HIGH SCHOOL EDITION
Journal of Student Research
Volume 10 Issue 1 (2021)

Predicting Loan Defaults Using Logistic Regression

Selena Zhao[1] and Jiying Zou[2]

[1]Lynbrook High School, San Jose, CA, USA
[2]Polygence Research Academy, Stanford, CA, USA

ABSTRACT

We used anonymized data from a loan company to analyze correlations between loan defaults and other characteristics of loans or borrowers of loans. We performed an exploratory data analysis of the different factors and how they correlated with loan defaults. Using observations made in the EDA, we proceeded to use logistic regression to predict the odds of loan defaults with several loan characteristics as predictor variables. Different models were evaluated and cross-validated using AIC, AUC, and predicted accuracy. Weighted accuracy was also measured because the loan dataset was a stratified sample. We concluded that the interest rate most accurately predicted the odds of a loan default and that the most useful model was both simplistic and accurate. Research was limited by the variables that were not analyzed during EDA, the limited variables the loan dataset contained, and the modeling technique used.

Introduction

Before the 20th century, the process of money lending was fairly subjective, and potential borrowers were often judged by how trustworthy their character seemed. [1] Naturally, this process was subject to bias, which was why credit scores were created. Today, lenders are able to use tools such as FICO Scores to quantify how trustworthy potential borrowers are and make lending decisions. [2]

Predicting default rates is a significant part of moneylending because lenders must predict whether giving out a loan will result in profit or loss. Most loans are successfully repaid, [3] but sometimes a borrower will default, which is both a betrayal of the moneylender's trust and a risk to the moneylender's business. Thus, it is important that the lender can gauge the likelihood of a borrower defaulting before lending. [4]

Given the high number of factors that might affect borrower default rate, it may be infeasible to come up with good estimates heuristically or by hand. The goal of this project is to explore whether we can employ statistical and machine learning models to better predict the risk of borrower default. By analyzing variables that describe loans and the financial situations of their borrowers, we may determine key relationships between default rates and a few other variables. Along the way, we will investigate key relationships between loan default risks, loan characteristics, and buyer behaviors.

Data Description

For this project, we use anonymized data from a lending company. The data contains historical information on details of the loan itself and characteristics of the lender. In practice, a small percentage of loans are defaulted on [5], but we upsample this group to one default in every three loans to better extract signals on what might lead to loan default. Some feature names are also anonymized to protect sensitive information. Of the variables in the original data file, we will target the following variables as points of interest:

Variable	Description
Default	A binary variable representing whether the buyer defaulted on the loan. Default rates will be the focus of this project so that we can analyze how they are related to other variables. The data set contains 1,000 loans that had been defaulted and 2,000 that had not.
Reason	A categorical variable representing the reason the loan was taken out. Reasons for taking out a loan have been coded as the following: for the purchase of a boat, for a business, for credit cards, for an event, for a holiday, for the purchase of a home, for medical bills, for home re-location, for home renovation, for the installation of solar panels, for transport, and for other reasons.
Amount	A continuous variable representing the amount of money that was loaned out.
Interest	A continuous variable representing the amount of interest charged on the loan.
Term	A categorical variable representing the length of time the loan lasts. In this data set, loan terms are either 3 or 5 years.
Annual Income	A continuous variable representing the amount of money that the borrower earned last year.
Employment	A categorical variable representing the length of time the borrower has been employed, ranging from < 1 year to 1 year to 10+ years.
Credit Balance	A continuous variable representing the amount of money that the borrower spent on credit last year. Used in tandem with income, this could give us an estimate of the borrower's financial standing.
Credit Ratio	A continuous variable representing the ratio of the credit the borrower has used to the credit line. Because values are expressed as percentages, the ratio is multiplied by 100. Although credit used does not typically surpass the credit line, a few borrowers have credit ratios greater than 100. Such data points are particularly interesting to analyze with regards to loan defaults and other credit variables.
v5 and **v6**	Anonymized continuous variables. Although we may not know what they represent, we can observe their correlations with the default rate.

Exploratory Data Analysis

Independent Variables

We will first examine the distributions of and between some characteristics of the loan or the borrower of the loan. This will help us determine which predictor variables may have interesting patterns and where we should be concerned about multicollinearity, which is when the model breaks down because multiple variables are too correlated.

