# Analyzing Credit Risk Using Regression Models

Jay Suresh Singhvi, Benjamin Correia

2024-05-14

## 1.0 Introduction

**1.1 Project Description** In this project we used machine learning to analyze a loan default dataset and evaluate the greatest predictors of lenders defaulting on their loans. Using these predictors, we trained a logistic regression model to predict a loan applicant's likelihood to default. We also used a random forest model to analyze the most significant predictors of loan defaults.

**1.2 Background** When creditors (money lenders like banks) are considering who to lend money to, they must do so cautiously. Haphazardly offering loans to whoever requests them is not great business practice, as many debtors (money borrowers, in this case people) are not fully equipped to pay back money that is lent to them with interest.

In order to maximize the probability of debtors paying them back, creditors must be able to predict the probability of loan applicants to pay them back using financial data.

The goal of this analysis was to use a data set of debtors' financial data and loan statuses to find out which predictors creditors should take into account when deciding on who to lend money to.

## 2.0 Data Description

The dataset we used is called the Credit Risk Dataset, obtained from Kaggle. It contained 32,581 rows of 12 variables. It's important to note that the dataset was a simulation of credit bureau data made for educational purposes. A quick analysis of these variables will follow:

person_age - numeric: Borrower's Age in years

person_income - numeric: Borrower's income in dollars

person_home_ownership - categorical: Borrower's home ownership status

person_emp_length - numeric: Borrower's length of most recent employment in years

loan_intent - categorical: Borrower's reason for taking out the loan.

loan_grade - categorical: loan grade

loan_amnt - numeric: loan amount in dollars

loan_int_rate - numeric: loan interest rate

loan_status - categorical: levels 1 & 0. 0 - Borrower Did not default 1 - Borrower defaulted

loan_percent_income - numeric - calculated as loan_amnt/person_income

cb_person_default_on_file - categorical: Y if person has a loan default on file, N otherwise

cb_person_cred_hist_length - Length of borrower's credit history in years

We started with 32,581 rows of data before cleaning. First, we removed any rows that contained an empty value. This left us with 28,638 observations.

Just by a glance of the data person_age, person_emp_length, and person_income had some hefty outliers. The box plots below display how much variance existed some of our data.

**Fig 2.1: Box Plot for Income**

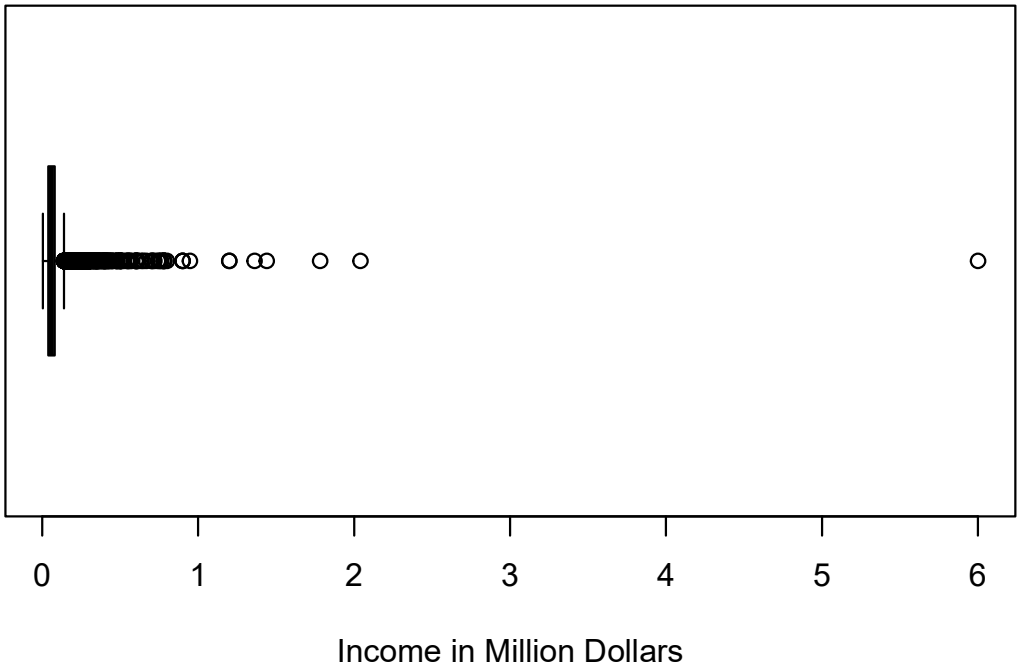
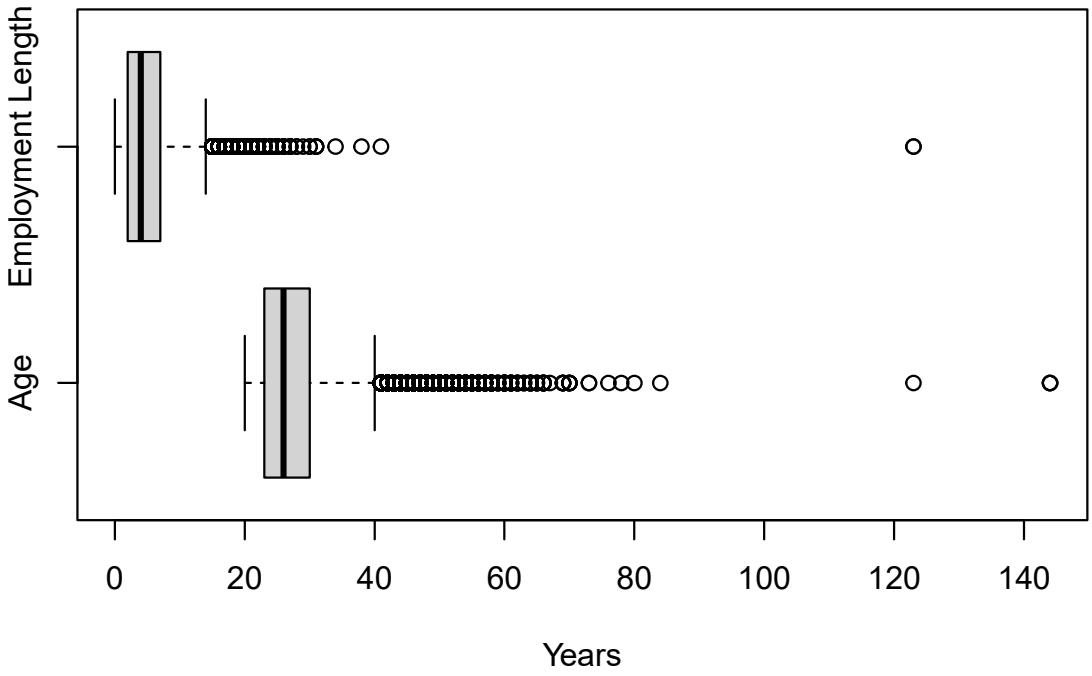


