Group 32 Final Report

V. Mullins (vsm2ey), H. Shaikh (hrs4zc), J. Harwood (jeh5rhk), & R. Funderburk (rpf2fh)

2023-12-02

Regression Question

Executive Summary

Using the data in the AES CreditCard data set, we wanted to create a regression model that predicts the average monthly credit card expenditure of a given individual. By exploring the prediction of average monthly credit card expenditure, valuable insights can be obtained regarding consumer behavior. In turn, this information allows stakeholders like financial institutions to better understand and anticipate their customers' spending habits based on their application characteristics. Credit card companies can then assess applications to find applicants who will use more credit and thus generate more revenue for credit card companies. Furthermore, this predictive model can assist various stakeholders in identifying high-risk individuals or potential defaults by flagging unusual spending patterns before accepting an applicant who may default on their credit.

Through our extensive regression analysis, we determined that the most important factors in predicting average monthly credit card expenditure were the applicant's annual income, their number of outstanding derogatory reports, and their age. While income and derogatory reports make logical sense in their relation to expenditure, it was surprising to see the importance placed on an applicant's age. Considering the model took into account predictors such as the number of dependents or the number of active credit accounts an applicant has, seeing a rather rudimentary statistic perform so strongly was a shock. This may be evidence to support a further analysis into the between one's age and their credit habits.

With this data, stakeholders like financial institutions and credit card companies can implement findings like these into their preexisting information repositories. With evolving data systems, data scientists and engineers at these companies can place greater weight on variables such as income and age when designing their approval techniques. With improved background checking capabilities, operations could be streamlined and result in quicker approvals, providing for an overall better customer experience. Conversely, this predictive model can assist credit card companies in identifying high-risk individuals or potential defaults by flagging unusual spending patterns before accepting an applicant who may default on their credit.

Data and Variable Description

Data Description & Sources

The AES CreditCard data set we are using is obtained from a built-in package within R-studio. Originally this data set was built by W.H. Greene from New York University's Stern School of Business in 1992. Specifically, the data set consists of 1,319 observations on 12 different variables all on the credit history and demographics for a sample of applicants applying for a type of credit card. The variables contain information about applicants' lifestyle, financial responsibilities, and credit history which are then evaluated by the credit card company to determine if the applicant will be approved for the specific card. Specific variables used within our model can be seen below.

Predictors

Table 1. *Data dictionary of variables used in EDA and further Regression Modeling*

Variable Name	**Description**	**Type**
reports	Indicates the number of major derogatory reports against an applicant	Quantitative
age	Age of the applicant in years plus twelfths of a year (number of months)	Quantitative
income	Yearly income of the applicant in USD	Quantitative
expenditure**	Average monthly credit card expenditure	Quantitative
owner	Indicates if an applicant owns their home	Categorical- levels: yes, no
selfemp	Indicates if an applicant is self-employed	Categorical- levels: yes, no
majorcard	Indicates if the applicant has any other major credit cards	Categorical- levels: yes, no
dependents	Number of dependents supported by the applicant	Quantitative
months	Months, the applicant has been living at the current address	Quantitative
active	Number of active credit accounts associated with the applicant	Quantitative

**** indicates response variable**

The data required slight modifications and cleaning to be used for analysis. First of all the **income** column was multiplied by a factor of 10000 to keep units regulated across all monetary-based variables for the sake of interpretability, they are now all in dollar units. Second, all of the categorical variables that were originally coded as string variables, were changed to factor variables. Thirdly, we converted **majorcards** to a factor variable because the data set only contained unique values of zero and one, which made the data set only representative of a sample that either is in possession of a major credit card or not in possession of a major credit card, so we viewed it as categorical. Lastly, we removed **card** from the predictors because

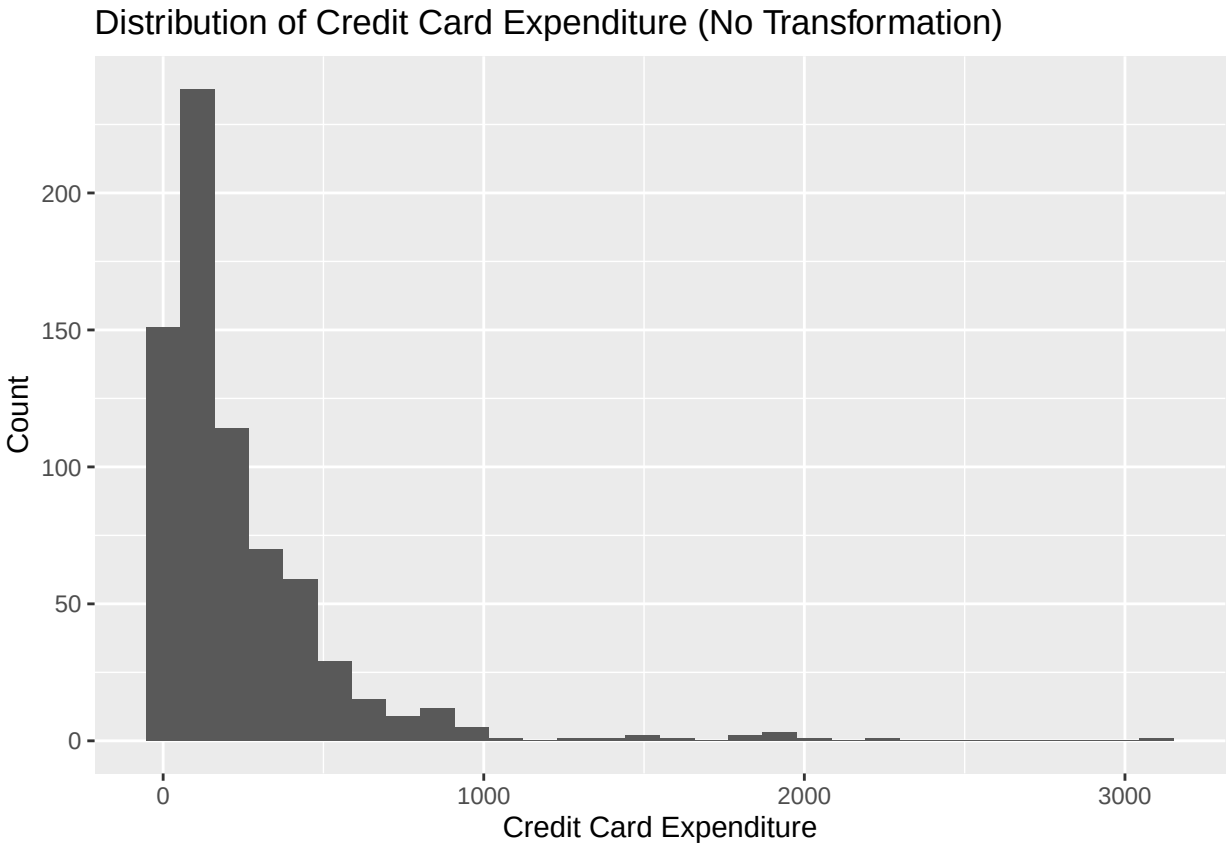
only individuals who were approved for credit cards have expenditure values hence it is a perfect predictor as to whether or not the applicants use their card as they either have one or don't.

