Final Report

Group 2 (Ronghua Ni, Arushi Shah, Clarence Chen, Pranav Ravichandran)

Part 1: Executive Summary for Regression Question

(a)

The regression question of interest is how does age, work experience, education level, credit to debt ratio, and other debts of the customer change the yearly income.

(b)

The question is significant because it allows for prediction of a customer's income using their bank and financial data, which is available to a lot of banks. This prediction is significant because it allows banks to be able to use that to in turn evaluate the risk they are taking with a customer of whether they are going to default or not, allowing banks to make more informed decisions on loans. So, using regression methods to predict income levels, banks are the relevant stakeholders who can assess the risk on a specific customer.

(c)

The analyses carried out help answer this question of interest because they include various methods to see if we can predict the yearly income using the other factors and with what accuracy we can predict. Using the different models is useful because we can compare the results of the different models and see what model yields the best prediction accuracy. We first dove into the question of interest itself to derive that regression would be useful by seeing what variables we have. We then built different regression models to get an accurate prediction for yearly income. It included an exploratory data analysis to show the variables relationships with each other and yearly income to assure that the variables are proper predictors of yearly income. Then, there were shrinkage methods, regression trees, and random forests models created to predict the yearly income. The various models had different benefits such as being more specific versus reducing overfitting, but in the end, the best model based on error rate was the random forests model.

(d)

Using these analyses, I would suggest banks to really pay attention to a customer's other debts and years of experience working to get a scope of a customer's income and how it may reflect on their likelihood of defaulting.

Part 2: Data and Variable Description for Regression

(a)

The dataset is about various financial metrics and personal information and characteristics for customers of an unnamed leading bank. For each customer, work experience, age, yearly income, credit to debt ratio, other debts, and education level are included in the data set.

(b)

We obtained our dataset from Kaggle. Kaggle describes it as a dataset with a variety of variables on customers of an unnamed leading bank. They most likely got the data from the bank directly.

(c)

Predictor Variables:

- Work Experience (employ): Year of work experience that the customer has had.
- Age (age): Age in years of the customer.
- Credit to debt ratio (creddebt): Measure of how much of their available credit a customer is using. It is calculated by dividing total credit balance by credit limit.
- Other debts (othdebt): Total amount of other debts owed by the customer divided by 1,000
- Education Level (ed): The level of education of the customer. 1 means high school. 2 means undergraduate. 3 means master. 4 means phd. 5 means higher education.

Response variable:

- Income (income): Yearly income of the customer divided by 10,000.

Part 3: Regression Question

3.1 Exploratory Data Analysis for Regression Question

a) Data Cleaning and Processing:

Handling Missing Data:

We removed rows with missing values in the 'default' variable since it lacked meaningful information.

Excluding Variables:

We decided not to include the 'debt to income ratio' variable due to its direct correlation with the target variable (yearly income). This ratio primarily measures an individual's debt relative to their income, and its inclusion could inadvertently leak information about the target variable. Additionally, the 'address' variable, represented numerically without specific regions, was deemed irrelevant as a predictor for yearly income and was consequently excluded from the analysis.

Converting Variables:

We converted the variables 'ed' (education level) and 'default' from numeric numbers to factors. Education levels were coded as follows: 1 for high school, 2 for undergraduate, 3 for master, 4 for Ph.D., and 5 for higher education, with code 5 serving as the reference category. The 'default' variable was coded as 0 for non-default and 1 for default.

Histograms for Quantitative Variables:

Histograms were created for three quantitative variables: 'income,' 'creddebt' (credit-to-debt ratio), and 'othdebt.' The purpose of these histograms was to assess the normality of the data distribution. Skewed data distributions can negatively impact the performance of regression models. As a result, we transformed the 'income', 'creddebt', and 'othdebt' variables by taking the natural logarithm.

These steps were undertaken to enhance data quality and its relevance in the context of the analysis. The decision to exclude specific variables and conduct histogram analysis aimed to optimize the data for regression modeling.

b) Relevant summaries

Variable Summaries:

```
      age                          ed           employ           income     
 Min.   :20.00   high school     :188   Min.   : 0.000   Min.   : 14.00  
 1st Qu.:28.00   undergraduate   : 99   1st Qu.: 3.000   1st Qu.: 24.00  
 Median :34.00   master          : 43   Median : 7.000   Median : 33.00  
 Mean   :34.56   ph.d            : 16   Mean   : 8.074   Mean   : 43.23  
 3rd Qu.:40.00   higher education:  4   3rd Qu.:12.000   3rd Qu.: 52.75  
 Max.   :55.00                          Max.   :31.000   Max.   :253.00  
    creddebt           othdebt        
 Min.   : 0.01483   Min.   : 0.04558  
 1st Qu.: 0.32416   1st Qu.: 1.05246  
 Median : 0.77620   Median : 2.02279  
 Mean   : 1.36168   Mean   : 2.83945  
 3rd Qu.: 1.73800   3rd Qu.: 3.63434  
 Max.   :15.01668   Max.   :18.26913  
```

These summaries cover key statistics for each variable, excluding ‘default’ due to its categorical nature. They provide minimums, maximums, quartiles, means, and medians, aiding in understanding the range and differences between variables. This helps standardize and accurately compare features before model building.

Correlation Matrix:

```
           age     ed employ income creddebt othdebt
age      1.000  0.002  0.574  0.474    0.259   0.335
ed       0.002  1.000 -0.147  0.201    0.113   0.114
employ   0.574 -0.147  1.000  0.631    0.356   0.424
income   0.474  0.201  0.631  1.000    0.532   0.596
creddebt 0.259  0.113  0.356  0.532    1.000   0.659
othdebt  0.335  0.114  0.424  0.596    0.659   1.000
```

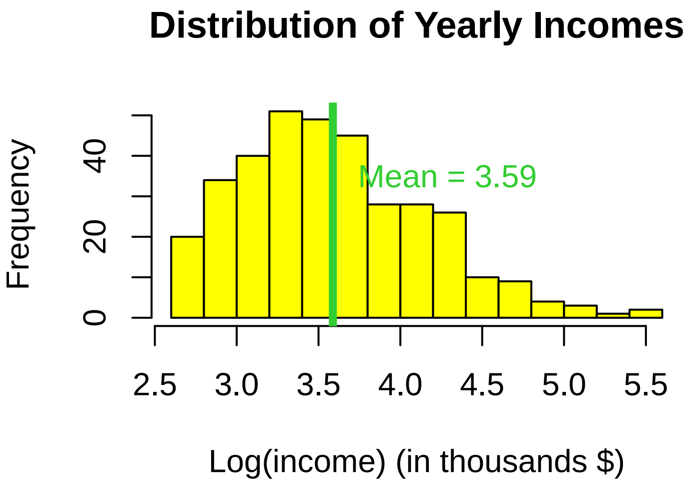
The correlation matrix shows that income has a strong positive relationship with employment, credit to debt ratio, and the other debts. Income has a slightly weak relationship with age and education level.

```
[1] 2.858047
```

