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

Final Report

1. Executive Summary for Regression Question

a. **Regression Question of Interest**: 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?

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.

c. The analyses carried out in previous milestones sought to better understand the relationship between various financial factors and the interest rate granted by LendingClub.com, and more specifically, whether or not a borrower having a higher FICO credit score would result in them receiving a lower interest rate on their loan. From the analyses, it was determined that when predicting the interest rate, the most important factors are the borrower's FICO credit score and the monthly installment owed by the borrower if the loan is funded. Higher FICO credit scores are associated with a lower interest rate, and lower monthly installments are also associated with a lower interest rate. This seems intuitive, as a high FICO credit score is an indicator of a borrower's trustworthiness, and LendingClub.com likely has good motivation to assign a higher interest rate to large loans.

While FICO credit score and the monthly installment are two of the most important factors when determining the interest rate for a loan granted by LendingClub.com, other financial factors still have an influence. In fact, each of the following additional factors can be useful when predicting the interest rate: the annual income of the borrower, the debt-to-income ratio of the borrower, the number of days the borrower has had a credit line, the borrower's revolving balance, the borrower's revolving line utilization rate, the borrower's number of inquiries by creditors in the last six months, the number of times the borrower was over thirty days past due on a payment in the past two years, and the borrower's number of derogatory public records. This indicates that a large number of factors are taken into consideration by LendingClub.com when determining the

interest rate for a loan. To directly answer the research question, a higher FICO credit score does lead to a lower interest rate in loans granted by LendingClub.com, and while FICO credit score is one of the most important predictors, several other financial factors impact the granted interest rate.

d. As mentioned above, this information is valuable to borrowers who may wish to improve specific financial metrics to obtain loans at a lower interest rate. Borrowers should pay careful attention to their FICO credit score, as well as the size of the loan that they are seeking to obtain. Raising one's FICO credit score is likely a good way to obtain a loan at a lower interest rate from LendingClub.com, and refraining from requesting a loan with large installments will also likely lead to a lower interest rate. For investors, understanding the relationship between these various financial factors and the granted interest rate can aid in assessing the risk associated with different interest rates. Loans with high interest rates are likely to have large installments and to belong to a borrower with a low FICO credit score, potentially making these loans riskier investments. Investors should reconsider investing in loans with high interest rates, as this could potentially lead to better risk-adjusted returns.

2. Data and Variable Description for Regression

a. The data used in this report 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 whether each customer meets the credit underwriting criteria for the company.

b. 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.

c.