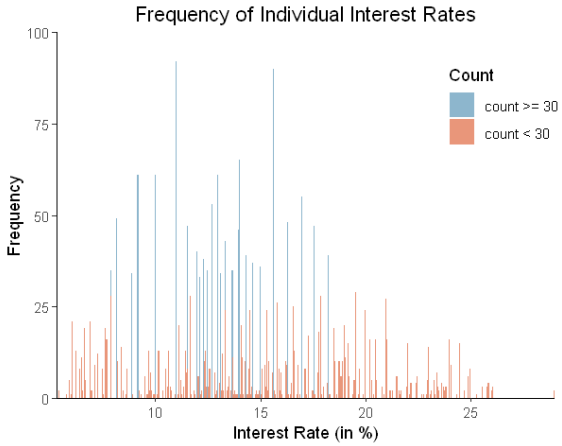

Figure 1

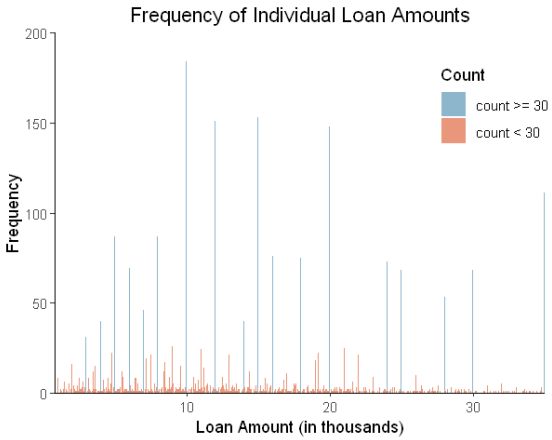

Figure 2

It is interesting to note that the values of interest rate and loan amount have varied frequency, with some values occurring over 30 times in the data set and others occurring only once (Figure 1 & 2). This is probably because some interest rates and amounts are more popular as parts of standard loan packages.

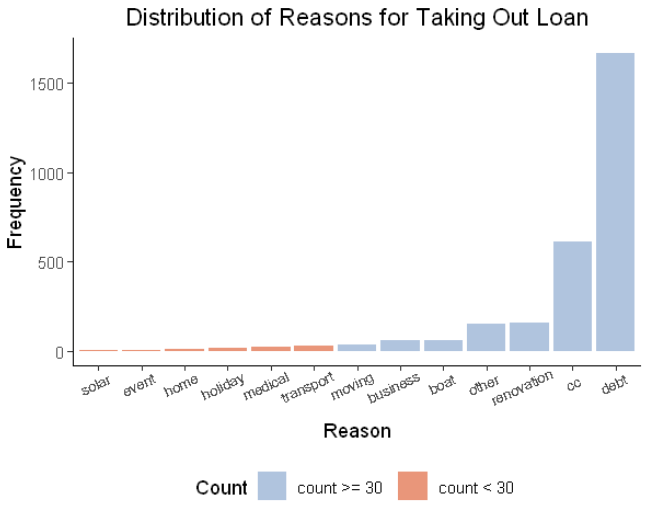

Figure 3

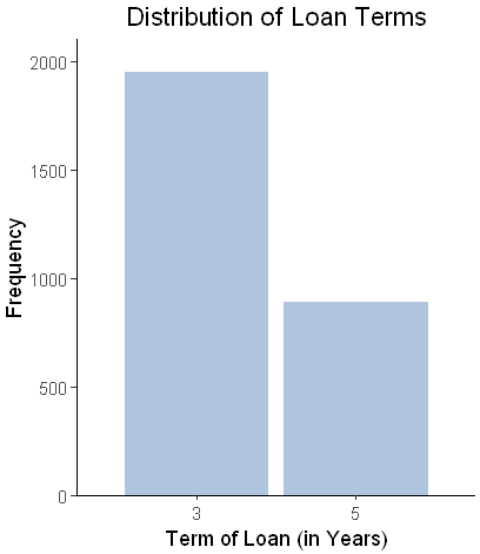

Figure 4

It is clear in Figure 3 that debt and credit cards are the most common reasons that borrowers take out loans. This is probably because people take out loans to pay off pre-existing debts or to pay off credit card bills. It is important to note that there are 6 categories with sample sizes of less than 30, so if we intend to use reasons in our model, we should be cautious about them due to high variability.

Figure 4 shows that there are far more short-term loans than long-term loans. Both loan terms have enough entries that sample size is not a concern.

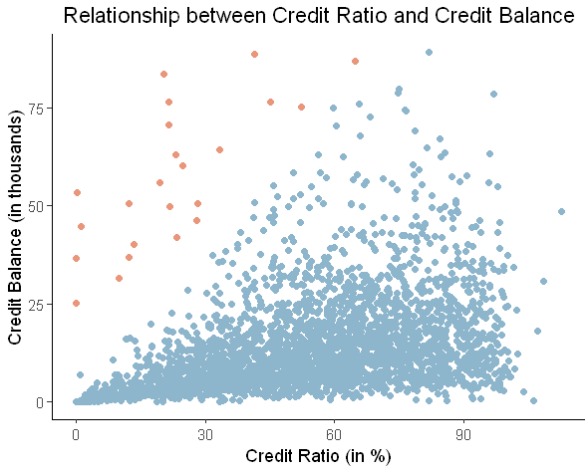

Figure 5

Figure 6

The relationship between credit ratio and credit balance (Figure 5) is positive and linear but not very strong. This makes sense intuitively because people who spend more on credit are also likely to be closer to maxing out their credit limits, thus having a higher credit ratio However, this relationship is not extremely strong, so we will be able to include both variables in the model without worrying about multicollinearity. In Figure 5, the points in red are visual outliers, where the credit balance is over 9,000 dollars greater than 1,200 times the credit ratio.

