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

Milestone 3: Classification

1. Introduction

a. Research Question: 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. Understanding the factors that contribute to meeting the credit underwriting criteria serves multiple purposes. 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. Also, in particular, FICO credit score is one of the simplest, most common ways to judge a person's financial standing and ability to get loans, and having this variable as the focus of our question will help us explore if FICO credit score is as simple and accurate in describing a person's financial status and opportunities as it seems.

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. Data Description

For this classification problem, the following are the variables used:

Variable Name	Description	Type
credit.policy (Response)	Indicates if the customer meets the credit underwriting criteria (1: Yes, 0: No)	Categorical Levels: 0, 1
int.rate	Interest rate of the loan as a	Quantitative

	proportion	
log.annual.inc	Natural logarithm of the borrower's self-reported annual income	Quantitative
dti	Debt-to-Income ratio, representing the portion of the borrower’s income that goes towards debt payments	Quantitative
fico	FICO credit score of the borrower	Quantitative
revol.bal	Borrower's revolving balance in dollars, which is the amount unpaid at the end of credit card billing cycle	Quantitative

These variables have been selected to build a robust classification model that aims to answer the question of interest effectively.

2. Exploratory Data Analysis

a. Data cleaning

The dataset was well-structured upon download, eliminating the need for additional data cleaning or processing before generating graphical summaries.

b. Graphical Summaries and Interpretations

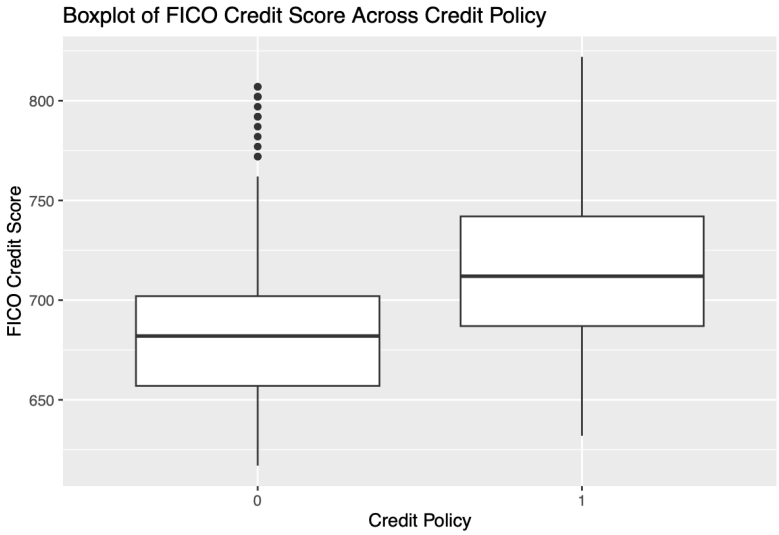

Figure 1 - Boxplot of Credit Score Across Credit Policy

From Figure 1, 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.