**Fig 2.2: Box Plot for Employment Length and Age**

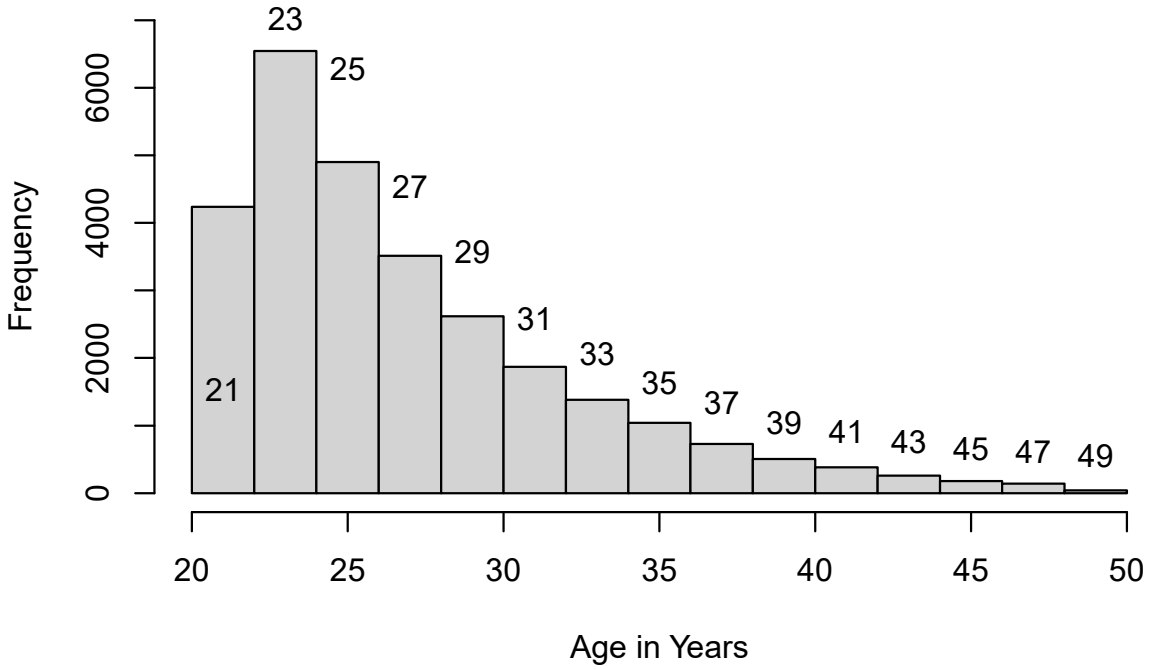


A typical standard for removing outliers is removing any values over 1.5 interquartile ranges above Q3 or below Q1.

