Patrick Burns, Michael Harder, David Okposio
Dr. Maadooliat
MSSC 5750, Computational Statistics
5 May 2025

MSSC 5750 Final Project Report

I. Abstract:

With increasing market volatility, secure investments and financial strength are more important than ever. According to Moody's [10], corporate default rates have more than doubled since 2021. To that end, it is more critical than ever that banks heavily scrutinize which personal loan applications they approve. In this paper, we propose multiple linear classification models to help assess the risk of default for individual loans and seek to explain which factors make a person more or less trustworthy. We employ lasso regression for dimensionality reduction, logistic regression and linear discriminant analysis to predict successful repayment, and then assess variable importance using SHAP values. We had moderate success in model accuracy, however, the interpretation of variables was counterintuitive to what we had expected.

II. Introduction

Banking institutions are a cornerstone of the American economy, not only for protecting assets but for assessing credibility and issuing loans. Taking on loans is critical for most Americans whenever they need to make a major purchase. Whether purchasing a house, a new car, tuition at a university, or preparing for a wedding, loans make these major purchases possible. While it is very helpful that banks can offer this service, it does not simply come from the goodness of their hearts and trust in their borrowers. While the interest may seem to be lucrative, defaults on loans pose a large risk as well, so a bank must assess how likely borrowers are to pay back their loans. In this paper, we explore several different classification models that can help determine this risk so banks can decide which applicants should be allowed to borrow.

III. Background

Banks consider several factors when deciding whether or not to approve a loan. Many people will talk about the "5 C's of credit", which outlines the five important factors that are considered [11]. Those Cs being Character, Capacity, Capital, Collateral, and Conditions. Character is an assessment of a person's reputability and trustworthiness, as evidenced by things like their career and credit score. Capacity is going to be estimating how capable you are of

repaying based on existing debt and current income. Capital assesses your assets beyond just the salary you are earning. Collateral is what you can offer back in the case that you cannot make payments, and Conditions would be any other information, such as what the loan is intended for or the current market volatility that could impact your ability to pay them back.

1. Logistic Regression

Overview: Logistic Regression is a widely used statistical method for binary classification problems, such as predicting loan default.

Strengths:

- **Interpretability:** Provides clear insights into how each input variable affects the probability of default, making it suitable for regulatory compliance.
- **Simplicity:** Easy to implement and computationally efficient.
- **Baseline Performance:** Often serves as a strong baseline model in credit scoring tasks.

Weaknesses:

- **Linearity Assumption:** Assumes a linear relationship between independent variables and the log odds of the dependent variable, which may not capture complex patterns.
- **Sensitivity to Multicollinearity:** Performance can degrade when independent variables are highly correlated.
- **Interpretability:** While usually a strength, in cases with class imbalance (like ours), estimation bias can be magnified [3].

2. Linear Discriminant Analysis

Overview: LDA is a classification technique that finds a linear combination of features that best separates two or more classes [8].

Strengths:

- **Efficiency:** Can perform well when class distributions are approximately normal with equal covariance matrices.
- **Dimensionality Reduction:** Reduces feature space while preserving class-discriminatory information.

Weaknesses:

- **Assumption Sensitivity:** Performance deteriorates if the assumption of equal class covariances is violated.

- **Less Flexible:** Not suitable for capturing non-linear relationships.

3. Lasso Regression

Overview: LASSO (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization to enhance prediction accuracy [7].

Strengths:

- **Feature Selection:** Automatically selects important features by shrinking less important feature coefficients to zero.
- **Overfitting Reduction:** Helps prevent overfitting, especially in high-dimensional datasets.

Weaknesses:

- **Bias Introduction:** Can introduce bias by shrinking coefficients, potentially affecting model accuracy.
- **Variable Selection Instability:** Selection can be unstable when variables are highly correlated.

4. SHAP Values (SHapley Additive exPlanations)

Overview: SHAP values are a unified approach to explain the output of any machine learning model, based on cooperative game theory [1][6].

Strengths:

- **Model-Agnostic:** Can be applied to any machine learning model to interpret predictions.
- **Consistent Feature Attribution:** Provides consistent and locally accurate feature contributions.
- **Transparency:** Enhances model transparency, aiding in regulatory compliance and trust.