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

3. Regression Question

3.1 Exploratory Data Analysis

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

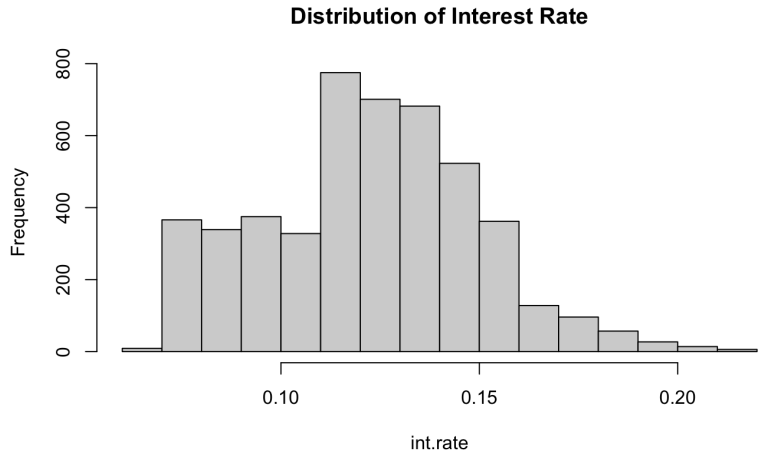

Figure 1: Distribution of Response Variable

	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

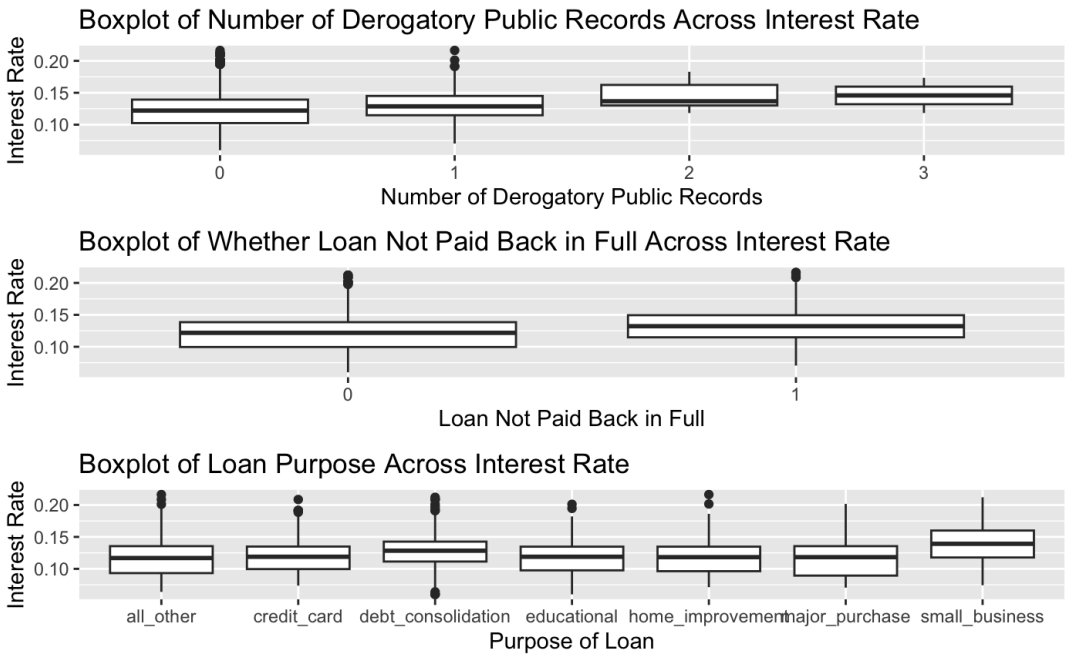

Figure 3: Boxplots of Interest Rate Across Categorical Predictors

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.

3.2 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.

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.

b. 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}

3.3 Regression Trees

a. We chose to present the tree built with recursive binary splitting because, after applying the pruning process, the number of terminal nodes suggested was the same as the one obtained from the initial recursive binary splitting. This outcome indicated that the tree constructed through recursive binary splitting had already achieved an optimal level of complexity. Essentially, the pruning process did not find any redundant nodes to remove.

b. 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.