From Figure 6, it appears that credit balance is not significantly affected by how long the borrower was employed. It is interesting to note that credit balance is slightly higher for borrowers who have been employed for at least 10 years, which makes sense, as people with more consistent incomes have more spending freedom.

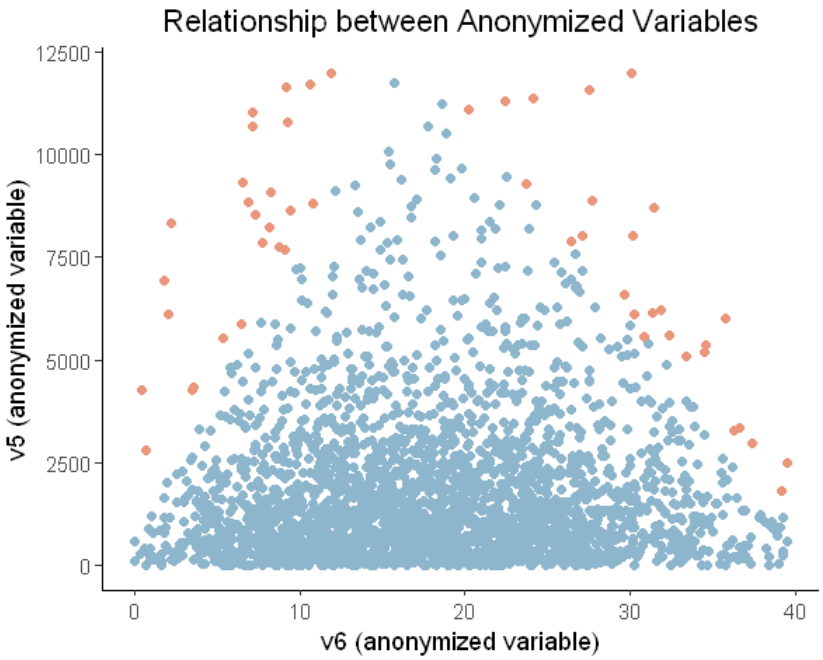

Figure 7

While the distribution of v6 seems normally distributed (Figure 7), the distribution of v5 is strongly skewed to the right. There does not appear to be a relationship between the two variables, so we can use both variables in a model without worrying about multicollinearity. The points in red are visual outliers.

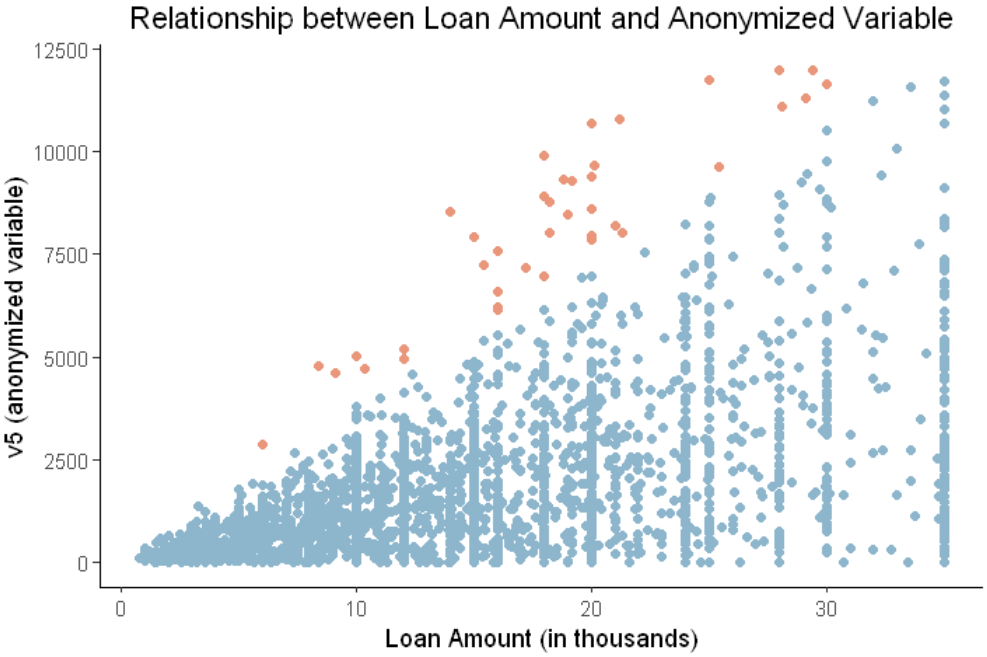

Figure 8

There is a positive linear relationship between loan amount and v5 (Figure 8), but it is relatively weak. This may be because v5 is a variable that depends on or is related to the loan amount. The points in red are visual outliers, where v5 is over 500 units greater than 0.35 times the loan amount.