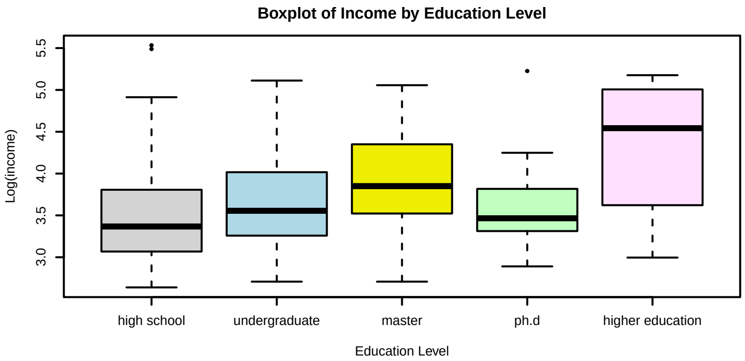
```
[1] 3.726225
```

```
[1] 2.365104
```

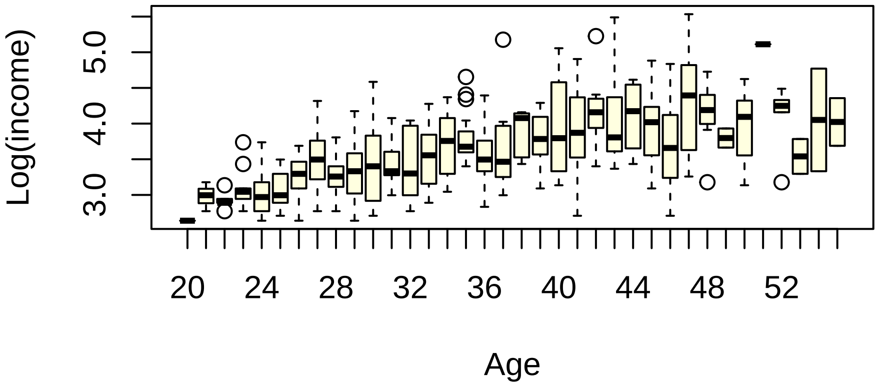
The skewness values for income, creddebt and othdebt are 2.858047, 3.726225, and 2.365104 respectively.



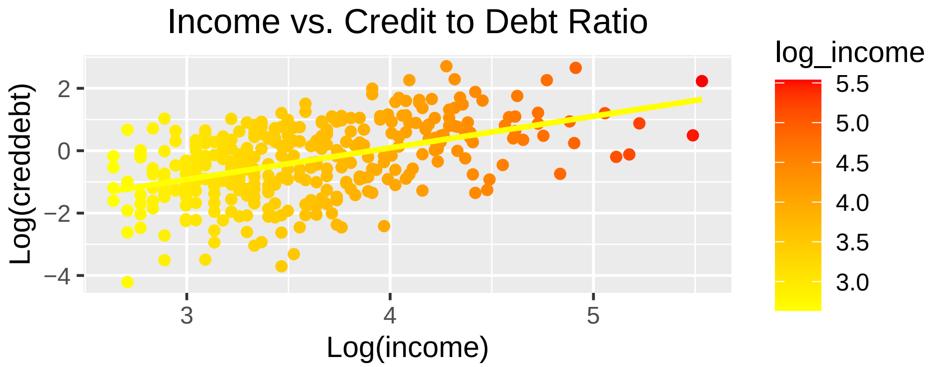
The average yearly income is $36,234, with the highest frequency occurring is between approximately $24,500 and $30,000.

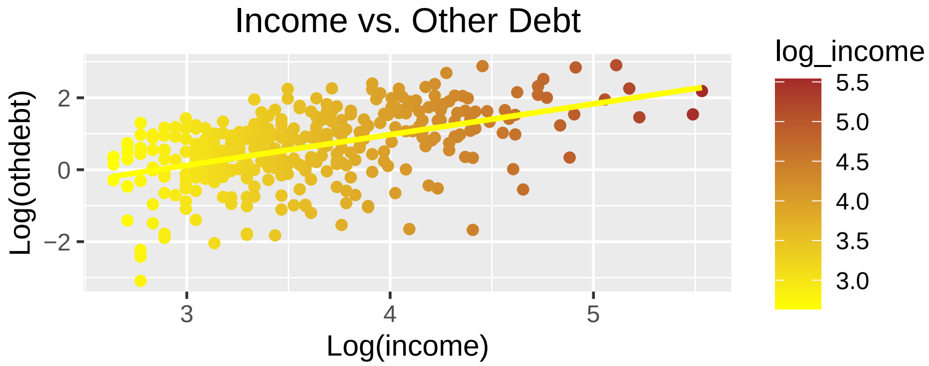


The average annual income for high school, undergraduate, master, Ph.D,and higher education is approximately $30,000, $37,000, $49,000, $33,000, and $99,000,respectively. We can see the average income for Ph.D. holders seems lower compared to master and undergraduate. (averages might not fully capture individual experiences and that some Ph.D. holders do achieve high incomes)

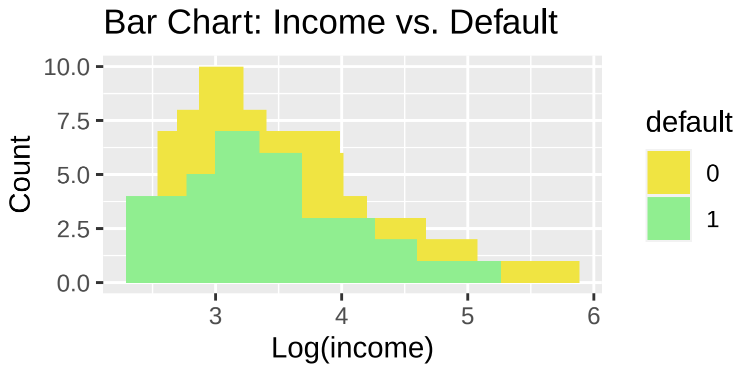


There is a slightly strong positive relationship between income and age.

