Credit Card Approval Model

UGA Data Science Competition
April 15, 2021

Will Matray, Patrick Nercessian, Hayden Dessommes, Thomas Hills

1 Introduction

In order to apply for a loan, banks require individuals to provide things like proof of income, proof of assets, and collateral. This is because it is not in the bank's best interest to loan money to people who are likely to default. However, even with ensuring that they only loan money to the most responsible people, it is still not uncommon for people to default. If there was a model that predicted whether someone would default based on data on their credit history, banks could distribute money for loans more effectively. The goal of this project is to do just that, build a model to predict whether or not a person will default on their loan based on their credit history data.

The datasets consist of historical data from a hypothetical bank XYZ in the Southeastern US. We were given a test dataset (5,000 records), a training dataset (20,000 records), and a validation dataset (3,000 records), each of which contains 20 predictor variables and one response variable. The response variable is Default_ind, an indicator variable that equals 1 if the approved account defaulted, and equals 0 if not defaulted. The predictor variables and their descriptions are shown in Table 1:

Table 1: Predictor Variables

Variable	Description
tot_credit_debt	Total debt on all credit products
avg_card_debt	Average debt on all credit cards over last 12 months
credit_age	Age (months) of first credit product
credit_good_age	Age (months) of first credit product currently in good standing
card_age	Age (months) of first credit card
non_mtg_acc_past_due_12_months_num	# of non-mortgage credit products past due in last 12 months
non_mtg_acc_past_due_6_ months_num	# of non-mortgage credit products past due in last 6 months
mortgages_past_due_6_ months_num	# of mortgages past due in last 6 months
credit_past_due_amount	Total amount of money past due on all accounts
inq_12_month_num	# of credit inquiries in last 12 months
card_inq_24_month_num	# of credit card inquiries in last 24 months
card_open_36_month_num	# of credit cards opened in last 36 months
auto_open_36_month_num	# of auto loans opened in last 36 months

uti_card	Utilization on all credit card accounts
uti_50plus_pct	% of open credit products with >50% utilization
uti_max_credit_line	Utilization of credit product with highest credit limit
uti_card_50plus_pct	% of open credit cards with >50% utilization
ind_acc_XYZ	Indicator: 1 if applicant has an account with XYZ Bank; 0 otherwise
rep_income	Self-reported annual income
States	Residence state (AL, FL, GA, LA, MS, NS, SC)

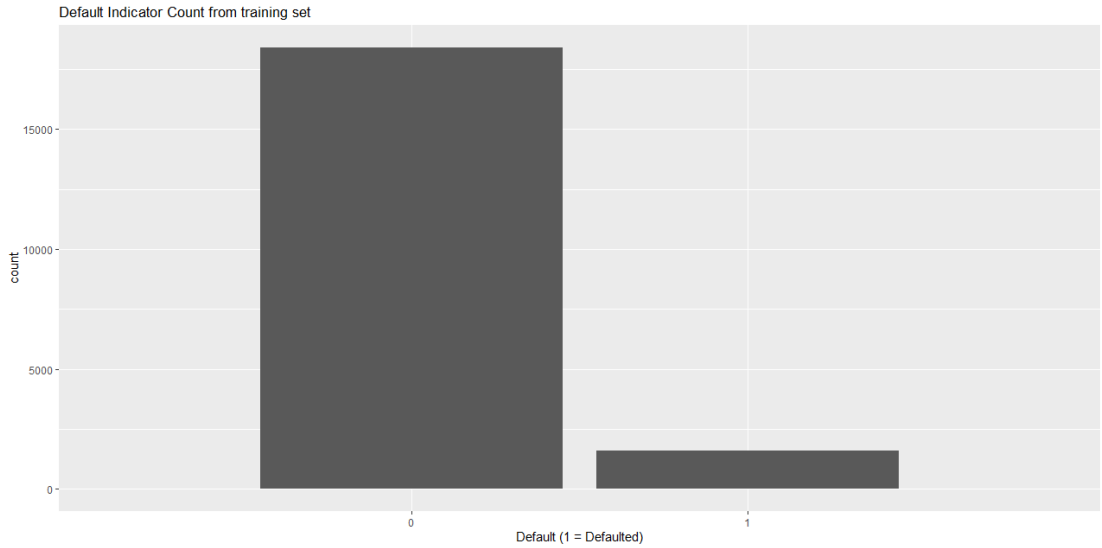
The 20 independent variables will be used to predict the response. Two types of models will be built, a logistic regression model, and a random forest model.