Relationship to Defaults

The following bar graphs explore the correlations between some loan/borrower characteristics and whether the loan was defaulted on. The following characteristics seem to have the most influence on default rates.

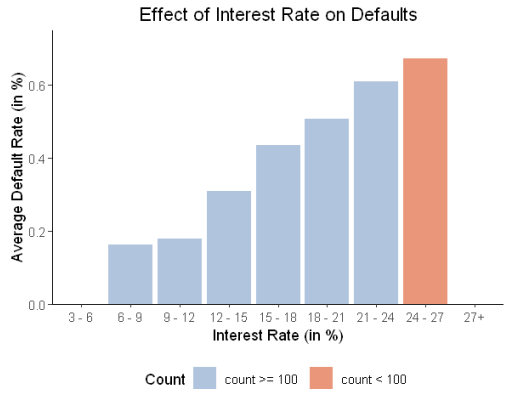

Figure 9

Looking at Figure 9, we observe that as interest rate increases, so does the average default rate. This makes sense, because higher interest rate means the loan is harder to pay back. It is also worth noting that the 30 loans with interest rates from 3% to 6% do not appear on the graph because none of them were defaulted.

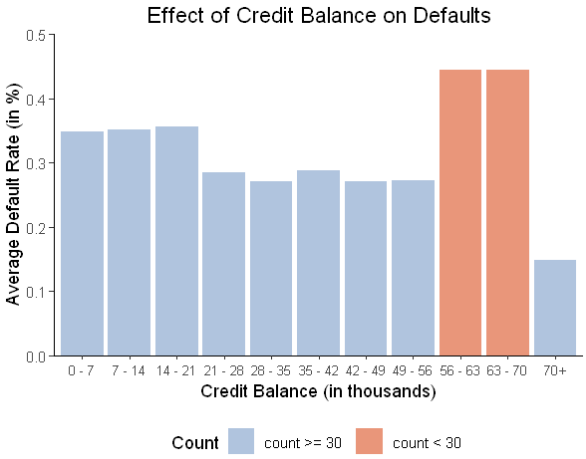

Figure 10

Average default rates generally decrease as credit balance increases (Figure 10). This may be because credit balance is correlated with socioeconomic status, so those who are able to spend more are also more capable of paying off loans. A brief look at loans for borrowers with credit balances above $70,000 shows that default rates continue to decrease as credit balance increases, although the sample sizes are small, so we need to be careful about our evaluations.

Between $55,000 and $70,000, average default rates get very high. This may be because people overspend and cannot pay back their loans. However, the sample sizes in that range are also very small, so further research would be required to make a conclusion.

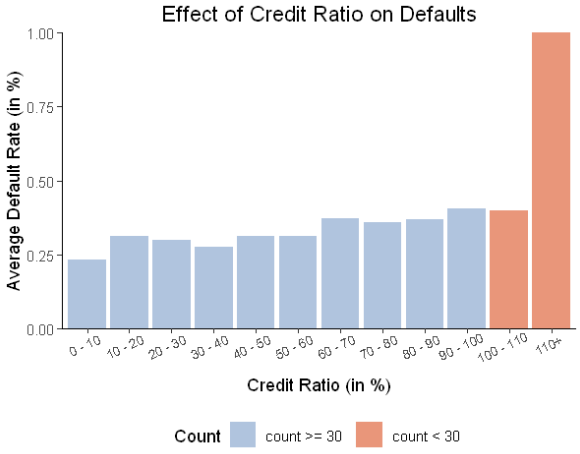

Figure 11

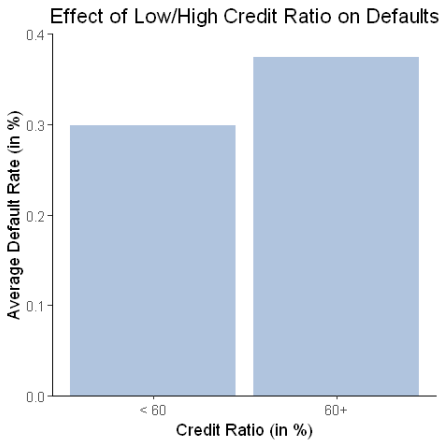

Figure 12

Overall, default rates appear to slowly increase as credit ratio increases (Figure 11). When we compare low credit ratio loans with high credit ratio loans in Figure 12, it is clear that borrowers with low credit ratio tend to default less. This may be because people who are cautious about spending are more responsible about loans. One thing to note is that borrowers within our data set with credit ratios above 110 always default. A borrower with a credit ratio above 100 has overcharged his/her credit card, so it makes sense for the borrower to be equally irresponsible with

loans or less able to pay back loans due to other outstanding debts. However, there are also few samples in this category, so we must be careful not to overfit the model.