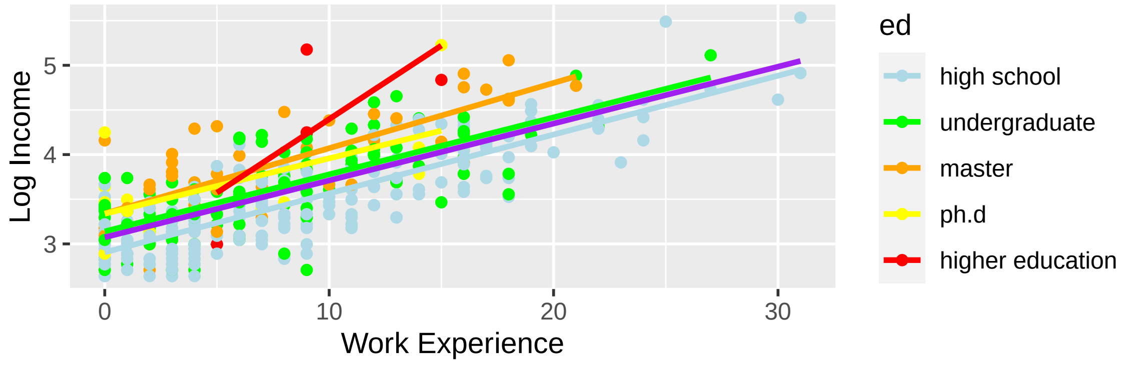




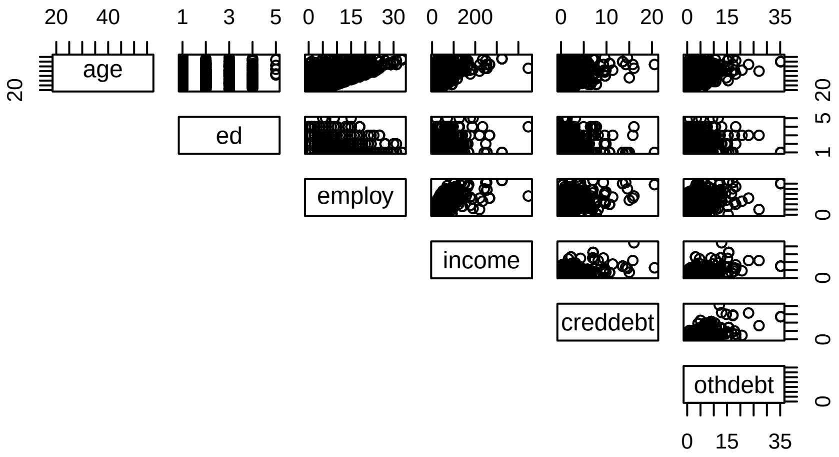
There is a moderately strong positive relationship between income, credit-to-debt ratio, and other forms of debt. The data shows that individuals with higher incomes tend to have a more favorable credit-to-debt ratio and higher levels of other debts. This positive correlation implies that as income increases, both the credit-to-debt ratio and other forms of debt tend to rise.



This bar chart illustrates that individuals with an income below $13,000 tend to default, while those with an income higher than $200,000 are less likely to default. The proportions for default and non-default are 0.26 and 0.74, respectively.The observed data does not provide sufficient evidence to establish a statistically significant relationship between income and default. Further analysis may be required to draw more robust conclusions in this regard.



The purple line represents the regression line. It reveals a strong positive link between work experience and income. It also shows that, on average, Ph.D. holders tend to have higher incomes when their work experience is factored in.



The correlation matrix reveals that pairs of predictors involving “work experience,” “other debt,” and “credit-to-debt ratio" exhibit high correlations. In contrast, pairs involving education and age show weaker correlations. Hence, there seems to be a degree of multicollinearity among certain predictors.

c) These graphical summaries were explored to gain an initial understanding of the data, identify trends and relationships, assess variable relevance, and prepare for subsequent regression analysis aimed at answering research question about the factors affecting annual income.

d) In our analysis of a subset comprising 350 individuals, we utilized graphical summaries to gain insights into the determinants of income. The findings align with established knowledge within the realm of credit risk assessment. Notably, individuals with advanced degrees, extensive work experience, and a higher credit-to-debt ratio generally exhibit elevated yearly incomes. Additionally, age plays a role in income dynamics. While age alone doesn't follow a straightforward trend of 'the older you get, the higher your annual income,' our analysis reveals a nuanced relationship. Age interacts with other factors, such as education and work experience, influencing income levels. Conversely, individuals with higher yearly incomes often carry increased levels of other forms of debt, a trend consistent with economic principles, where higher-income individuals may leverage their financial position for additional investments or expenditures. These insights emphasize the critical role of considering such factors, including age, in the evaluation of creditworthiness and lending decisions, shedding light on potential risk factors associated with loan defaults."

3.2 Shrinkage Methods

(a)

The predictors that are included are:

- Work experience (employ): years of work experience that the customer has had
- Age: age in years of the customer
- Log of credit to debt ratio (logcreddebt): log measure of how much of their available credit a customer is using. It is calculated by taking the log of dividing total credit balance by credit limit
- Log of other debts (logothdebt): log of total amount of other debts owed by the customer divided by 1,000
- Education Level (ed): the level of education of the customer. 1 means high school. 2 means undergraduate. 3 means master. 4 means phd. 5 means higher education. Note that this variable is converted into dummy variables, with each dummy variable representing a level of the original categorical variable.

The predictors that are excluded are:

- Address (address): address by itself does not correlate directly with income unless it's categorized in a way that reflects socio-economic divisions or regions. In this dataset, the address is simply a number which does not represent any region, so an address would not be a meaningful predictor for yearly income.
- Debt to Income Ratio (debtinc): Debt to Income Ratio is removed due to its direct correlation with the variable "yearly income". The debt to income ratio is essentially a measure of how much debt someone has in comparison to their income. Since we are trying to predict income, this ratio might inadvertently leak the target information.

(b)

To choose the value of the threshold used in the glmnet() function, we firstly try the value 10^{-7}:

```
## 9 x 2 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept) 2.53306535 2.63304026
## employ      0.04757750 0.04922534
## age         0.01143306 0.01114976
## logcreddebt 0.03648648 0.06928511
## logothdebt  0.03336981 0.10634697
## ed2         0.23267300 0.22364325
## ed3         0.42494000 0.43025666
## ed4         0.42662537 0.43156677
## ed5         0.80844086 0.79667912
```

The estimated coefficients from ridge regression are slightly different from the estimated coefficients from OLS, so the threshold may need to be smaller, especially if multicollinearity is present. We can try setting the threshold to be 10^{-23}:

```
## 9 x 2 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept) 2.53306535 2.63297199
## employ      0.04757750 0.04922347
## age         0.01143306 0.01115176
## logcreddebt 0.03648648 0.06926560
## logothdebt  0.03336981 0.10636100
## ed2         0.23267300 0.22364850
## ed3         0.42494000 0.43025704
## ed4         0.42662537 0.43157198
## ed5         0.80844086 0.79665847
```

To address the previous comments, there is a comment on "you made the threshold smaller. But looks like the coefficients are still not the same. Did you try even smaller threshold?". We can try setting the threshold to be 10^{-100}:

```
## 9 x 2 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept) 2.53306535 2.63297199
## employ      0.04757750 0.04922347
## age         0.01143306 0.01115176
## logcreddebt 0.03648648 0.06926560
## logothdebt  0.03336981 0.10636100
## ed2         0.23267300 0.22364850
## ed3         0.42494000 0.43025704
## ed4         0.42662537 0.43157198
## ed5         0.80844086 0.79665847
```

We can see that the estimated coefficients from ridge regression are still slightly different from the estimated coefficients from OLS after we change the threshold from 10^{-23} to 10^{-100}, and surprisingly, the estimated coefficients are the same for the threshold of 10^{-23} and 10^{-100}. This indicates that beyond a certain point, further reducing the lambda value does not significantly change the coefficients. This is because the impact of the penalty term has already become so minimal at 10^{-23} that going even lower to 10^{-100} does not materially change the outcome. Thus, we will still choose the value of 10^{-23} to be the threshold.