2 Exploratory Data Analysis

2.1 Distribution of Variables

Our dataset contains several different types of variables. The response is a binary indicator variable (only takes on values of 0 or 1). Figure 1 shows the distribution of the response variable from the training set so that we can see the distribution of 0's and 1's that appear.

Figure 1: Distribution of default indicator from training set.

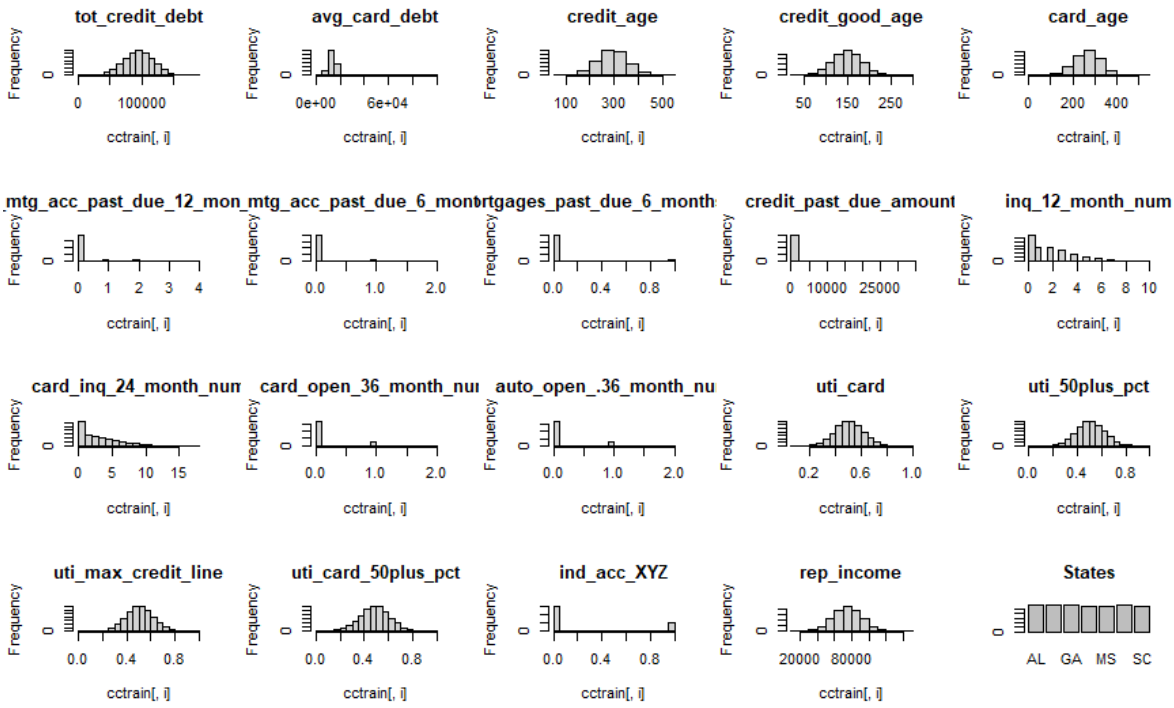


The majority of applicants did not default on their loans. Of the 20,000 records in the training set, only 1,586 of them (7.93%) defaulted. There is obvious class imbalance in the

response variable. If possible, this may need to be addressed in the creation and selection of our predictive models.

For the predictors, there are several different classes of variables in the dataset. Most of the variables are continuous. However there are a few countable numeric variables, an indicator variable, and a categorical variable. Figure 2 shows the distributions of all 20 independent variables.

Figure 2: Distributions of independent variables.



Most of the numeric variables appear to be normally distributed, and several are skewed to the right. The only variable that appears to have clear outliers, based on the histogram, is average card debt. These values will need to be investigated to see if they are erroneous and need to be thrown out. For states, the only categorical variable, the data are basically evenly distributed across all seven states on the dataset. For the account indicator variable, there appears to be some class imbalance as well. The proportion of 1's in the XYZ indicator variable is 0.2586, so there is not nearly as much class imbalance as we saw in the response indicator variable.