Weaknesses:

- **Computational Complexity:** Can be computationally intensive, especially with large datasets or complex models.

- **Interpretation Challenges:** While providing detailed explanations, interpreting SHAP values requires statistical expertise.

IV. Data

The data that we examined came from Kaggle and is structured one row per person, with the outcome variable being whether or not that person successfully paid off their loan (loan_status = 1). The data set is sizable, with 13 attributes and 45,000 rows. It is moderately imbalanced, with only 22% of people successfully paying back their loans. Unfortunately, the dataset is synthetic, so it may not be a perfect representation of reality.

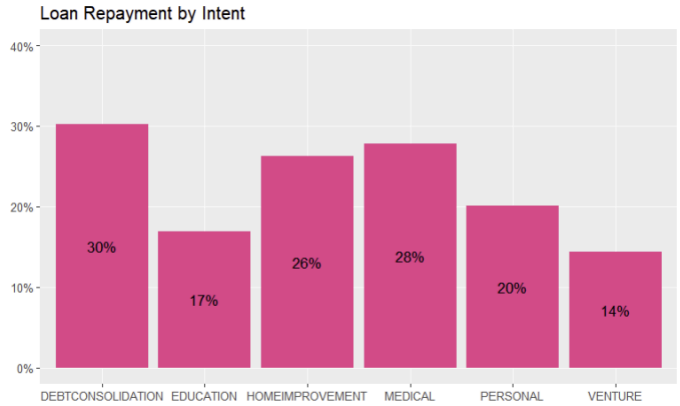
Many of the variables included in our dataset can be directly tied to the 5 Cs. For example, included is each person's credit score and years of experience working for their company, which would fall under Character. We have income and amount of loan which falls under capacity, home ownership would be Capital/Collateral, and loan intent would be a Condition. So we have very good coverage across all the main areas considered for loan approval.

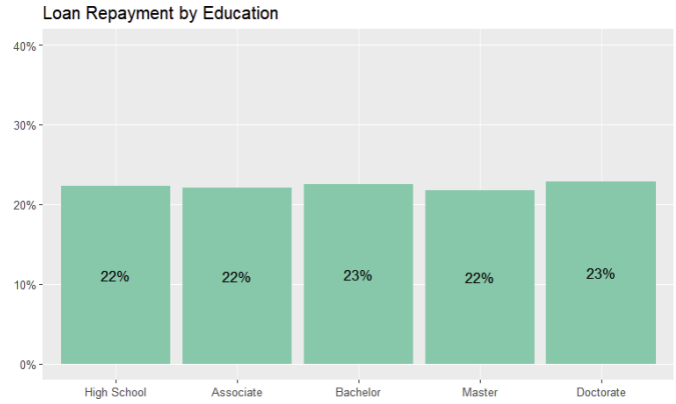
Upon conducting some basic data exploration, we were fortunate that there was no missing data, however, we did find some suspicious patterns. When looking at the age of applicants, we found that there were 7 records above age 100. While this is not necessarily impossible, taking a closer look at those people, they were ages 109, 116, two people were 123, and three were age 144. Certainly, nobody was applying for a loan at that age, so we needed to do some cleaning. Given our knowledge of loans, and that the majority of people applying for loans will be young adults to middle age, there is an expected right skew where the majority of our data falls to the left. On top of this, we aren't necessarily expecting a linear effect. On average, a person aged 20 is much less trustworthy than someone aged 30, but further down, is someone aged 55 much more trustworthy than 45? Probably not. To account for this and clean up our outliers, we bucketed age to <25, 25-30, 31-44, 45-59, and 60+. Similar to age, the prior years of experience (at their job) had a similar pattern with extreme outliers over 100 years. This was much less severe, so instead we just capped this variable at 50 years of experience.

One major anomaly we found in our data set was with the variable indicating whether or not a person had previously defaulted on a loan. What was odd here was that it formed a clean class separation with our target variable, loan status. For every single data point where a person had previously defaulted, they also had a loan status = 0. This does not mean everybody who did not default had loan_status = 1, but still, with such a strong pattern, it likely will not make sense to include this variable in our models.