3.3 Regression Trees

(a)

Recursive binary splitting is used to build the regression tree because it take into account the various feature interactions and they also are very easily interpreted due to the layout of a tree. However, pruning is used over the recursive binary splitting tree to assure that the tree avoids overfitting by making sure that features are not overused in the tree. From both trees, the pruned tree results in a lower error rate, so that model is used to present the tree.

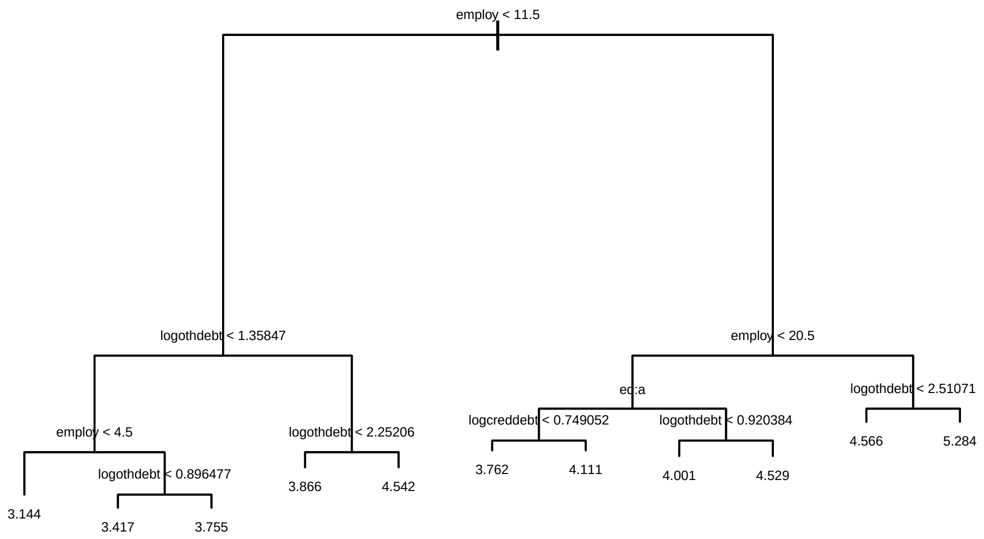
(b)

```
## 
## Regression tree:
## tree::tree(formula = logincome ~ age + ed + employ + logcreddebt + 
##     logothdebt, data = loans)
## Variables actually used in tree construction:
## [1] "employ"      "logothdebt"  "ed"          "logcreddebt"
## Number of terminal nodes:  11 
## Residual mean deviance:  0.1239 = 141.1 / 1139 
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -0.97400 -0.25320 -0.02421  0.00000  0.22430  1.57100 
```

(c)

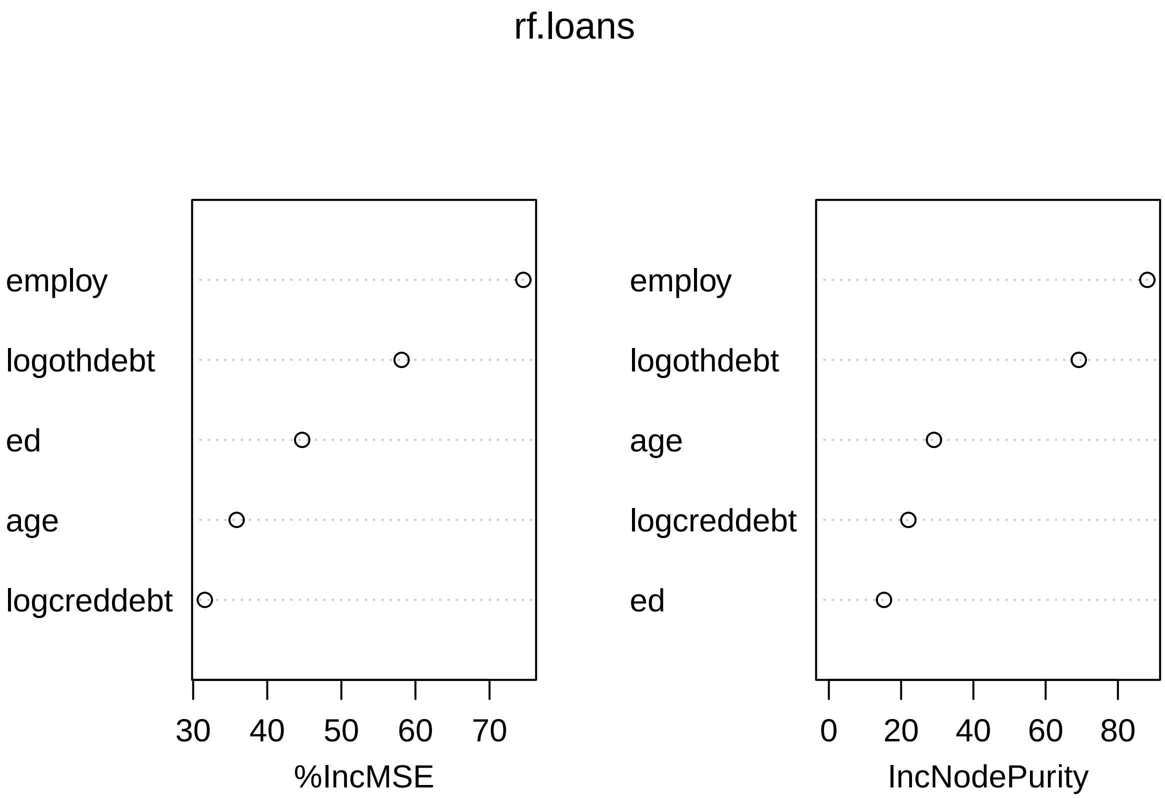
The pruned tree has 11 terminal nodes.

(d)



(e)

```
##             %IncMSE IncNodePurity
## age           35.87         29.09
## ed            44.71         15.27
## employ        74.56         88.15
## logcreddebt   31.58         22.01
## logothdebt    58.13         69.17
```



3.4 Summary of Findings

(a)

```
##                     Model   Test_MSE
## 1       Linear Regression 0.62757060
## 2        Ridge Regression 0.10722490
## 3        Lasso Regression 0.10664060
## 4 Pruned Regression Tree 0.12211410
## 5          Random Forests 0.08685707
```

(b)

The values of the test MSE's are relatively decent, except for linear regression. The random forests MSE is the best one at a 0.0868 because it shows that the mean of the squared errors between the predictions and actual values of the income (log transformed) is pretty low because it is only 0.08. Even for the other models, the shrinkage methods yield a 0.10 MSE and the pruned regression tree yields a 0.12 MSE which is relatively good because it could go all the way up to 1 so it is considered small. All in all, excluding linear regression, the models only have about a 10% mean squared error rate, which is good.