The only categorical predictor variable is the state indicator, which we will convert into binary indicators for each state. Since there are 7 states, this means the states variable will be converted into six indicator variables that take on a value of either 0 or 1. When fitting the predictive models going forward we will use GA as default and drop it as a predictor.

2.2 Missingness and Erroneous Data

Three of the independent variables contain missing values. In the case of average card debt it appears that there were many cases with value $99,999. This is likely erroneous and will

be treated as missing going forward. Table 2 shows the variables with missingness and their respective amounts of missingness in the training data.

Table 2: Variables with missingness in the training data.

Variable	# of Missing Records	% of Missingness
uti_card_50plus_pct	2055	10.28
rep_income	1570	7.85
avg_card_debt	1586	7.93

None of the variables listed in Table 2 contain a concerningly large proportion of missingness, so there is no need to consider removing either of them. Additionally, according to Figure 2, all of these variables appear to be normally distributed. We now have to determine the pattern of missingness so that we can choose an appropriate imputation method.

Figure 3: Card Utilization 50 Plus Missingness Correlations

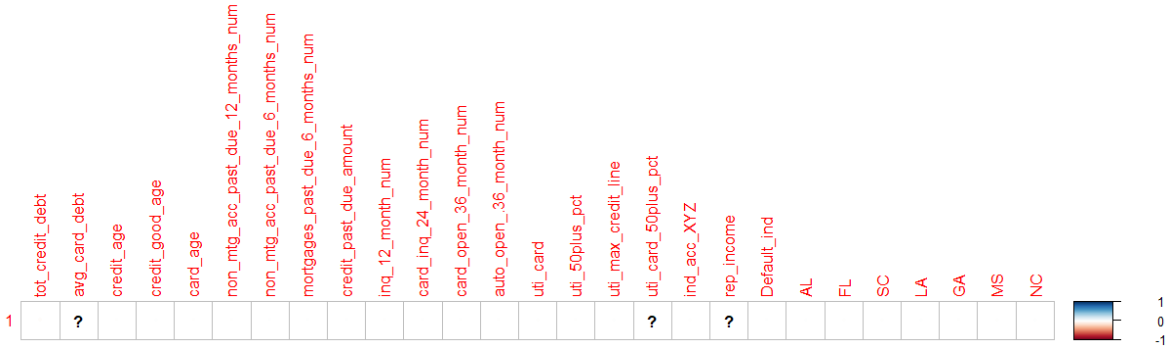
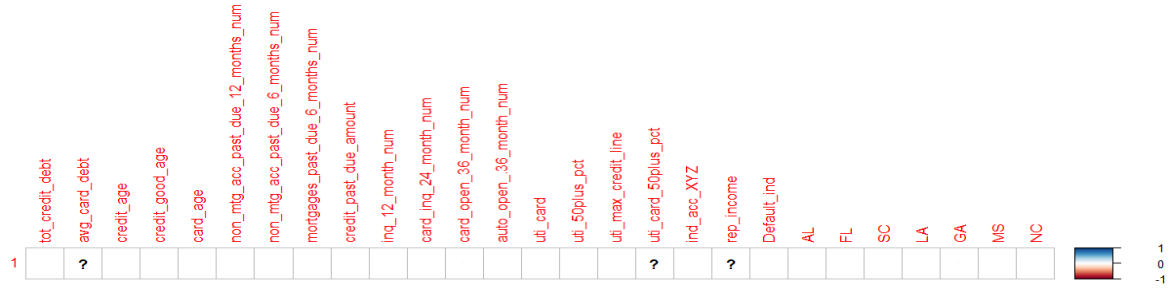
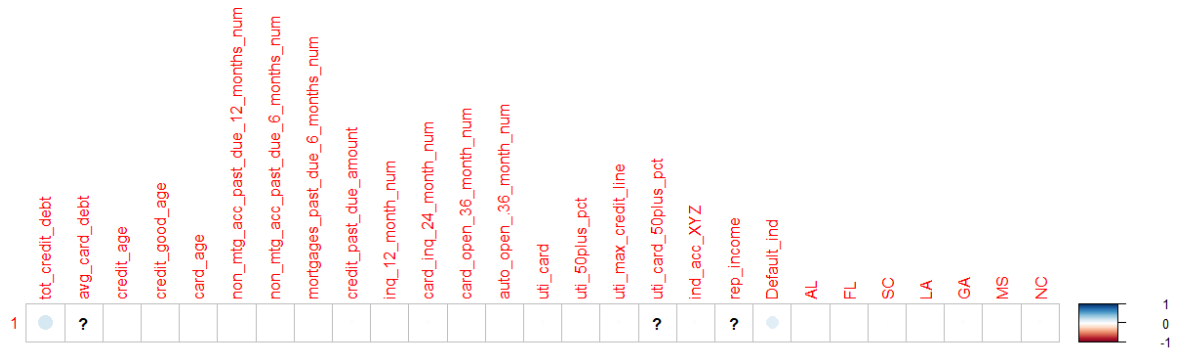