We decided not to strictly follow this standard while cleaning our data, as we had quite a large set of data and following this standard would remove over half of it. Most of our observations, such as age and income, have quite a large variance, as one would expect when encountering people in the real world. Instead, we elected to inspect each column that contained outliers and remove data intuitively, judging by where certain values become rare.

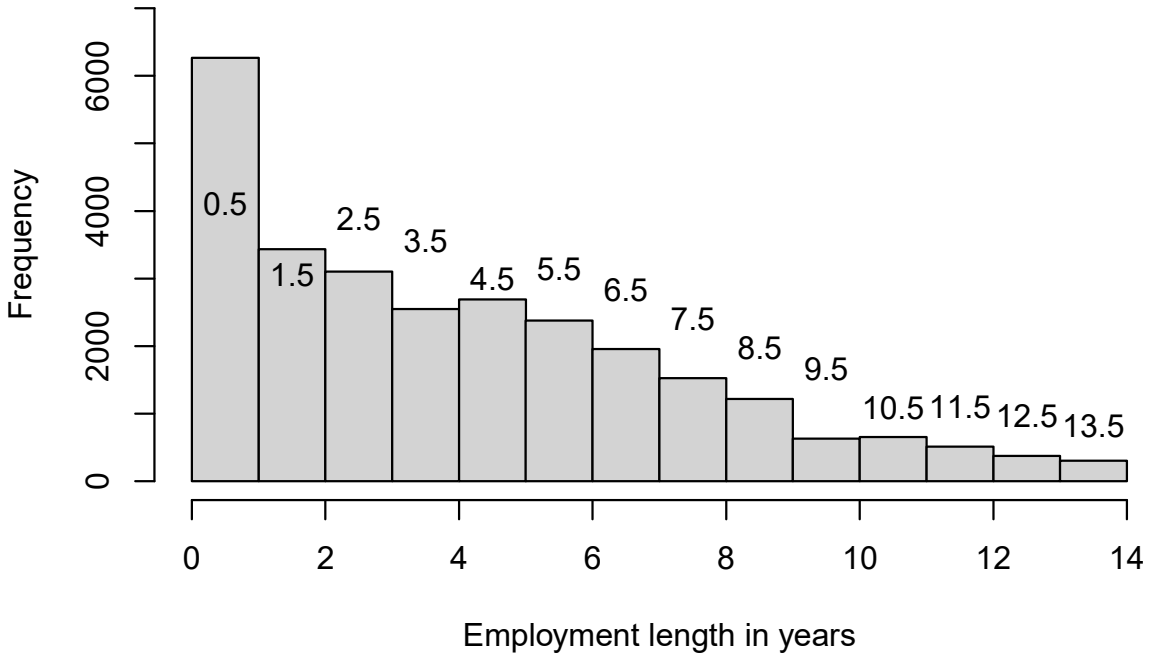
As for age, we wanted to remove the incredibly large outliers because we simply didn't have enough data on elderly applicants to make sufficient conclusions about them. As seen in Figure 2.3, some observations were outright unrealistic. Many of these outliers, however, were representative of a true variance in the ages of the population. In fact, over 1000 observations exist within the 40-50 y/o range. We used a cutoff of fifty years old as a maximum age; everything higher was removed from our data set. 28340 observations remain.