Looking at some of our categorical variables, we immediately could see some differing relationships to our outcome variable. When we examine loan repayment by education level, there is almost no difference in success rate. However, when we examine repayment by loan

intent, there are very sizeable differences. This suggests that education may not be a great predictor, but intent could play a larger role.



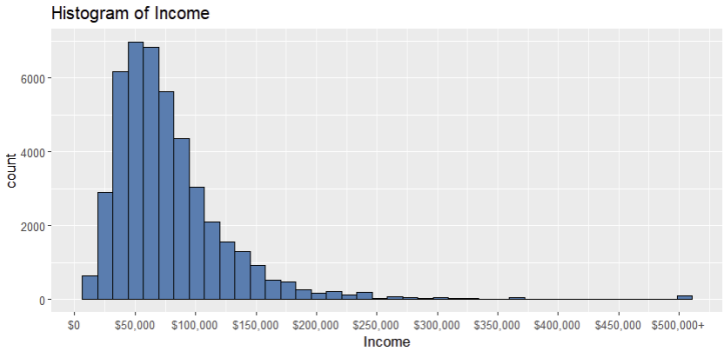


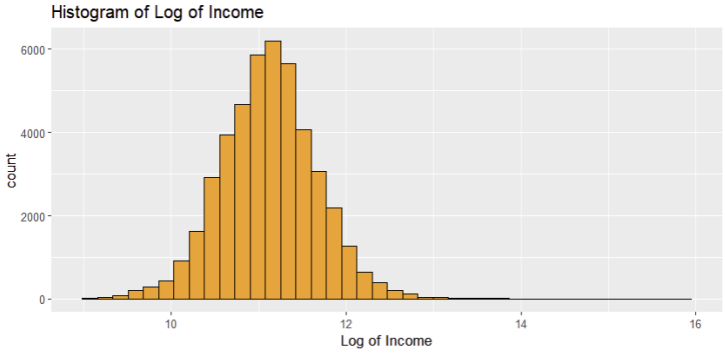
Another oddity in our data was that we did not see any immediate relationship between credit score and loan repayment. We have a good range of credit scores, but if we look at the simple correlation between credit score and loan status, there is almost no correlation, with a coefficient of -0.01. If we just compare the lowest quartile of credit scores to the highest, 1% more of the lowest quartile successfully repaid. This suggests that credit score does not have a very strong relationship with trustworthiness, which is counterintuitive and may be indicative that this data set is not very realistic.

V. Methodology

i. Data Transformations

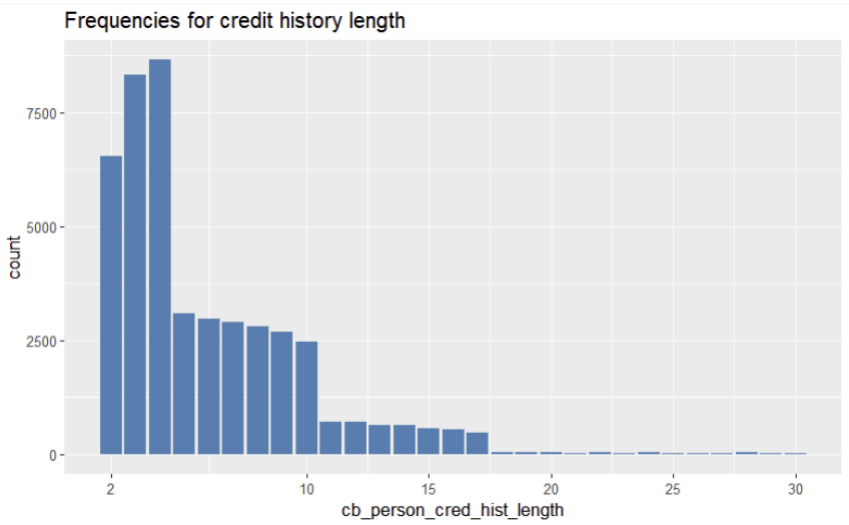
Before putting our variables through any kind of linear regression, we needed to do some further data cleaning to prepare categorical variables, scale numeric variables, and prevent multicollinearity. For our categorical variables, we followed a one-hot-encoding approach to transform the variables, removing the first of each category as our reference. These variables were gender, education, home ownership, loan intent, and age, and our reference categories became female, high school education, mortgage, loan intent of debt consolidation, and the <25 age bucket.