(c)

The findings from these models answer the regression question because they are able to predict the yearly income of customers based on the given variables with a relatively low error rate. Especially, as the table shows, the random forests model was very successful in predicting the yearly income based on the test set mean squared error, which shows that if banks were to use the data of the other debts, years of experience, credit to debt ratio, education level, and age of customers, they would be able to get a relatively decent estimate on a customer's income level if they don't disclose it. This could really help banks see if they trust a customer enough to loan them money. If they make a good amount of money yearly, it also is indicative of a customer's ability to be able to pay the bank back the money. Obviously, this cannot speak for everyone, but overall using these models, we can see that the income prediction is pretty reliable, depending on the level of risk banks are willing to take.

(d)

The random forests method was the best in answering the question of interest because its predicted value gives the lowest error rate on average so that is the model that banks would be recommended to use to minimize error. However, the pruned tree is a pretty decent model to use for the question because the purpose of this analysis is to see if a customer would default or not and the tree output gives a good look at exactly what factors and cutoffs for those factors(variables) are significant to output a certain prediction income value for the customer. This model was the best out of all the models that have been used and attempted to predict the income, so perhaps increasing the data size or the variables could improve the random forests model, or finding a better fitting dimensional model could address the question better by accounting for very specific predictor values

3.5 Address Previous Comments

We have successfully addressed the comments from previous milestones regarding the regression question.

Part 4: Executive Summary for Classification Question

(a)

The purpose of this analysis is to determine whether certain variables can provide relevant information about whether a customer has or has not defaulted in the past. In particular, the question of interest we want to answer is as follows: Can we utilize different borrower characteristics– specifically credit history, debt history, age, income, and work experience–to predict whether a customer has defaulted in the past?

(b)

Relevant stakeholders include: Lending institutions, borrowers, and financial regulators

Identifying which customers are more likely to default is important for a number of reasons. By identifying high-risk borrowers, lenders can reduce their risk of loan defaults. This can lead to lower interest rates and fees for borrowers, as well as a wider range of loan products and services. It can also help lenders by allowing them to be more informed in their lending practices. This can help to ensure that borrowers are not overextended and that they are able to repay their loans. In addition to helping make better practices, lenders can use this information to make tailored loan products and services that satisfy the needs of specific borrower segments. This information can also be used to help financial regulators develop sounder financial policies. For example, regulators may use this information to set capital requirements for banks or to develop new regulations to protect consumers.

(c)

The top three variables that were useful in predicting whether a customer has defaulted in the past are years of employment, credit to debt ratio, and other debts held by that customer.

(d)

In order to predict a customer's likelihood of defaulting, one should mainly look at that customer's years of employment, credit to debt ratio, and other debts. Specifically, more years of employment, less credit to debt ratio, and low other debts will indicate a customer as not likely to default.

Part 5: Data and Variable Description for Classification

Predictors:

- Work Experience (employ): Year of work experience that the customer has had.
- Age (age): Age in years of the customer.
- Income (income): Yearly income of the customer divided by 10,000.
- Credit to debt ratio (creddebt): Measure of how much of their available credit a customer is using. It is calculated by dividing total credit balance by credit limit.
- Other debts (othdebt): Total amount of other debts owed by the customer divided by 1,000
- Education Level (ed): the level of education of the customer. 1 means high school. 2 means undergraduate. 3 means master. 4 means phd. 5 means higher education. Note that this variable is converted into dummy variables, with each dummy variable representing a level of the original categorical variable.

Response Variable:

- Defaulted In Past (default): Is a 0 if the customer has not defaulted in the past and 1 if the customer has defaulted in the past.

Part 6: Classification Question

6.1 Exploratory Data Analysis for Classification

a) Data Cleaning and Processing:

Handling Missing Data: We removed rows with missing values in the ‘default’ variable as they lacked meaningful information.

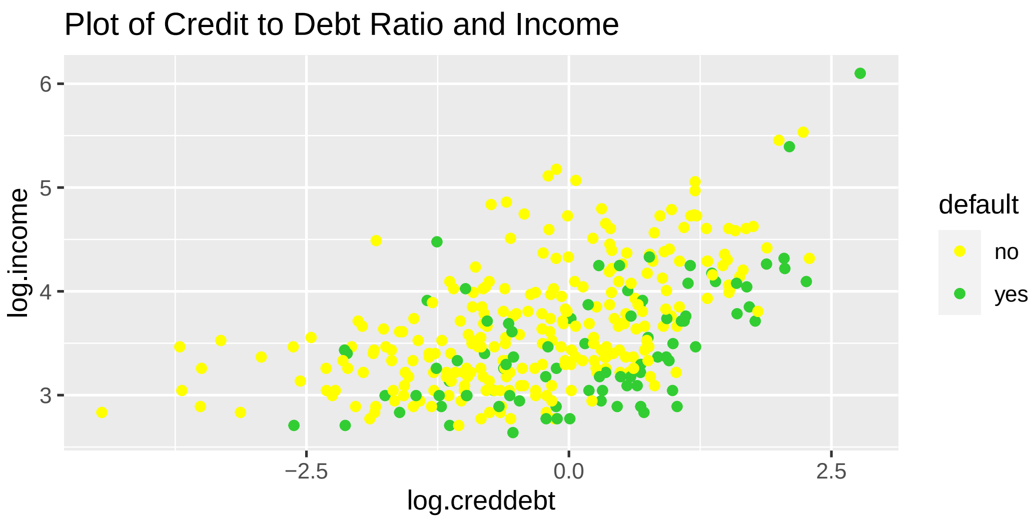
Excluding Variables: The ‘debt-to-income ratio’ variable was omitted due to potential multicollinearity with income or credit-to-debt ratio. Additionally, education level, being a categorical predictor unsuitable for discriminant analysis, was excluded. The ‘address’ variable, lacking specificity regarding regions, was considered irrelevant as a predictor for default and was consequently removed from the analysis.

Converting Variables: Numeric-to-factor conversion was applied to the ‘default’ variable, coded as 0 for non-default and 1 for default.

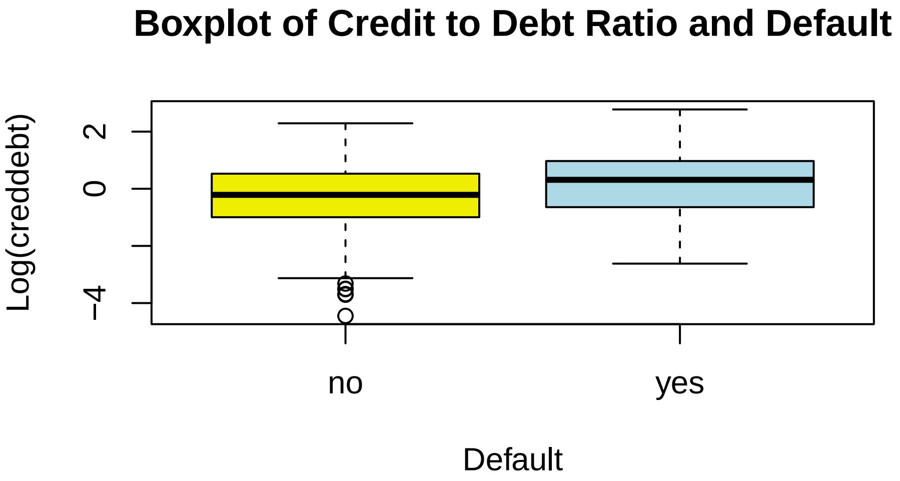
Histograms for Quantitative Variables: Histograms were generated for three quantitative variables: ‘income’, ‘creddebt’, and ‘othdebt’ These aimed to assess data distribution normality. Skewed distributions can impact regression models negatively. To address this, we transformed ‘income,’ ‘creddebt,’ and ‘othdebt’ variables using the natural logarithm.

```
    yes
no    0
yes   1
```