Figure 4: Annual Income Missingness Correlations



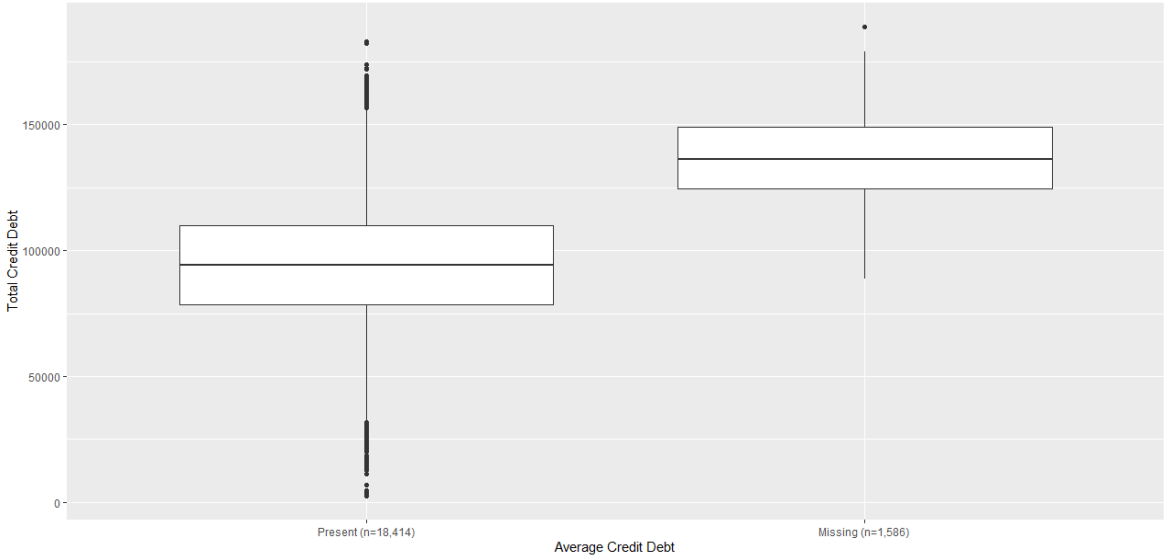
According to Figures 3 and 4, there does not appear to be any linear correlations between the missing values of rep_income (self-reported annual income) or uti_card_50plus_pct (the percent of open credits with over 50% utilization). We can conclude that the values for these two predictors are missing at random.

Figure 5: Average Monthly Debt Missingness Correlations



The missing values of average credit card debt are positively correlated with total credit card debt and the default indicator, with correlations of 0.185 and 0.143, respectively. This may indicate that the data is not missing at random and warrants further analysis. Figure 6 shows the distribution of average card debt broken down by whether it is missing or present.

Figure 6: Distribution of Total Credit Debt by Average Credit Debt Missingness



We can see from Figure 6 that the distributions of total credit debt by average credit debt missingness both appear normal but differ in mean and standard deviation. The samples have means of $136,626 and $94,113 for missing and present, respectively. We can see if the two samples have a significant difference by checking with a two sample t-test for differing means.

