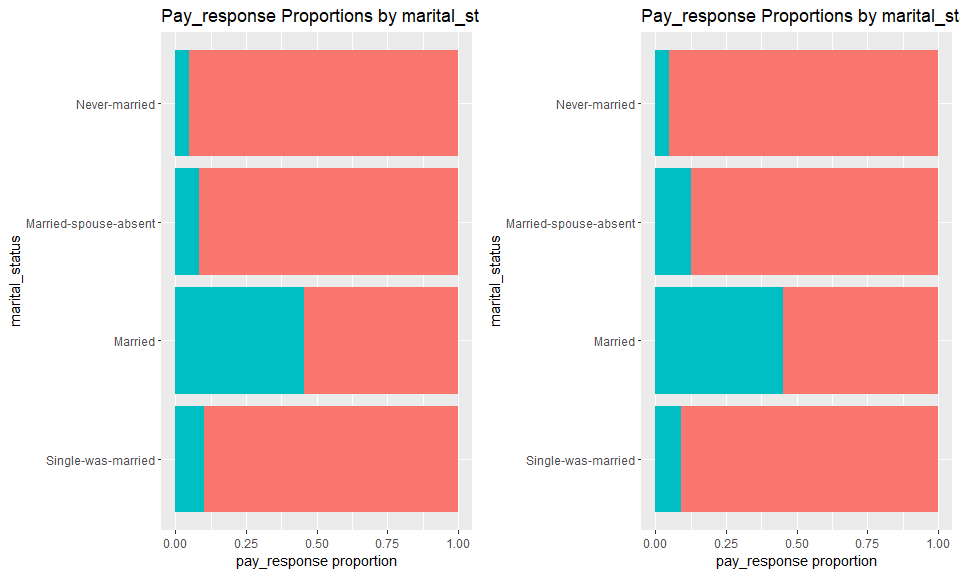
**Appendix 3.10 Pay response for Occupation**



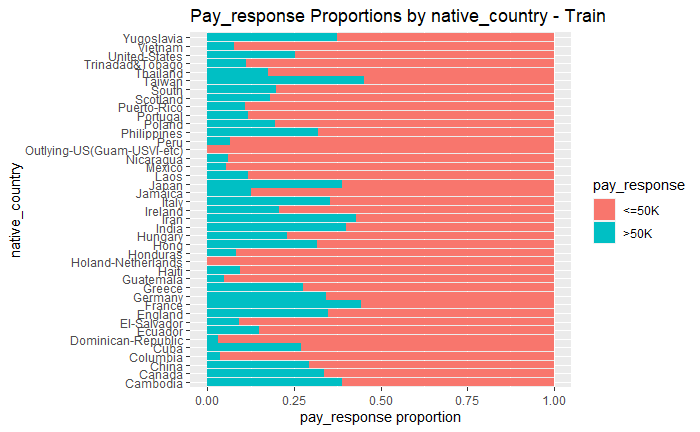
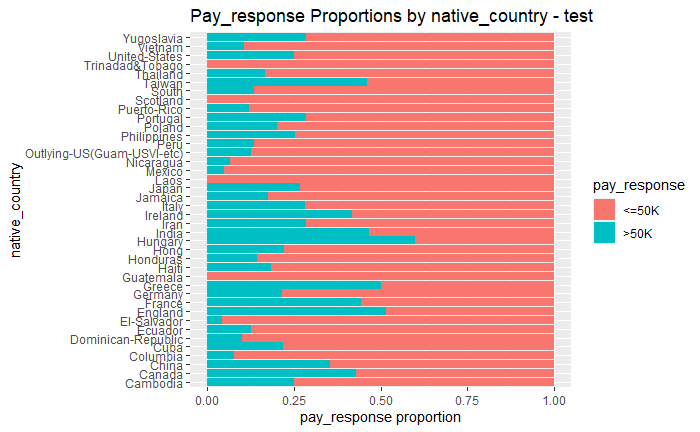
**Appendix 3.11 Pay response for Marital status pre transformation (Training and Test Sets)**