Regression Question

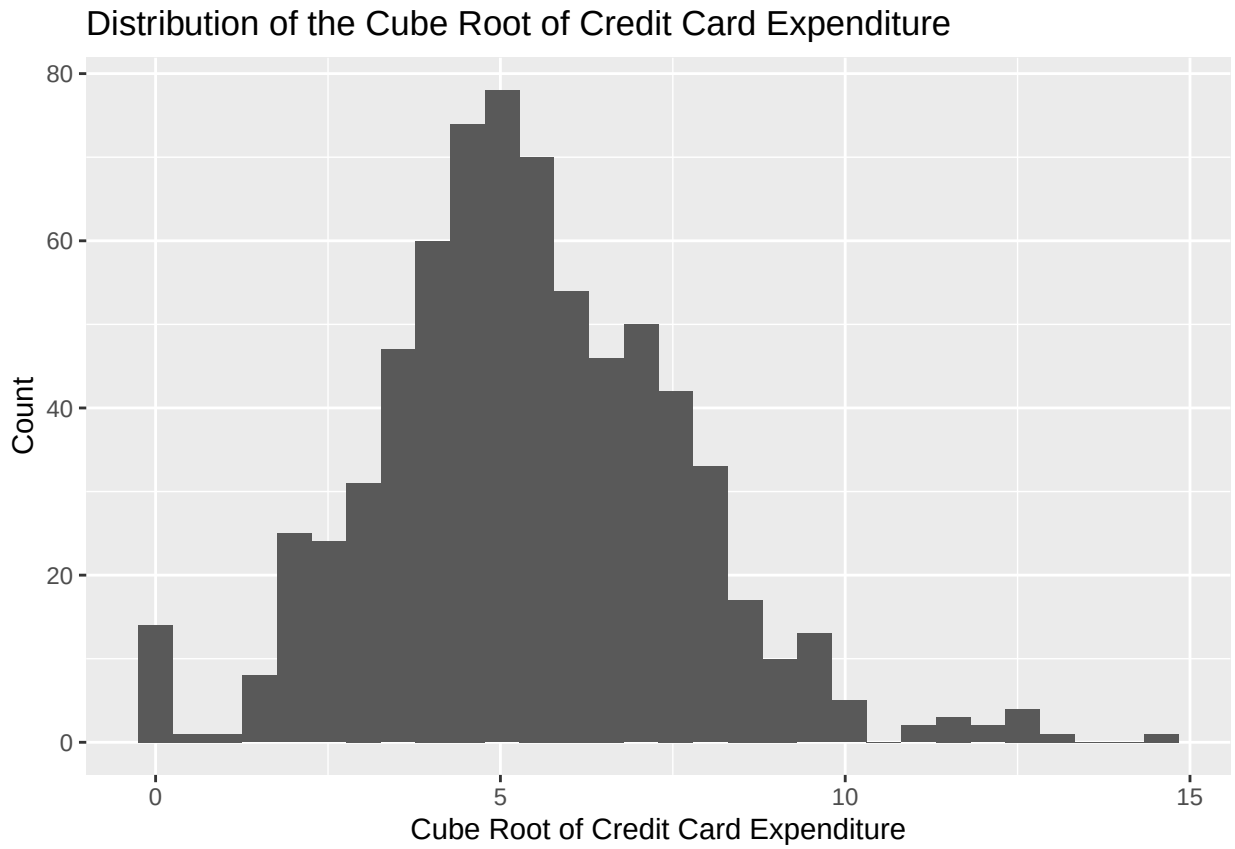
Exploratory Data Analysis

Histogram to Assess Distribution of Response Variable

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

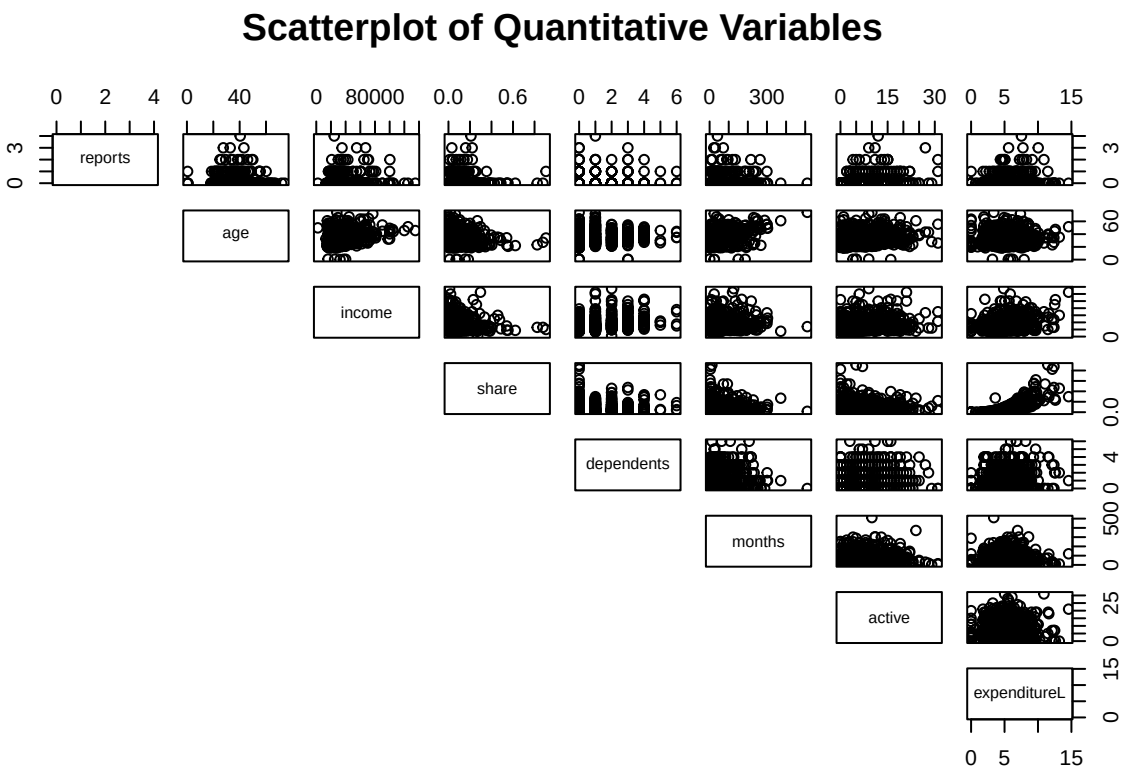



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



First, we created a histogram of our response variable, credit card expenditure to view an overview of the distribution of our data and display any patterns or skewness that needs to be attended to. Based off of earlier iterations, the original data was heavily right skewed, thus a cube root transformation was performed to achieve a more normal distribution. This is important as tree based methods are heavily dependent on and affected by the mean value of the response variable when creating models that are true to the data and skewed data creates challenges.

Scatter Plot & Correlation Matrix of all Quantitative Predictors

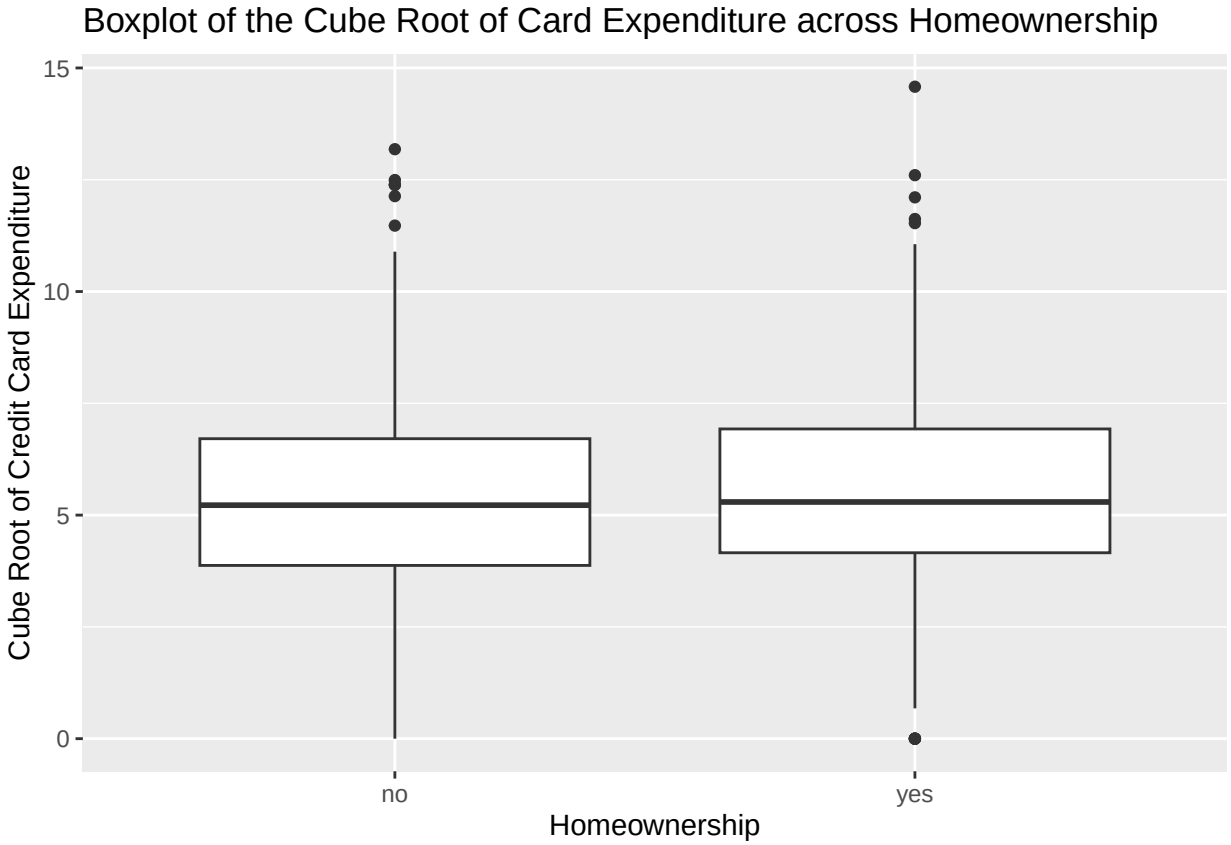



```
##              reports    age income  share dependents months active expenditureL
## reports        1.000  0.083  0.117  0.067      0.002  0.097  0.213        0.127
## age            0.083  1.000  0.355 -0.142      0.235  0.443  0.184       -0.023
## income         0.117  0.355  1.000 -0.111      0.355  0.124  0.191        0.267
## share          0.067 -0.142 -0.111  1.000     -0.085 -0.070 -0.070        0.782
## dependents     0.002  0.235  0.355 -0.085      1.000  0.082  0.139        0.084
## months         0.097  0.443  0.124 -0.070      0.082  1.000  0.125       -0.018
## active         0.213  0.184  0.191 -0.070      0.139  0.125  1.000        0.020
## expenditureL   0.127 -0.023  0.267  0.782      0.084 -0.018  0.020        1.000
```