As previously discussed, some of our variables had a very large right skew with many far-right outliers. To prepare these variables for modeling, we used a log transformation to help normalize their distributions. Income, for example, ranged from $8,000 to over $7,000,000. In the attached histogram, you can see the distribution before and after the log transformation, successfully correcting the skew and removing extreme outliers. Similar to income, we had a lot of skew in the prior experience a person had working for their company. To correct this, we first limited extreme answers to a very large but still realistic value (50 yrs), and then took a log transformation of that, as 75% of our data fell in the range of 0-8 years. We did have to add a constant (0.5) before transforming, as this variable started at zero.

We also had a right skew to each person's credit history length. Looking at the frequency of the length, we can see we have a lot of people falling in the 2-10 range, a fair amount in the 11-17 range, but then very, very few 18 and above. Since the frequencies are so small, we set anyone who has a credit length of 18 or more years to 18.



ii. Previous Loan Defaults

As previously discussed, we discovered that in every situation where a person had previously defaulted on a loan, their loan status was 0, as if they had defaulted again. While this seems like a strong predictor, it is going to throw off our model's coefficient interpretability. To account for this, we split our models into a two-step process. We train them all without including previous loan deferrals. We evaluate that accuracy, but then afterwards we take any situation where a record had a previous loan deferral, and adjust the model prediction to class 0. Afterwards, we re-evaluate accuracy. This approach allows us to still utilize this information (even though it seems unrealistic), without breaking any model assumptions.

iii. Lasso Regression

Before doing any other types of regression, we first created a lasso regression model using all of our variables (after cleaning and transforming) to reduce the dimensionality of our data. Our data is not exceptionally wide, but as we have seen in early data exploration, many variables appear to have a weak relationship to our outcome, and lasso can help us reduce some of the noise. We choose lasso regression for this purpose because of the term it introduces to penalize complexity, such that weaker predictors are given a coefficient of zero. Once the lasso has eliminated some of the noisy variables, we can use this to inform which variables we will

pass on to other models. For our Lasso model, we use a cross-validation technique, so we are passing our full dataframe through rather than a training subset.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = SSE + \lambda\sum_{j=1}^{p}|\beta_j| \quad (1)$$

Where $\lambda \geq 0$ *is the tuning parameter.*

The β_i*'s are the coefficients of each feature considered.* As (1) is minimized some of the coefficients are reduced to zero thereby yielding dimensionality reduction and avoiding overfitting.

iv. Logistic Regression

We take three different approaches to logistic regression. First, we create a model using all of our variables, just like we had for our Lasso model. Second, we create a model that uses only the variables selected by Lasso. The idea here is that we can compare accuracy between the two and see if there appears to be any information lost by eliminating variables. Finally, we create a third logistic regression model, this time trained on a rebalanced dataset. For the first two methods, we train the models on a training dataset made up of 80% of the data, then evaluate accuracy on the holdout 20%. For our third method, we do the same, except we conducted undersampling on the training data such that we had about 8000 records belonging to each of our classes. By rebalancing our classes, this third logistic model should be able to better identify our positive class, while the others may try to improve overall accuracy based on the negative class being more frequent.

For our binomial response variable, we have y~ Binomial(p_i)

$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ is the linear predictor across the real line. Since we want predictable probabilities on (0,1) we have:

$$p_i = \frac{exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

v. Linear Discriminant Analysis

We next take a look at the numerical values in our data set and work to conduct a linear discriminant analysis using loan status as our response variable. The data set is filtered into two groups namely loan defaulters and non-loan defaulters and we obtain the sample covariance matrix for both groups.