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.

```
## 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

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

d.

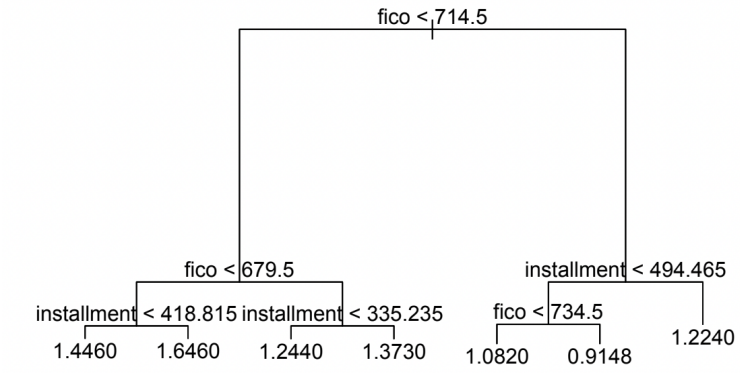

Figure 9: Graphical Output of Regression Tree

e. 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.

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

Figure 10: Predictor Importance Using Random Forests

3.4 Summary of Findings

a.

Table 1: Methods vs Test MSE

Method	Test MSE
Linear Regression	0.0002624469
Ridge Regression	0.0002663066
Lasso Regression	0.0002623285
Recursive Binary Splitting	0.02874148
Random Forests	0.01812283

b. From Table 1, we can observe that the test MSEs of using shrinkage methods 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. In contrast, the Recursive Binary Splitting and Random Forests methods show higher MSE values (0.02874 and 0.01812, respectively), indicating less precision in prediction. The small test MSEs for the regression models suggest that they perform well in predicting the response variable, which is likely due to their ability to capture linear relationships

effectively. 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.

c.

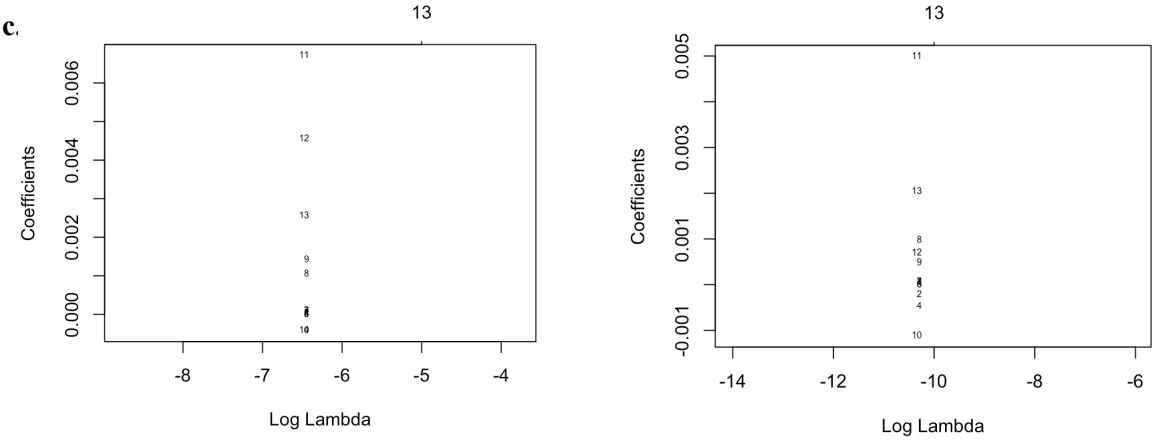

Figure 11: Coefficient Path of Ridge (Left) and Lasso Regression (Right)

Based on Figure 11, both plots confirm that the signs of the coefficients are consistent across a range of lambda values, providing a reliable interpretation of the direction of associations.

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.

In addition to the FICO score, our analysis revealed several other significant predictors. 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. 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.

While the direction of the relationships (positive or negative) between these factors and interest rates is evident, it is crucial to note that the magnitude of the coefficients obtained from Lasso and

Ridge regression should be interpreted with caution. The regularization process inherent in these methods alters the coefficients to prevent overfitting, which can affect the interpretability of their absolute values. Therefore, our conclusions from the shrinkage methods primarily focus on the presence and directional influence of these variables, rather than the exact coefficient sizes.

The insights gained from the recursive binary splitting model and random forests further solidify the prominence of the FICO score as a predictor. 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. 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.

In summary, our research investigated that while various financial factors influence the interest rates of loans offered by LendingClub.com, the FICO score emerges as the most significant predictor. It is followed by other factors like loan installment amounts, DTI ratios, and credit inquiries, which also play significant roles but are overshadowed by the predominance of the FICO score in the lender's assessment.

d. For the purpose of prediction, the Lasso Regression model emerged as the most suitable choice. This model yielded the lowest test Mean Squared Error (MSE) among the evaluated methods, indicating its superior predictive accuracy. However, our research question is focused on understanding the significance and influence of individual predictors, Random Forests take precedence. This method excels in offering a clear perspective on variable importance, a critical aspect of our research goal. Random Forests, through their ensemble approach, offer robust insights into how different predictors interact and contribute to the overall prediction. The method's ability to rank variables in terms of their importance provides clarity on which factors are most influential in determining interest rates.

3.5 Address Previous Comments

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

4. Executive Summary for Classification Question

a. **Classification Question of Interest:** 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 and potentially leading to better risk-adjusted returns.

c. The analyses we carried out in previous milestones sought to find the relative importance of each financial factor in terms of how it contributed to the likelihood of a borrower being approved for a loan by LendingClub.com. We initially focused on a borrower's FICO credit score as a potentially important contributing factor, for FICO credit score is perhaps the most widely recognized national metric for evaluating whether a borrower will pay back a loan. Our analyses indeed revealed that out of the many factors we explored, FICO credit score is one of the two most important factors in determining whether a borrower will be approved for a loan, and higher FICO credit scores generally lead to an increased likelihood of a borrower meeting the credit underwriting criteria. Our most advanced model revealed that while controlling for the other financial factors, for every point by which a borrower's credit score was higher, the odds of them being approved for a loan increased by 4.8%. The fact that FICO credit scores are measured from 300 to 850, covering a range of 550 possible points, reveals the substantial importance of credit score.