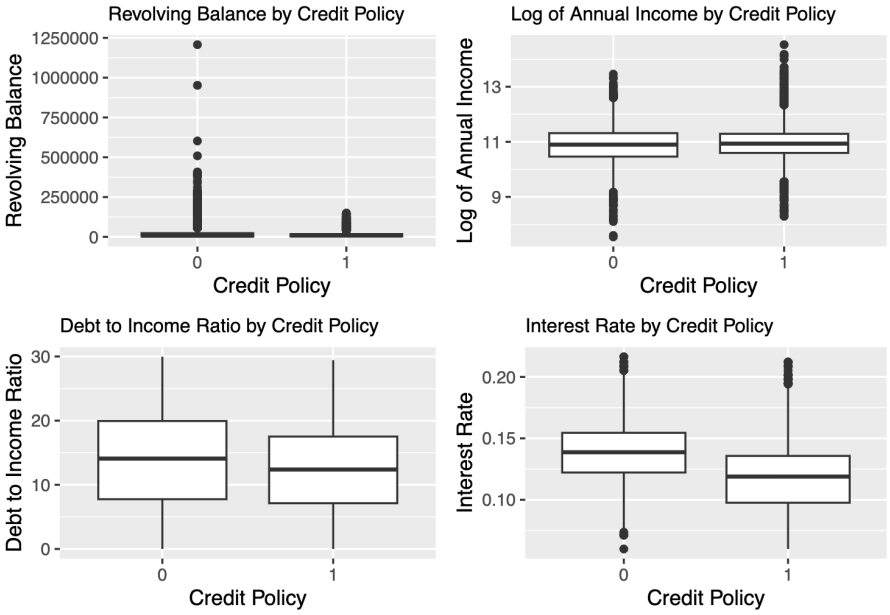

Figure 2 - Boxplot Matrix Other Four Predictors Across Credit Policy

Figure 2 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	log.annual.inc	dti	fico	revol.bal
credit.policy	1.00	-0.29	0.04	-0.10	0.35	-0.17
int.rate	-0.29	1.00	0.05	0.23	-0.71	0.08
log.annual.inc	0.04	0.05	1.00	-0.07	0.13	0.35
dti	-0.10	0.23	-0.07	1.00	-0.25	0.19
fico	0.35	-0.71	0.13	-0.25	1.00	-0.02
revol.bal	-0.17	0.08	0.35	0.19	-0.02	1.00

Figure 3 - 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.

c. Contextual Interpretations

While these observations align with conventional financial knowledge, it is important to establish statistical significance. Further statistical tests will be conducted to confirm these preliminary insights.

3. Model Building

a. Data cleaning

No data cleaning was necessary to build the models.

b. Given that discriminant analysis assumes predictors to come from a multivariate normal distribution, categorical predictors are not suitable for inclusion in the model. Therefore, we chose to use some of the quantitative predictors in our data set that seemed most likely to be influential. The predictors selected for both the logistic regression and linear discriminant analysis (LDA) models are:
   - Interest Rate (int.rate)
   - Log of Annual Income (log.annual.inc)
   - Debt-to-Income Ratio (dti)
   - FICO Credit Score (fico)
   - Revolving Balance (revol.bal)

c. **Logistic Regression Model**

```
##
## Call:
## glm(formula = credit.policy ~ int.rate + log.annual.inc + dti +
##     fico + revol.bal, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4266   0.1426   0.3971   0.6947   1.7795
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.534e+01  1.655e+00 -15.313  < 2e-16 ***
## int.rate       -5.015e+00  2.230e+00  -2.249   0.0245 *
## log.annual.inc  4.747e-01  7.591e-02   6.254 3.99e-10 ***
## dti             1.335e-02  6.242e-03   2.138   0.0325 *
## fico            3.215e-02  1.962e-03  16.386  < 2e-16 ***
## revol.bal      -2.295e-05  1.753e-06 -13.091  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4812.7  on 4788  degrees of freedom
## Residual deviance: 3871.9  on 4783  degrees of freedom
## AIC: 3883.9
##
## Number of Fisher Scoring iterations: 5
```

Figure 3 - Logistic Regression Model

The logistic regression model is:

$$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -25.34 - 5.02 * int.rate + 0.47 * log.annual.inc + 0.013 * dti + 0.032 * fico - 0.00002295 * revol.bal$$

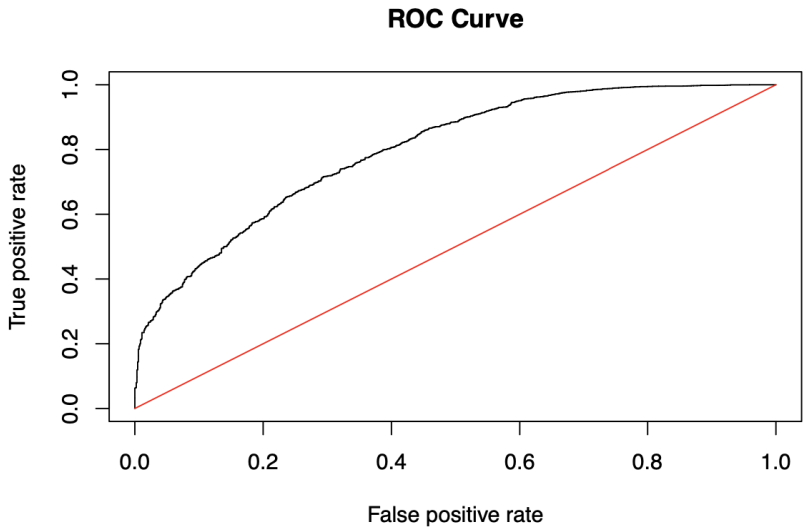

Figure 4 - ROC Curve of the Logistic Model

Figure 3 contains the output of the logistic regression model fit using the selected predictors. The summary indicates that all predictors have some degree of statistical significance. Figure 4 contains the ROC curve fit to the above logistic regression model. The ROC curve lies significantly above the diagonal line of random guessing, indicating that the model has good discriminative power. The AUC value of approximately 0.798 further confirms the model's ability to distinguish between the classes effectively.