So we have: $y_{11}, y_{12},\ldots y_{1n_1} \sim N(\mu_1, \Sigma)$ and $y_{21}, y_{22},\ldots y_{2n_2} \sim N(\mu_2, \Sigma)$

$z_{1j} = a'y_{1j} = \sum_{j=1}^{n_1} z_{1j}$ and $z_{2j} = a'y_{2j} = \sum_{j=1}^{n_2} z_{2j}$

Using the mean $\overline{z_1}$ =a`$\overline{y_1}$ and $\overline{z_2}$ =a`$\overline{y_2}$ and the pooled sample covariance matrix,S_{pl} it is easy to obtain the Mahalanobis' squared distance which is maximized to separate between the two groups.

$D^2 = \frac{(z_1 - \bar{z}_2)^2}{s_z^2}$

Where $S_z^2 = a'S_{pl}a$ and

$S_{pl} = \frac{1}{n_1+n_2-2}[(n_1 - 1)S_1 + (n_2 - 1)S_2]$

S_1 and S_2 are the respective covariance matrices for each group.

Upon implementing the linear discriminant function in R we found that loan_percent_income is the most significant variable in separating between the two groups. Loan interest rate was a not so close second, and income ranked quite low. In making sense of this data we think about the level of leverage a borrower has. i.e the percentage of income spent servicing debt is a major factor in predicting their potential for default. A highly leveraged borrower has little wiggle room for unexpected shocks, such as a sudden job loss, an unexpected expense etc.

```
person_age                   0.0112
person_income                0.0000
person_emp_exp              -0.0118
loan_amnt                   -0.0001
loan_int_rate                0.2328
loan_percent_income         13.0457
cb_person_cred_hist_length   0.0012
credit_score                -0.0002
```

vi. Accuracy evaluation

After building our different models, we compare their efficacy in a few different ways. We generate a confusion matrix for each, and first examine overall accuracy, which percent of the records that were classified correctly. In addition to this, we evaluate both sensitivity and specificity. Sensitivity will give us an idea of how well the model performs on our

positive/minority class (loan repayment), while specificity tells us how well it performs on the negative class (loan default). In the real world, mislabeling people who default is going to be more harmful to a bank than mislabeling those who would have repaid, so we should pay close attention to our specificity.

vii. Shapley Values

After creating each of the above models and evaluating their accuracy, we close the analysis with the generation of Shapley values for our balanced logistic regression model. SHAP values are incredibly helpful here because they give us easily interpreted, model-agnostic interpretations of variable importance [4]. Our goal with calculating these is to take an exploratory approach to the problem and build an understanding of what factors play the largest role in loan repayment.

VI. Results

i. Lasso Variable Selection

In the snippet to the right, we examine the coefficients selected by our Lasso model. All of the variables with a period suggest that the coefficient was reduced to zero; it was not important to our outcome, and so we went on to drop these from our log-lasso model. Interestingly, we dropped the entirety of our age and gender variables, telling us that neither of these is relevant to whether or not somebody will repay their loan. Similar to what we found with our early exploration, credit score and education did not help to predict loan repayment either, and were also dropped.

```
> coef(lasso)
25 x 1 sparse Matrix of class "dgCMatrix
                                      s1
(Intercept)                    4.1470900
loan_amnt                      .
loan_int_rate                  0.2969195
loan_percent_income            8.8790399
cb_person_cred_hist_length     .
credit_score                   .
log_person_income             -0.9462135
log_person_emp_exp             .
`person_age_25-30`             .
`person_age_31-44`             .
`person_age_45-59`             .
`person_age_60+`               .
person_gender_male             .
person_education_Associate     .
person_education_Bachelor      .
person_education_Master        .
person_education_Doctorate     .
person_home_ownership_OTHER    .
person_home_ownership_OWN     -1.0656434
person_home_ownership_RENT     0.7243193
loan_intent_EDUCATION         -0.5302459
loan_intent_HOMEIMPROVEMENT    0.0885264
loan_intent_MEDICAL            .
loan_intent_PERSONAL          -0.2704400
loan_intent_VENTURE           -0.7266126
```

ii. Model Accuracy without the previous default variable

When we evaluate each of the 5 models created, we see fairly consistent results across the board. We saw the highest overall accuracy for our lasso model, closely followed by our full logistic regression and the logistic regression with the lasso-selected variables. While these models stood out for overall accuracy, they had very poor sensitivity, sitting under 50%. This means that for over half of the applicants who did repay their loans, we are predicting they would not, and they would never receive any money. While less important than specificity, this is probably an unacceptably low rate. As expected, the logistic regression model trained on the balanced dataset performed much better here, but it came at the cost of the high specificity and overall accuracy that the first models had. Interestingly, even though LDA was not trained on a rebalanced dataset, it performed very similarly to our balanced logistic model.