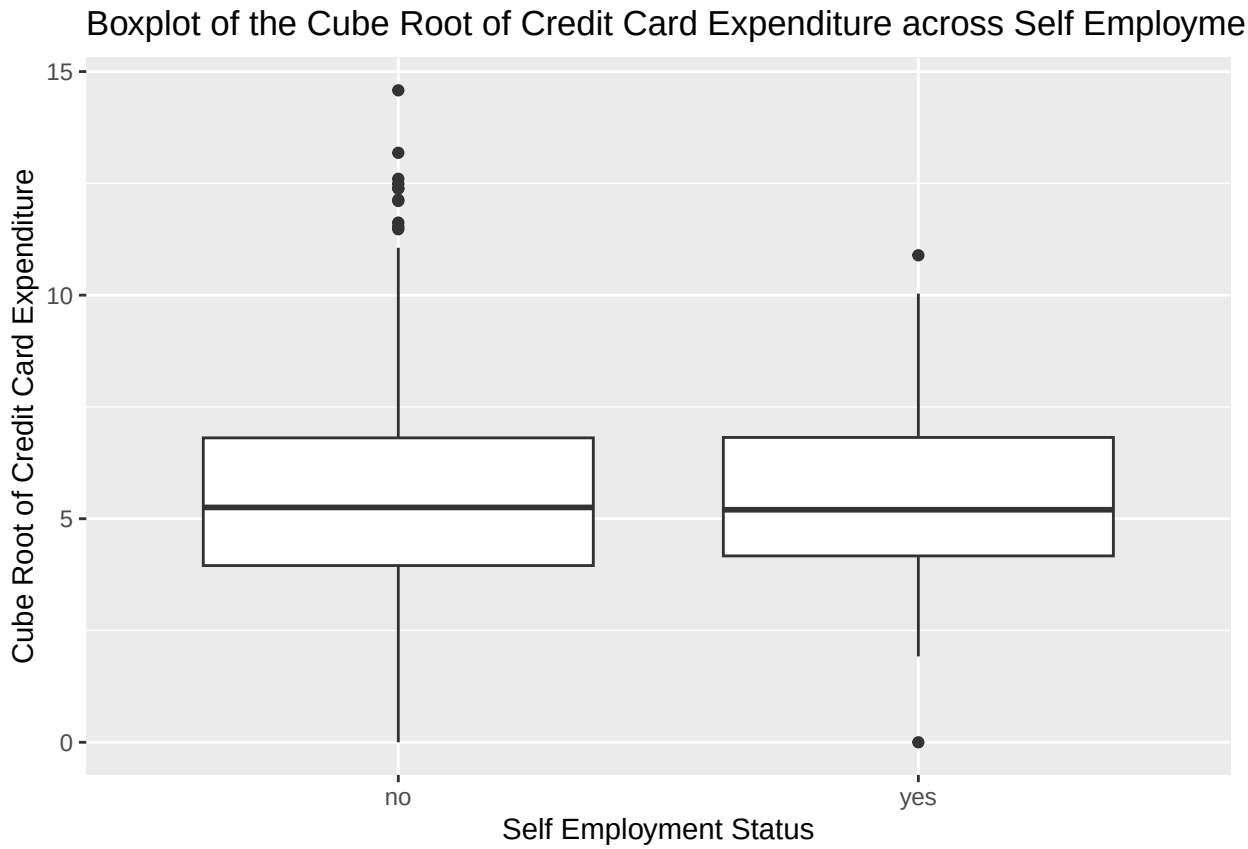
Next, we created a scatter plot matrix of all the quantitative predictors to better understand the relationship between the variables and their associated outcome with the quantitative response variable, credit card expenditure. Simultaneously, we included a correlation matrix indicative of which quantitative predictors have a weak, moderate, or strong positive/negative association with credit card expenditure.

The scatter plot matrix displayed above of all the quantitative predictors highlights a strong positive linear association for **share**, a moderate positive linear association for **income**, a weak positive linear association for **reports**, **dependents**, and **active**, and a weak negative linear association for **age** and **months** with credit card expenditure corroborated by their correlation values. Based on these summaries we decided to remove **share** as it is a linear function of **income** and **expenditure**, and could contribute as a source of multicolinearity.

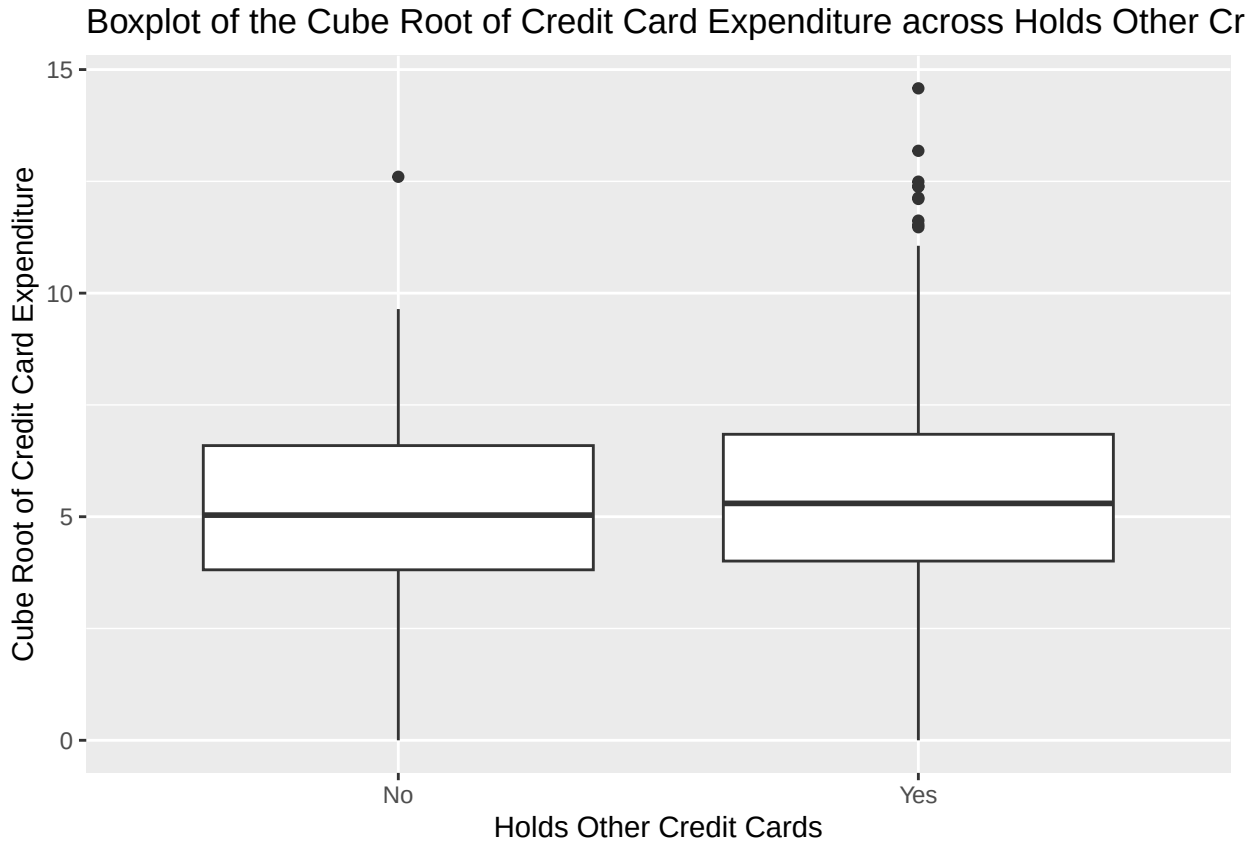
Box Plots for Categorical Variables



From this box plot, there is barely any variation between the distributions of credit card expenditure for those with ownership of a home or not. With the table of the medians there is evidence that the distribution of the expenditures difference is very minimal (0.116326) with credit card expenditure being higher for those owning a home. Further these findings may go against prior knowledge that home ownership entails several financial responsibilities such as maintenance, mortgage, appliances, taxes, insurance, etc and questions if home ownership is a plausible predictor for the average monthly credit card expenditure of a given individual.



From this box plot, there is a slight variation between the distributions of credit card expenditure for those that are self employed or not. This variation is extremely minimal (0.331401).



From this box plot, there is a slightly larger variation between the distributions of credit card expenditure for those that hold other cards or not. Within the table of the medians those who hold other cards have a higher credit card expenditure than those who do not. These differences in the distribution could possibly suggest that **majorcards** is a plausible predictor for the average monthly credit card expenditure of a given individual.

Shrinkage Methods

Predictors

For shrinkage methods, the categorical variables needed to be dummy coded, so sorted the data set into a model matrix and response vector. The response variable was transformed by taking the cube root as described above as well.

The predictors excluded from our model were **card** and **share**. **Share** was excluded from the model because it is calculated using income and expenditure. **Share** is the percentage of income taken up by the observations monthly expenditures, so it is directly related to the **income** and **expenditure** variables. **Card** wasn't included in the model because to control for normality assumption we subsetted the data based on the true values of the card variable. Therefore, all values of the **card** in the model would be true and it would be useless for prediction purposes. Meaning our model includes **reports**, **age**, **income**, **owner**, **selfemp**, **majorcard**, **dependents**, **months**, and **active** to predict the cube root of **expenditure**.