**Fig 2.3: Histogram of Loan Recipient's Ages**



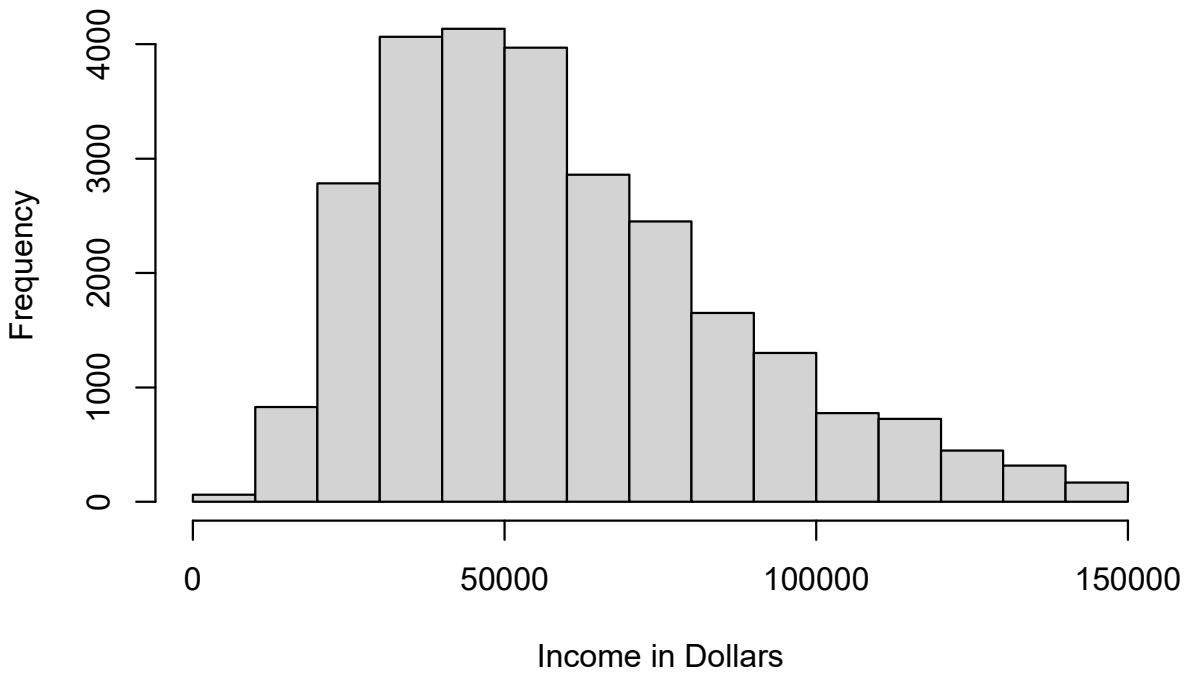
Employment length is another variable for which it probably would not be great practice to indiscriminately remove all of the outliers, as there can be a great degree of natural variance. By making a boxplot of the data the massive outliers are clear in Fig 2.2. We ultimately elected remove observations above 15 years, as that is when their frequency became sparce relative to those below. 27599 Observations Remain.

**Fig 2.4: Histogram of Loan Recipient's Length of Current Employme**



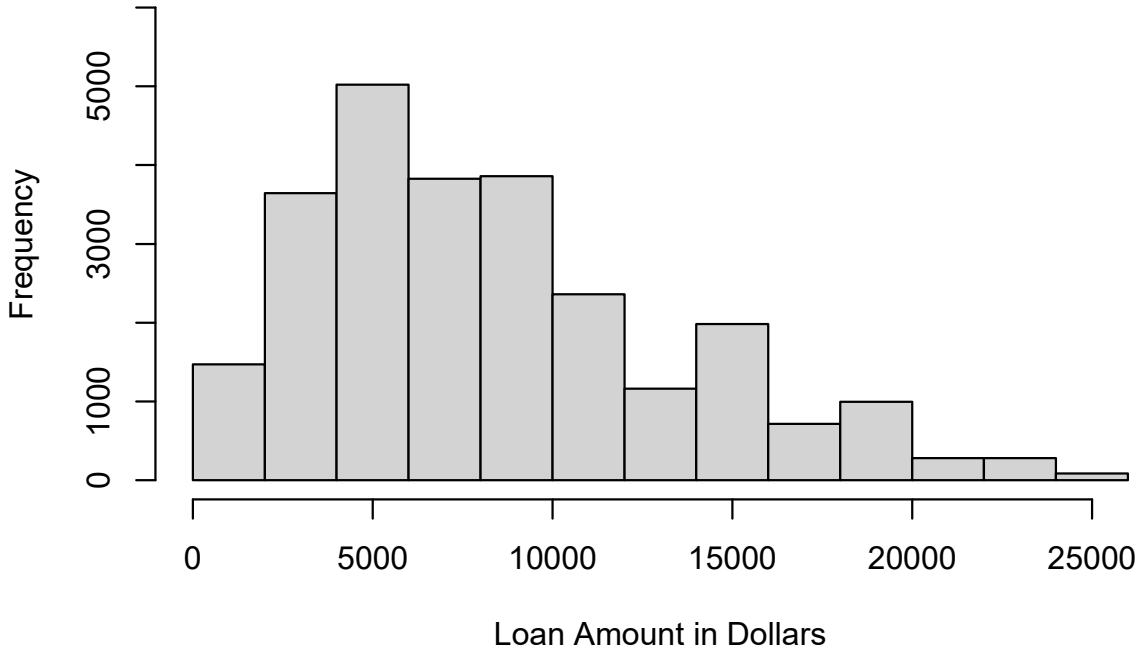
Income is, of course, the most likely of our variables to vary greatly. First we removed the greatest few outliers that we could see from Fig 2.1. Afterwards, making a labeled histogram showed us that income is relatively normally distributed, skewed to the right, and has a mean around the 50,000 mark. There is a sharp dropoff in the number of observations with income above 150,000, which should be a reasonable point upon which we stopped considering measurements to base our model upon and that is shown in Fig 2.5. 26536 remain

**Fig 2.5: Histogram of Loan Recipient's Income**



Like most of our numeric data, Loan Amount is normally distributed and skewed to the left. There is a hard dropoff in the number of loans above 25,000, so we naturally removed those from the dataset and plot a histogram shown in Fig 2.6. 25701 remain.