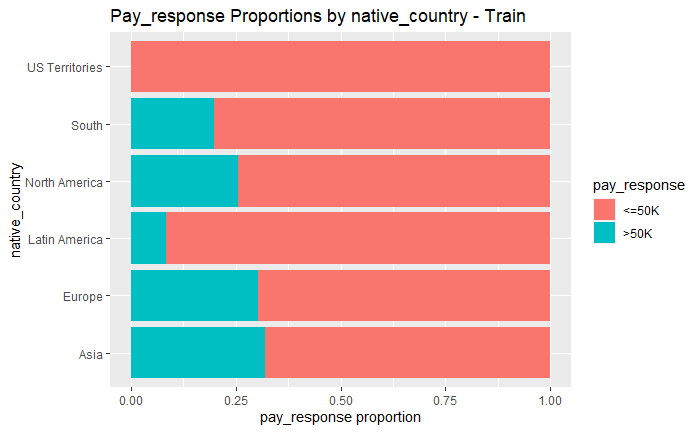
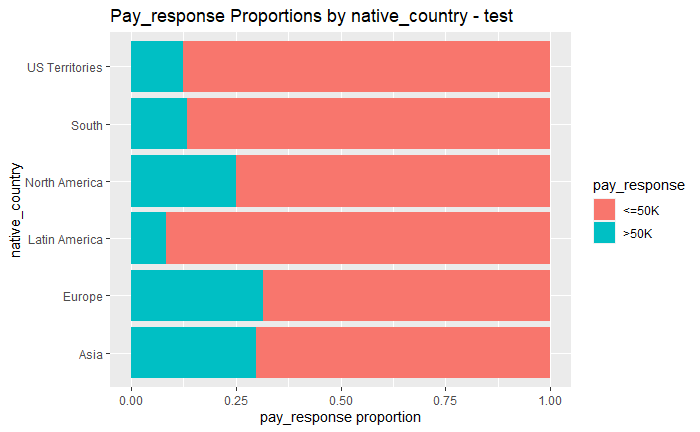
**Appendix 3.12 Pay response for Marital status post transformation (Training and Test Sets)**



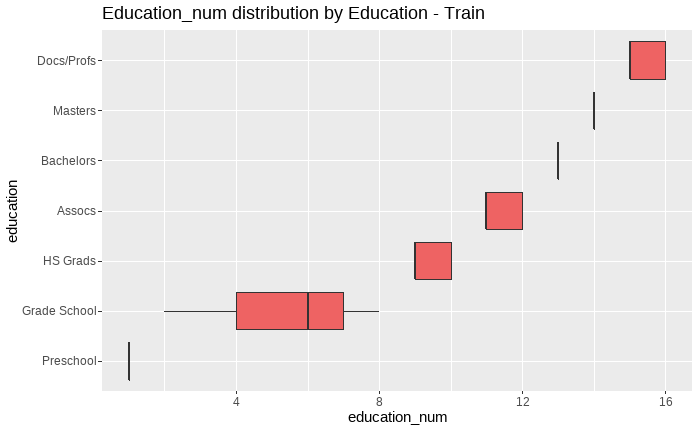
**Appendix 3.13 Pay response for Native Country pre transformation (Training and Test Sets)**