K-fold cross-validation was performed with both k=5 and k=10, and the estimated test error rates were found to be approximately 0.19 and 0.192, respectively. These consistent results across different k-values suggest that the model is stable and not sensitive to the particular folds used for validation. The actual test error rate was 0.147.

Linear Discriminant Analysis

We are performing two tests for each credit policy result, one for yes and one for no. For each test, the null hypothesis is that the distribution of the predictors is consistent with a multivariate normal distribution, and the alternative hypothesis is that the distribution of the predictors is not consistent with a multivariate normal distribution. According to Figure 5, for both tests for both results, we reject the null hypothesis, so we don’t have evidence the assumption for discriminant analysis is met. While the assumption of multivariate normality is important for LDA, its violation doesn’t necessarily mean that LDA cannot be used for classification; it just may not be the optimal choice for our data.

```
##
##  Multivariate Normality Test Based on Skewness
##
## data:  data_1
## U = 5815.5, df = 5, p-value < 2.2e-16
```

```
ICS::mvnorm.kur.test(data_1)
```

```
## Warning in pchisqsum(n * W.stat, df = dfs, a = chi.fac, method =
## "integration"): Probable loss of accuracy

##
##  Multivariate Normality Test Based on Kurtosis
##
## data:  data_1
## W = 36227, w1 = 0.73469, df1 = 14.00000, w2 = 1.14286, df2 = 1.00000,
## p-value = 2.143e-11
```

```
ICS::mvnorm.skew.test(data_0)
```

```
##
##  Multivariate Normality Test Based on Skewness
##
## data:  data_0
## U = 9256.8, df = 5, p-value < 2.2e-16
```

```
ICS::mvnorm.kur.test(data_0)
```

```
## Warning in pchisqsum(n * W.stat, df = dfs, a = chi.fac, method =
## "integration"): Probable loss of accuracy

##
##  Multivariate Normality Test Based on Kurtosis
##
## data:  data_0
## W = 488099, w1 = 0.73469, df1 = 14.00000, w2 = 1.14286, df2 = 1.00000,
## p-value = 2.143e-11
```

Figure 5 - Test of Multivariate Normality

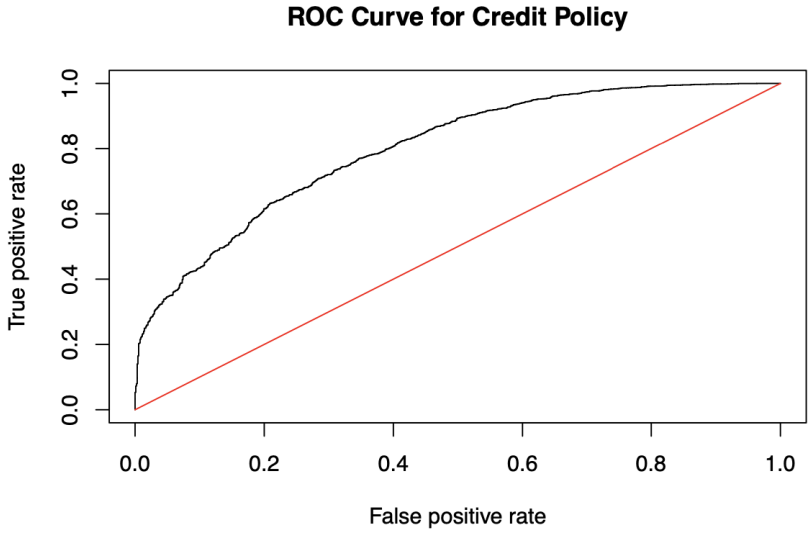

Figure 6 - ROC Curve for LDA

For the LDA model, the ROC curve also lies above the diagonal, and the AUC value is approximately 0.799, which is very close to that of the logistic regression model. These metrics suggest that both models are comparable in their predictive performance. The estimated test error rate using 5-fold cross-validation is approximately 0.167, and with 10-fold, it's around 0.166. However, the LDA model shows a slightly higher actual test error rate, 0.162, than that of logistic regression, indicating it might be a less accurate model for this particular problem; when coupled with the fact that the multivariate normality assumption is not met, the logistic regression model seems the better choice for our data.