   What our analyses revealed that we did not initially theorize was the substantial importance of three other key financial factors: annual income, the number of days a borrower has had a credit line, and the number of times creditors have made inquiries about a borrower in the past six months of a borrower's history. Similar to credit score, higher values of annual income and the number of days a borrower has had a credit line led to substantially improved chances of being approved for a loan, although to a lesser extent than higher credit scores. The number of inquiries by creditors in the last six months, however, was the only factor that we found to be more important than credit score. In fact, in one of our models, the number of credit inquiries was almost twice as important as a borrower's credit score in determining whether or not that borrower was approved for a loan. We found that having a high number of inquiries was a serious demerit on a borrower's record, and having multiple inquiries in the past six months could immediately offset having a high credit score, a large annual income, and a large number of days with a credit line. In our most advanced model, while controlling for the other financial factors, every additional inquiry a borrower had led to the odds of them being approved for a loan decreasing by around 182%. Multiple inquiries within six months means multiple loan applications within a relatively short period of time and heavily detracts from the credibility of a borrower.

   Aside from these key financial factors, we found that numerous other factors have meaningful effects on the likelihood of a borrower meeting the credit underwriting criteria, although not as large of an effect as the four variables that were just explained. In general, we found that higher monthly loan payments and higher proportions of the credit that a borrower uses compared to the credit they have available, as well as lower revolving credit balances and lower loan interest rates,

led to a somewhat higher likelihood of being approved for a loan. Relative values of these factors, either high or low as explained above, seemed fitting of less risky borrowers, and it did not surprise us that they made a borrower more likely to receive a loan, with one exception; it surprised us that higher proportions of a borrower's available credit that they use led to a higher likelihood of being approved, but this could potentially be explained by outside factors that we did not explore, such as a borrower using more credit because of their confidence in paying it off.

d. From our analyses, we recommend that potential borrowers do their best to maintain a reputable FICO credit score, as well as not apply for multiple loans within too short of a time frame, as a low credit score or too many loan applications could hurt their chances of receiving a loan. We also recommend that adults who have not yet started a credit line do so as soon as they are able, as having a longer credit history may help their chances in receiving loans in the future. And while a higher annual income leads to a greater chance of being approved for a loan, it is not practical for us to recommend that borrowers try and make more money, so we instead recommend that borrowers only apply for loans that are reasonable when compared to their current income. For lenders, we recommend placing great importance on a borrower's credit score and credit history, including recent credit inquiries as well as when their credit line began, for lower credit scores, more inquiries, and a shorter credit line history are telling of a riskier borrower. We also advise that lenders be wary of borrowers who apply for loans that seem quite large when compared to their annual income.

5. Data and Variable Description for Classification

a. The data used in this report 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 whether each customer meets the credit underwriting criteria for the company.

b. 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.

c.

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

6. Classification Question

6.1 Exploratory Data Analysis

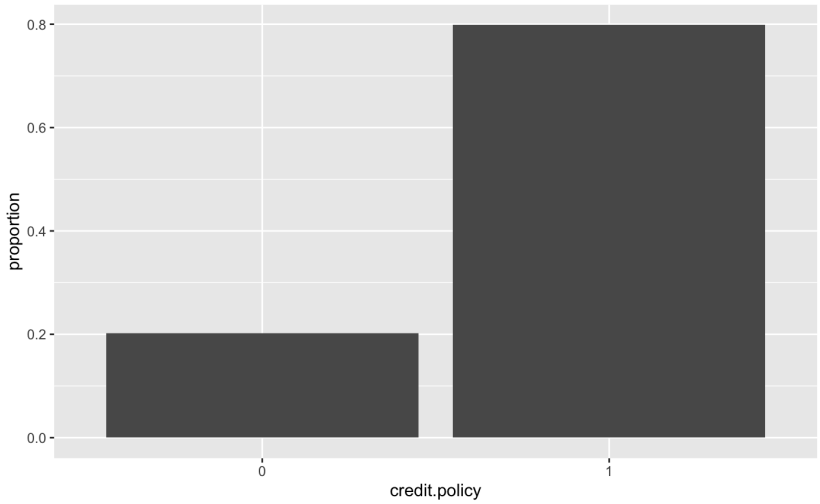

Figure X - Distribution of Credit.Policy

Figure X displays the proportion of observations for each class in the binary response variable, credit policy. Nearly 80% of the observations belong to class 1, which represents the customers who met the underwriting criteria of LendingClub.com, while the other 20% belong to class 0, which represents the customers who did not meet the underwriting criteria of LendingClub.com. These skewed proportions indicate that our data is unbalanced, which could affect the performance of LDA. The logistic regression model, rather than the LDA model, will be presented in this section, but if the LDA model were being presented, additional metrics would need to be checked. Unbalanced data necessitates checking the confusion matrix, false positive rate, and false negative rate when performing LDA, as the test error rate, ROC curve, and AUC curve may look fine, but these other metrics might not.