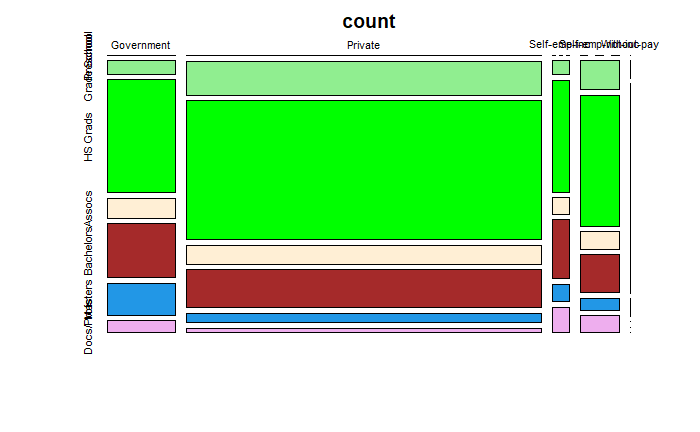
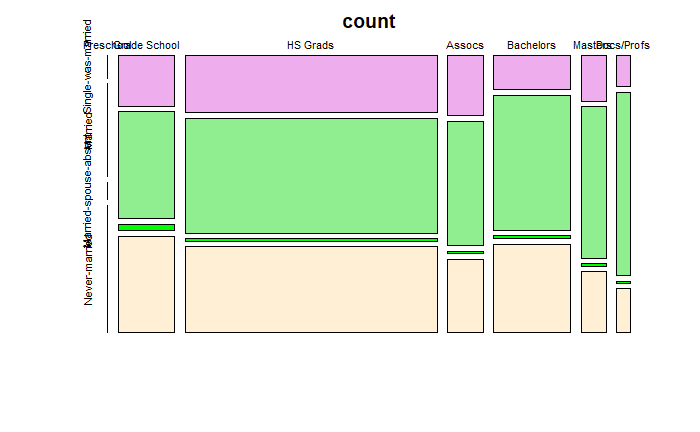
**Appendix 3.14 Pay response for Native Country post transformation (Training and Test Sets)**

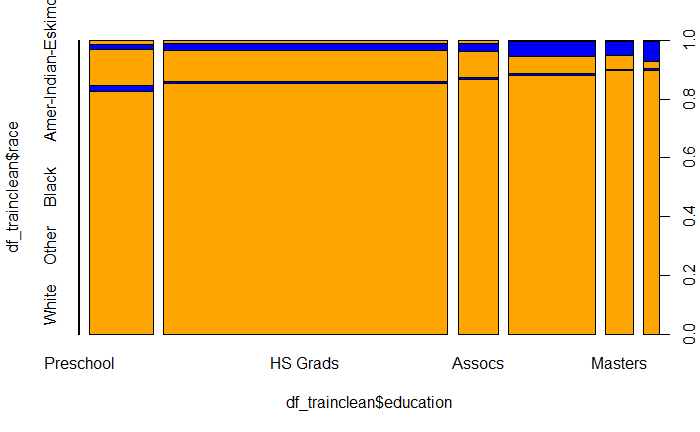
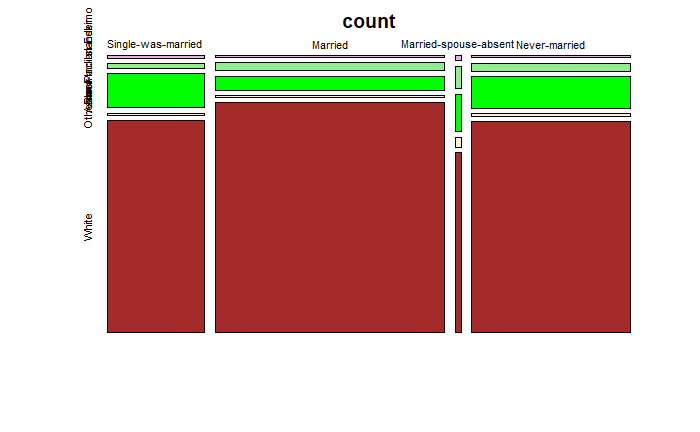
 