**Fig 2.6: Histogram of Loan Amount**




After cleaning our data we were left with 25,701 observations. 6,880 observations was been removed from our data set. Now that we cleaned our data, we were ready to begin modeling it.

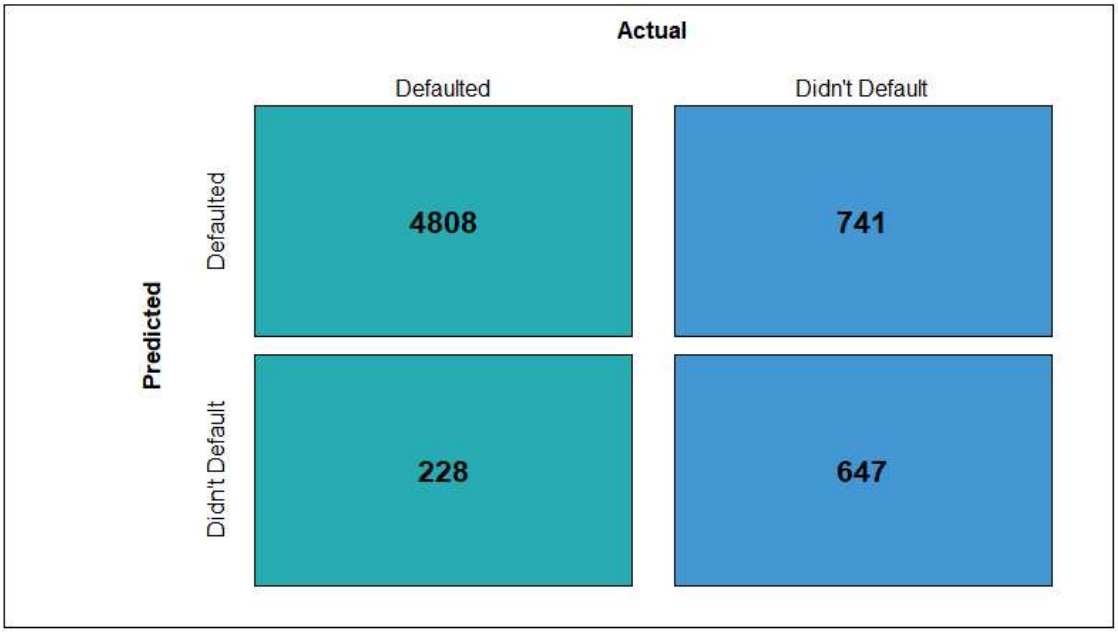
## 3.0 Analysis

First, we partitioned our data into training and testing sets. 75% of our data, or 19,277 observations, were used for testing. The remaining 6424 observations went into our training set.

After partitioning our data we trained a logistic regression model, which we will refer to as LR1, using the training set. Default status (whether or not a debtor defaulted) was our dependent variable, and everything else except for loan grade was used as a predictor. We elected not to use loan grade as a predictor because it is a grading system lenders often use to categorize loans by risk. Although it is a great predictor of risk, it is one that was not standardized and was the outcome of statistical analysis on credit risk that has already been carried out.

After our model was built, we used it to generate predictions of the default status of our testing observations. The confusion matrix below compares our predictions to their actual observed values.

**Fig. 3.1: LR1 Confusion Matrix**




**Confusion matrix statistics**

| Sensitivity | Specificity | Precision | Recall | Balanced Accuracy |
|---|---|---|---|---|
| 0.95 | 0.47 | 0.87 | 0.95 | 0.71 |

| Accuracy | Kappa |
|---|---|
| 0.85 | 0.49 |

Our model, LR1, performed with 84.92% accuracy. Unfortunately our sensitivity was 46.61%, which is lower

than if we were to predict randomly. Sensitivity is a measure of how accurately our model predicts true positives. In order to decrease risk, our main goal for our model's sensitivity was to be very high, as to prevent lending to borrowers who were expected not to default, but do.