Threshold value

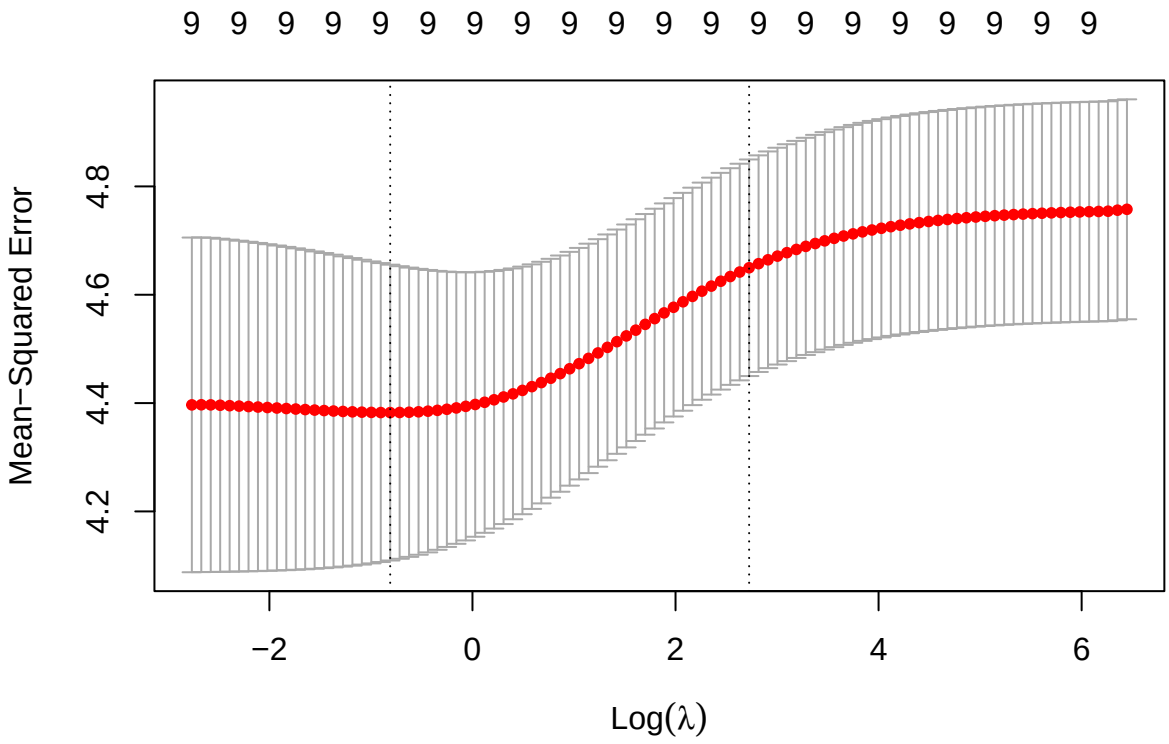
```
## 10 x 3 sparse Matrix of class "dgCMatrix"
```

```
##                                    s0           s0
## (Intercept)    4.723836e+00  4.723836e+00  4.723836e+00
## reports        4.778134e-01  4.778134e-01  4.778134e-01
## age           -2.859613e-02 -2.859613e-02 -2.859613e-02
## income         3.879208e-05  3.879208e-05  3.879208e-05
## owneryes       1.061240e-01  1.061240e-01  1.061240e-01
## selfempyes    -1.411818e-01 -1.411818e-01 -1.411818e-01
## dependents     1.186622e-01  1.186622e-01  1.186622e-01
## months         7.663469e-05  7.663469e-05  7.663469e-05
## majorcardsYes  3.041830e-01  3.041830e-01  3.041830e-01
## active        -3.579432e-02 -3.579432e-02 -3.579432e-02
```

A threshold value of 1eˆ-18 was used in the glmnet() function. We then moved on to applying Ridge and Lasso Methods to the data.

Apply Ridge Regression to the Training Set

```
## [1] 0.444738
```




```
##
## Call:  glmnet(x = train.x, y = train.y, alpha = 0, lambda = best.lam,      thresh = 1e-12)
##
##   Df  %Dev Lambda
## 1  9 11.23 0.4447
```

Based on the 10 fold cross validation for Ridge Regression the best tuning parameter lambda is 0.444738. The Number of predictors left in the model based on the value of lambda chosen by cross validation is 9 as the ridge fails to remove any predictors and all are left in the model.

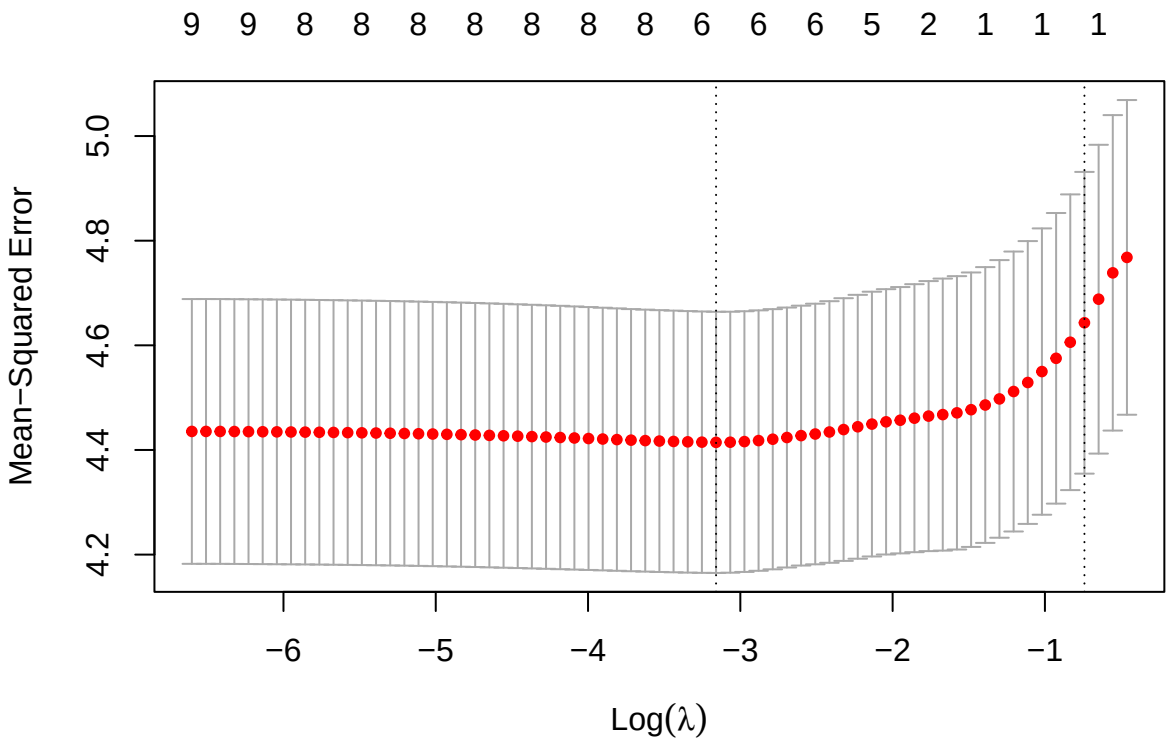
Actual test MSE for Ridge Reression

```
## [1] 4.730406
```

The actual test MSE of Ridge regression based on the model using the value of lambda chosen by CV is 4.7304057.

Apply Lasso Regression to the training set

```
## [1] 0.0424524
```




```
##
## Call:  glmnet(x = train.x, y = train.y, alpha = 1, lambda = best.lam1,      thresh = 1e-18)
##
##   Df  %Dev  Lambda
## 1  6 11.36 0.04245
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                             s0
```

```
## (Intercept)    4.683305e+00
## reports        3.400841e-01
## age           -2.129893e-02
## income         3.610213e-05
## owneryes       .
## selfempyes     .
## dependents     8.992174e-02
## months         .
## majorcardsYes  1.762989e-01
## active        -2.313542e-02
```

Value of the tuning parameter lambda based on 10-fold cross-validation on the training data is 0.0424524. The number of predictors left in the model based on the value of lambda chosen by CV is 6. The predictors left in the model after Lasso Regression are **reports**, **age**,**income**,**dependents**,**majorcards**, and **active**.

Actual test MSE for Lasso

```
## [1] 4.694509
```

The actual test MSE based on the model using the value of lambda chosen by CV is 4.694509.

Actual test MSE for OLS regression

```
## [1] 4.72794
```

The actual test MSE for OLS regression is 4.72794.

Regression Trees

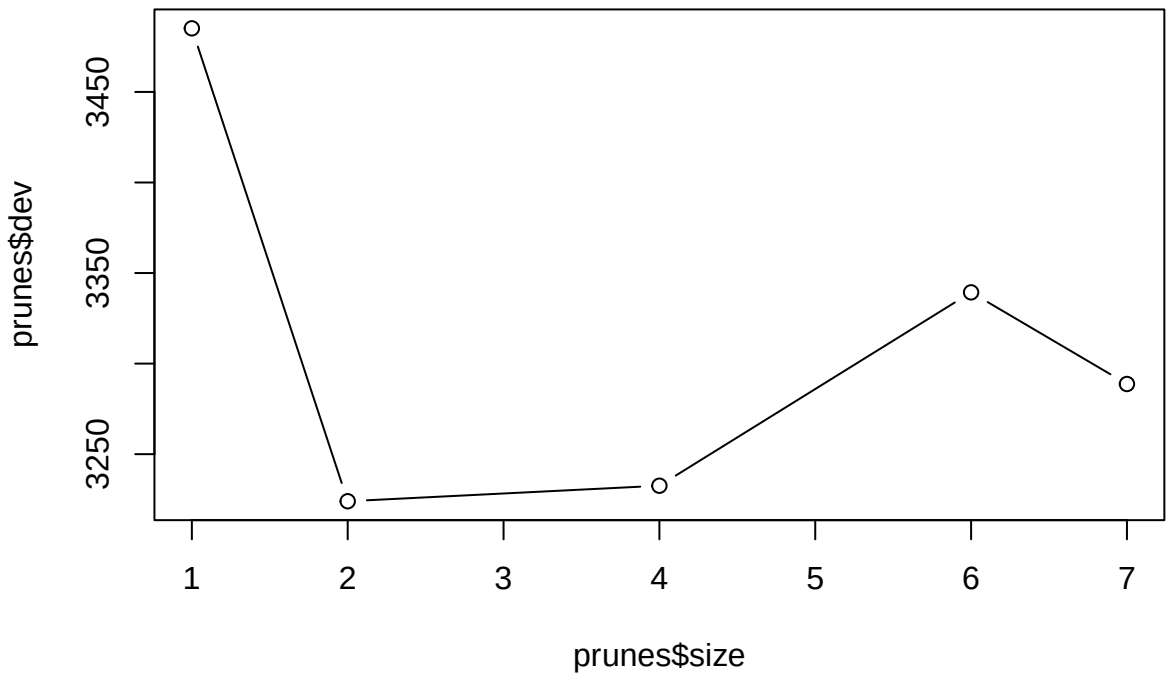
Reasoning for Pruning

For the Regression tree we used all plausible predictors that remained after our data cleaning including **reports**, **age**, **income**, **owner**, **selfemp**, **majorcard**, **dependents**, **months**, and **active** to predict the cube root of **expenditure**.

The pruned regression tree has a slightly lower test MSE that the tree that is a result of Recursive Binary splitting. Because we are concerned with accurately predicting applicant expenditure, a lower test MSE is key. The Output of the summary() function is below.