**Appendix 3.15 Box plot graphs of the redundancy of education and education\_num variables**

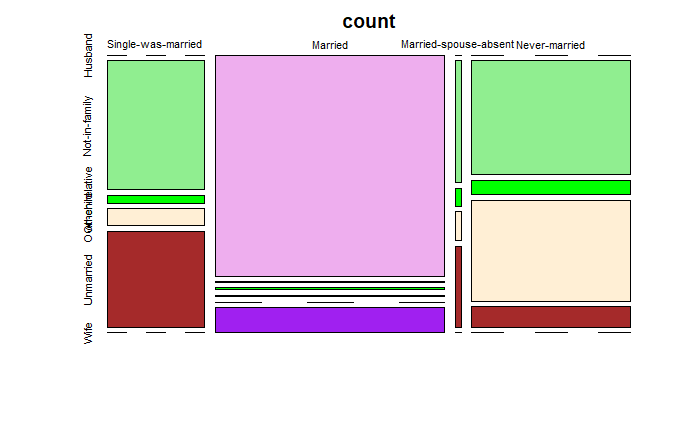


**Appendix 3.16 Mosaic plots to check for collinearity among categorial variables (minimal to no evidence)**

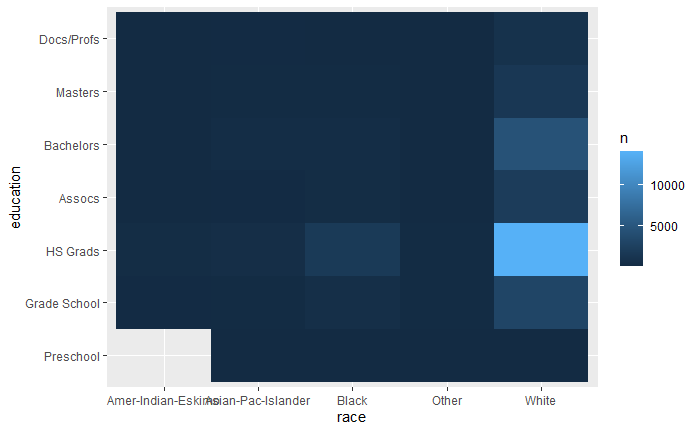
 

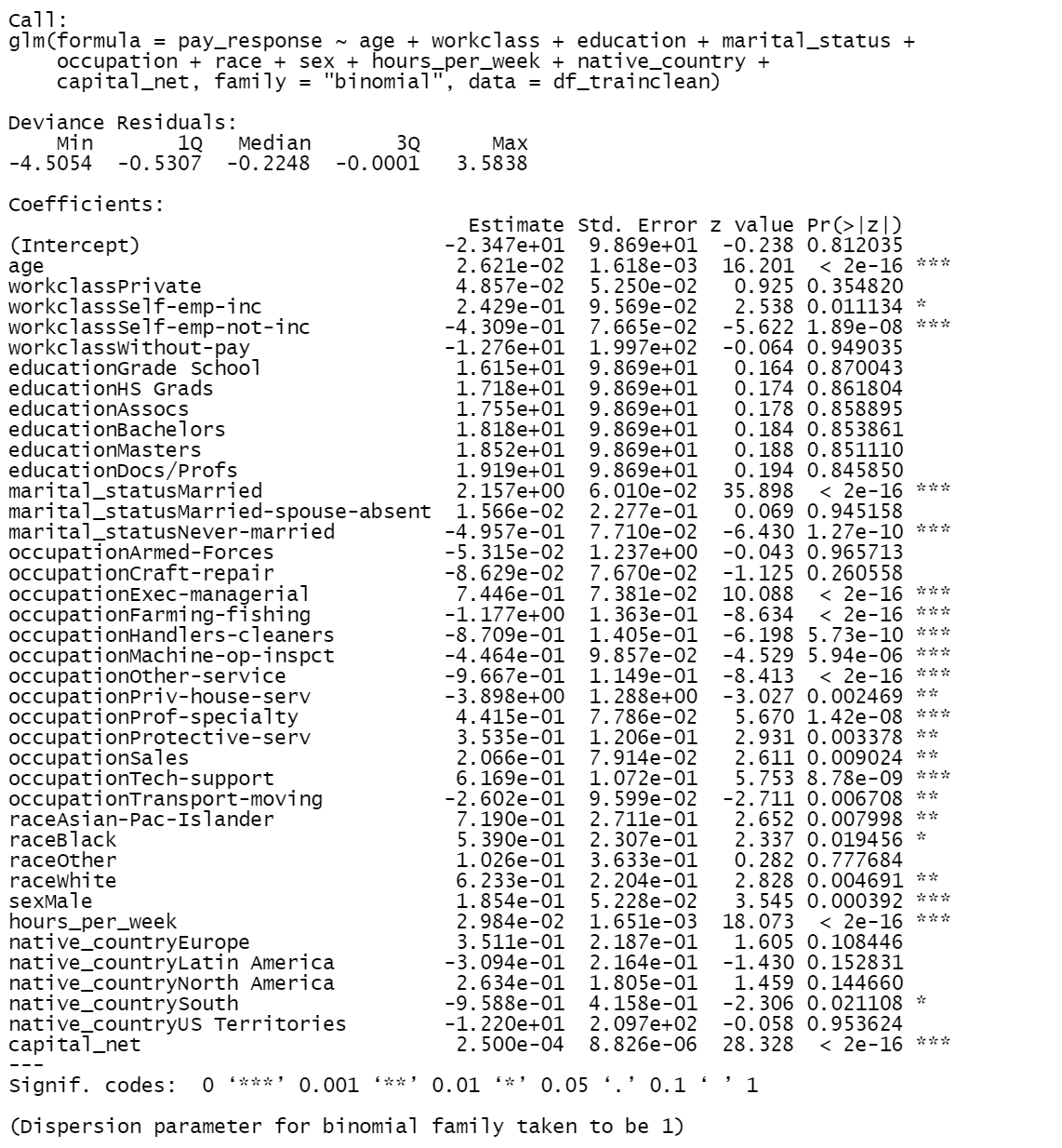
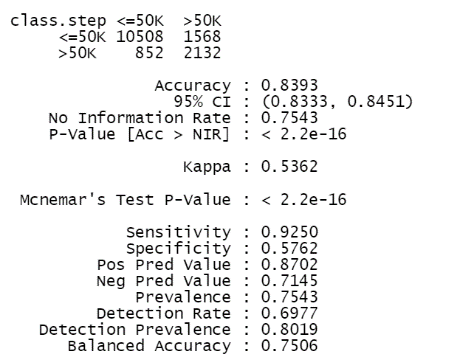
**Appendix 3.17 Mosaic plots to check for collinearity among categorial variables (Strong evidence)**