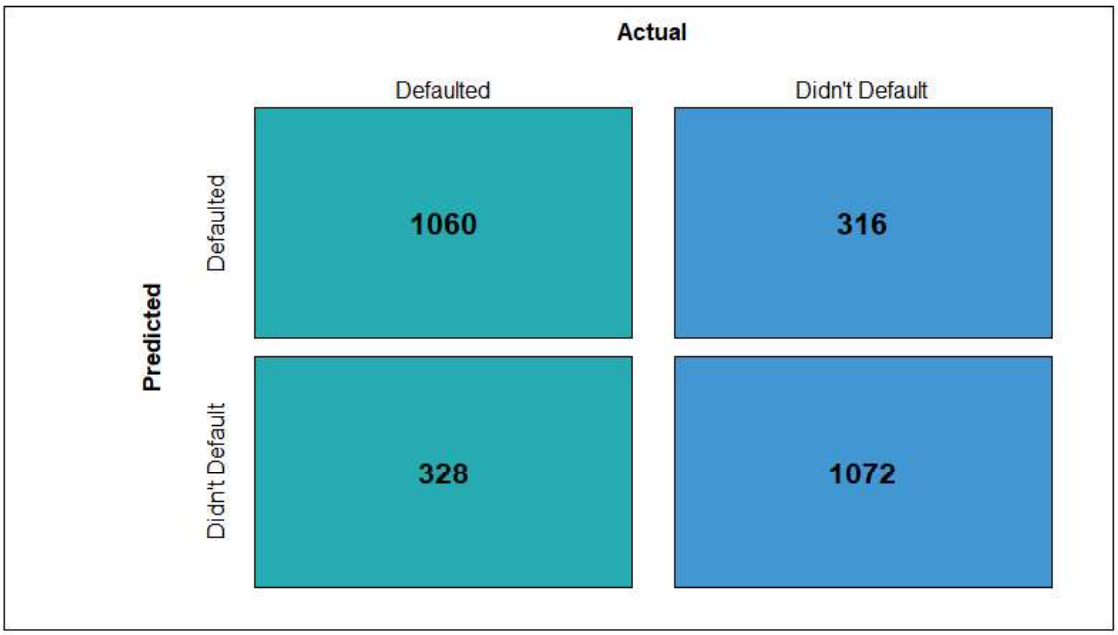
One possible reason for our model's high sensitivity was the imbalance of the data. Of our entire cleaned data set, 20,147 did not default on loans while 5,554 did. This means we have a minority class, default, which is only 21% of the data. While training a regression model, this imbalance can hinder our ability to predict the minority class. We used undersampling, a technique where we discard observations from our majority class until our dataset is balanced, in order to balance our data. Afterwards we are left with a total of 11108 observations, 5554 in each class.

After undersampling we partition our data once again into training and testing sets, where 75%, or 8331 observations go towards testing and 2777 towards training.

After undersampling we trained yet another basic logistic regression model using the same predictors, which we will label LR2.

After balancing our data LR2 made predictions with a significantly lower accuracy of 76.8%, but our sensitivity almost doubled to 77.2%. Remember, our goal was to decrease risk by increasing sensitivity. So far just by balancing our data we made a massive stride towards this goal despite it being at the cost of accuracy.