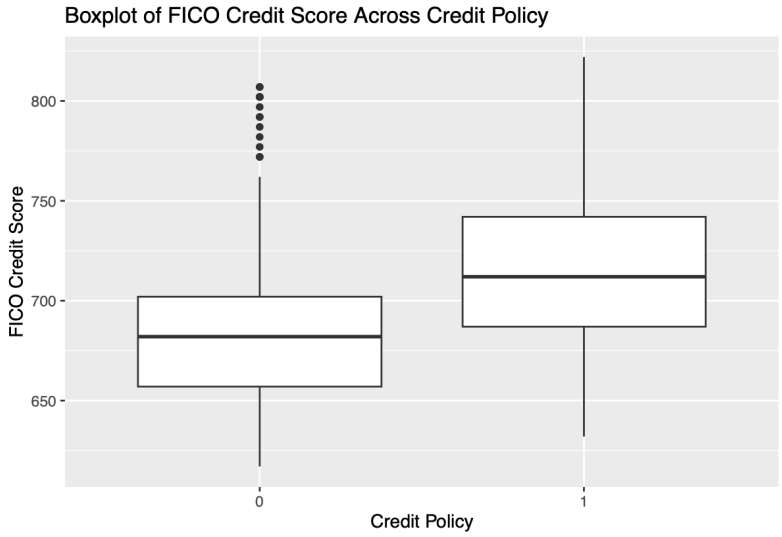

Figure X - Boxplot of Credit Score Across Credit Policy

Figure X is a boxplot of FICO credit scores across whether or not a borrower met the credit policy, and exploring this relationship is a key part of answering our research question. From Figure X, it is evident that individuals with higher FICO Credit Scores are more likely to meet the credit underwriting criteria, as

the median and IQR of the FICO credit score distribution are noticeably higher for those who do meet the criteria versus those who do not. This makes sense and aligns with our previous knowledge of the subject matter as a higher FICO score is generally seen as an indicator of financial trustworthiness and evidence that a person will pay back his or her loans. However, further testing is required to determine the statistical significance of this difference, as well as the impact that this difference may have on the classification problem at hand.

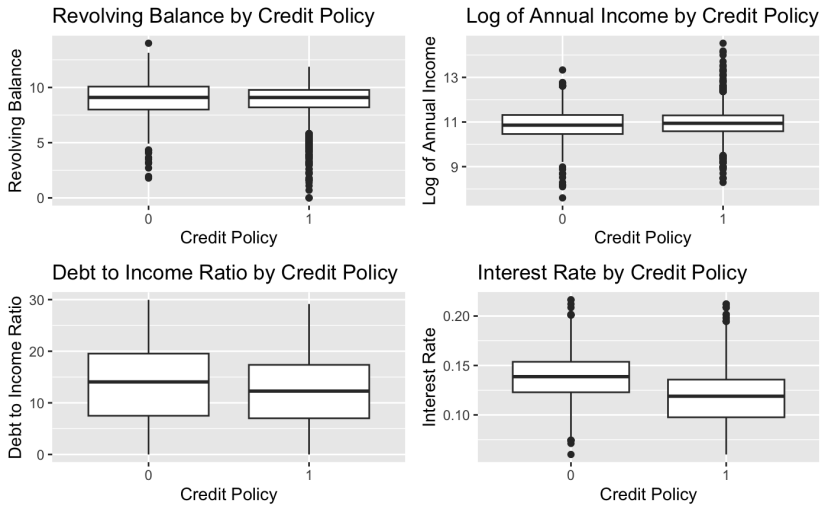

Figure X - Boxplots of Selected Variables Across Credit Policy

Figure X contains boxplots of other selected predictors across credit policy. Figure 3 shows that loans with higher interest rates are generally associated with a lower likelihood of the borrower with that loan meeting these criteria, for the median interest rate for those who meet the credit underwriting criteria is noticeably lower than the median interest rate of those who do not. This is intuitive and aligns with our previous knowledge of the subject matter, for riskier loans are given higher interest rates in order to minimize the extent of potential losses. For other variables such as revolving balance, annual income, and debt-to-income ratio, the boxplots do not indicate significant differences between the two categories of credit policy compliance, and further testing is needed to explore the relationship between these variables and likelihood of meeting the credit underwriting criteria.

	credit.policy	int.rate	installment	log.annual.inc	dti	fico	days.with.cr.line	revol.bal	revol.util	inq.last.6mths	delinq.2yrs
credit.policy	1.00	-0.29	0.08	0.04	-0.09	0.36	0.11	-0.16	-0.09	-0.53	-0.08
int.rate	-0.29	1.00	0.27	0.06	0.22	-0.71	-0.14	0.09	0.47	0.20	0.16
installment	0.08	0.27	1.00	0.44	0.05	0.10	0.18	0.24	0.08	-0.02	0.00
log.annual.inc	0.04	0.06	0.44	1.00	-0.07	0.12	0.33	0.35	0.06	0.03	0.03
dti	-0.09	0.22	0.05	-0.07	1.00	-0.25	0.05	0.19	0.33	0.03	-0.02
fico	0.36	-0.71	0.10	0.12	-0.25	1.00	0.27	-0.03	-0.55	-0.18	-0.22
days.with.cr.line	0.11	-0.14	0.18	0.33	0.05	0.27	1.00	0.23	-0.02	-0.03	0.07
revol.bal	-0.16	0.09	0.24	0.35	0.19	-0.03	0.23	1.00	0.21	0.02	-0.03
revol.util	-0.09	0.47	0.08	0.06	0.33	-0.55	-0.02	0.21	1.00	-0.01	-0.04
inq.last.6mths	-0.53	0.20	-0.02	0.03	0.03	-0.18	-0.03	0.02	-0.01	1.00	-0.01
delinq.2yrs	-0.08	0.16	0.00	0.03	-0.02	-0.22	0.07	-0.03	-0.04	-0.01	1.00

Figure X - Correlation Matrix of All Variables

Based on our correlation matrix, there seems to be little correlation between each of the explanatory variables we want to explore, with exception to fico and int.rate. This exception is intuitive and is consistent with our previous knowledge about the subject matter, for those with higher FICO scores are deemed more likely to pay their loans back and thus more likely to receive lower interest rates on their loans.

The correlation of -0.71 between fico and int.rate will not have much effect on the interpretation of our models. The VIF of this correlation is calculated to be around 2, which is not high enough for concern; additionally, this is our highest VIF, meaning that we will see low multicollinearity overall in our model, which increases confidence in the results of our hypothesis tests and interpretations.

6.2 Logistic Regression Model

a. For both our initial logistic regression model and classification tree model, all of the available predictors were included. They are as follows:

- 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

We chose to include all predictors in our initial logistic regression model and then perform stepwise selection to select the most important predictors. Similarly, we chose to include all predictors in our classification tree, as natural variable selection occurs with pruning and only the most important predictors will remain in the tree.

b. Below is the output from our final logistic regression model. As mentioned previously, some of the predictors mentioned above are not in this output because stepwise selection was used for our final logistic regression model. While our initial logistic regression model included all available predictors in our data, stepwise selection removed some of these variables.