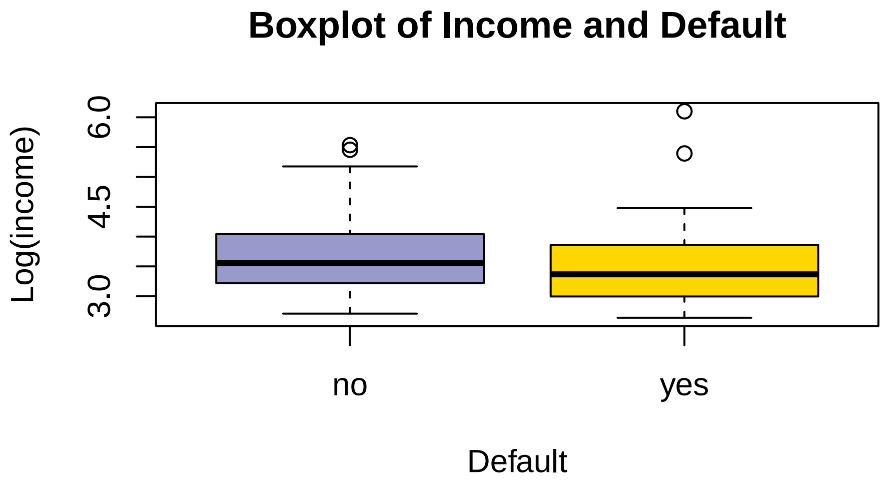
```
       no       yes
0.7385714 0.2614286
```



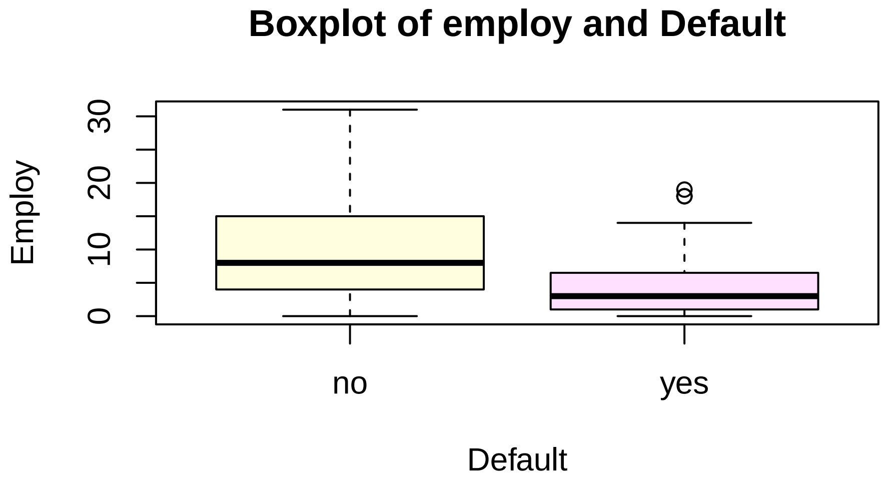
We created a scatter plot that compares annual income and monthly credit card to debt ratio for a subset of 350 individuals. Defaulted individuals are represented in green, while non-defaulted individuals are in yellow. The plot shows that those who defaulted generally had lower credit to debt ratios and lower incomes compared to those with higher incomes who did not default.



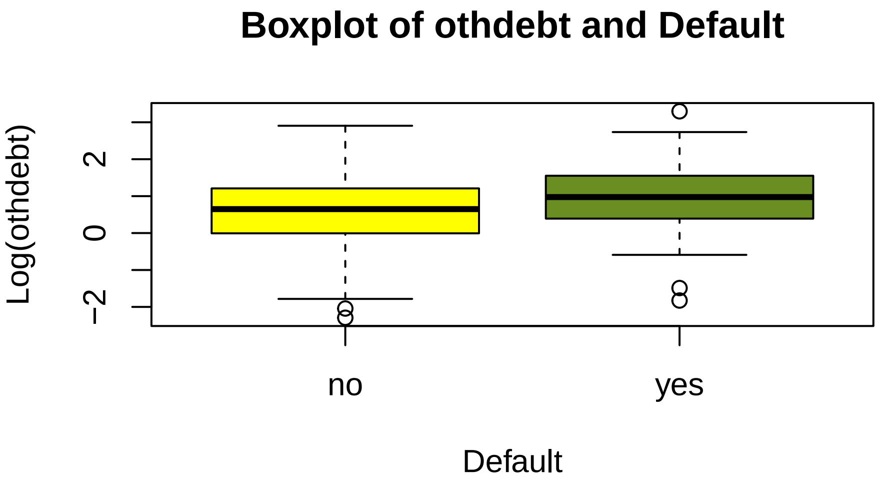
The plot indicates that individuals who have defaulted on their financial obligations typically exhibit a higher average credit-to-debt ratio when compared to those who have not defaulted. This suggests a correlation between a lower credit-to-debt ratio and a higher likelihood of defaulting on financial commitments.



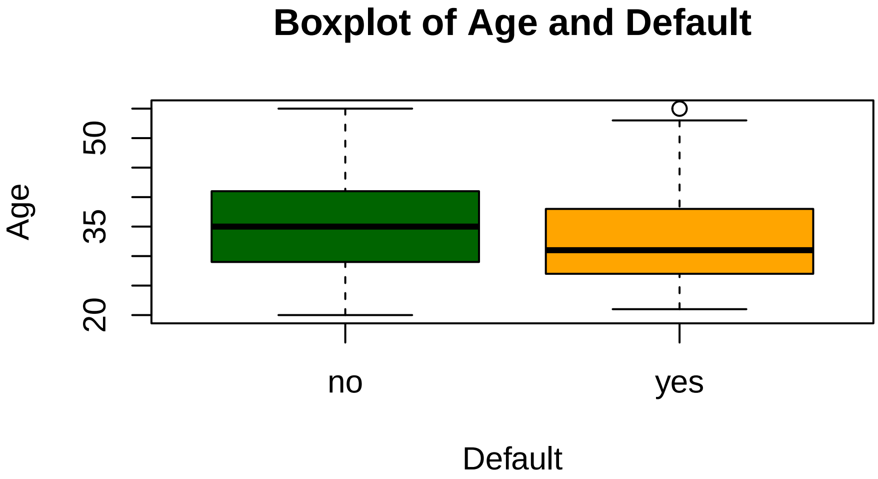
The boxplot reveals that individuals who have defaulted on their financial obligations generally exhibit lower average incomes in comparison to those who have not defaulted. Furthermore, it is noteworthy that there is a higher occurrence of outliers within the non-defaulting individuals' group.



