Group 22
Jack Griffin: uuf7tk
Joshua Fu: xsk6jc
Sarah Child: bee6tg
Jiayi Niu: nwb4td

Milestone 4 (Shrinkage Methods and Tree-Based Methods)

1. Introduction

a. **Regression Problem** : Does a higher FICO credit score lead to lower interest rates in loans granted by LendingClub.com, and to what extent do other financial factors such as debt-to-income ratio, annual income, and the number of recent credit inquiries impact this relationship?

Classification Problem: Does a higher FICO credit score make a borrower more likely to meet the credit underwriting criteria in loans granted by LendingClub.com, and to what extent do other financial factors such as revolving balance, annual income, debt to income ratio, and the number of recent credit inquiries impact this relationship?

b. Our regression problem is of particular interest because understanding the factors that will influence the interest rate can offer insights into the underwriting process of LendingClub.com. This is valuable for both borrowers and investors. On the borrower's side, they can work on improving specific financial metrics to secure loans at lower interest rates. For investors, understanding these factors can help in assessing the risk associated with different interest rates, thereby aiding investment decisions, potentially leading to better risk-adjusted returns.

Our classification problem is worth exploring because it functions as an invaluable risk assessment. For investors, this serves as a preliminary filter to identify potentially risky loans, thereby aiding in more informed investment decisions. For borrowers, understanding the factors that contribute to meeting these criteria can guide them in improving their financial standing. Moreover, for the platform itself, such a predictive model can streamline the loan approval process, making it more efficient and reliable.

c. The dataset for this study is sourced from LendingClub, a company that offers peer-to-peer personal and business loans, and it is publicly available on Kaggle. It contains various financial metrics and loan information for an array of customers and their loans, focusing on the lending process and borrower characteristics, and it also reveals (in a variable) whether each customer meets the credit underwriting criteria for the company.

d.

Variable description:

Variable Name	Description	Type
int.rate (response)	Interest rate of the loan as a proportion	Quantitative
installment	The monthly installments owed by the borrower if the loan is funded	Quantitative
log.annual.inc	The natural log of the self-reported annual income of the borrower	Quantitative
dti	The debt-to-income ratio of the borrower	Quantitative
fico	The FICO credit score of the borrower	Quantitative
days.with.cr.line	The number of days the borrower has had a credit line	Quantitative
revol.bal	The borrower's revolving balance	Quantitative
revol.util	The borrower's revolving line utilization rate	Quantitative
inq.last.6mths	The borrower's number of inquiries by creditors in the last 6 months	Quantitative
delinq.2yrs	Number of times borrower was 30+ days past due on a payment in the past 2 years	Quantitative
pub.rec	The borrower's number of derogatory public records	Categorical
not.fully.paid	If loan was fully paid	Categorical
purpose	Purpose of loan	Categorical

2. Exploratory Data Analysis for Regression Question

a. Data cleaning

The dataset was mostly well-structured upon download, eliminating the need for any major data cleaning prior to the exploratory data analysis. It was necessary to convert some categorical variables into factors. For example, pub.rec, a variable indicating a borrower's number of derogatory public records, was initially considered numeric. Given its categorical nature and potential non-linear relationship with the interest rate, we decided to treat it as a categorical variable and convert it into a factor. Pub.rec initially took on the values 0, 1, 2, 3, 4, and 5, but upon closer inspection, it only took on the values 4 or 5 two times in our data set of 9,576 observations. To prevent the introduction of noise and potential bias into our predictive model due to their sparse presence, we excluded the two observations where pub.rec equals 4 or 5. Similarly, it was also necessary to convert the variable not.fully.paid, which indicates if a borrower did not fully pay back their loan, into a factor, as this variable should be treated as categorical. These steps were necessary to create the boxplots shown in our exploratory data analysis, and these steps were also necessary for the construction of our models in later sections.