Pruned Tree summary() output

```
##
## Regression tree:
## tree(formula = expenditureL ~ ., data = trainR)
## Variables actually used in tree construction:
## [1] "income"     "months"     "age"        "dependents"
## Number of terminal nodes:  7
## Residual mean deviance:  4.065 = 2882 / 709
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -5.69200 -1.30700 -0.02962  0.00000  1.30400  7.76300
```



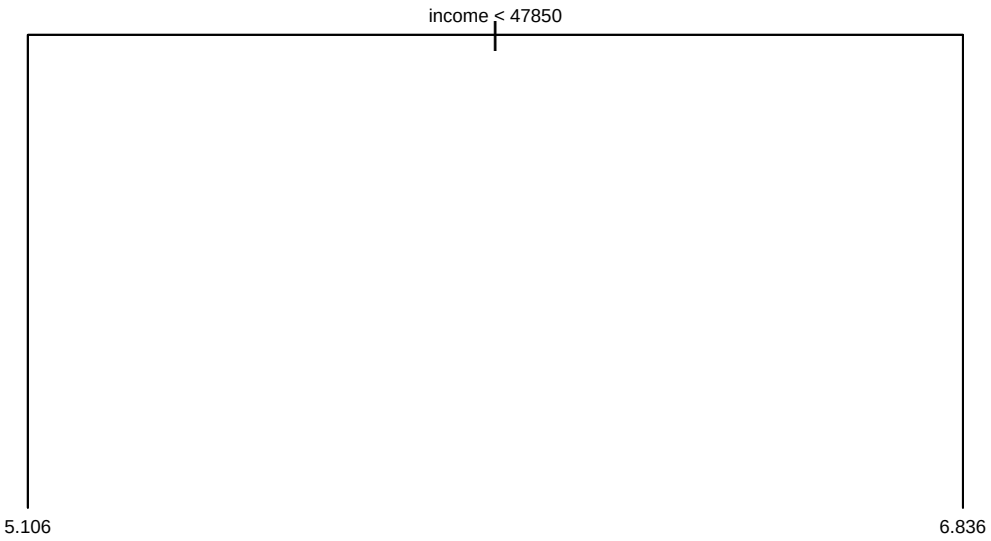

```
## [1] 2
```

```
##
## Regression tree:
## snip.tree(tree = exp.tree, nodes = 3:2)
## Variables actually used in tree construction:
## [1] "income"
## Number of terminal nodes:  2
## Residual mean deviance:  4.431 = 3164 / 714
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -6.83600 -1.27900 -0.05579  0.00000  1.33200  7.74400
```

Terminal Nodes

There are 2 terminal nodes in pruned tree, compared to 7 terminal nodes in the original which included **income**, **months**,**age**, **dependents**. The variable used in the pruned tree is **income** as seen in the graphical summary below. These classification tree allows us to understand which predictor(s) are most significant in helping to predict an applicant's credit card expenditure. This tree shows that **income** is the most important predictor in determining **expenditure**, while also combating overfitting on the train data.

Graphical Output



This regression tree, shows that **income** is the most important factor in determining an applicant's credit card expenditure of all the explanatory variables.

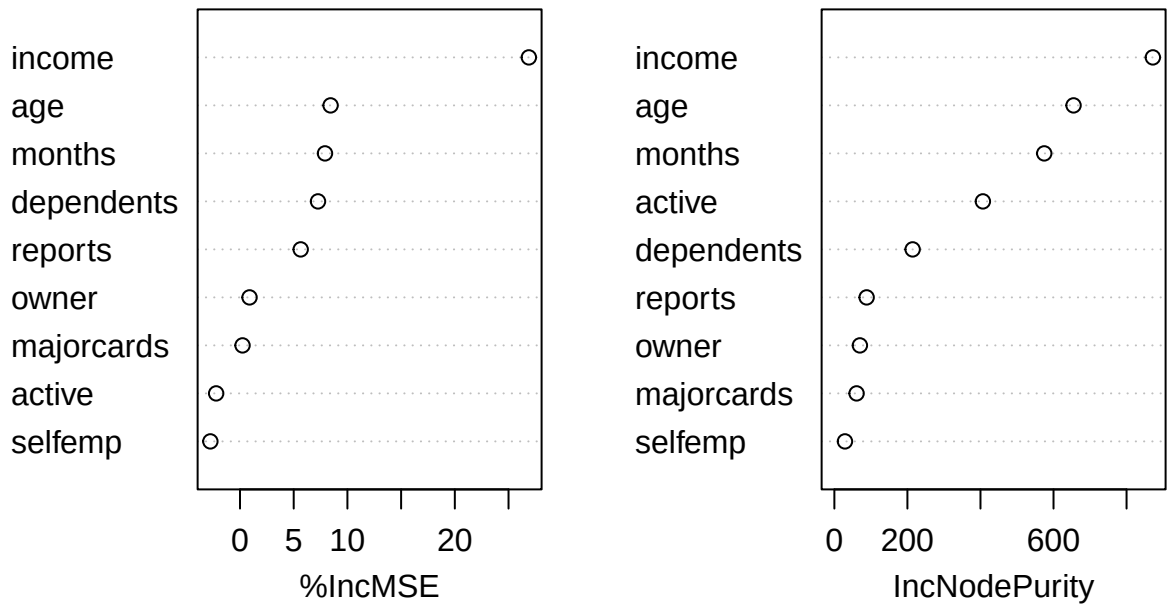
Test MSE for pruned tree

```
## [1] 4.246427
```

The test MSE of the pruned tree is 4.246427.

Random Forest varImpPlot() function

```
## [1] 4.377839
```



In random forest with m=3, the predictors that were found to be most important were **income**, **age**, and **month**. And the test MSE was 4.377839.

Summary of Findings

Test MSE Comparison Table

```
##   ols.mse ridge.mse lasso.mse prune.tree.mse   rf.mse
## 1 4.72794  4.730406  4.694509       5.639214 4.377839
```

MSE Commentary

The values of all the test MSEs indicate strong performance from each of our regression methods. As seen in the tree and shrinkage methods, none of the methods produced a test MSE that exceeded 5.0 with the cube root transformation. This small size (relative to the transformed response variable) indicates that these regression methods provided reliable results when addressing our question of interest.

Findings & Question of Interest

We aimed to explore the prediction of average monthly expenditure using a set of predictor variables. This is important to our question of interest as, it gives credit card companies models and trees that can help them better understand the effect of different applicant predictors on potential expenditure after they are approved for a credit card. Overall, all of the models predict a transformed response variable and provide estimated coefficients that show the direction and weight of the variable's influence. Looking at the results of Lasso

Regression, the predictors left in the model are **reports**, **age**, **income**, **dependents**, **majorcards**, and **active** after minimizing the coefficients. Futhermore, based off of the pruned regression tree, it was found that **income** was the most significant predictors in determining **expenditure** for applicants. This gives us a better understanding of which predictors are most influential in predicting an applicants expenditure.

Best Method for Question of Interest

In terms of shrinkage methods Lasso Regression had the best fit of our data with the lowest test mean square error. This result is not surprising because we had only a few important predictors in the model, we eliminated about half of the available predictors and lasso performs best when only a few factors are significant. The variables this model deemed important were **reports, age, income, majorcards, and active. It is surprising that an applicant's age and the months an applicant has been living at a current address had more importance in predicting monthly expenditure than variables like self employment and dependents, especially if an applicant is supporting more dependents you would assume they pay more monthly for the more people they support.

Overall, looking at the result of Prunning and Random Forests, the Pruned Regression Tree seems to be most helpful in making sense of the data and offering credit card companies guidance on making sense of applicants potential spending habits. The pruned tree has the lowest MSE but is very simplified, suggesting that income is the most important predictor and should be the only one considered when predicting one's expenditure.

If a credit card company is looking for a model to determine approved applicants spending Lasso Regression is best in the context of the question of interest, however if they are looking for guidlines to implement in decision-making processes trees are a best.

Addressing Previous Comments

In terms of addressing previous comments all formatting comments have been resolved and the original histogram with the non-transformed response variable has been included for reference and justification for the transformation.

Due to the nature and formatting of the data, this analysis was conducted on the 1023 observations/applicants that were approved for credit cards. Because the credit card company that populated this data set could not further track applicant expenditure if they were denied a card only approved applicants have populated **expenditure** values. This leaves this analysis only interpretable for approved applicants and fails to consider denied applicants. It was recommended that a different subset of the data was used to run the regression portion of the project. However with further EDA, it was found that reasonable subsets of the data including applicants with 2+ derogatory reports, or above average incomes created insufficiently small subsets of the data (72 and 385) which significantly decrease the size of the data frame at the cost of including only a few denied applicants.

Furthermore, when contextualizing this decision, it makes significantly more sense that credit card companies would only be concerned with the spending habits of approved applicants who are using their credit card to make purchases. Once a decision is made to deny an applicant, credit card companies no longer have interest in learning about their potential spending habits that they could default on as it is a waste of resources and time. In context, this analysis can be implemented after the approval process is mainly completed to potentially predict overall company profits and further trends in spending etc.

In further iterations of this project and model building, a subset of the data that could better predict expenditure of denied and approved applicants is applicants with over 3 active credit cards, and above average monthly income in 1992 which is around $2158.58 according to the Bureau for Labor Statistics. This leaves 131 denied applicants, and 624 approved applicants in the data set. These characteristics allow for us to make decisions about credit card applicants overall, with the trade off that we would be looking at higher income and previously pre-approved applicants. However, in the context of our question of interest, we decided that only looking at approved applicants was more useful for potential stakeholders.

```
##
##  no yes
## 178 757
```

Classification Question

Executive Summary

The dataset consists of over one thousand observations on different variables relating to credit history and demographics. Using this data, the question of interest for our group regards whether a given individual is approved for a credit card. To do this, we wanted to create a model that accurately classifies whether said individual is approved for a credit card, given the available data. Our hope was to obtain valuable insights about consumer behavior through the exploration of average monthly credit card expenditure.