**Fig. 3.2: LR2 Confusion Matrix**



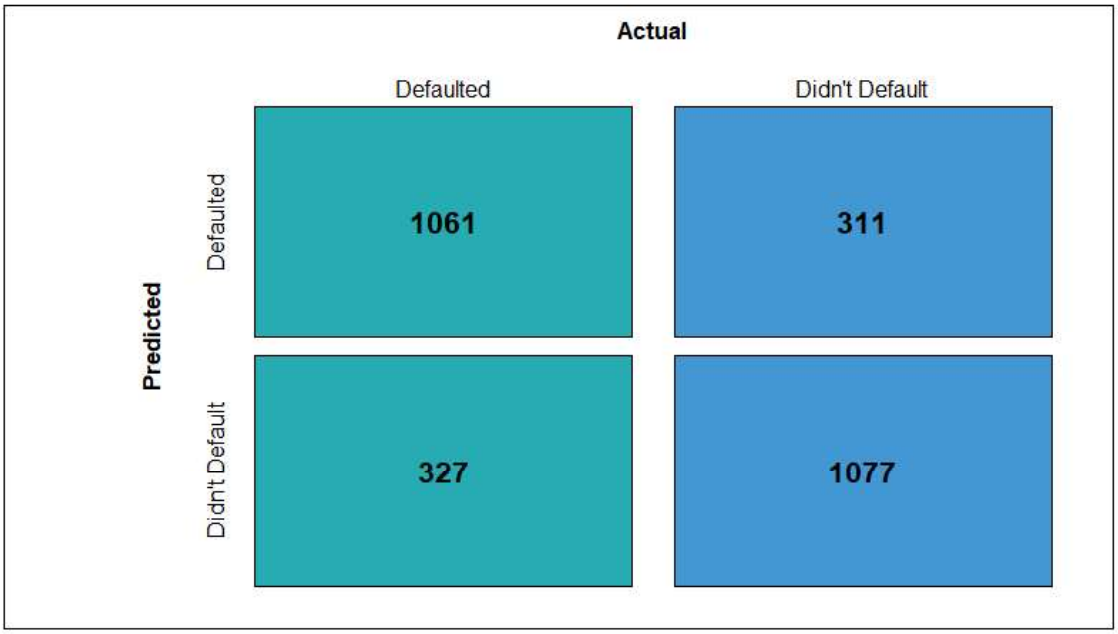

**Confusion matrix statistics**

| Sensitivity | Specificity | Precision | Recall | Balanced Accuracy |
|---|---|---|---|---|
| 0.76 | 0.77 | 0.77 | 0.76 | 0.77 |

| Accuracy | Kappa |
|---|---|
| 0.77 | 0.54 |

After using a new sampling technique we decided to improve our model by removing the predictors that LR2 considered insignificant, i.e. their p-value was above 0.05. Our new model, LR3, was only trained on home ownership, loan intent, loan amount, interest rate, and the loan to income ratio.

**Fig. 3.3: LR3 Confusion Matrix**




**Confusion matrix statistics**

| Sensitivity | Specificity | Precision | Recall | Balanced Accuracy |
|---|---|---|---|---|
| 0.76 | 0.78 | 0.77 | 0.76 | 0.77 |

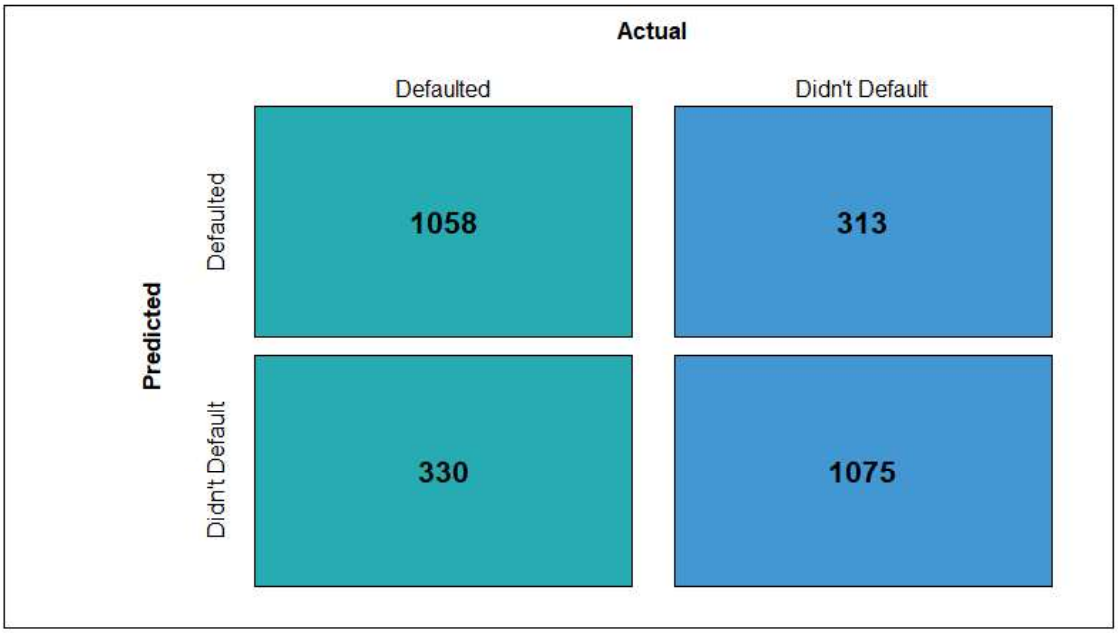
| Accuracy | Kappa |
|---|---|
| 0.77 | 0.54 |

Model LR3 had an accuracy of 77% and a sensitivity of 77.6%. These are marginal increases, but the true improvement of LR3 was that it uses many fewer predictors than LR2.

Rather than continuing to remove predictors, we decided to try another method for choosing them; stepwise regression. Our next model, LR4, used stepwise regression to iteratively decide which predictors to use.

Using stepwise regression told us that age, home ownership, loan intent, loan amount, loan interest rate, and loan to income ratio were the most significant predictors of loan defaults. Using these predictors our model LR4 performed with 76.8% accuracy and 77.5% sensitivity, a slight step down from LR3's stats.