b. Graphical summaries

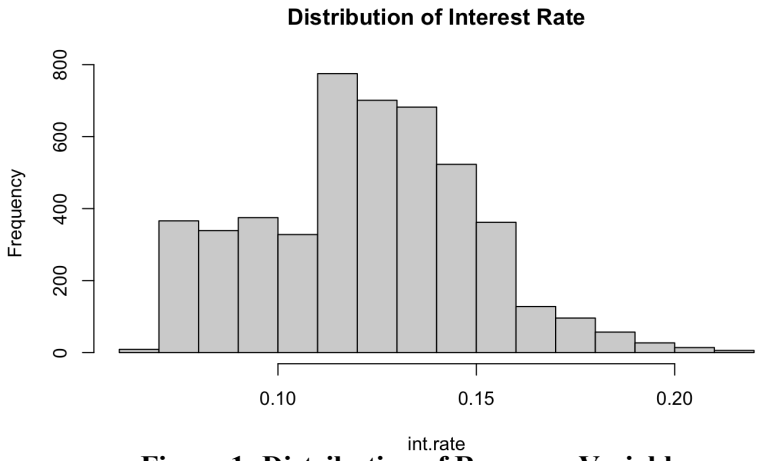

Figure 1: Distribution of Response Variable

	int.rate	installment	log.annual.inc	dti	fico	days.with.cr.line	revol.bal	revol.util	inq.last.6mths	delinq.2yrs
int.rate	1.000	0.273	0.059	0.223	-0.706	-0.137	0.092	0.466	0.196	0.156
installment	0.273	1.000	0.445	0.046	0.100	0.184	0.238	0.083	-0.022	0.000
log.annual.inc	0.059	0.445	1.000	-0.069	0.120	0.328	0.348	0.058	0.029	0.031
dti	0.223	0.046	-0.069	1.000	-0.248	0.054	0.188	0.334	0.030	-0.023
fico	-0.706	0.100	0.120	-0.248	1.000	0.271	-0.026	-0.546	-0.176	-0.217
days.with.cr.line	-0.137	0.184	0.328	0.054	0.271	1.000	0.232	-0.024	-0.035	0.069
revol.bal	0.092	0.238	0.348	0.188	-0.026	0.232	1.000	0.209	0.019	-0.034
revol.util	0.466	0.083	0.058	0.334	-0.546	-0.024	0.209	1.000	-0.010	-0.043
inq.last.6mths	0.196	-0.022	0.029	0.030	-0.176	-0.035	0.019	-0.010	1.000	-0.008
delinq.2yrs	0.156	0.000	0.031	-0.023	-0.217	0.069	-0.034	-0.043	-0.008	1.000

Figure 2: Correlation Matrix of Quantitative Predictors

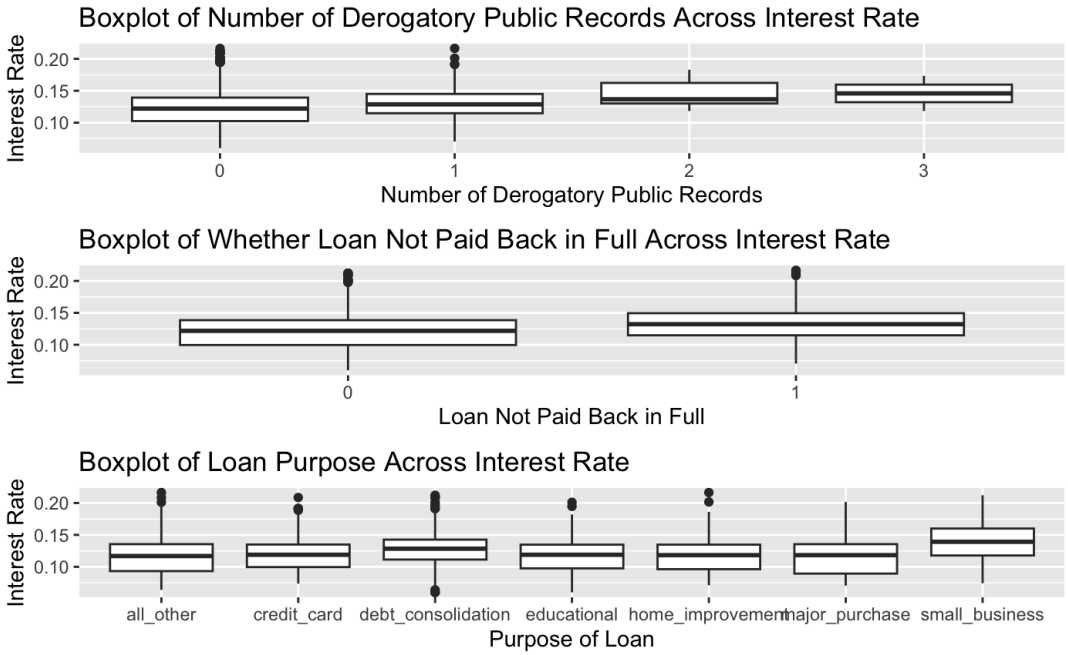

Figure 3: Boxplots of Interest Rate Across Categorical Predictors

c. Figure 1 displays the distribution of our response variable, int.rate. Since this milestone involves shrinkage and tree-based methods, our response variable should have an approximately normal distribution. From Figure 1, it can be seen that the distribution of our response variable is roughly normal, which means that it is appropriate to apply varying types of regression to our data as desired.

Figure 2 was created to provide insight into the relationship between our quantitative predictors, as well as the relationship between all of our variables and the response variable, int.rate. Exploring the relationships between quantitative predictors in our dataset is important because two predictors that are strongly correlated could potentially be related to one another, making it inappropriate to include both predictors in the model. Observing the correlations in Figure 2 assured us that this was not the case with any of our predictors, as none of our quantitative variables are highly correlated. In addition, the correlations in Figure 2 are useful for identifying the strength and direction of the relationships between our quantitative predictors and response variable. Understanding which predictors are highly correlated with int.rate can provide us with insight as to which predictors should be included in our regression model.

Figure 3 displays boxplots of the categorical predictors across the response variable. These boxplots compare the distribution of int.rate across pub.rec, not.fully.paid, and purpose, respectively. These boxplots allow us to see if the mean interest rate appears to be significantly different across various groups within each of these predictors. If a boxplot shows that the mean interest rate does appear to be significantly different across various groups within a certain predictor, this provides evidence that the predictor should be included in our regression model.

d. From our exploratory data analysis, it appears that fico is the only predictor that is strongly correlated with the response variable, int.rate. While it is intuitive that a borrower's FICO credit score could have a strong impact on the interest rate at which they are offered a loan, it is surprising that more predictors do not have a strong correlation with int.rate. For example, it would seem intuitive that a borrower who has been late on payments several times would receive a loan at a much higher interest rate than a borrower who has never been late on payments, but the correlation between delinq.2yrs and int.rate is only 0.156. Similarly, Figure 3 shows that the value of int.rate does not vary significantly across different levels within a categorical predictor. This could indicate that on their own, none of our categorical predictors are significant when predicting interest rates.