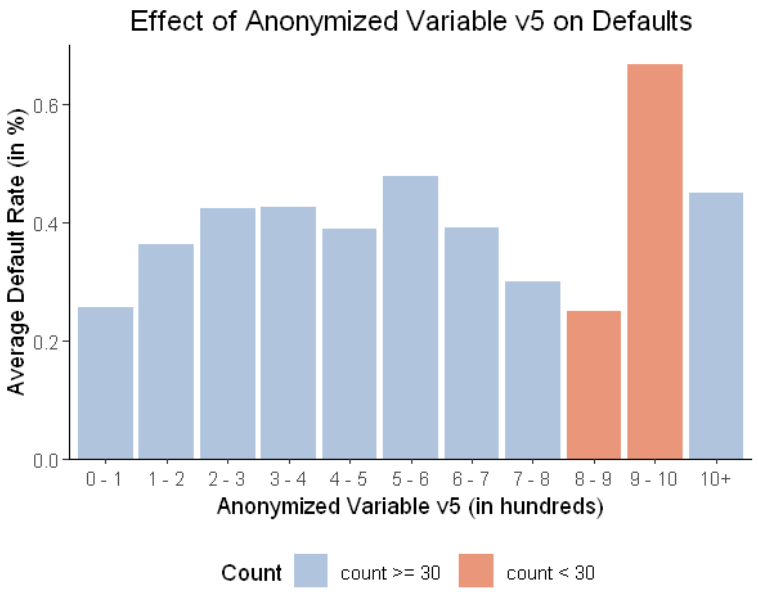

Figure 13

Default rates appear to be higher in the middle range of the anonymized variable v5 (Figure 13). Additionally, the default rate for loans with v5 between 900 and 1,000 seems to be out-of-place. This may have something to do with what the variable represents, but it could also be because the sample size is small.

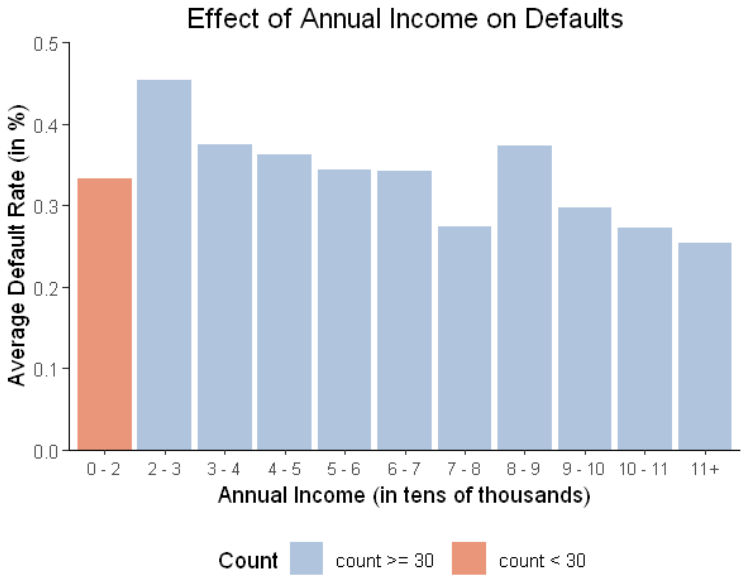

Figure 14

Generally, there is a slight downward trend in average default rates as annual income increases (Figure 14). However, defaults seem to spike for borrowers with annual incomes of around $80,000. It is unclear why this is.

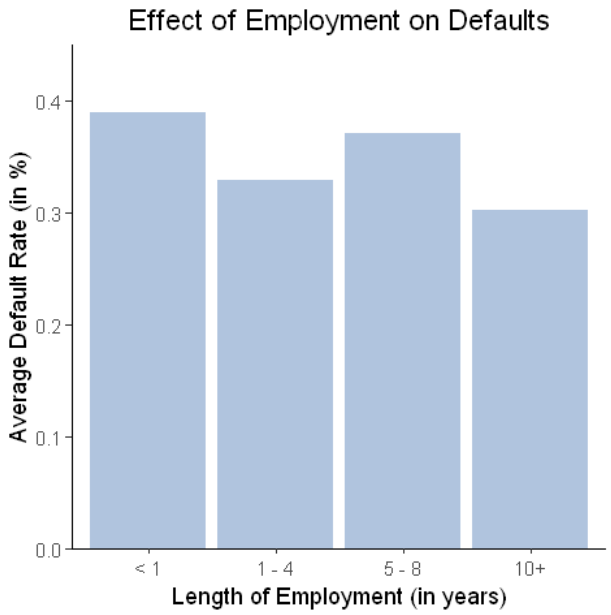

Figure 15

For employment (Figure 15), it appears that default rates are highest for the recently employed and surprisingly also those who have already been employed for a while. Perhaps the initial difficulty with paying back loans is because people struggle to pay when they are unemployed. The higher default rates in later years may be because people take out loans when they start a new job—perhaps upon graduation or when moving to a new city—and the loans are not due until 3 or 5 years later.