```
Call:
glm(formula = credit.policy ~ installment + log.annual.inc +
    fico + days.with.cr.line + revol.bal + revol.util + inq.last.6mths +
    delinq.2yrs + not.fully.paid, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8687   0.0392   0.1748   0.4012   2.6591

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.529e+01  2.063e+00 -17.103  < 2e-16 ***
installment       1.085e-03  3.089e-04   3.511 0.000446 ***
log.annual.inc    6.901e-01  1.071e-01   6.446 1.15e-10 ***
fico              4.383e-02  2.392e-03  18.325  < 2e-16 ***
days.with.cr.line 1.417e-04  2.663e-05   5.321 1.03e-07 ***
revol.bal        -4.183e-05  2.542e-06 -16.459  < 2e-16 ***
revol.util        1.082e-02  2.234e-03   4.842 1.28e-06 ***
inq.last.6mths   -9.868e-01  3.711e-02 -26.591  < 2e-16 ***
delinq.2yrs      -1.958e-01  7.883e-02  -2.484 0.012991 *
not.fully.paid1  -2.636e-01  1.294e-01  -2.037 0.041662 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4815.1  on 4787  degrees of freedom
Residual deviance: 2377.3  on 4778  degrees of freedom
AIC: 2397.3

Number of Fisher Scoring iterations: 7
```

Figure X - Logistic Regression Output

6.3 Classification Trees

a. The classification trees produced using pruning and recursive binary splitting were identical, so the tree presented is a result of both recursive binary splitting and pruning.

b.

```
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 X - Classification Tree Output

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

d.

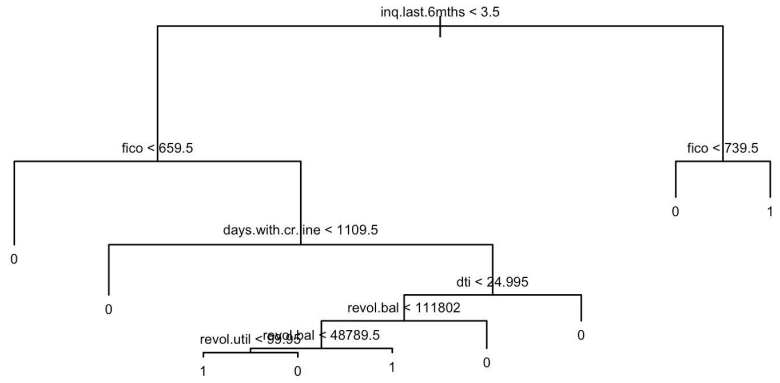

Figure X - Graphical Output of Classification Tree

e.