We believe that this question is worth exploring because this information could allow financial institutions to better understand and anticipate their customers' spending habits. For example, exploring the prediction of average monthly credit card expenditure could provide valuable insight into customer behavior and streamline approval processes. These institutions can also use this knowledge to identify high-risk individuals or potential defaults by flagging unusual spending patterns before accepting an applicant who may default on their credit. Additionally, individuals looking to apply for a credit card may find value in these findings as they highlight the areas for which an individual can improve their financial habits. Ultimately, stakeholders like banks, brokerage firms, credit unions, and a number of additional financial institutions may find great value in the answers to our research question.

After conducting extensive analysis via classification methods into this dataset, it was revealed that the number of derogatory reports an applicant previously received was the biggest determining predictor in terms of classifying an applicant as approved for a credit card or not. More specifically, any one individual with more than one outstanding derogatory report had a very minimal chance of being approved. Additionally, our analyses indicated that the variable active, which is the number of active credit accounts associated with any one applicant, played a significant role in classification. In general, individuals with several active credit accounts were seldom classified as being denied a credit card, which makes sense when considering prior experience with credit accounts should have a strong positive correlation with future approval.

Stakeholders such as financial institutions and large banks may want to implement findings like these into their preexisting information systems. As AI/machine learning continues to expand and become more intertwined with the financial world, it is recommended that these institutions implement such findings to improve operational efficiency and streamline company workflows. Furthermore, everyday citizens - specifically those planning to apply for a credit card in the near future - may want to internalize these results and consider actions they can take in their own life to mitigate the probability of credit card rejection (i.e. avoid derogatory reports, maintain multiple credit accounts). Although factors such as age and income are difficult for any one individual to change, it is important to be aware of the controllable adjustments one can make as it regards to their day-to-day spending habits.

Variable Description for Classification

Using the variables in the AER CreditCard data set, **card** was used as the response variable, indicating the approval status of the applicant. The following variables were used in classification techniques to answer the question of interest:

Table 2. *Data dictionary of variables used in EDA and further Classification Modeling*

Variable Name	**Description**	**Type**
card**	Indicates if an application is approved for a credit card	Categorical- levels: yes, no
reports	Indicates the number of major derogatory reports against an applicant	Quantitative
age	Age of the applicant in years plus twelfths of a year (number of months)	Quantitative
income	Yearly income of the applicant in USD	Quantitative
dependents	Number of dependents supported by the applicant	Quantitative
months	Months applicant has been living at current address	Quantitative
active	Number of active credit accounts associated with the applicant	Quantitative
majorcard	Indicates if the applicant has any other major credit cards	Categorical- levels: yes, no
owner	Indicates if an applicant owns their home	Categorical- levels: yes, no
selfemp	Indicates if an applicant is self-employed	Categorical- levels: yes, no

** indicates the response variable

Very similarly to Regression Question, the data required slight modifications and cleaning in order to be used for analysis. First the **income** column was multiplied by a factor of 10000 to keep units regulated across all monetary based variables for the sake of interpretability, all monetary values are now in dollars. We also removed **share** from the predictors because it is a function of **expenditure** and **income**, as there are concerns about covariance and interaction becoming issues in the models. Similarly, **expenditure** was removed as it is a perfect preditor of **card** becuase the data set is formatted to have unapproved applicants have 0 dollars in expenditure. Following that, all of the categorical variables that were originally coded as string variables were changed to factor variables. When previously considering Linear Discrimination Analysis, we removed the **owner** and **selfemp** variables from the predictor variables as they're categorical and Linear Discriminant Analysis is unable to properly account for them, however we considered them in Logistic Regression and they are included in this analysis.

Classification

Exploratory Data Analysis

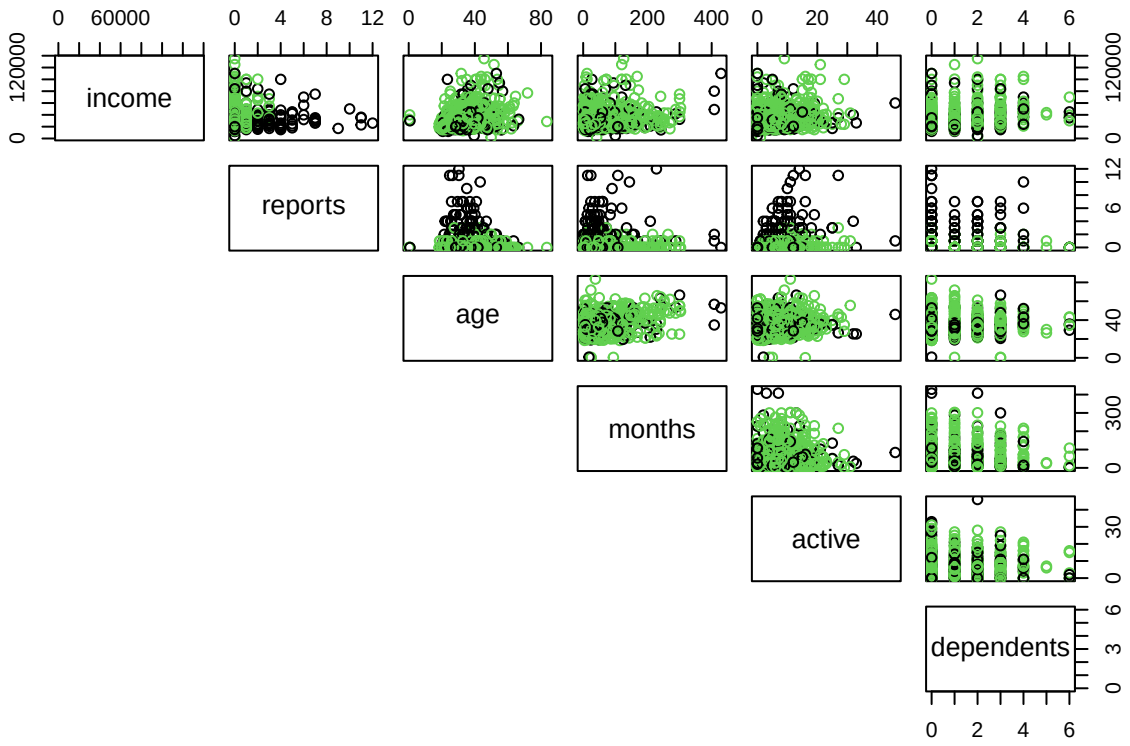
Next, graphical summaries on the training data were conducted to learn more about the relationships between the predictors and response classification levels. The graphs and their interpretations can be seen below.

```
##
##  no yes
## 211 712
```

```
##
##        no       yes
## 0.2286024 0.7713976
```

First, to get a better sense of the training data a proportion table was created to understand the distribution of approved and unapproved applications. We have about a 20/80 split of denied applicants to approved applicants in our training data set.

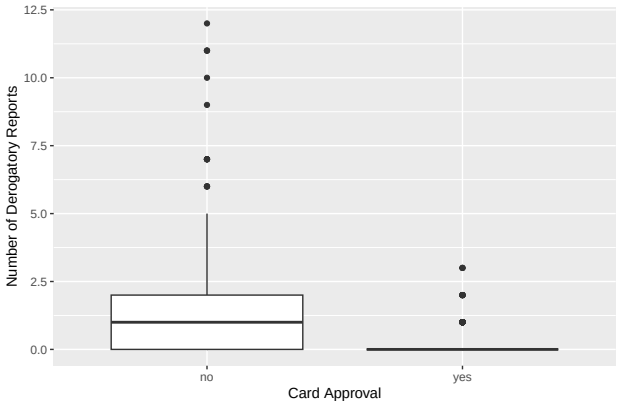
Scatter Plot Matrix Next a scatter plot matrix of the quantitative predictors was run to better understand the relationship between the variables and their associated outcome.



The scatter plot matrix of the quantitative predictors for our LDA analysis doesn't show the greatest separation between the approved and denied credit card applicants; this may lead to some error or difficulty in our classification methods. The plots for the variable **reports** show denied applicants as generally having

more derogatory reports than approved applicants. The plots for **income** show a little more approved applicants at higher values of income. Based on the scatter plot matrix, box plots are included below to further investigate certain variables that are of particular interest and promise to our question of interest.

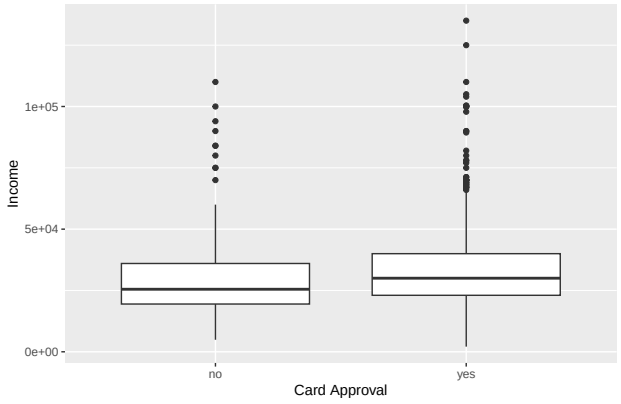
Figure 2: Distribution of Number of Reports by Card Approval



Boxplots of Interest /newline

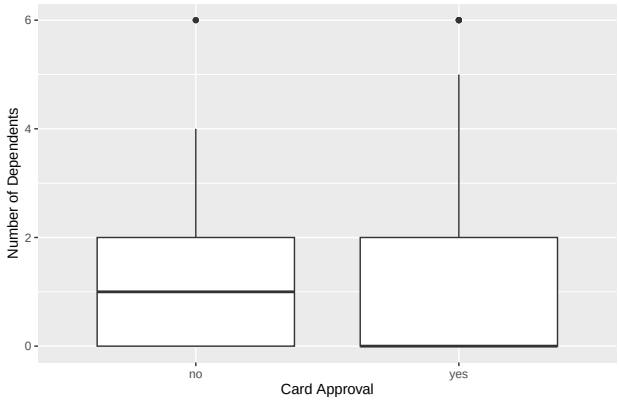
From the box plots, applicants who are approved for a credit card have a much lower number of derogatory reports than those that do not. There is a large difference in the ranges of values between those approved and those not. These differences in the distribution might suggest that **reports** is a plausible predictor for card approval or not.