Methods

With a basic understanding of the correlation between independent variables such as interest and the dependent variable, loan default, we now look to predict the probability of default given several independent variables. Since default is a binary variable—loans are either defaulted or not defaulted—we will use logistic regression, a modeling technique used to predict probability for dependent variables that exist in pass/fail (binary) form. The formula for logistic regression is:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

where p is the probability that the target variable is 1 (loan defaulted), and the variables on the right side are predictor variables. [6] Continuous predictor variables contribute one independent variable to the equation, while categorical variables may be slightly more complicated. For example, given a variable with four categories, one category becomes the base, while the other three contribute three binary, mutually exclusive independent variables. By consequence, we would interpret the resulting changes in log odds in relation to the base category. [7]

To evaluate the accuracy of these logistic regression models, we will analyze the following performance measures: AUC, AIC, predicted accuracy, and weighted accuracy. AUC measures the area under the ROC Curve, a graph with the False Positive Rate (FPR) as its x-axis and True Positive Rate (TPR) as its y-axis. Given that FN and TN refer to false negatives and true negatives respectively, the formulae for TPR and FPR are written as follows:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Predicting true positives more accurately in the model will thus increase the area under the ROC and maximize the performance measure. [8]

The Akaike information criterion (AIC) approximates the difference between the predicted model and a true model, so a lower AIC suggests better accuracy. We use this measure to compare the quality of each model against each other. The basic formula for AIC is

$$\text{AIC} = -2(\text{log-likelihood}) + 2k,$$

where log-likelihood represents the fit of the model and k represents the number of parameters in the model. [9]

We will also compare predicted accuracy by calculating the proportion of loans that were accurately predicted to have been defaulted/not defaulted. However, the data set did not accurately reflect the actual distribution of defaulted loans, since the proportion of defaulted loans in the data set was approximately 33% while the proportion of defaulted loans tends to be far lower in practice. [10] Weighted accuracy accommodates for this imbalance by putting more value in defaulted loans that are predicted accurately. The formula for calculating weighted accuracy is as follows:

$$\text{weighted accuracy} = \sum_{k=1}^{|G|} w_i \sum_{x:g(x)=k} I\big(g(x) = \hat{g}(x)\big). \text{ [11]}$$

We will also cross-validate our models to ensure that the model can adapt to different loan data sets. Using a train-test split at an 80:20 ratio will give the model enough data to train with while still leaving some for it to test with.

We will also compare the models built with a null, or “coin toss,” model. This model randomly predicts defaults for loans based on the proportion of defaulted loans in the data set. Comparing the null model with other models will help us gauge the impact of predictor variables.

After evaluating different models that used different predictor variables, I noticed that of all the independent variables, interest predicted default rates most accurately. Thus, interest rate was used to predict default rates for all the models included in the results. Other characteristics of loans or borrowers of loans that proved to be useful for predicting default were annual income and loan amount.

Results

The first two models in the table (Figure 16), Models 1 and 2, were simple models to start off with. The next four, Models 3 through 6, were more complex models that performed slightly better according to the evaluation metrics. The last model is a random "coin toss" model that predicted around 1 defaulted loan for every 2 loans not defaulted.

Model	Formula	Weighted Accuracy	Predictive Accuracy	AUC	AIC
1	interest, amount, interest * amount	0.335	0.685	0.688	2672.031
2	interest, amount, income	0.340	0.685	0.691	2668.237
3	interest, amount, term, employment	0.362	0.684	0.697	2665.888
4	interest, amount, income, term, interest * amount	0.365	0.694	0.695	2648.982
5	interest, amount, income, term, employments_1, interest * amount	0.362	0.695	0.699	2647.627
6	interest, income, reasons, employments, high_bal, high_ratio, v5	0.355	0.692	0.695	2657.505
random	null model	0.3895	0.508	0.561	inf

Figure 16: Evaluation Metrics for Different Models
An asterisk (*) signifies an interactive effect.