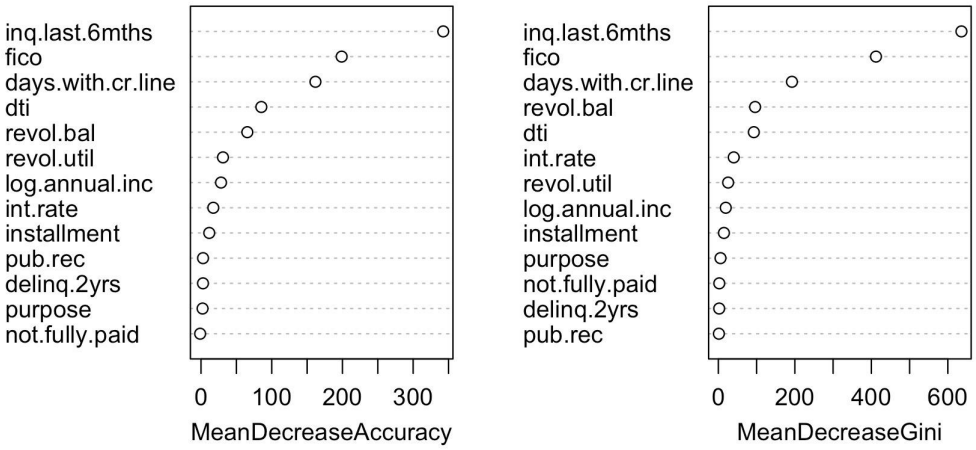

Figure X - Variable Importance Using Random Forests

6.4 Summary of Findings

a.

	FALSE	TRUE
0	601	300
1	143	3744

Figure X - Confusion Matrix for Logistic Regression

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

Figure X - Confusion Matrix for Classification Tree

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

Figure X - Confusion Matrix for Random Forest

b.

	Logistic Regression	Classification Tree	Random Forests
Test Error	0.0925	0.0165	0.0165
False Positive Rate	0.3330	0.0677	0.0677
False Negative Rate	0.0368	0.0046	0.0046

c. The threshold should not be adjusted for logistic regression and the classification tree, but it should be adjusted for random forests. Initially, 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. For logistic regression, even slightly raising the threshold results in a large increase in the false negative rate, mitigating the reduction of the false positive rate. For the classification tree and random forests, raising the threshold has no effect until it is raised to 0.91. This threshold is quite high, and for the classification tree, the resulting increase in the overall test error rate and the false negative rate is not worth the small reduction in the false positive rate. However, for random forests, raising the threshold to 0.91 results in a large reduction in the false positive rate and an acceptable increase in the overall test error rate and the false negative rate. This indicates that raising the threshold to 0.91 for random forests is ideal, while the threshold should remain unchanged from 0.5 for the other two methods.

d.

y.test	FALSE	TRUE
0	891	10
1	136	3751

Figure X - Confusion Matrix for Random Forests with Threshold of 0.91

e.

	Logistic Regression	Classification Tree	Random Forests
Test Error	0.0925	0.0165	0.0305
False Positive Rate	0.3330	0.0677	0.0111
False Negative Rate	0.0368	0.0046	0.0350

f. The findings from our logistic regression model indicated that installment, log.annual.inc, fico, days.with.cr.line, revol.bal, revol.util, inq.last.6mths, delinq.2yrs, and not.fully.paid are all significant predictors of whether or not a borrower meets the credit underwriting criteria of LendingClub.com. Looking at the values of the coefficients of each predictor in the model can give a sense of the predictors' relative importance. The exact importance of each predictor is hard to quantify as each one has a different distribution. However, log.annual.inc, inq.last.6mths, delinq.2yrs, and not.fully.paid have some of the largest coefficients, suggesting that these predictors are of particular importance. The coefficients of installment, log.annual.inc, fico, days.with.cr.line, and revol.util are positive, suggesting that larger values of these predictors are associated with a higher probability of meeting the credit underwriting criteria of LendingClub.com. Similarly, the coefficients of revol.bal, inq.last.6mths, delinq.2.yrs, and not.fully.paid are negative, indicating that larger values of these predictors are associated with a lower probability of meeting the credit underwriting criteria of LendingClub.com.