Figure 3: Distribution of Income by Card Approval



From the box plots, applicants who are approved for a credit card have a higher average income than those that did not. The range of incomes is larger for approved applicants. These differences in the distribution might suggest that **income** is a plausible predictor for card approval or not.

Figure 4: Distribution of Number of Dependents by Card Approval



From the box plots, applicants who are not approved for a credit card have a higher number of dependents than those that did not. The range of dependents is larger for approved applicants. These differences in the distribution might suggest that **dependents** is a plausible predictor for card approval or not.

Logistic Regression Model

Predictors

In all of the tests for credit card applicants approved and denied, the predictors are not consistent with the assumption of multivariate normality for Linear Discriminant Analysis. The p-values for all the tests for kurtosis and skewness were significantly less than .05, so the null hypothesis that the data is consistent with the multivariate normal distribution is rejected. Thus interpreting LDA output must be done carefully and with the assumptions in mind, for this reason the conclusion was made that a Logistic Regression Model would be the best fit for the data set.

For the logistic regression model, we used **reports**, **age**, **income**, **dependents**, **months**, **active**, **owner**,**selfemp**, and **majorcards** to predict the binary response **card**.

Logistic Regression Model Performance & summary() output

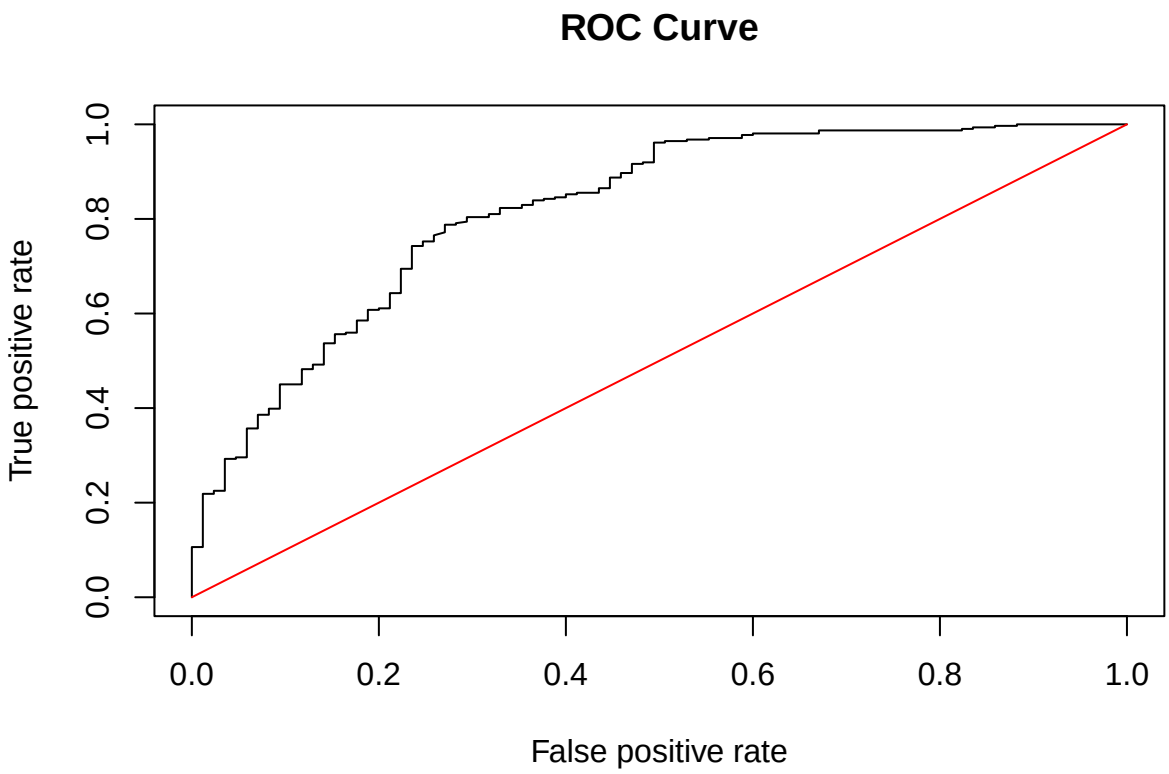
```
##
## Call:
## glm(formula = card2 ~ income + dependents + reports + age + months +
##     active + owner + selfemp + majorcards, family = binomial,
##     data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4381   0.1589   0.4113   0.5920   2.2146
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.014e-01  4.069e-01   0.495 0.620674
## income       2.979e-05  8.310e-06   3.585 0.000337 ***
## dependents  -3.108e-01  8.441e-02  -3.683 0.000231 ***
## reports     -1.725e+00  1.652e-01 -10.444  < 2e-16 ***
## age         -3.617e-03  1.255e-02  -0.288 0.773165
## months      -1.527e-03  1.803e-03  -0.847 0.396907
## active       1.167e-01  2.302e-02   5.072 3.93e-07 ***
```

```
## owneryes     5.218e-01  2.393e-01   2.180 0.029232 *
## selfempyes  -2.266e-01  3.905e-01  -0.580 0.561697
## majorcards1  6.685e-01  2.289e-01   2.920 0.003497 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 992.38  on 922  degrees of freedom
## Residual deviance: 684.53  on 913  degrees of freedom
## AIC: 704.53
##
## Number of Fisher Scoring iterations: 6
```

After running the Logistic Regression with all of the variables, **age**, **selfemp**, and **months** were found to be insignificant, they were removed and the Logistic regression was run again and all of the remaining predictors were found to be significant. The summary statistics are shown above.

```
##
## Call:
## glm(formula = card2 ~ income + dependents + reports + active +
##     owner + majorcards, family = binomial, data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4316   0.1629   0.4121   0.6015   2.2491
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.408e-02  2.849e-01   0.190 0.849421
## income       2.837e-05  8.113e-06   3.496 0.000471 ***
## dependents  -3.073e-01  8.321e-02  -3.693 0.000222 ***
## reports     -1.740e+00  1.657e-01 -10.498  < 2e-16 ***
## active       1.164e-01  2.293e-02   5.075 3.88e-07 ***
## owneryes     4.538e-01  2.255e-01   2.013 0.044149 *
## majorcards1  6.763e-01  2.282e-01   2.964 0.003035 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 992.38  on 922  degrees of freedom
## Residual deviance: 686.19  on 916  degrees of freedom
## AIC: 700.19
##
## Number of Fisher Scoring iterations: 6
```

All of the remaining predictors werre significant so the improved Logistic Regression Model including **income**, **dependents**, **reports**, **active**, **owner**, and **majorcards** was then assessed on its performance




```
## [[1]]
## [1] 0.8196141
```

The ROC curve shows that the model is better than random guessing because the curve is above the red line which is representative of random guessing. They are also very similar suggesting the model does well at predicting the outcome of **card**. The AUC for Logistic Regression is 0.8196141 this value is closer to 1 & it indicates that the model does better than random guessing in classifying observations.

Classification Trees

Reasoning

In the Initial Milestone, the pruned classification tree had the lowest false positive rate and a relatively manageable overall error rate. With these results and the ability to correct for overfitting on the test data is was decided that the pruned tree was the most reasonable option to present in the context of the Question of Interest. When choosing if an application should be approved or not trees are particularly helpful as they act as a guide to follow n evaluation, but they an quickly get confusing so a more simple yet powerful version (pruning) is key for making sucessful decisions.

Output from the summary() function for the tree

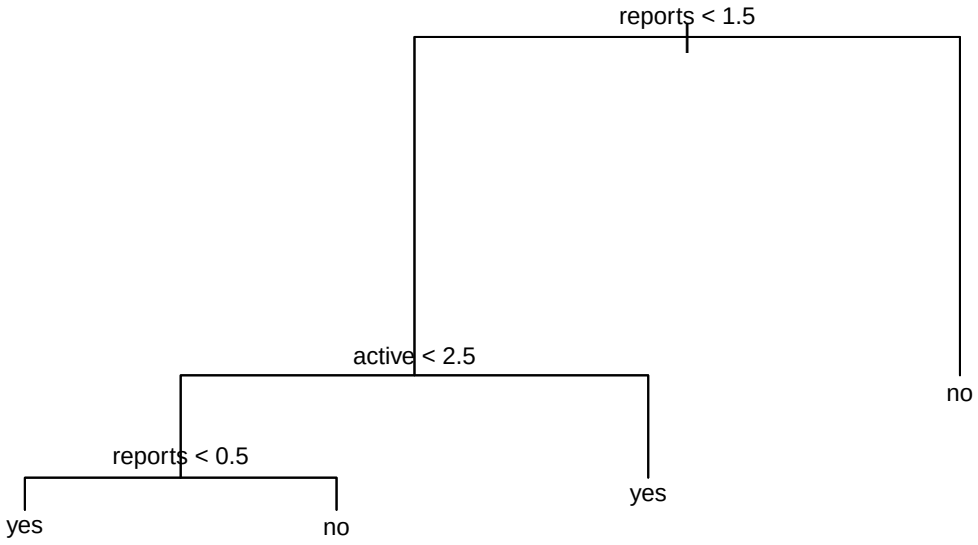
```
## [1] 4
```

```
##
## Classification tree:
## snip.tree(tree = tree.class.train, nodes = c(3L, 5L, 8L))
## Variables actually used in tree construction:
## [1] "reports" "active"
## Number of terminal nodes:  4
## Residual mean deviance:  0.8111 = 745.4 / 919
## Misclassification error rate: 0.1517 = 140 / 923
```