**Fig. 3.4: LR4 Confusion Matrix**




**Confusion matrix statistics**

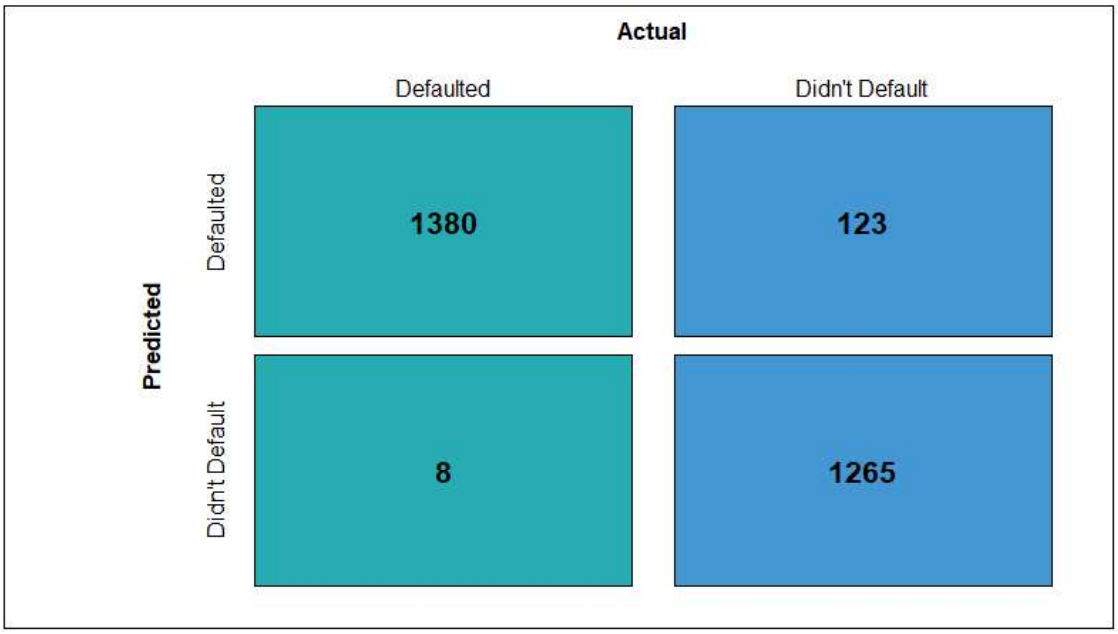
| Sensitivity | Specificity | Precision | Recall | Balanced Accuracy |
|---|---|---|---|---|
| 0.76 | 0.77 | 0.77 | 0.76 | 0.77 |

| Accuracy | Kappa |
|---|---|
| 0.77 | 0.54 |

After analyzing different logistic regression models and selecting the predictors generated by multiple logistic regression (LR2), we built a random forest model of 500 trees with 6 independent variables each to predict loan defaults. Because random forests perform well despite imbalanced data, we trained it on the entire training set rather than the undersampled set.

As seen in figure 3.5, the random forest performs incredibly. It has an accuracy of 95.2% and a sensitivity of 91.1%. It outperforms logistic regression by a massive degree, and we are able to use it to make lending decisions with a great deal of safety and accuracy.

**Fig. 3.5: Random Forest Confusion Matrix**




**Confusion matrix statistics**

| Sensitivity | Specificity | Precision | Recall | Balanced Accuracy |
|---|---|---|---|---|
| 0.99 | 0.91 | 0.92 | 0.99 | 0.95 |

| Accuracy | Kappa |
|---|---|
| 0.95 | 0.91 |

The results of our random forest model suggested that loan to income ratio, interest rate and home ownership status were the most significant predictors of loan defaults among the predictors used in the formula returned by model LR3.

## 4.0 Conclusions

The results of our testing continuously showed that home ownership, loan intent, loan amount, interest rate, loan_percent_income, and credit history length are the most significant predictors of loan defaults.

The most surprising result to us was that having a default in one's credit history is a very weak predictor, with a p-value of over 0.8.

This experiment was also a massive lesson on the importance of great standards when it comes to cleaning and sampling data. The greatest improvement we made to our logistic regression models was not from changing predictors, but from balancing our data set.

Another great insight was the use of multiple modeling methods. The random forest model we created was miles ahead of logistic regression in terms of accuracy and precision, the exact things we were looking to improve in our model.

## 5.0 References

https://www.kaggle.com/datasets/laotse/credit-risk-dataset/discussion

https://campus.datacamp.com/courses/fraud-detection-in-r/imbalanced-class-distributions?ex=5

https://medium.com/sfu-cspmp/surviving-in-a-random-forest-with-imbalanced-datasets-b98b963d52eb#:~:text=Random%