Table 3: Two Sample T-Test of Differing Means

t-score:	34.029
Degrees of Freedom:	218.54
p-value:	<0.00001

95% Confidence Interval for the difference:	[40051, 44976]

With a p-value less than 0.05 we have sufficient evidence to reject the null hypothesis that the two samples have equivalent means, and we can conclude that the two samples have significantly different means. Next we will examine the relationship between average card debt missingness and the default rate.

Table 4: Cross Tabulation of Default Rate and Average Card Debt Missingness

Default\Missing	Present	Missing
Non-defaulted	18,298	1,490
Defaulted	116	96

We can see that the rate of default is much higher for cases that are missing annual income. The odds of defaulting are 10.2 times higher for cases missing average credit debt than those with the value present. We can check for association with a chi square test.

Table 5: Chi Square Test

X^2:	404.32
Degrees of Freedom:	1
p-value:	<0.00001

With a p-value less than 0.05 we have sufficient evidence to reject the null hypothesis that defaulting and missing average credit debt are independent, and we can conclude that there is a significant association between the two variables.

Given the significant relationships that average card debt has with total credit debt and the default indicator, we can conclude that the missing values are not missing at random. We will include an indicator for average card debt missingness as its own predictor going forward.

The variables that were determined to be missing at random can be imputed from the existing data. Since we have found strong evidence of multicollinearity we will use the random forest method of imputation because it can account for interactions between variables. For average card debt, which we determined to be not missing at random, we cannot make assumptions about the distribution of the missing values. We can however use the missingess as a predictor, and we will substitute all missing instances with the sample mean.

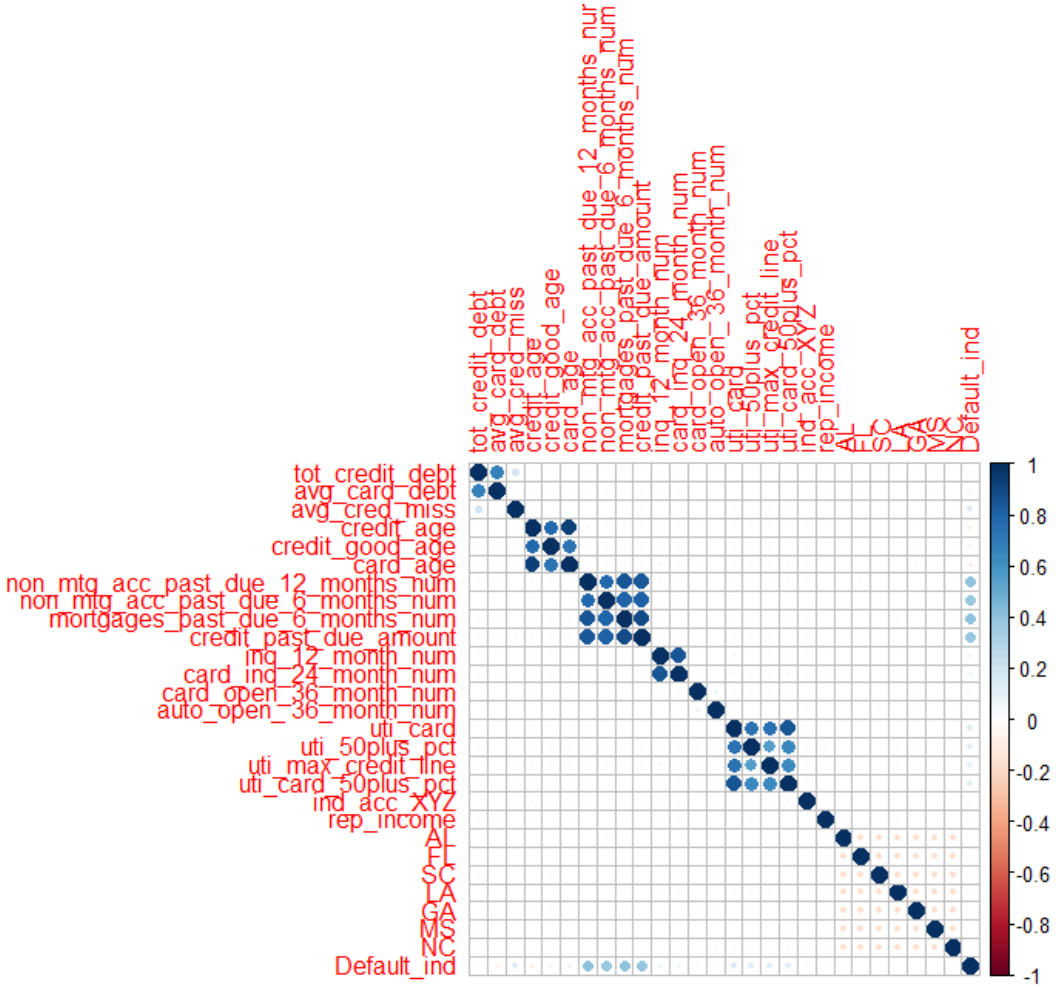
Random forest imputation is an imputation method which begins by inserting an initial guess for missing values. The mode is selected for categorical and binary variables while the median is selected for numeric variables. Next, a random forest is created for the data, and each instance's similarity to other instances is determined and used to refine the initial guess.