Model <chr>	accuracy <dbl>	sensitivity <dbl>	specificity <dbl>
Lasso	0.850	0.482	0.955
Log	0.839	0.488	0.944
LogLasso	0.836	0.476	0.944
LogBalanced	0.774	0.770	0.775
LDA	0.775	0.737	0.786

iii. Model Accuracy, including the previous default variable

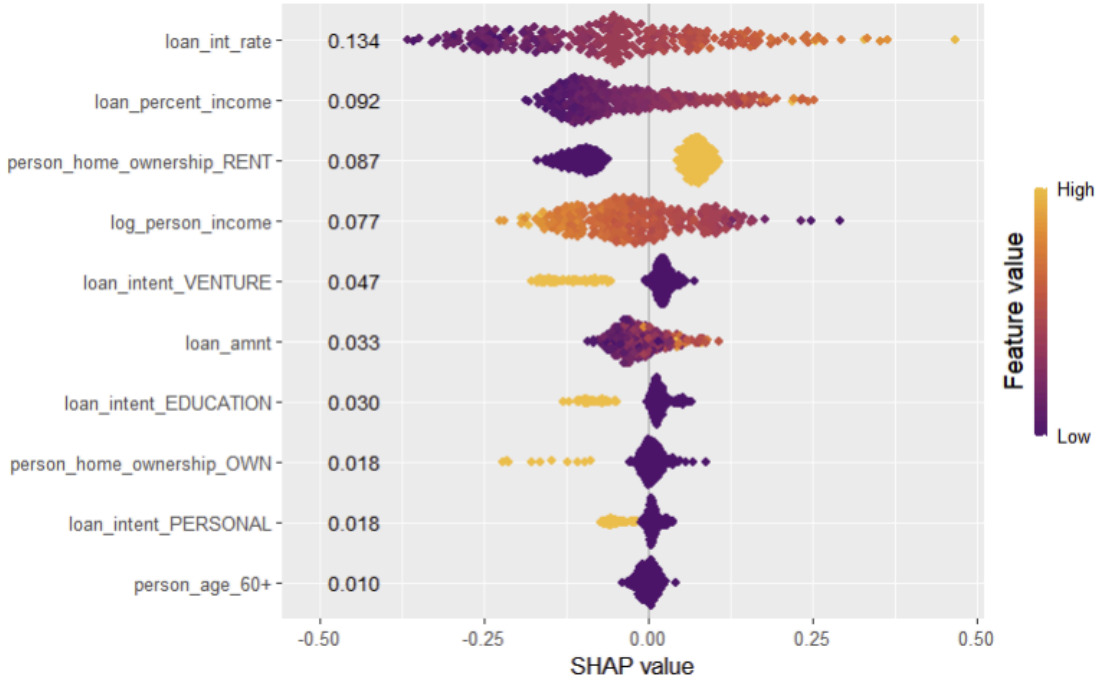
Using the same models we trained for the section above, we are now introducing a second layer prediction, which is very simply changing all predictions for records that had a previous default to a predicted class of 0. This is based on the fact that our prior knowledge (in this data set, not necessarily reality) tells us that this will always be the case. Once we do this, we see a slight improvement in our Lasso and first two logistic regression models, but then a major increase in overall accuracy and specificity for our rebalanced and LDA models. After doing this, it appears that our rebalanced logistic model is the strongest in terms of overall accuracy and sensitivity, while still maintaining a strong specificity of 92%.