3. Shrinkage Methods

a. The steps that were taken to prepare the dataset for the exploratory data analysis carried over to the construction of the ridge and lasso regression models, so all of the data cleaning described in part (a) of the exploratory data analysis section would apply here as well.

b. Our analysis centers on the response variable int.rate, which signifies the loan's interest rate as a proportion. This rate reflects the risk level assessed by LendingClub.com, with higher rates typically linked to higher perceived risk. Our regression analysis seeks to discern the determinants that influence the interest rate assigned to a borrower. We will employ glmnet() for our regression analysis, which necessitates numerical predictors. Here is the revised list of predictors we will include:

- installment: Included because it reflects the borrower's regular payment obligations, which could influence the lender's interest rate due to its impact on the borrower's cash flow.
- log.annual.inc: Included because it is a log-transformed measure of income that can affect the interest rate by indicating the borrower's ability to service the loan.
- dti: Included as it represents the borrower's debt-to-income ratio, a key indicator of financial health that lenders use to determine interest rates.
- fico: Included because it is a direct indicator of credit risk, which is often inversely related to the interest rate offered.
- days.with.cr.line: Included because a longer credit history can suggest reliability and potentially lower interest rates.
- revol.bal: Included as it represents the borrower's outstanding balance on revolving accounts, which can impact their credit utilization and overall credit profile.
- revol.util: Included because it shows how much of the available credit the borrower is using, a high utilization can suggest higher risk and thus a higher interest rate.
- inq.last.6mths: Included as frequent inquiries can indicate credit-seeking behavior, which may lead to higher interest rates due to potential risk.
- delinq.2yrs: Included because a history of delinquency can significantly increase the interest rate as it indicates higher credit risk.

- pub.rec dummy variables: Included to capture the effect of derogatory public records on the interest rate, with each dummy variable representing a level of the original categorical variable.
- not.fully.paid dummy variable: Included because it captures the potential influence of a borrower's previous loan repayment behavior on the interest rate assigned to a new loan.

We have decided to exclude the purpose variable from our model despite its potential relevance. The reason for this exclusion is the complexity it introduces to the model due to its many levels (seven distinct categories). Including such a variable with numerous categories could lead to overfitting, especially if some categories have very few observations.

By focusing on these numerical predictors, we ensure compatibility with the glmnet() function and maintain the integrity of our regression analysis. For now, we proceed with the numerical variables that are most relevant to predicting the credit.policy outcome.

c. The chosen value of the threshold used in the glmnet() function was 10^{-23}. Using the default threshold with ridge regression, the estimated coefficients from ridge regression are different from the estimated coefficients from ordinary least squares regression.

```
## 14 x 2 sparse Matrix of class "dgCMatrix"
##                                           s0
## (Intercept)        4.580523e-01  4.580889e-01
## installment        4.394130e-05  4.394325e-05
## log.annual.inc    -8.958053e-04 -8.961261e-04
## dti                5.064837e-05  5.057550e-05
## fico              -4.855693e-04 -4.856140e-04
## days.with.cr.line  9.263501e-08  9.276456e-08
## revol.bal         -7.974015e-09 -7.964015e-09
## revol.util         5.959987e-05  5.957142e-05
## inq.last.6mths     9.489351e-04  9.488451e-04
## delinq.2yrs        4.783106e-04  4.775591e-04
## pub.rec1          -7.131262e-04 -7.137176e-04
## pub.rec2           2.465113e-03  2.463845e-03
## pub.rec3           2.662266e-03  2.660968e-03
## not.fully.paid1    1.675392e-03  1.674949e-03
```

Figure 4: Coefficients of the OLS model with Threshold at 10^{-7}

The glmnet() function uses a method based on coordinate descent in the minimization procedure, and the procedure stops when a certain threshold is met. The default is 10^{-7}, but the threshold might have to be smaller if multicollinearity is present. We tried a threshold of 10^{-23}, and refit the ridge regression, which resulted in estimated coefficients from ridge regression and ordinary least squares regression that were almost identical.

```
## 14 x 2 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)         4.580523e-01  4.580523e-01
## installment         4.394130e-05  4.394130e-05
## log.annual.inc     -8.958053e-04 -8.958053e-04
## dti                 5.064837e-05  5.064837e-05
## fico               -4.855693e-04 -4.855693e-04
## days.with.cr.line   9.263501e-08  9.263501e-08
## revol.bal          -7.974015e-09 -7.974015e-09
## revol.util          5.959987e-05  5.959987e-05
## inq.last.6mths      9.489351e-04  9.489351e-04
## delinq.2yrs         4.783106e-04  4.783106e-04
## pub.rec1           -7.131262e-04 -7.131262e-04
## pub.rec2            2.465113e-03  2.465113e-03
## pub.rec3            2.662266e-03  2.662266e-03
## not.fully.paid1     1.675392e-03  1.675392e-03
```

Figure 5: Coefficients of the OLS Model with Threshold at 10^{-23}

d. **i.** Based on 10-fold cross-validation on the training data, the optimal tuning parameter is $\lambda = 0.001883365$.

ii. The below plot shows the estimated test MSE against log(λ).