This process, creating a new random forest and using a similarity to refine the imputed value, is repeated multiple times.

2.3 Multicollinearity

A key assumption that must be met when interpreting the effects of a predictive model is that all of the predictor variables are independent from each other. This means that we should be able to analyze the effect of each predictor in the model while holding the others constant. If multiple predictors are highly correlated with each other, it becomes difficult to meet this assumption and therefore difficult to accurately interpret the results of the model. Figure 7 shows the correlation matrix between all of the numeric predictor variables in our dataset.

Figure 7: Correlation matrix of numeric predictor variables



As shown in Figure 7, there are several “blocks” of variables that seem to be highly correlated with each other. This alone does not always indicate a presence of problematic multicollinearity, but we can also look at VIF scores from a logistic regression model containing all of the predictors. A VIF score is a metric that quantifies the severity of multicollinearity, where

scores close to 1 no multicollinearity, and scores above 5 indicate problematic multicollinearity which should be addressed. Table 6 shows the VIF scores which are greater than 5 out of the complete set of predictors.

Table 6: Problematic VIF scores from the complete set of predictors.

Variable	VIF Score
credit_age	9.868
card_age	8.279
mortgages_past_due_6_months_num	7.189
credit_past_due_amount	6.342
uti_card	6.101

As shown in Table 6, there are five predictors whose VIF scores indicate problematic multicollinearity. All of these variables correspond to those which showed high correlation with other predictors in Figure 7.

The best way to begin to deal with the multicollinearity shown amongst our predictors is to interpret them in context and decide if any of them can essentially be explained by other variables in the dataset. If you look closely at the blocks of correlated variables in Figure 7, it appears that the variables in each block are all conveying similar information. For example, there is a group of variables that relates to utilization of accounts, and another group that has to do with past due accounts. To deal with the multicollinearity in our data, we can interpret the variables and decide which are redundant and can be removed from the data.

Specifically, the number of non-mortgage credit products past due in the last 6 months and 12 months, number of mortgages past due in the last 6 months, and the total amount of credit past due are highly correlated variables that all measure credit that is past due. These variables contain redundant information because an individual's past due credit can essentially be broken down into all of their past due accounts. Additionally, mortgages_past_due_6_months_num has a very high VIF score according to Table 6. Since the variables relating to the number of accounts past due carry essentially the same information as the total past due credit amount for an individual, we will keep credit_past_due_amount and drop the other three aforementioned variables.

Another block of highly correlated variables is the block containing credit_age, credit_good_age, and card_age. These all have to do with the longevity of individuals' credit cards or other credit products. The two variables that seem most related which we will focus on first are credit_age (the age in months of the first credit product obtained by the applicant), and credit_good_age (the age in months of the first credit product obtained by the applicant that is currently in good standing). After checking the data to see if any of the applicants' first credit product was still in good standing (checking if credit_age == credit_good_age), the data showed that this was true for none of the applicants. This means that all 20,000 applicants in the dataset

have missed at least one payment on their first credit product. Thus, the variable credit_age does not contain any additional information than credit_good_age for any of the applicants. Additionally, credit_age has a very high VIF score, so it will also be dropped.