d. Improved Logistic Model

After performing stepwise selection on the logistic regression model, we dropped the predictor 'dti' from our model, as it was found relatively insignificant, and the following new predictors were added:

Variable Name	Description	Type
installment	The monthly payments in dollars owed by the borrower if the loan is funded	Quantitative
days.with.cr.line	The number of days the borrower has had a credit line	Quantitative
revol.util	The amount of credit used relative to the total credit available to the borrower	Quantitative
inq.last.6mths	The number of times creditors have made inquiries about the borrower in the last 6 months	Quantitative
delinq.2yrs	The number of times the borrower has been 30+ days past due on a payment in the past 2 years	Quantitative

These are predictors that could provide additional nuances to the model that might not have been captured in the initial logistic regression. By making these changes, we aimed to improve the model's predictive accuracy while also capturing a broader range of factors that could influence the credit underwriting criteria.

```
##
## Call:
## glm(formula = credit.policy ~ int.rate + 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
## -4.0903   0.0362   0.1672   0.3837   2.8308
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.769e+01  2.727e+00 -13.821  < 2e-16 ***
## int.rate           6.808e+00  3.386e+00   2.010  0.04439 *
## installment        9.472e-04  3.427e-04   2.764  0.00571 **
## log.annual.inc     6.602e-01  1.108e-01   5.959 2.54e-09 ***
## fico               4.685e-02  3.097e-03  15.127  < 2e-16 ***
## days.with.cr.line  1.492e-04  2.766e-05   5.394 6.87e-08 ***
## revol.bal         -4.404e-05  2.587e-06 -17.026  < 2e-16 ***
## revol.util         9.470e-03  2.305e-03   4.109 3.97e-05 ***
## inq.last.6mths    -1.036e+00  3.885e-02 -26.656  < 2e-16 ***
## delinq.2yrs       -1.405e-01  9.191e-02  -1.529  0.12629
## not.fully.paid    -3.263e-01  1.321e-01  -2.469  0.01353 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4812.7  on 4788  degrees of freedom
## Residual deviance: 2281.1  on 4778  degrees of freedom
## AIC: 2303.1
##
## Number of Fisher Scoring iterations: 7
```

Figure 7 - Summary of the Improved Logistic Model

The improved logistic model is:

$$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -37.59 + 6.81 * int.rate + 0.00095 * installment + 0.67 * log.annual.inc + 0.047 * fiico + 0.00015 * days.with.cr.line - 0.00004 * revol.bal + 0.009 * revol.util - 1.04 * inq.last.6mths - 0.14 * delinq.2yrs$$

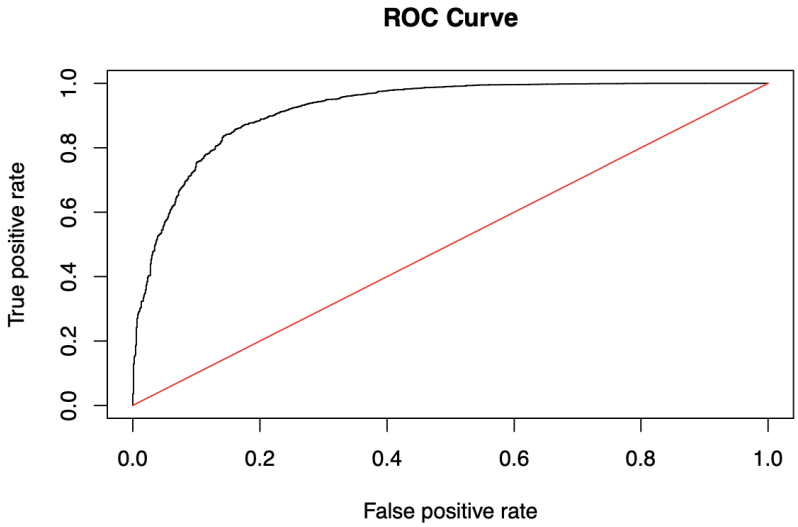

Figure 8 - ROC Curve for the Improved Logistic Model

Figure 7 contains the output of the improved logistic regression model fit using the selected predictors. The summary indicates that all predictors have some degree of statistical significance, except for delinq.2yrs. Figure 8 contains the ROC curve fit to the above logistic regression model. The ROC curve in Figure 8 lies even further above the diagonal line of random guessing than it does in Figures 4 and 6, indicating that the model has good discriminative power and is an improvement over the previous logistic regression model. The AUC value of approximately 0.921, compared to an AUC value of approximately 0.798 for the previous logistic regression model, further confirms this improvement.