Model <chr>	accuracy <dbl>	sensitivity <dbl>	specificity <dbl>
Lasso	0.873	0.482	0.984
Log	0.868	0.488	0.983
LogLasso	0.866	0.476	0.983
LogBalanced	0.887	0.770	0.922
LDA	0.884	0.737	0.926

iv. SHAP Values

To interpret our SHAP values, we create a beeswarm plot for 150 of the records in our test dataset. Each dot represents one of these records, the color represents the value for that variable, and how far left or right it falls represents how much of an impact that variable had on the predicted class for the records. When we view this all together in a beeswarm plot, we get a clear picture of the most important variables and values.

At the top, we can see that the most important variables are the interest rate and what percentage of each person's income that is. We see that higher values fall towards the right, which tells us that loan applicants with a high interest rate loan and applicants that will be using a large share of their income to pay it back get positive SHAP values, suggesting that they are more likely to repay their loan. We also observe that renters are more likely to repay loans, and following a similar pattern to interest rates and income, we surprisingly see that higher-income individuals are more likely to default. Continuing down the next most important variable was intending to use your loan for a venture, followed shortly by intending to use it for education. We see that this makes individuals significantly more likely to default, which makes sense as this is a more risky behavior. Any of the variables that did not fall within the visual below had very little impact on the predicted class for these records. For the most part, these variables agree with the variables and their coefficients selected by the Lasso model above.



VII. Discussion

i. Discussion

Two variables stood out as significantly more influential to whether or not a person would repay their loans, and those were the loan interest rate and the loan percentage of income. This can be acknowledged in both the lasso coefficients as well as the beeswarm plot. Oddly, these both show the opposite relationship with repayment than we would expect. Similarly, for home ownership, we saw that renting meant someone was more likely to repay, versus homeownership, which suggested they would default. This goes against the Cs of capital/collateral, where home ownership should have strengthened a person's trustworthiness. Perhaps most strange of all, credit score had no significant role in our models, which is typically the quickest and most straightforward way that people will assess trustworthiness. We were not able to properly assess the importance of a person having a previous default on file. In this data set, it seems to be extremely important, but because we have zero instances of people with prior defaults still getting approved for a loan, we know this is not a representation of reality.

At first, we assumed that we had overfit our models, but after looking at the raw data and visualizations of the distributions and their relationship to our target variable, we had confirmation that this was the pattern in our data. Likely, this data is only loosely based on reality.

We were quite happy with the results of our lasso regression. This was very helpful for quickly narrowing to a much smaller set of variables, and when we compare our full logistic

regression against the logistic regression with only the lasso-selected variables, we saw almost the same accuracy measures. This meant that those variables were truly adding very little to the model, and we were able to make a much simpler model that was just as strong.

ii. Future Work

One major limitation of this project is that the banking data available to us was synthetic. While this was nice in that no data was missing, a lot of our variables' relationships to our target variable seem questionable, and so whatever insights we draw about the relationship of these different factors to loan repayment should be taken with a grain of salt. Nonetheless, one could take a similar approach to what we have already started here on a better data set and find their results.

One potential problem we may be running into here is some multicollinearity. In future work, it would be worthwhile to continue to explore how independent these variables are from one another, perhaps taking a look at VIF values. In our model, we included the log of income, loan amount, and loan percent of income. Looking back, it may have been wiser to drop percent income, as that information is already included across our income and loan amount variables. This could be negatively impacting our results, although hopefully, by transforming our income and then taking the log, we have successfully mitigated at least part of that problem.

As discussed, we do believe that many of the irregularities we are seeing here can be attributed to the fact that this dataset is synthetic. However, loan repayment may be contingent on more complex relationships than what we investigated here. Perhaps people are only trustworthy if they are doing well across 4 or more of the 5 Cs of credit, and they cannot be carried by their income and credit score alone. While our linear models were a good first attempt, it would be interesting to introduce a more complex model, like a neural network, that might pick up on these hidden relationships.

VIII. Conclusion

In this project, we set out to explore the predictive capabilities of multiple linear classification techniques on bank loan data and discover which factors play the most important role in determining whether or not somebody is going to repay their loan. We used lasso regression for variable selection, performed undersampling for rebalancing data, trained logistic regression and linear discriminant classification models, and finally interpreted variable importance using SHAP values.