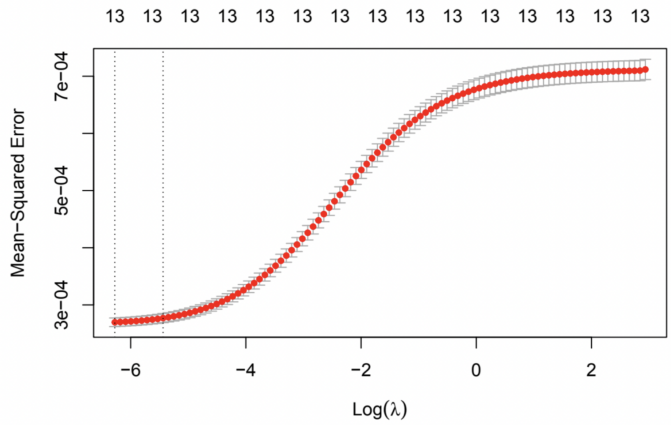

Figure 6: Estimated Test MSE Against Log(λ) of Ridge Regression

iii. Based on the value of λ chosen by cross-validation, there were 13 predictors left in the model. The nature of ridge regression is that all predictors are left in the model.

iv. The predictors that were left in the model are:

- installment
- log.annual.inc
- dti
- fico
- days.with.cr.line

- revol.bal
- revol.util
- inq.last.6mths
- delinq.2yrs
- pub.rec1
- pub.rec2
- pub.rec3
- not.fully.paid1

v. Based on the value of λ chosen by cross-validation, the actual test MSE is 0.0002663066.

e. **i.** Based on 10-fold cross-validation on the training data, the optimal tuning parameter is $\lambda = 4.453197 * 10^{-5}$.

ii. The below plot shows the estimated test MSE against log(λ).

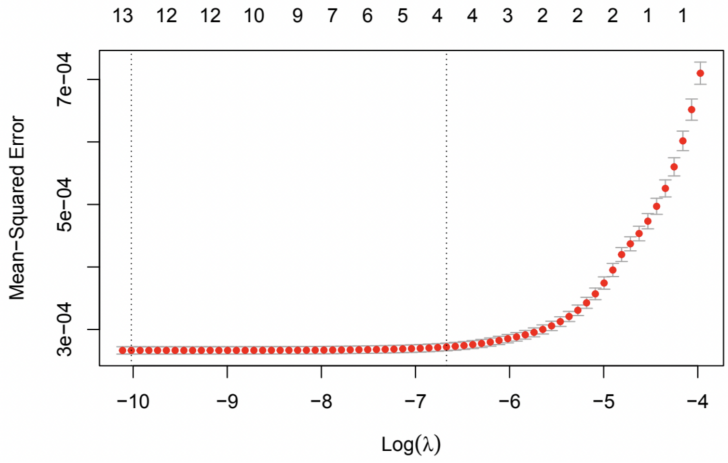

Figure 7: Estimated Test MSE Against Log(λ) of Lasso Regression

iii. At the optimal value for the tuning parameter, there are 13 predictors in the model, and this number will decrease to 1 if λ keeps increasing.

iv. The predictors that were left in the model are:

- installment
- log.annual.inc
- dti
- fico
- days.with.cr.line
- revol.bal
- revol.util
- inq.last.6mths
- delinq.2yrs
- pub.rec1

- pub.rec2
- pub.rec3
- not.fully.paid1

v. Based on the value of λ chosen by cross-validation, the actual test MSE is 0.0002623285.

f. The test MSE with ordinary least squares regression is 0.0002624469.

g. **i.** Conclusions:

Table 1: Regression Methods vs Test MSE

Method	Test MSE
OLS	0.0002624469
Ridge Regression	0.0002663066
Lasso Regression	0.0002623285

Table 2: Regression Coefficients vs Regression Methods

Variable	OLS	Ridge	Lasso
Intercept	4.454562e-01	4.035612e-01	4.451401e-01
installment	4.425298e-05	3.986583e-05	4.387682e-05
log.annual.inco	-3.034306e-04	4.603604e-05	-2.162741e-04
dti	6.986956e-05	1.019640e-04	6.468986e-05
fico	-4.764898e-04	-4.228421e-04	-4.770450e-04
days.with.cr.line	-4.167793e-08	-2.302713e-07	-2.905039e-08
revol.bal	-1.922623e-08	-1.537157e-08	-1.749391e-08
revol.util	5.908547e-05	8.904752e-05	5.723178e-05
inq.last.6mths	9.891708e-04	1.047089e-03	9.672777e-04
delinq.2yrs	5.709248e-04	1.415668e-03	4.734329e-04
pub.rec1	-1.316600e-03	-4.216380e-04	-1.119951e-03
pub.rec2	6.091018e-03	6.711963e-03	4.988505e-03

pub.rec3	2.935671e-03	4.554293e-03	6.965944e-04
not.fully.paid1	2.127164e-03	2.558923e-03	2.036276e-03

From Table 1, we can observe that the test MSEs are very close, with lasso regression having the lowest MSE, followed by OLS and then ridge regression. These results suggest that all three models have a strong predictive performance, with lasso regression offering a slight advantage. It is important to note that the small MSE values in our models are partly due to the nature of our response variable. Interest rates are constrained within a narrow range, and thus are inherently less volatile than other variables, such as income. It is necessary to address this fact to appropriately understand and evaluate our model performance.

ii. A key observation from our analysis is the minimal lambda value used in the ridge and lasso models. This small lambda value implies that the penalty applied to the regression coefficients is not significant. Consequently, this allows us to properly interpret the coefficients of these models, as the regularization effect is minimal.