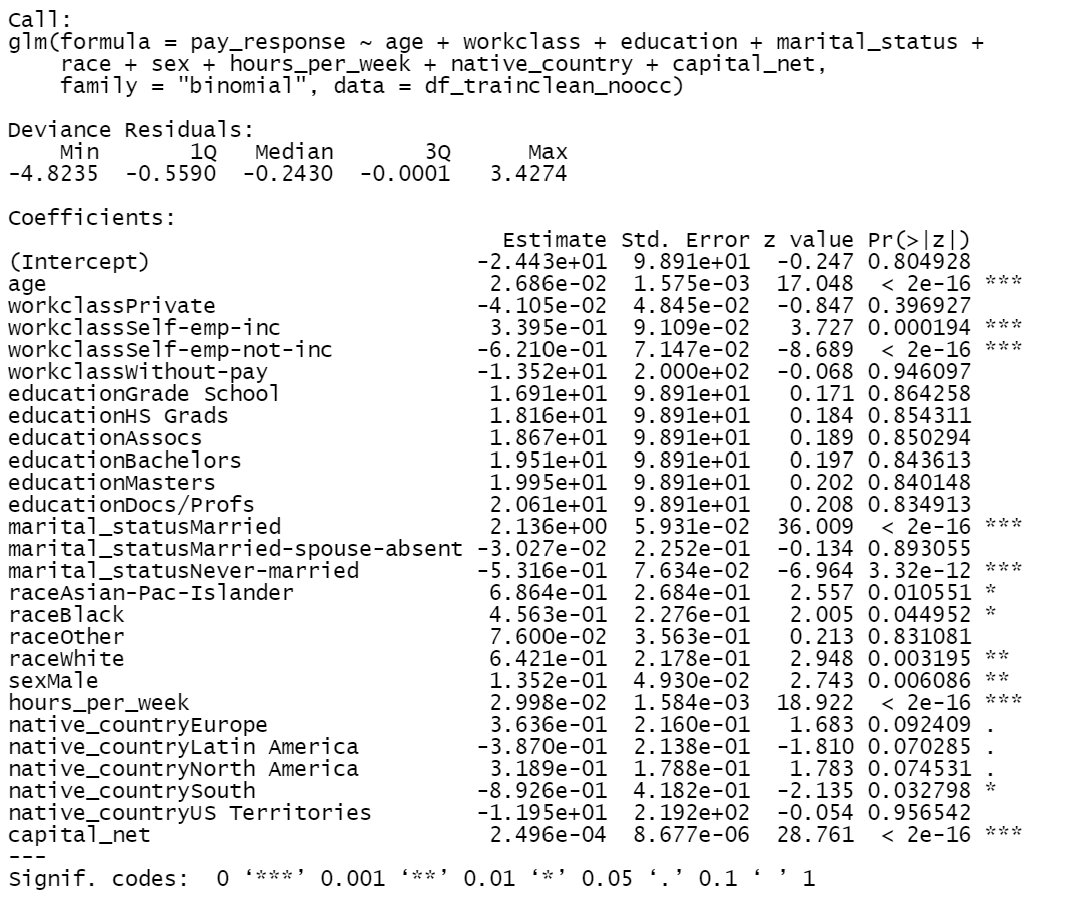
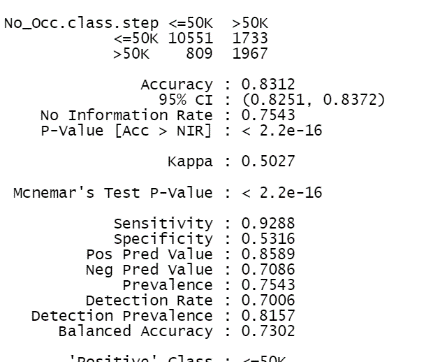
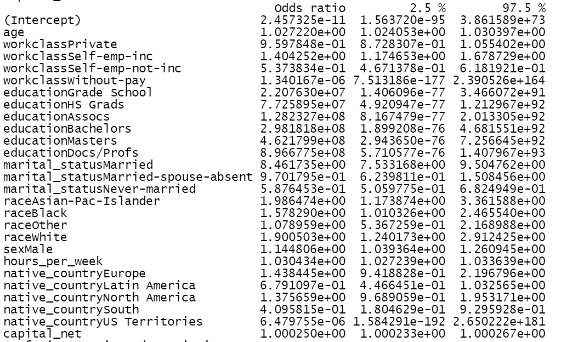
**Appendix 3.18 Race by education heat map count**