Variables such as the borrower's length of employment, reason for borrowing, credit ratio, and credit balance were categorized differently in some of the models. In Model 3, for example, the employment predictor variable remained unchanged, so that there were 11 different categories ranging from less than a year of employment to at least 10. In Model 5, these categories were grouped into 3 categories—less than 3 years, 3 to 9 years, and at least 10 years—while in Model 6, they were grouped into 4 categories—less than a year, 1 to 4 years, 5 to 8 years, and at least 9 years.

Also, in Model 6, the "reasons" independent variable narrowed the many different reasons for taking out a loan down to business, renovation, cc, debt, and all others. The "high_bal" variable was binary, true for any borrower with a credit balance above $15,000, and the "high_ratio" variable was binary and true for any borrower with a credit ratio above 60%.

While the null model performed better in terms of weighted accuracy, Models 1 through 6 have higher AUCs and scored around 20% higher in terms of actual (predictive) accuracy, so ultimately, our efforts in modeling paid off. Figure 17 is a graph of the AUC curves for Model 1, Model 4, Model 6, and the null model. The closer the curve is to the top left area, the greater its AUC, and thus, the better it performs. Even the simplest models such as Model 1 seem to perform drastically better than the null model.

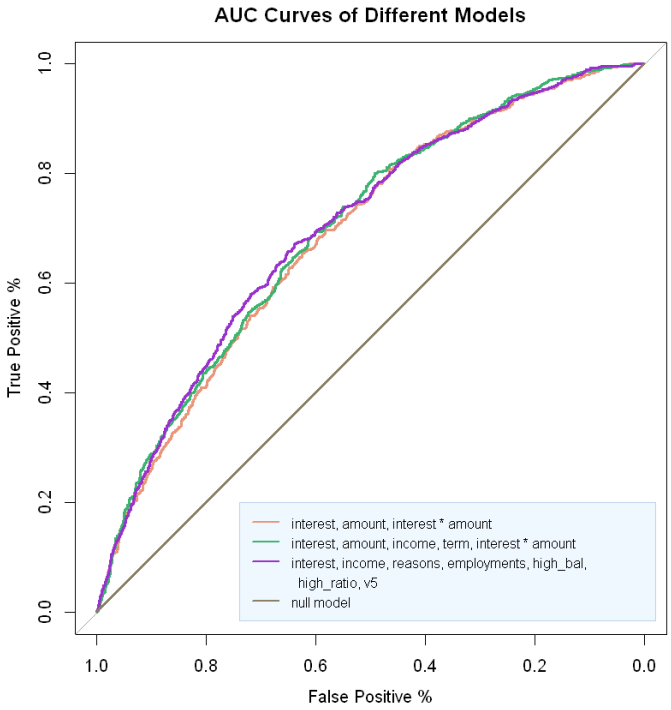

Figure 17

Although Model 5 tended to be the most accurate of all the models listed, it may be argued that the best model is Model 4, which uses amount, income, interest, term, and an interaction between interest and amount to predict default rates. It is nearly as accurate as Model 5 and performed better in terms of weighted accuracy, but it is simpler, using fewer independent variables. Additionally, all the variables had significant effects ($p < 0.05$) on the default rate, so this model is both explanatory and predictive to a good degree.

Figure 18 shows the coefficients of predictor variables in Model 4. As an equation, Model 4 is:

$$log\left(\frac{p}{1-p}\right) = -1.398 - (4.156 * 10^{-5}) \times \text{amount} + 0.133 \times \text{interest} - 0.270 \times \text{term} - (3.994 * 10^{-6}) \times \text{income} + (2.914 * 10^{-6}) \times \text{amount} * \text{interest}.$$

```
Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.9175  -0.8867  -0.6372   1.1358   2.2330

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.398e+00  3.510e-01  -3.982 6.84e-05 ***
amount         -4.156e-05  1.937e-05  -2.146  0.03191 *
interest        1.327e-01  2.108e-02   6.293 3.12e-10 ***
term           -2.696e-01  5.533e-02  -4.873 1.10e-06 ***
income         -3.994e-06  1.218e-06  -3.278  0.00105 **
amount:interest 2.914e-06  1.196e-06   2.436  0.01486 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 3615.9  on 2839  degrees of freedom
Residual deviance: 3297.6  on 2834  degrees of freedom
AIC: 3309.6

Number of Fisher Scoring iterations: 4
```