Our group's question of interest centers on identifying the determinants that influence the interest rate assigned to loans by LendingClub.com. The results from our regression models consistently show that the FICO credit score is a significant predictor of interest rates. In all models, the coefficient for the FICO score is negative, indicating that as the FICO score increases, the interest rate on loans tends to decrease. This aligns with the general understanding in credit scoring that higher credit scores are associated with lower credit risk, and thus, borrowers with higher scores are typically offered loans at lower interest rates. A higher DTI is associated with higher interest rates, suggesting that borrowers with a higher proportion of debt relative to their income are considered to be at a higher risk of default and thus face higher borrowing costs. An increase in the number of recent credit inquiries is associated with higher interest rates. This suggests that borrowers who have been seeking credit more frequently are perceived as higher risk, which is reflected in the terms of their loans. In conclusion, fico score is indeed indicative of the interest rate. Higher fico scores are associated with lower interest rates. Other variables such as loan installments, debt-to-income ratio, credit utilization, the number of recent credit inquiries, history of delinquency, and 2 derogatory public records exhibit a positive correlation with interest rates, indicating higher rates with increases in these factors. Conversely, variables like annual income, credit history time, balance on revolving accounts, and the presence of a single derogatory public record show a negative association, suggesting lower interest rates as these values rise.

iii. An interesting insight from the analysis is that all predictors were retained in the lasso regression model, which typically performs variable selection. This suggests that each predictor contributes unique information to the determination of interest rates, which is a valuable finding for understanding the complexity of financial risk assessment. In the ridge and lasso regression model, the coefficients for pub.rec1 are negative, while pub.rec2 and pub.rec3 are positive and larger in magnitude, suggesting that while one derogatory public record might not significantly increase the interest rate, two or more such records lead to a noticeable increase in the interest

rates offered to borrowers. Interestingly, there are discrepancies regarding the coefficients for income between lasso and ridge regressions. As for the OLS and lasso regression, the coefficient is negative, indicating an increased income will result in decrease in interest rate when controlling for other factors. However, for the ridge regression, the coefficient of income is positive, suggesting an increased income will lead to higher interest rate. This is counterintuitive as one might expect higher income to be associated with lower interest rates. This could be due to complex interactions among the variables or other unmodeled factors. These coefficients must be interpreted with caution, as they are conditional on the other variables in the model being held constant. Additionally, the unexpected negative coefficient for pub.rec1 in the models could be due to other interacting variables.

iv. Challenges faced by our group:

- Understanding the implications of lasso's variable selection was challenging, especially since it did not shrink any predictor to 0. This required us to consider the relevance and contribution of each predictor more closely.
- The similarities in MSE values made it difficult to draw strong conclusions about the superiority of one model over the others. This necessitated a further and nuanced discussion about model selection that goes beyond mere predictive performance.

In conclusion, the analysis has demonstrated that the selected predictors are all significant in determining the interest rate, as evidenced by the lasso model not excluding any variables. The marginal improvement in MSE by the lasso model over OLS and ridge suggests that while the predictors are all relevant, the regularization effect of lasso provides a slight enhancement in predictive accuracy. This finding underscores the importance of including all relevant variables in the model and the benefits of regularization methods in predictive modeling.

4. Regression Tree

a. The steps that were taken to prepare the dataset for the exploratory data analysis carried over to the construction of the regression tree, so all of the data cleaning described in part (a) of the exploratory data analysis section would apply here as well. In addition, the variable int.rate was converted into a percentage to aid in interpretability and to make it a similar magnitude to the other variables.

b. For the regression tree, all predictors were included. All variables were included because none of them were highly correlated with one another, and it seemed plausible that each predictor could provide unique information towards predicting interest rates. Furthermore, natural variable selection occurs later with pruning, so initially including all predictors seemed justified. Here is the list of predictors we will include:

- installment
- log.annual.inc
- dti

- fico
- days.with.cr.line
- revol.bal
- revol.util
- inq.last.6mths
- delinq.2yrs
- pub.rec dummy variables
- not.fully.paid dummy variable
- Purpose

One additional variable, purpose, is included in this section that is not included in the previous section. Purpose represents the purpose of the borrower's potential loan, categorized into seven classes. This variable was omitted from the shrinkage methods section due to concern that it would lead to overfitting, but as mentioned previously, this is less of a concern in the regression tree section, as natural variable selection occurs with pruning.

c. **i.**

```
## Regression tree:
## tree(formula = int.rate ~ ., data = train)
## Variables actually used in tree construction:
## [1] "fico"        "installment"
## Number of terminal nodes:  7
## Residual mean deviance:  0.02911 = 139.2 / 4781
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.84610 -0.11480 -0.02076  0.00000  0.08003  0.73810
```

Figure 8: Regression Tree Output

ii. There are 7 terminal nodes in the regression tree.

iii. The predictors that were used in the regression tree are fico and installment.

iv.