The last block of highly correlated variables has to do with utilization. There are four variables in the dataset that all have to do with credit utilization, so it is likely that at least one of the variables is redundant. It is difficult to interpret which of these four variables is redundant based on their interpretations, but if we look back at Table 6 we can see that uti_card is the only one of them with a problematic VIF score. Therefore we can assume that this variable carries redundant information and should also be dropped.

After removing the redundant variables, we can recalculate the VIF scores and see if multicollinearity is still a problem. Table 19 in the appendix shows the updated VIF scores of the complete set of predictors after excluding the dropped variables. All of the predictors now have a VIF score under the threshold of 5, meaning that the multicollinearity in the dataset has been essentially neutralized and we can continue with our modelling. Removing redundant predictors also reduces the dimensionality, and therefore increases the interpretability, of our models.

3 Logistic Regression Modelling

Since our response is binary, our records are independent, and our variables have little multicollinearity, all of the assumptions for logistic regression have been met and we are ready to fit and interpret a model with the default indicator as the response. We will use the variables shown in Table 7 so that we have the most effective interpretation of the coefficient estimates.

Using the variables shown in Figure 7, we fit a simple logistic regression model with the default indicator as the response. Table 7 shows the significant predictors from the model along with some additional information about the interpretation of the coefficient estimates.

Table 7: Estimated Coefficients of Significant Predictors

Significant Variable	Change in Odds of Defaulting for every 1 unit change in predictor	Standardized Change in Log Odds of Defaulting
Total Credit Debt ($1000)	0.99445	-3.159
Average Card Debt ($1000)	0.94674	-3.709
Average Card Debt Missing Indicator	23.514	18.007
Age in Months of First Credit Card	0.99417	-7.8
Credit Past Due Amount ($1000)	1.4335	30.994
Number of Credit Inquiries in the past 12 Months	1.21086	5.926

Number of Credit Cards Opened in the past 36 Months	1.1621	2.065
Percent of Open Credit Products with >50% Utilization	23.895	4.747
Utilization of Credit Product with Highest Credit Limit	4.9724	4.314
Percent of Open Credit Cards with >50% Utilization	23.895	8.528
XYZ Bank Account Indicator	0.77257	-3.519

Before we assess the effectiveness of the model, we must interpret the odds ratios to see each variable's effect on the model. We can say that for every 1 unit increase, or $1000, in an applicant's amount of credit past due, the odds of a default are 1.4 times greater. Alternatively, an increase in the age of an applicant's first card by 1 month will change the odds of a default by 0.994 times, in other words a decrease of about 1% per month.

To compare the effect size of each variable we can consider the standardized coefficient estimate, also called the z-value. The standardized coefficient represents the change in log odds given a one standard deviation sized change in the predictor. A high absolute value z-value means that the coefficient estimate is large and precise enough to be significantly different than zero.

Due to the imbalance in the response we must choose the metrics for measuring effectiveness carefully. To get a good overall understanding of the model we will examine all aspects of the prediction confusion matrix. For the predictions we will use a cutoff of 0.0793, which is the percentage of the response that are positive.

Table 8: Confusion Matrix of Test Data Labels

		Observed	
		0	**1**
Predicted	**0**	3,584	116
	1	1,015	285

Table 9: Prediction Metrics of Test Data

Metric	**Value**
Precision	0.2192
Recall	0.7107
Accuracy	0.7738
F1 Score	0.3351

From table 9, our precision is 0.22, which means the model over-predicts positive results. With a recall value of 0.71, the model identifies actual cases of default fairly well. Our F1-Score is a more balanced metric which considers both precision and recall, and it is a good indication of the model as a whole. Our F1 score is fairly low at 0.34, showing that overall the

model has poor effectiveness. However, in this scenario a false negative is far more costly than a false positive. We would rather deny a loan application that would not default than approve a loan which will end up defaulting. For this reason, our model's high recall is evidence that it would still be very useful for considering a loan application.

To be careful of overfitting we should examine the effectiveness of the model on the training set as well.