Figure 18: Summary of Model 4

At the intercept, when amount, interest, and income are (hypothetically) \$0, and the term is 3 years long, the log odds is −1.398. This means the odds of defaulting are $e^{-1.398} \approx 0.247$, where odds are defined as the ratio of the probability of loan default to the probability of successful loan payment. When the term of the loan is 5 years instead of 3, the log odds decrease by 0.270, so the odds of defaulting decrease by

$$1 - e^{-0.2696} \approx 23.6\%.$$

It seems that a borrower is more likely to default on a shorter loan than on a longer one. When income is \$10,000 higher, the odds of defaulting decrease by

$$1 - e^{-0.0399} \approx 3.9\%.$$

When interest is fixed at a constant percentage, and the amount of the loan increases by \$1,000, the associated log odds are expected to decrease by 0.0416, so the odds of defaulting decrease by

$$1 - e^{-0.0416} \approx 4.07\%.$$

However, if interest is raised as well as amount, then the log odds of default increases by $2.914 \cdot 10^{-6}$ for every additional unit increase in amount. For example, a \$1,000 increase in amount would decrease log odds by 0.0416, but a \$1,000 increase in amount alongside an increase in interest would decrease log odds by $0.0416 - 0.0029 = 0.0387$, so the odds of defaulting would decrease by

$$1 - e^{-0.0387} \approx 3.80\%.$$

In other words, as loan interest starts to rise, the lowering effect of higher loan amounts on default rates starts to diminish.

Conversely, if loan amount increases, the log odds of default increases by $2.914 \cdot 10^{-6}$ for every additional unit increase in interest. When the loan amount is fixed and the interest rate increases by 1 percent, the log odds are expected to increase by 0.133, so the odds of defaulting increase by

$$1 - e^{0.133} \approx 14.2\%.$$

However, if loan amount is rising, interest rate increasing by 1 percent would cause the log odds to increase by

$$0.133 + 2.914 \cdot 10^{-6} \approx 13.3\%.$$

Discussion

The main goal of this research project was to examine the relationship between loan default and predictor variables such as interest and amount at a basic level. Using logistic regression, we discovered which properties were most directly related to the chance of loan default and which properties could be used in conjunction to predict defaults.
In Model 4, it was surprising that larger loan amounts could cause loan defaults to decrease. This may be because borrowers who take out larger loans are more cautious or plan it out more carefully. However, if loan interest is increased as well, then the lowering effect of higher loan amounts on probability of default diminishes. Thus, this model implies that loans with large amounts and low interest rates minimize risk.

Generally, Model 4, which used the predictor variables of amount, interest, term, income, and an interaction between amount and interest, performed the best because it balanced simplicity and performance. It was accurate without being overly complex, and every predictor variable contributed a significant effect on the probability of default. In terms of future research, combining predictor variables from Model 4 with other variables left unexplored in this paper could yield a better model.

Using other modeling techniques would also allow for different interpretations of the same variables. This study was limited to logistic regression models; as a result, variables that did not have a one-directional trend did not predict as well as linear predictors. For example, interest was successful in logistic regression models because it had a linear relationship with default rates. However, variables such as credit balance or loan amount have more complicated trends, so exploring other modeling techniques could yield more accurate models in future research.

Conclusion

Through Exploratory Data Analysis, we discovered correlations in and between predictor variables that would guide us in building our model. We were able to conclude that the probability of a loan default may be predicted by loan interest rates, loan amount, and borrower income, among other factors. We also proved the credibility of our models with evaluation metrics that measured accuracy and error. The predictor variable that best suited logistic regression was interest because of its linear correlation with default. To further improve on this research, different predictor variables or types of models may be examined.

Acknowledgments

This project was conducted under the mentorship of Jiying Zou. I would like to express my sincere gratitude towards her and the research program, Polygence, for all their guidance and counsel throughout the process.

References

[1] Kaufman, R. (2018, August 21). *The History of the FICO® Score*. Retrieved from https://www.myfico.com/credit-education/blog/history-of-the-fico-score.

[2] Konsko, K. (2014, August 12). *The Origin of the Credit Score*. Retrieved from https://www.nerdwallet.com/blog/finance/origin-credit-score-history/.

[3] Kagan, J. (202, December 1). *Default Rate*. Retrieved from https://www.investopedia.com/terms/d/defaultrate.asp.

[4] Kagan, J. (2020, October 19). *Default Risk.* Retrieved from https://www.investopedia.com/terms/d/defaultrisk.asp.

[5] Kagan, J. (202, December 1). *Default Rate*.

[6] Statistics Solutions. (n.d.). *What Is Logistic Regression?* Retrieved from https://www.statisticssolutions.com/what-is-logistic-regression/.

[7] UCLA: Statistical Consulting Group. (n.d.). *FAQ: How do I Interpret Odds Ratios in Logistic Regression?* Retrieved from https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/.

[8] Google Developers. (n.d.). *Classification: ROC Curve and AUC.* Retrieved from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[9] Glen, S. (2015, September 7). *Akaike's Information Criterion: Definition, Formulas*. Retrieved from https://www.statisticshowto.com/akaikes-information-criterion/.

[10] Kagan, J. (202, December 1). *Default Rate*.

[11] Döring, M. (2018, December 4). *Performance Measures for Multi-Class Problems*. Retrieved from https://www.datascienceblog.net/post/machine-learning/performance-measures-multi-class-problems/.