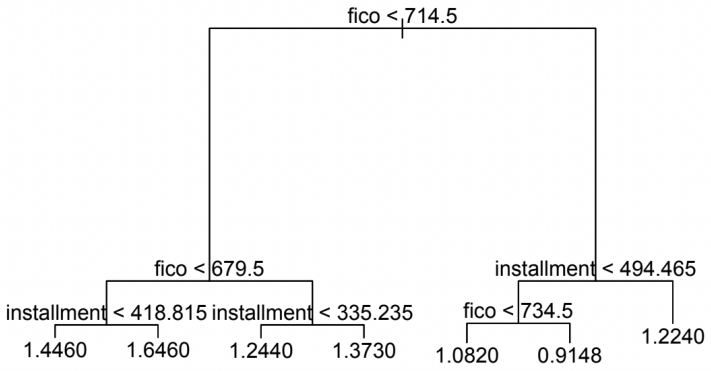

Figure 9: Graphical Output of Regression Tree

v. From the graphical output of the regression tree, it can first be noted that there are two main predictors that seem to be significant when predicting interest rates. This does not disqualify the importance of considering other predictors, but it does indicate that fico and installment are likely two of the most important. We can also note the order of importance - fico seems to be a better predictor of int.rate than installment based on the ordering of the nodes in the tree. Also, the regression tree indicates that those with higher values of fico seem to have lower values of int.rate, which seems intuitive. Alternatively, those with lower values of installment seem to have lower values of int.rate.

vi. The test MSE is 0.02874148.

d. Calculating the best size of the tree using cross-validation, we see that the optimal number of terminal nodes is 7, which is the same as the tree created from recursive binary splitting in the previous part. The pruned tree is the same as the tree from recursive binary splitting.

```
set.seed(4630)
cv.class<-tree::cv.tree(tree.model, K=10)
trees.num.class<-cv.class$size[which.min(cv.class$dev)]
trees.num.class
```

```
## [1] 7
```

Figure 10: Optimal Pruned Tree Size

e. **i.** The predictors that were found to be most important in random forests were fico and installment. These two predictors were found to be by far the most important predictors.

```
##                     %IncMSE IncNodePurity
## credit.policy      16.723155      6.475459
## purpose            52.999532     15.495860
## installment       160.788688     56.047846
## log.annual.inc     21.489497      9.872516
## dti                48.205030     16.315549
## fico              259.363607    156.529760
## days.with.cr.line  22.350833     13.328742
## revol.bal          30.505916     11.470514
## revol.util         53.406530     35.977280
## inq.last.6mths     35.233319      9.644210
## delinq.2yrs         1.647614      1.376042
## pub.rec             2.531865      0.643877
## not.fully.paid      5.387613      1.186260
```

Figure 11: Predictor Importance Using Random Forests

ii. The test MSE for random forests is 0.01812283.

f. **i.**

Method	Test MSE
Recursive Binary Splitting	0.02874148
Pruning	0.02874148
Random Forests	0.01812283

ii. One of the main strengths of decision tree models is their interpretability, especially the interpretability of predictor importance. In regards to the recursive binary splitting model, we see that fico is the most important predictor, followed by installment. With random forests, we can take this interpretation one step further and see that fico is more than 50% more important than installment when looking at changes in the test MSE when fico is not included. Installment is also as important as the next three most important predictors combined.

Our research question seeks to answer if a higher FICO credit score leads to lower interest rates in loans granted by LendingClub.com, and from our regression trees, we can see that not only does a higher FICO score result in lower interest rates, but it is also one of the most important predictors. We are also able to observe that installment is another important financial factor, while other financial factors might not be as relevant.

iii. It was interesting to observe that only fico and installment were considered by our regression tree, as our group thought that more predictors would be included. This provides interesting insight about our data, because while LendingClub.com might have access to a large number of financial variables about potential borrowers, some of those variables appear to be considered much more important than others. The random forest model had the lowest test MSE and did the best overall, which is not surprising as our singular decision tree likely had high variance.

iv. Our results from this section were straightforward, so we did not face any major challenges.

5. Classification Tree

a. The steps that were taken to prepare the dataset for the exploratory data analysis carried over to the construction of the regression tree, so all of the data cleaning described in part (a) of the exploratory data analysis section would apply here as well. Also, as it was for our regression tree, the variable int.rate was multiplied by 100 to convert it from a proportion to a percentage, which makes interpretation easier and also makes the variable a similar numerical magnitude to the other predictors.

b. Our goal in this section is to explore to what extent a higher FICO credit score, as well as other financial factors, make a borrower more likely to meet the credit underwriting criteria in loans granted by LendingClub.com. We included all variables in the dataset as predictors, as none of the variables are highly correlated with each other, and it seemed possible for each one of the variables to have a significant impact on the likelihood of a borrower meeting the criteria; all of

them make sense in the context of our question of interest. Furthermore, variable selection occurs naturally with classification tree pruning, which we will explore in this section, so initially including all variables seems like the best choice. Below is the list of variables used:

- installment
- log.annual.inc
- dti
- fico
- days.with.cr.line
- revol.bal
- revol.util
- inq.last.6mths
- delinq.2yrs
- pub.rec
- not.fully.paid
- purpose

One variable is included in this section that was not included in section 3; that variable is purpose, which represents the purpose of the borrower's loan and is categorized into 7 classes. It is a categorical variable, and each level of purpose has a sufficient number of observations. This variable was omitted from the shrinkage methods section due to concern that it would lead to overfitting, but this is less of a concern in the classification tree section, as natural variable selection occurs with pruning.

c. i.

```
Classification tree:
tree::tree(formula = credit.policy ~ ., data = train)
Variables actually used in tree construction:
[1] "inq.last.6mths"    "fico"              "days.with.cr.line" "dti"
[5] "revol.bal"         "revol.util"
Number of terminal nodes:  9
Residual mean deviance:  0.09678 = 462.5 / 4779
Misclassification error rate: 0.01149 = 55 / 4788
```

Figure 12: Classification Tree Output

ii. There are 9 terminal nodes in the classification tree.

iii. The predictors used in the tree are inq.last6mths, fico, days.with.cr.line, dti, revol.bal, and revol.util.

iv.