K-fold cross-validation was performed with both k=5 and k=10, and the estimated test error rates were both found to be approximately 0.24. These consistent results across different k-values suggest that the model is stable and not sensitive to the particular folds used for validation. The actual test error rate was 0.096. While the K-fold cross-validation errors for the improved logistic regression model are higher than those for the previous logistic regression model, the actual test error rate is significantly lower. Given that the previous logistic regression model performed better than our LDA model, and the improved logistic regression model performed better than our original logistic regression model, the improved logistic regression model performed better than our LDA model as well. The ROC curve and the AUC of the improved logistic regression model are better than those of our LDA model, and while our k-fold error rates for the improved model were higher than those of our LDA model, the test error rate was much lower.

e. In our original logistic regression model, all of our predictors were significant at the 0.05 significance level. This is not surprising, as we chose the predictors we thought would be most influential on the probability of meeting credit underwriting criteria based on our previous finance knowledge and understanding of credit risk. Our initial theory that these predictors would be significant was correct.

In the improved logistic regression model, all predictors were significant at the 0.05 significance level, with the exception of delinq.2yrs. Despite being a new addition, it did not prove to be statistically significant in the model, which was surprising given that past financial delinquencies, or being past due on loan payments, are often considered a strong indicator of credit risk. The other predictors being significant was not surprising, as most of the new variables measure risky attributes about a borrower, and we thought they would be significant in predicting if a borrower meets the credit underwriting criteria or not.

Looking at the values of the coefficients of each predictor in the model, we can get a sense of how important each predictor is. The exact importance of each of the predictor variables is hard to quantify as the distributions of each of them are different. However, we can make an estimate for interpretability.

For an average person being evaluated, the predictors that have the most impact on the chance that they'll be approved for a loan are the days.with.cr.line, fico, and log.annual.inc variables. On the other hand, we see that inq.last.6mths is a serious demerit, and having multiple inquiries in the past six months can immediately offset having high values of days.with.cr.line, fico, and log.annual.inc. We can estimate the importance of each variable by multiplying the median of the values of the predictor by its calculated coefficient in the model to get the average increase in log odds that the predictor provides.

```
## Call:
## lda(credit.policy ~ int.rate + log.annual.inc + dti + fico +
##     revol.bal, data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.2015034 0.7984966
##
## Group means:
##    int.rate log.annual.inc      dti     fico revol.bal
## 0 0.1381846       10.89136 13.86980 683.9534  28147.03
## 1 0.1189532       10.94583 12.20506 717.4145  13638.74
##
## Coefficients of linear discriminants:
##                            LD1
## int.rate       -7.646116e+00
## log.annual.inc  3.119122e-01
## dti             1.469626e-02
## fico            2.128408e-02
## revol.bal      -1.579542e-05
```

Figure 9 - LDA Model Output

f. Below is an interpretation of the coefficients of each linear discriminant:

- int.rate: The negative coefficient suggests that higher interest rates are associated with a lower likelihood of meeting the credit underwriting criteria. This makes intuitive sense as higher-risk borrowers are less likely to meet credit policy standards.

- log.annual.inc: The positive coefficient indicates that higher annual income on a logarithmic scale is associated with a higher likelihood of meeting the credit policy standards.

- dti (Debt-to-Income Ratio): The positive but small coefficient suggests that a higher debt-to-income ratio is associated with a slightly higher likelihood of meeting the credit policy standards.

- fico: The positive coefficient for FICO scores suggests that higher credit scores are associated with a higher likelihood of meeting the credit policy standards, which is expected.

- revol.bal: The negative but very small coefficient suggests that a higher revolving balance is associated with a slightly lower likelihood of meeting the credit policy standards.