Terminal Nodes

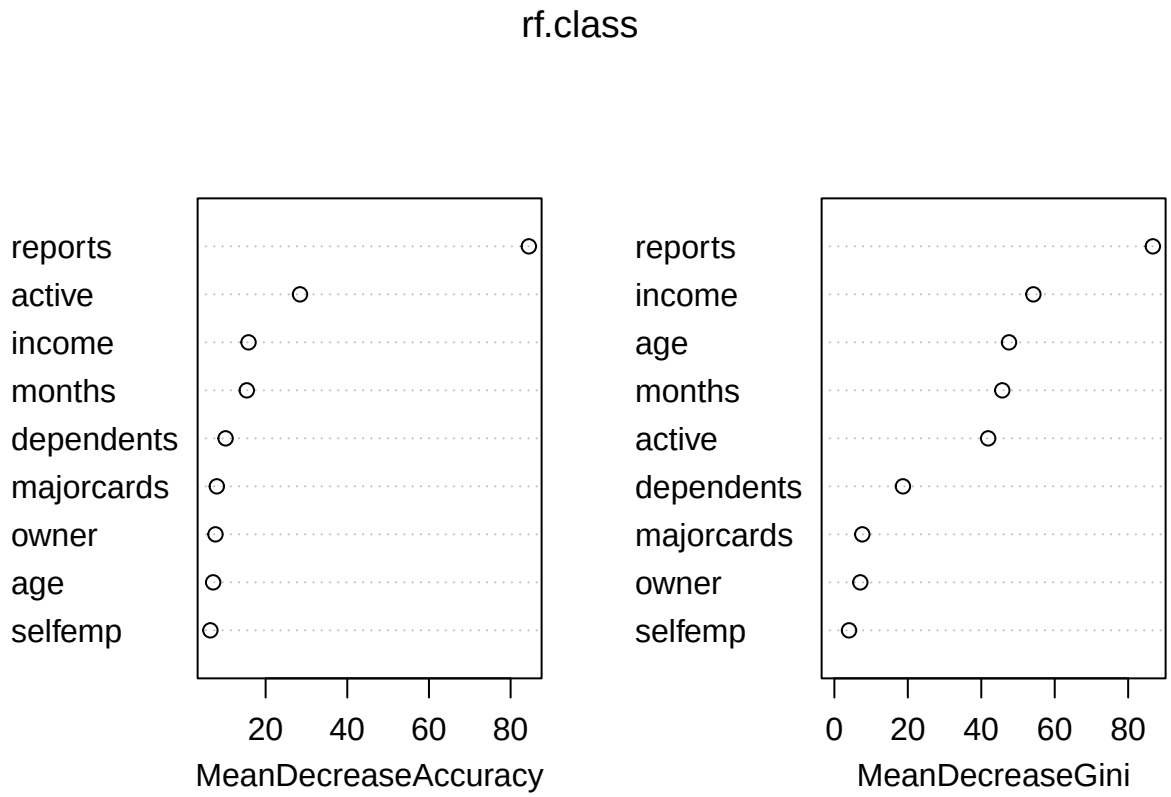
The Classification tree with pruning has 4 nodes.

Graphical output of the Regression Tree



This tree shows **reports**,and **active** as the most important predictors. if an applicant has more that 1.5 derogatory reports the tree suggests they should be denied for previous bad behavior. If they have less than 1.5 derogatory reports having other active credit cards (meaning they were previously approved by another company) is the next best indicator.

Random Forest varImpPlot() function output



Off Random Forests, **reports**, **income**, and **active** were found to be the most influential variables in indicating if an applicant is approved.

Summary of Findings

Confusion matricies at 0.5 Threshold

Logistic Regression:

```
##
##     FALSE TRUE
##   0    35   50
##   1     8  303
```

Pruned Classification Tree:

```
##        tree.pred.prune
## y.test3  no yes
##     no   33  52
##     yes   6 305
```

Random Forests:

```
##        rf.pred
## y.test3  no yes
##     no   41  44
##     yes  14 297
```

Table of Test Error, FPR & FNR Based on a 0.5 Threshold

Findings at 0.5 Threshold

Model Type	Test Error Rate	False Positive	False Negative
Logistic Regression	0.1464646	0.5882353	0.02572347
Classification Tree	0.1464646	0.6117647	0.0192926
Random Forests	0.1464646	0.5176471	0.04501608

Threshold Adjustment

Based on the context of our Question of Interest, a False Positive would entail approving an applicant that would not be a suitable candidate for a card and may default on credit payments, leading the company to lose revenue. On the other hand, a false negative would be denying a worthy candidate which does not intrinsically put the credit card company in danger of extending credit to someone who will default on loans. A false negative on the other hand would entail not approving someone who was qualified for a credit card. In this context the goal is to minimize the false positive rate and be overly cautious. The original 80/20 split in the training and test data contributes to this issue as well. Thus, the error rate threshold is increased and new false positive and false negative rates are found:

Confusion matricies at 0.8 Threshold

Logistic Regression:

```
##
##     FALSE TRUE
##   0    62   23
##   1    68  243
```

Pruned Classification:

```
##
## y.test3 FALSE TRUE
##     no     64   21
##     yes    84  227
```

Random Forests:

```
##
## y.test3 FALSE TRUE
##     no     59   26
##     yes    85  226
```

Table of Test Error, FPR & FNR based on a 0.8 Threshold

Findings at 0.8 Threshold

Model Type	Test Error Rate	False Positive	False Negative
Logistic Regression	0.229798	0.2705882	0.2186495
Classification Tree	0.2651515	0.2470588	0.2700965
Random Forests	0.280303	0.3058824	0.2733119

Increasing the threshold raises the False Negative Rate significantly but not to too concerning of a level, while minimizing the false positive rate and controlling for the uneven distribution in our original data set.

Finding and Question of Interest

The findings from the classification methods help us to understand how the models trained on the training data perform on classifying the credit card approvals using the test data. The goal of our question is to create a model that best classifies approved applicants, as we are taking the perspective of the creditor and want to mitigate our risk when providing credit. The improved logistic model gives estimated coefficients that show the predictors' relationship with the response variable. The coefficients are the increase in log odds of the response given a one unit increase in the predictor, which can be converted to probability by taking the exponential of the log odds and simplifying.

Best Method for Question of Interest

The model that best answers our question is the Classification tree created through pruning because it has the lowest False Positive Rate(FPR) of 24.71% at the increased threshold of .8. This model only has 4 nodes but still contains the important predictors reports and **active**. From this tree we can predict classifications for new observations based on their values for **reports** and **active**. Specifically, given this tree a **reports** value of greater than 1.5 would automatically result in a classification of no. Otherwise, if the observation's value for **active** is greater than 2.5 they are classified as yes, if not their **reports** need to be less than .5 to be classified as yes. All other values, **reports** greater than .5 and less than 1.5 are classified as no.

The formatting of the Tree also allows creditors a sort of decision tree that they can follow, which is a helpful visual tool.

Addressing Previous Comments

We were able to make the necessary adjustments to resolve the issues that were previously plaguing our report. We made the necessary adjustments to the error rate threshold for the confusion matrices, and we feel as though our analysis is contextualized in the context of our question of interest. Categorical predictors were included as potential predictors for out improved logistic model, some of which were statistically significant and contributed to a stronger model. And finally, the significance of the multivariate normality assumption was not met, so we chose to further investigate our Logistic Regression Model as it performed better overall in our Milestone 4 and the assumptions of the model are met.

Further Work

If we had more time to work on this project, we would have like to have explored more potential explanatory varibles that were not included in the original data set, and employ some degree of unsupervised learning to discover interesting characteristics with those new variables if approval status or expenditure values were not available.

Because our data set is built into an existing R package it would be challenging to be able to find more data for these specific individual applicants that would be reliable. However, using our existing analyses as a foundation and conducting a more substantial data collection to find more potential predictors like credit score and mortgage payment among others could allow us the opportunity to run Principal Component Analysis to better understand groupings of application characteristics. Furthermore because the Data set is somewhat older, we think there could be more modern predictors that could be of use because they are better indicators of peoples lifestyle and habits today. Overall, while this analysis of this data set is interesting in substantial it could be modernized to better reflect modern algorithms that are used by credit card companies to make decisions about applicants.

Reflecting on Learning

As a group, this project has helped reinforce our learning in this class by allowing the opportunity for us to get hands on experience with the methods and models we have learned about in class. While learning these topics in class has been rewarding and given us a plethora of new approaches to statistical machine learning and problem solving, many of the examples in class are streamlined to not include the outside noise that often comes with real world problem solving. This project provided us with challenges beyond just homeworks, quizzes, and labs because we were having to make decisions about imperfect real world data as issues arose throughout the milestones. The methods that we explored through this project solidified our understanding of the course content, but arguably, the Data manipulation, exploratory data analysis, and applied decision making allowed us to better understand these models im a deeper way.

Furthermore, working in a group with new people challenges us to grow our team working skills, and mimicked a professional setting that many of us will be in after graduation, this project was great for implementing the collaboration, communication, and time management skills that are necessary for group woke to be successful. Each of us learned something about the ways we can best contribute to a groups, and ways that we can continue to grow.

Overall, this project allowed for us to get hands on experience with data that forced us to make the best critical real world decisions in the frame work of machine learning, where there were variables outside of our control. We as a group were able to grow our collaborations skills, get direct experience implementing machine learning frameworks, and grow our problem solving skills in the face of challenges that came up over the course of our project.