For our classification tree, only 6 predictors were important enough or had a significant enough effect on the likelihood of meeting the criteria to be included. 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.

The findings from our random forests indicated that inq.last.6mths, fico, and days.with.cr.line are the three most important predictors in terms of both MeanDecreaseAccuracy and MeanDecreaseGini. 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.

While the findings across our three methods were somewhat unique, each method indicated that higher FICO credit scores are in fact associated with a higher probability of meeting the credit underwriting criteria established by LendingClub.com. Additionally, inq.last.6mths and days.with.cr.line are two other important predictors of whether or not this criteria will be met. Having multiple credit inquiries in the past six months is a major demerit, while having had a credit line for a large number of days is viewed positively.

g. Our question of interest can be broken down into two parts: does a higher FICO credit score make a borrower more likely to meet the credit underwriting criteria of LendingClub.com, and to what extent do other factors impact this relationship? The classification tree did the best job answering the first part of this question, as it demonstrated not only the importance of fico, but also the positive relationship between fico and the probability of meeting the credit underwriting criteria. From the splits in the classification tree, it can be seen that higher values of fico were consistently associated with a higher probability of being classified as meeting the criteria.

To answer the second part of our question, to what extent do other factors impact this relationship, random forests proved to be the most useful. Random forests allowed us to see each variable's relative importance, and we were able to determine that inq.last.6mths and days.with.cr.line are two other variables that play a major role in determining whether a borrower meets the criteria. Additionally, random forests with the adjusted threshold had the lowest false positive rate, which is ideal.

Logistic regression had a significantly higher test error rate than both the classification tree and random forest, and additionally, the information provided by the logistic regression model was less helpful in answering our question of interest.

6.5 Address Previous Comments

The comments from previous milestones regarding the classification problem have been addressed in the above sections of this report. The relevant summary showing the proportion of observations in each class of our binary response variable is included in Section 6.1. We re-examined our initial and final logistic regression model and chose to initially include all predictors before running stepwise selection to

determine the optimal final model as described in Section 6.2. Although we chose to include the logistic regression model instead of the LDA model, we addressed the fact that our data is unbalanced and briefly discussed ways that we would address this if performing LDA in Section 6.1.

7. Further Work

There are several improvements can be made for future work regarding our research question:

- Incorporation of Additional Predictors:
  Future research could explore additional financial and non-financial predictors that may influence loan interest rates and creditworthiness. Variables such as employment history, educational background, and geographical location could offer deeper insights into the lending process.
- Incorporation of Time-Series Analysis:
  Considering the dynamic nature of financial markets and credit scoring, a time-series analysis could provide valuable insights. This would account for temporal changes in the economy, credit policies, and borrower behavior, offering a more complete understanding of the determinants of interest rates and credit decisions.

By pursuing these areas, future work can build upon the existing findings about the factors influencing lending practices and outcomes at LendingClub.com.

8. Reflection on Learning

Working on this project was helpful for our learning because it allowed us to explore a real-world application of the course material in depth and see how we might translate what we learned this semester into meaningful analyses we might perform in our careers, especially for those majoring or minoring in statistics. Homeworks, quizzes, and midterms are useful for learning, but we will no longer have to do such assignments after graduation, so the project was the assessment that will be most applicable to our future jobs. The format and instructions of the project required working in a team, analyzing and drawing conclusions on real data, and making actionable recommendations to relevant stakeholders, which is exactly what many statistics-related jobs such as consultants or actuaries require. Outside of applications directly to statistics, working in teams also helped us reinforce life skills that will help us in any career we may choose, such as communication, accountability, and organization.

Although we explored many worked examples in class that used non simulated data, such examples were relatively simple and were chosen in order to teach the foundational concepts of each topic. Each example presented a smooth, relatively simple analysis and did not present obstacles or inconsistencies that might arise with many data sets. For example, the pruning section of the learning unit on tree-based methods obviously showed a change from the original tree to the pruned tree in order to teach the process of pruning, but attempting to prune our tree in our project did not result in a change from our original tree. Furthermore, a different data set or topic was explored for almost every worked example, and only the project allowed us to apply numerous different learning methods on the same dataset or topic and see how those different methods compare or even contradict. The project was a comprehensive assignment that required us to use almost every method we have learned about throughout the semester, and we will probably better remember each one by applying it to a multi-stage, lengthy

project than we would by being tested on it with a final exam. There were several topics that we had to go back and refresh ourselves on in order to apply them to our project, for our recollection may have somewhat faltered after the initial assignments for that topic had passed, and the project should have served well to reinforce that material.