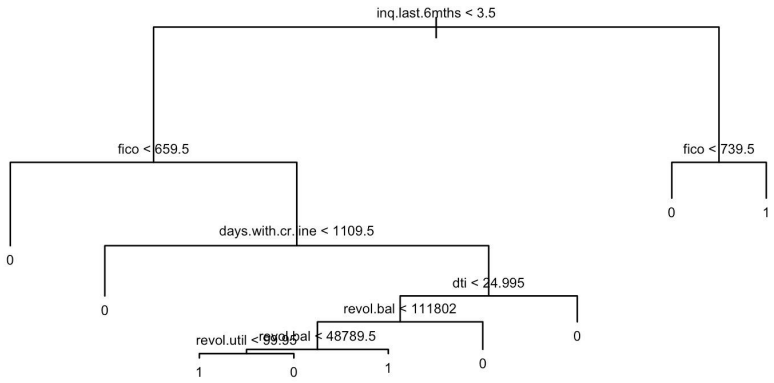

Figure 13: Graphical Output of Classification Tree

v. From the tree, we see that only 6 predictors were important enough or had a significant enough effect on the likelihood of meeting the criteria to be included in the classification tree. inq.last.6mths was the most important predictor, followed by fico, day.with.cr.line, dti, revol.bal, and revol.util. Higher FICO credit scores do appear to lead to a higher likelihood of meeting the criteria, as proposed in our research question, for in the splits with the fico variable, having a lower value than the cutoff leads to a classification of not meeting the criteria while having an equal or higher value than the cutoff leads to either a classification of meeting the criteria or a further set of splits needed to determine the outcome. From this classification tree, it appears that generally, higher values of days.with.cr.line, lower values of revol.bal, lower values of dti, higher values of fico, and lower values of revol.util generally lead to a higher likelihood of a borrower meeting the credit underwriting criteria, which are effects of financial factors that we wanted to explore with our question.

vi. Below is the confusion matrix for the classification tree fit on the test data and with the standard threshold of 0.5.

	rcTreeTest	
y.test	0	1
0	840	61
1	18	3869

Figure 14: Confusion Matrix with Threshold of 0.5

vii. The overall test error rate was 0.0165.

viii. The false positive rate was 0.0677.

ix. The false negative rate was 0.0046.

x. We decided to increase the threshold to try to get a lower false positive rate. In this situation, a false positive is likely worse than a false negative, as a borrower who is classified as meeting the criteria but actually does not runs the risk of getting a loan he or she is not prepared for, causing the lender to lose money and the borrower to further ruin his or her financial situation. A false negative only runs the risk of a borrower being denied a loan, and the borrower can likely find luck elsewhere. However, adjusting the threshold did not change any of our results until we adjusted it to 0.91. This threshold is quite high, and although our false positive rate decreased to 0.0366, our overall test error rate increased by almost as much to 0.0420, and our false negative rate increased by a greater amount to 0.0432. Because our overall test error rate and false negative rate were so low at the standard threshold, the decrease in our false positive rate with a threshold of 0.91 does not seem worth it for the corresponding increases in our false negative rate and overall test error rate, so the threshold should likely be left at 0.5. Because of the provided discussion behind adjusting the threshold, we still included the confusion matrix below.

y.test	FALSE	TRUE
0	868	33
1	168	3719

Figure 15: Confusion Matrix with Threshold of 0.9

d. After attempting to prune our tree, it turned out to be the same as the original tree. Below is an output of the summary() function on our pruned tree, which is the same as our original output.

```
Classification tree:
tree::tree(formula = credit.policy ~ ., data = train)
Variables actually used in tree construction:
[1] "inq.last.6mths"    "fico"              "days.with.cr.line" "dti"
[5] "revol.bal"         "revol.util"
Number of terminal nodes:  9
Residual mean deviance:  0.09678 = 462.5 / 4779
Misclassification error rate: 0.01149 = 55 / 4788
```

Figure 16: Pruned Classification Tree Output

e. **i.** We used a value of 5 for mtry, which we calculated by taking the square root of our number of predictors, 25.