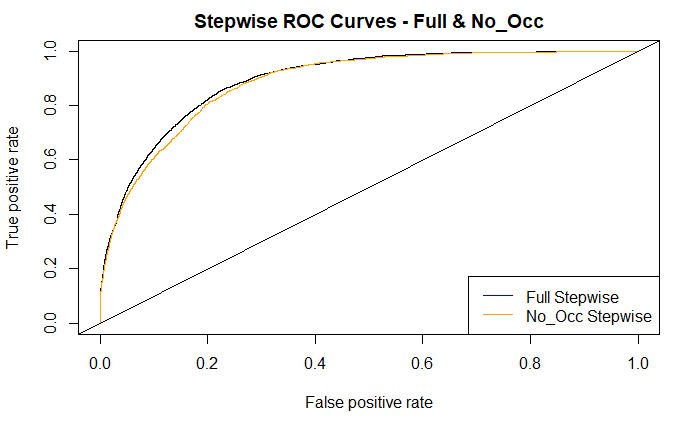
**Appendix 4.1 Step Wise Logistic Regression Output**

****

**Appendix 4.2 Step Wise Logistic Regression Output without occupation**

****

**Appendix 4.3 ROC for both stepwise test (one with occupation and one without)**



**Appendix 4.4 Cut off value of stepwise model excluding occupation**



**Appendix 5.1 Interaction Regression Output**

**Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated**

**Appendix 5.2 Interaction Regression Confusion Matrix**

Graphical user interface, text

Description automatically generated

**Appendix 5.3 Interaction Regression Accuracy / Cutoff Chart**

Chart, line chart

Description automatically generated

**Appendix 5.4 Interaction Regression ROC Curve**

Chart, scatter chart

Description automatically generated

**Appendix 6.1 shows the qqplot of the distribution with a response variable >50K.**

Graphical user interface, line chart

Description automatically generated with medium confidence

**Appendix 6.2 shows the qqplot of the distribution with a response variable <=50K.**

Chart, line chart

Description automatically generated

**Appendix 6.3 shows the scatter plot of the data indicating very little separation.**

**Chart

Description automatically generated**

**Appendix 6.4 shows the qqplot of the distribution with a response variable >50K after performing a log transformation.**

Graphical user interface, chart, line chart

Description automatically generated

**Appendix 6.5** shows the scatter plot of the data with better separation. However, the equal covariance assumption is violated.

Chart, scatter chart

Description automatically generated

**Appendix 6.6 shows the ROC plot of the LDA, QDA, and Stepwise models**

Chart, line chart

Description automatically generated

**Appendix 6.7 shows the output of the LDA model**

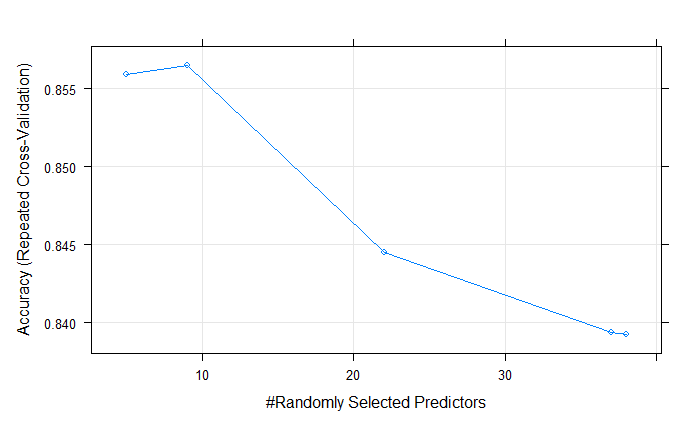
A screenshot of a computer

Description automatically generated with low confidence

**Appendix 6.8 shows the output of the QDA model**

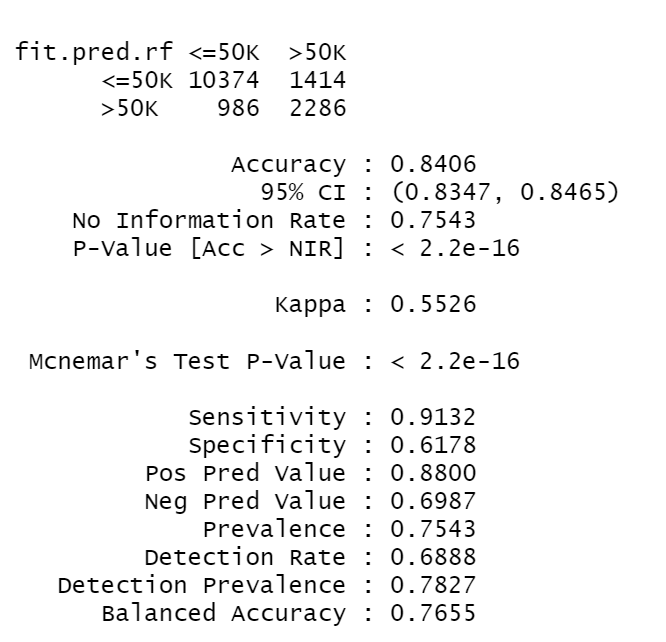
A picture containing text, receipt

Description automatically generated

**Appendix 6.9 MTRY generator output and Graph**

****

**Appendix 6.10 Random Forest Confusion Matrix for test results**

****

**R-Mark Down Code**

**To be added when code is finished**