The plot suggests that, on average, individuals who have defaulted tend to have less work experience than those who did not default.



The plot illustrates that, on average, individuals who defaulted have higher levels of other debt compared to those with less other debt.



The boxplot illustrates that, on average, individuals who have defaulted are younger in age compared to older individuals.

Summary:

In our analysis, we examined a subset of 350 individuals and created several graphical summaries to gain insights into factors associated with loan defaults. The findings are consistent with prior knowledge in the field of credit risk assessment. Individuals who defaulted generally exhibited higher credit card to debt ratios, lower incomes, less work experience, higher levels of other debt,and younger in age. These insights reinforce the importance of considering these factors in credit assessment and lending decisions, as they highlight potential risk factors associated with loan defaults.

6.2 Logistic Regression Model

(a)

The predictors that are included are:

- Work experience (employ): years of work experience that the customer has had
- Age (age): age in years of the customer
- Credit to debt ratio (creddebt): measure of how much of their available credit a customer is using. It is calculated by taking the log of dividing total credit balance by credit limit
- Other debts (othdebt): total amount of other debts owed by the customer divided by 1,000
- Income (income): Yearly income of the customer divided by 10,000.
- Education Level (ed): the level of education of the customer. 1 means high school. 2 means undergraduate. 3 means master. 4 means phd. 5 means higher education. Note that this variable is converted into dummy variables, with each dummy variable representing a level of the original categorical variable.

The predictors that are excluded are:

- Address (address): address by itself does not correlate directly with income unless it's categorized in a way that reflects socio-economic divisions or regions. In this dataset, the address is simply a number which does not represent any region, so an address would not be a meaningful predictor for yearly income.
- Debt to Income Ratio (debtinc): Debt to Income Ratio is removed due to its direct correlation with the variable "yearly income". The debt to income ratio is essentially a measure of how much debt someone has in comparison to their income. Since we are trying to predict income, this ratio might inadvertently leak the target information.

(b)

```
##
## Call:
## glm(formula = default ~ age + employ + income + creddebt + othdebt +
##     ed, family = binomial, data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.40939  -0.65839  -0.29462   0.02641   2.78906
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.080063   0.684216  -1.579  0.11444
## age          0.012566   0.021124   0.595  0.55192
## employ      -0.290466   0.047013  -6.178 6.47e-10 ***
## income      -0.016658   0.006456  -2.580  0.00988 **
## creddebt     0.673029   0.134886   4.990 6.05e-07 ***
## othdebt      0.251199   0.086542   2.903  0.00370 **
## ed2          0.601259   0.360289   1.669  0.09515 .
## ed3          0.293614   0.469614   0.625  0.53182
## ed4          0.088228   0.547714   0.161  0.87203
## ed5          1.499379   1.325957   1.131  0.25814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 394.73  on 349  degrees of freedom
## Residual deviance: 276.48  on 340  degrees of freedom
## AIC: 296.48
##
## Number of Fisher Scoring iterations: 6
```

6.3 Classification Trees

(a)

We chose to present the recursive binary splitting tree over the pruned tree because it had a false negative rate that was around ~0.1 lower than the pruned tree. This, however, came at a tradeoff as the RBS tree had ~0.1 higher false positive rate but, for our purposes, maintaining a lower false negative rate is more important than maintaining a lower false positive rate as predicting a customer as likely to not default when they are likely to default is more damaging for lending institutions.

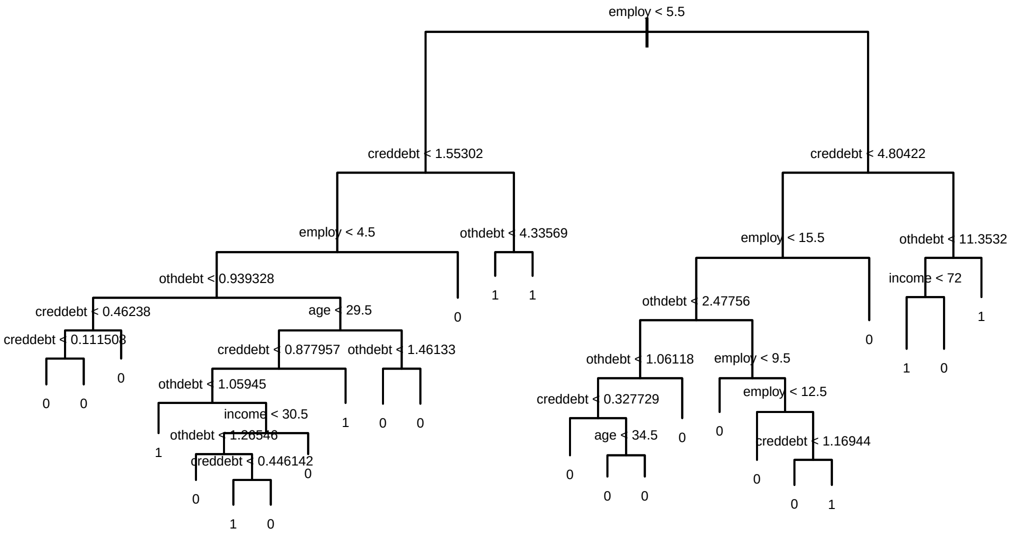
(b)

```
##
## Classification tree:
## tree(formula = default ~ ., data = train)
## Number of terminal nodes:  26
## Residual mean deviance:  0.5881 = 190.5 / 324
## Misclassification error rate: 0.16 = 56 / 350
```