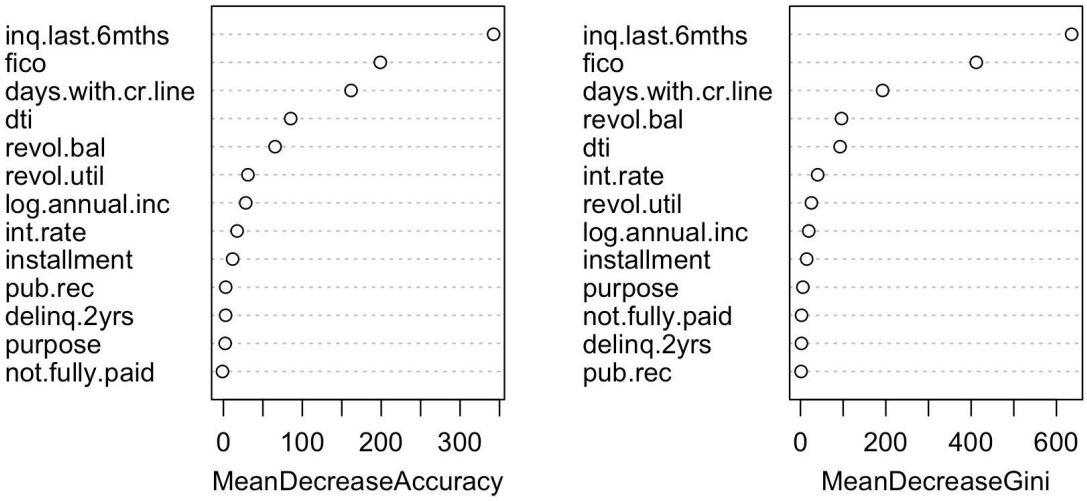

Figure 17: Predictor Importance Using Random Forests

From Figure 17, it can be seen that in terms of both MeanDecreaseAccuracy and MeanDecreaseGini, the three most important predictors were inq.last.6mths, fico, and days.with.cr.line. Dti and revol.bal, which were also used in the classification tree, are both in the top five for both measures of importance. However, although revol.util, which was the last predictor used in the classification tree, is sixth most important in terms of MeanDecreaseAccuracy, it falls to seventh behind int.rate on MeanDecreaseGini, which was not used in the classification tree. Inq.last.6mths, fico, and days.with.cr.line are substantially more important than the other predictors, for the predictors outside of the top three and especially out of the top 5 all have relatively similar values of each measurement.

ii. Below is the confusion matrix for the test data predictions using random forests and with the standard threshold of 0.5.

	rcTreeTest	
y.test	0	1
0	840	61
1	18	3869

Figure 18: Confusion Matrix Using Random Forests with Threshold of 0.5

iii. The overall test error rate was 0.0138.

iv. The false positive rate was 0.0610.

v. The false negative rate was 0.0028.

vi. We decided to increase the threshold to try to lower the false positive rate for the same reasons described in part c of this section. We decided to increase it to 0.91 to stay consistent with the previous section. Increasing the threshold increased our overall test error rate and our false negative rate, but it comparatively decreased our false positive rate by a much greater amount. The overall test error rate was 0.0305, the false positive rate was 0.0111, and the false negative rate was 0.0350. For random forests, because raising the threshold successfully lowered the false positive rate by a very large margin and resulted in acceptable increases to our overall test error rate and false negative rate that were smaller than when we changed the threshold of our classification tree, a threshold of 0.91 should be used.

y.test	FALSE	TRUE
0	891	10
1	136	3751

Figure 19: Confusion Matrix Using Random Forests with Threshold of 0.9

f. **i.** Recall that the pruned classification tree was identical to the classification tree produced using recursive binary splitting, so only the results from recursive binary splitting are included.

Method	Overall Test Error Rate	False Positive Rate	False Negative Rate
Recursive Binary Splitting (.5)	0.016499582	0.067702553	0.004630821
Recursive Binary Splitting (.91)	0.04197995	0.03662597	0.04322099
Random Forests (.5)	0.013784461	0.061043285	0.002829946
Random Forests (.91)	0.03049290	0.01109878	0.03498842

Figure 20: Test Error Rate, FNR, and FPR Across Different Methods

ii. Our overall goal was to explore if having a higher FICO credit score led to a higher likelihood of a borrower meeting the credit underwriting criteria of LendingClub.com, as well as to

understand the impact of other financial factors on this likelihood. The models produced in this section showed us which predictors are most important when predicting whether or not a borrower will meet the credit underwriting criteria of LendingClub.com, as well as which values of these predictors are generally associated with meeting the criteria. From the models built in this section, we were able to see that fico, inq.last.6mths, and days.with.cr.line are three of the most important predictors, and we were also able to see that higher values of fico and days.with.cr.line result in a higher likelihood of meeting the criteria.

iii. The information provided about our data by our models was largely already known. Our group suspected that fico would be one of the most important predictors, but the models did provide us with the additional information that inq.last.6mths and days.with.cr.line are two other important predictors. However, the directions of the relationships between our predictors and our response variable were intuitive, and our models only confirmed our intuition.

The classification tree built using random forests best addressed our question of interest, as this model had the lowest test error rate. From this model, the predictor importance of fico, inq.last.6mths, and days.with.cr.line was further confirmed.

iv. The biggest challenge faced by our group during this section was deciding what to set the threshold to for our confusion matrices. While the false positive rate seemed too high for our initial classification tree, we had to assess if a small decrease in the false positive rate was worth a much larger increase in both the false negative rate and the overall test error rate. Ultimately, for the initial classification tree, we chose to accept a slightly higher than desired false positive rate and leave the threshold unchanged. For the random forests, a large decrease in the false positive rate was associated with a smaller increase in the false negative rate and the overall test error rate. This change seemed desirable, so it was decided to change the threshold to 0.91.