  Something we found surprising from this output is that although the coefficient for the linear discriminant of dti is positive, meaning that a higher debt-to-income ratio is associated with a slightly higher likelihood of meeting the credit policy standards, the group in the training data who did meet the criteria have a lower mean debt-to-income ratio than the group who did not meet the criteria. This can likely be attributed to the fact that as mentioned above, our data violates the multivariate normality assumption of LDA.

4. Conclusions

a. Discussion

Through analysis of our models, we were able to provide direct answers to our question of interest. It does appear that a higher FICO credit score leads to a higher likelihood of meeting the credit underwriting criteria, due to the fico variable having a positive, significant relationship with our response variable credit.policy in all of our models and our models having good ROC curves, good AUCs, and good classification abilities.

We were also able to determine the impact of each of the other variables proposed in our question of interest, which were revolving balance, annual income, debt to income ratio, and the number of recent credit inquiries. In each of our models, these variables exhibited significant relationships with our response variable, with the exceptions of dti, which was significant in our original logistic regression model but was dropped in our improved model due to stepwise selection, and number of recent credit inquiries, which we only included in our improved logistic regression model but proved significant.

We were also able to determine that there was a significant relationship between all of the other variables we wanted to explore as they related to the likelihood of meeting the credit underwriting criteria and included in our improved logistic regression model, with the exception of delinq.2yrs. These other variables were important to our question of interest, as the question was worded as “other financial factors such as”, meaning there were more variables to explore than just those explicitly stated in our question.

Overall, with almost all of the variables that we explored in our models and mentioned or alluded to in our question of interest, we were able to determine that they had a significant relationship with the likelihood of a borrower meeting the credit underwriting criteria, and we were able to determine the nature of that relationship (whether higher values of that variable led to a greater or lesser likelihood) with models that were suitable for classification and inference (excluding LDA for inference), which is what we set out to do with this milestone.

b. Insights

In preliminary discussions about our variables in our group, all of the natures of the relationships between our predictor and response variables that we theorized turned out to be correct predictions from interpreting the results of our analysis, with the exceptions of dti and delinq.2.yrs.

Delinq.2.yrs was found to be insignificant after adding it to our improved logistic regression model, which was very surprising to us, for we theorized that it would be a great predictor of the likelihood of a borrower meeting the credit underwriting criteria or not. Banks decide not to give loans out if they do not think people will be able to pay them on time, and the number of financial delinquencies in the past two years is a direct record of how many times a borrower has been late on loan payments in the past two years, so we definitely theorized a significant negative relationship between the two. The relationship was negative, but it was not significant; this could simply be a result of a special case of the borrowers in our dataset, or it could be due to underlying outside factors not explored in our analysis.

For dti, or debt to income ratio, we theorized a significant negative relationship with the likelihood of meeting the credit underwriting criteria, for a higher ratio of debt to income

generally means less financial stability, which should mean a borrower is less likely to be granted a loan. The relationship did not turn out to be significant in our improved logistic regression model, but furthermore, even when it was significant in our original logistic regression and linear discriminant analysis models, the relationship was positive, meaning a higher debt to income ratio led to a higher likelihood of a borrower being classified as meeting the credit underwriting criteria. This directly countered our initial hypothesis, but again, it could be due to a special case of the borrowers in our dataset, or it could be due to outside underlying factors.

c. Challenges

One challenge that our group faced was the violation of the multivariate normality assumption in our linear discriminant analysis. We rejected the null hypothesis that the distribution of predictors is consistent with a multivariate normal distribution, meaning that the assumption is not met and we need to be cautious when interpreting our resulting model. This also potentially led to some weird results in our LDA output such as the coefficient of the linear discriminant of dti being positive not aligning with the differences in the group means between those who met the underwriting criteria and those who did not. For classification purposes however, this violation of the multivariate normality is not a problem.

Our improved logistic regression model performed a good deal better than our LDA model and would be a better choice for the main model of our project anyway, so this challenge discussed above does not present a significant obstacle in our project.

Another challenge that our group faced was a technical issue; for a few days, we could not get our k-fold validation code to work, for it was not returning numerical values. However, after discussing with Professor Woo and looking at the code in detail a bit more, we figured out that we had to convert our response variable to a factor, which solved the issue, and we were able to continue with our project.