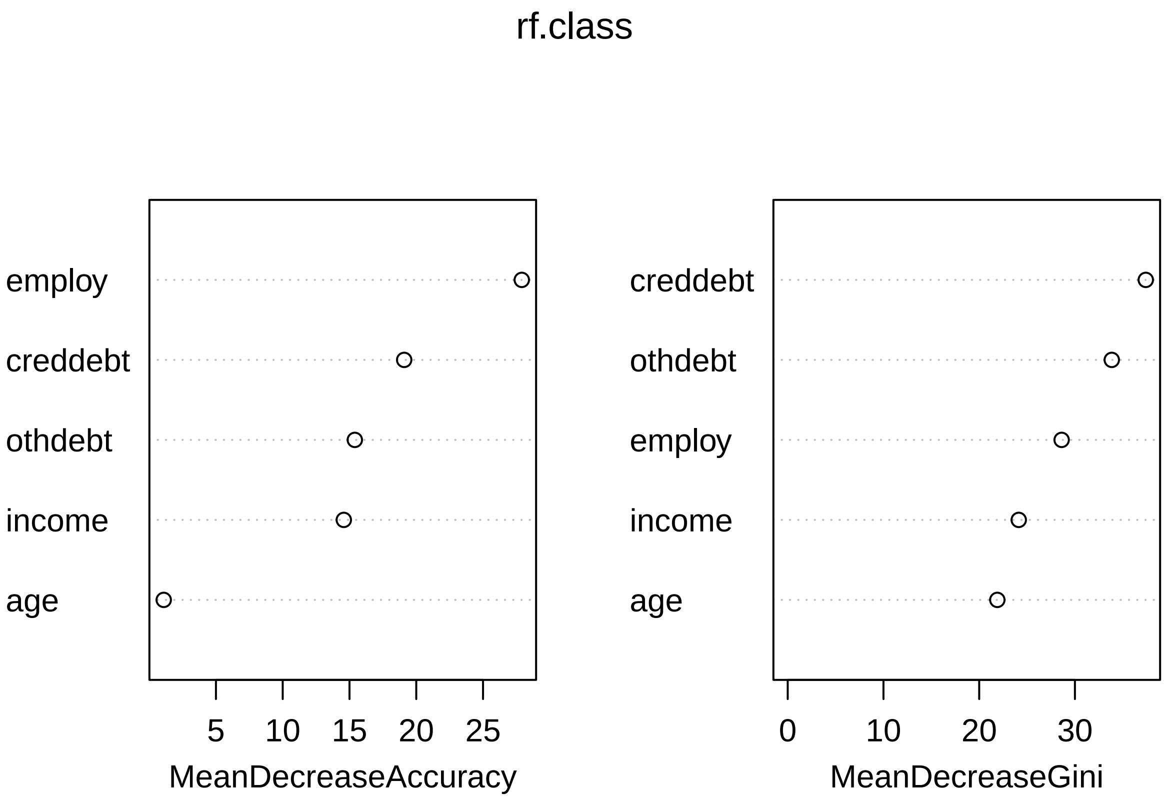
(c)

The tree has 26 terminal nodes.

(d)


employ < 5.5
creddebt < 1.55302
creddebt < 4.80422
employ < 4.5
othdebt < 4.33569
employ < 15.5
othdebt < 11.3532
othdebt < 0.939328
income < 72
1
1
othdebt < 2.47756
age < 29.5
creddebt < 0.46238
0
1
creddebt < 0.111508
0
creddebt < 0.877957
othdebt < 1.46133
employ < 9.5
othdebt < 1.06118
1
0
0
othdebt < 1.05945
employ < 12.5
creddebt < 0.327729
0
0
income < 30.5
1
0
0
0
othdebt < 1.26546
age < 34.5
0
creddebt < 1.16944
1
creddebt < 0.446142
0
0
0
0
0
0
1
0
1
0

(e)



6.4 Summary of Findings

(a)

Logistic Regression Matrix

```
##         log.pred.test
## y.test    0   1
##      0 241  30
##      1  45  34
```

```
##            Actual
## Predicted    0   1
##          0 212  67
##          1  59  12
```

Classification Tree (RBS) Matrix

```
##         tree.pred.test
## y.test    0   1
##      0 242  29
##      1  46  33
```

Random Forests Matrix

```
##       pred.rf
## y.test   0   1
##      0 243  28
##      1  45  34
```

(b)

```
##                                   Error Rate       FPR       FNR
## 0.5 Threshold RBS                  0.2142857 0.1070111 0.5822785
## 0.5 Threshold Logistic Regression  0.1828571 0.0996310 0.5696203
## 0.5 Threshold RF                   0.2085714 0.1048689 0.5696203
```

(c)

Logistic Regression Threshold Commentary

The logistic regression yields a FNR that is signficantly higher than the FPR so lowering the threshold to approximately 0.4 would equalize them. This may increase the overall error rate, but it would be more balanced and better so that the false negative rate can be decreased which is way worse because it would be detrimental to predict a customer to not default and then they default, making the bank lose more money than the other way around.

Classification Tree (RBS) Threshold Commentary

The FNR is significantly higher than the FPR and lowering the threshold to 0.4 would equalize them at ~.285. This, however, would increase the overall error rate but this is worth it as having a high false negative rate means means that we would classify a customer as not having defaulted when they did which is worse for our purposes than classifying a customer as having defaulted when they haven't.

Random Forests Threshold Commentary

Just like RBS, the false negative rate is significantly higher than the false positive rate so we decreased the threshold to .3 in order to even things out between FNR and FPR.

(e)

```
##                                   Error Rate       FPR       FNR
## 0.4 Threshold RBS                  0.2685714 0.2804428 0.2911392
## 0.4 Threshold Logistic Regression  0.2184897 0.2436279 0.2875843
## 0.3 Threshold RF                   0.2514286 0.2619926 0.2151899
```

(f)

The models we built show us that, of the predictors we used, there importance ranked from greatest to least is as follows: employ (years of employment) which had a negative correlation to default, creddebt (credit to debt ratio) which had a positive correlation to default, othdebt (other debt) which had a positive correlation to default, income which had a negative correlation to default, and age which had a negative correlation to default.

(g)

The logistic regression best answered our question because it had the lowest total error rate. Though, the random forest may be equally as good or arguable better for the question because it yielded the lowest false negative rate (which is crucial as users of this model would rather overshoot in their prediction rather than undershoot) and, as stated above, it found that employ, creddebt, and othdebt were the three most important factors in predicting customers who have defaulted in the past. This result isn't all that suprising as in our EDA we found all the predictors in our data set to be relatively valid factors to consider when predicting default, but it is interesting how age is the least important of the five as generally one would consider years of employment (the most important predictor) to be closely correlated to age.

6.5 Address Previous Comments

We have successfully addressed the comments from previous milestones regarding the classification question.

Part 7: Further Work

If more time was available, the project could have been extended to also include more features by researching other variables that could be indicative of a customer's income or their likelihood of defaulting. Another aspect that would have been interesting to dive into is clustering. Clustering of customers could be used to compare and especially visually depict customers who are similar in certain features that may make them more likely to default. This method would be visually appealing and makes sense due to the variety of numerical variables used and when thinking about the way people behave, it makes sense because in general, humans follow patterns of behaviors. For instance, people who are not organized with their finances and bills are generally more susceptible to defaulting due to their lack of organization. Clustering can help group behaviors like those to create predictions for income as well as likelihood of defaulting. Additionally, with more time, some tuning of hyperparameters would be a decent strategy to use to improve the models.

Part 8: Reflection on Learning

This project helps on learning, firstly by providing a practical application of the statistical and analytical concepts learned in class. By working on real data sets, we were able to apply our understanding of variables, data analysis, and graphical representation in a tangible way. The project connected classroom learning with real-world applications. By analyzing a dataset related to credit risk analysis for extending bank loans, we see the practical implications and applications of our learning, making the concepts more relevant and engaging. Additionally, the group project required managing time effectively and taking responsibility for individual contributions to meet deadlines. This aspect of the project mirrors real-world scenarios where time management and accountability are essential. Also, the project necessitated working as a group, which allowed for the exchange of ideas and collaborative problem-solving. This interaction led to a better understanding of concepts as we learn from each other's perspectives and approaches.