We found that linear approaches such as logistic regression and linear discriminant analysis could be successful, but it is important to do proper data cleaning, data transformations, and, in some cases, class balancing to prevent overfitting. Of all the approaches we took, we had the strongest results using a logistic regression model trained on a rebalanced training dataset.

Our results suggest that the most important factors to a person's reliability on paying back a loan were their loan interest rate, followed by the percentage of their income the loan takes up. However, these had the opposite impact than we had expected. This could be indicative that our dataset was not a good reflection of reality, or that there are more complex relationships that we need to discover. Regardless, this project was a good first step towards uncovering the potential of modeling approaches to inform banking institutions of a person's trustworthiness, but it still has room for improvement, both from data quality and more advanced modeling approaches.

IX. Appendix

As a quick follow up on the multicollinearity problem, we did rerun our lasso model without loan percent income. As you can see by the coefficients, the same patterns held true, and none of the odd patterns were resolved. Still, we could have some multicollinearity issues with other correlated variables. In the real world there will be strong correlations between income and a lot of these other variables, such as home ownership and education.

```
> coef(lasso)
24 x 1 sparse Matrix of class "dgCMatrix"
                                        s1
(Intercept)                 17.2373663588
loan_amnt                    0.0001162115
loan_int_rate                0.2816200329
cb_person_cred_hist_length   .
credit_score                 .
log_person_income           -2.0979173354
log_person_emp_exp           .
`person_age_25-30`           .
`person_age_31-44`           .
`person_age_45-59`           .
`person_age_60+`             .
person_gender_male           .
person_education_Associate   .
person_education_Bachelor    .
person_education_Master      .
person_education_Doctorate   .
person_home_ownership_OTHER  0.0589634732
person_home_ownership_OWN   -0.9862600368
person_home_ownership_RENT   0.7287980090
loan_intent_EDUCATION       -0.5255136154
loan_intent_HOMEIMPROVEMENT  0.0857098467
loan_intent_MEDICAL          .
loan_intent_PERSONAL        -0.2682827330
loan_intent_VENTURE         -0.7081086996
```

X. References

[1] An introduction to explainable AI with Shapley values — SHAP latest documentation. (2022). Readthedocs.io. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html?utm_

[2] Branzoli, N., & Fringuellotti, F. (2020). The Effect of Bank Monitoring on Loan Repayment. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3601944

[3] Chen, Y. Z., Calabrese, R., & Belén Martín-Barragán. (2024). Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 312(1), 357–372. https://doi.org/10.1016/j.ejor.2023.06.036

[4] Cooper, A. (2024, April 21). *Explaining machine learning models: A Non-Technical Guide to Interpreting Shap Analyses*. Aidan Cooper. https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/

[5] Hendrik, Schutte, W. D., & Helgard Raubenheimer. (2024, January 3). Shapley values as an interpretability technique in credit scoring. Journal of Risk Model Validation. https://www.risk.net/journal-of-risk-model-validation/7958697/shapley-values-as-an-interpretability-technique-in-credit-scoring?utm_

[6] Hjelkrem, L. O., & Lange, P. E. de. (2023). Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data. Journal of Risk and Financial Management, 16(4), 221. https://doi.org/10.3390/jrfm16040221

[7] Lasso Regression | Definition, Formula & Key Applications. (2025, February 25). Xenoss - AI and Data Software Development Company. https://xenoss.io/ai-and-data-glossary/lasso-regression?utm_

[8] Lee, S. (2025). Comparing QDA Versus LDA: Pros, Cons, and Real-World Applications. Numberanalytics.com. https://www.numberanalytics.com/blog/comparing-qda-versus-lda-real-world-applications?utm_

[9] mayer79. (2025). Shapviz. Github. https://github.com/ModelOriented/shapviz

[10] Moody's Asset Management Research team. (2025, March 4). US firms' default risk hits 9.2%, a post-financial crisis high. Moody's - credit ratings, research, and data for global capital markets. https://www.moodys.com/web/en/us/insights/data-stories/us-corporate-default-risk-in-2025.html

[11] What are the 5 C's of credit? Capital One. (n.d.). https://www.capitalone.com/learn-grow/money-management/five-cs-of-credit/