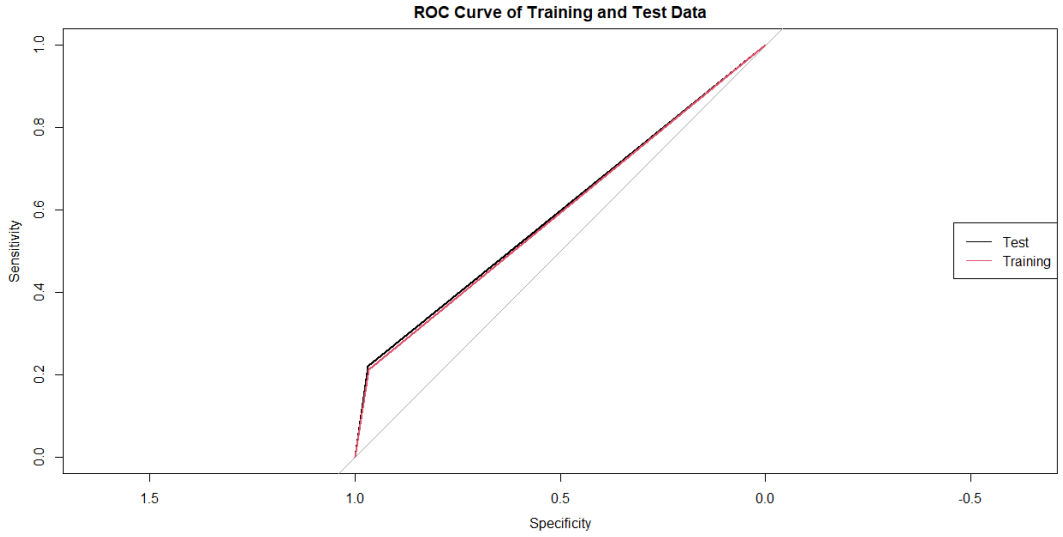
Table 10: Confusion Matrix of Training Labels

		Observed	
		0	**1**
Predicted	**0**	14,367	498
	1	4,047	1,088

Table 11: Prediction Metrics of Training Data

Metric	Value
Precision	0.2119
Recall	0.686
Accuracy	0.7728
F1 Score	0.3238

Figure 8: ROC Curve of Testing and Training Data



The effectiveness of the model predictions on the training data are very similar to the testing data. This indicates that there is no overfitting present in the model.

4 Random Forest

The machine learning algorithm we chose to use for this problem is the random forest, as opposed to a gradient boosting algorithm or a neural network. Each of these algorithms have their pros and cons, however a random forest is the optimal choice for the specific problem we are solving.

A random forest model is an ensemble machine learning algorithm based on decision trees, meaning that this complex model makes predictions based on the collective decision of

many simpler decision trees. Each decision tree in a random forest is created using a random sample (with replacement) of the dataset, also known as a bootstrapped dataset, and a randomly selected subset of the predictors. This ensures that each tree in the model is slightly different, which reduces the bias and error of the model. This also protects the model from overfitting the data because if the model overfits the training data it will be worse at making predictions on unseen data. Gradient boosting algorithms in particular are much more prone to overfitting. Additionally, a random forest model handles class imbalance in the response particularly well compared to the other machine learning algorithms, which is good for our data because there is significant class imbalance.

Perhaps the most important reason we decided to use a random forest model is its interpretability. This is important for the context in which we are working because if someone is denied a loan from XYZ Bank, they will most likely want to know why. This will also be helpful to the bank; they can use knowledge on the biggest predictors of credit worthiness to inform on future decisions. The random forest model is by far the simplest of the three machine learning algorithms, so it would be the easiest for a banker to explain to someone who may ask why they were denied for a loan. We believed that this was the most important criterion in which method we selected.

Ensemble machine learning algorithms like random forest have several different parameters that can be tuned to improve the accuracy and performance of the model. Table 12 provides explanations for the most important parameters for our random forest model.

Table 12: Random Forest Parameters

Parameter	Description
Number of estimators	The number of individual decision trees in the model that work as an ensemble to make predictions.
Subset of predictors	The maximum number of predictor variables considered in the construction of each tree.
Leaf/bag size	The number of records that must fall in each leaf node/bag of the individual trees.

To create an effective random forest model, we will need to start by fitting a model on the training set using the default values for the parameters. The next step is to use the validation set to tune the parameters to optimize the performance of the model. We will tune the parameters using the grid search procedure because it allows us to select multiple values for each parameter and iterate through each possible combination to find the optimal values for each. We use the F1 Score as our success metric for the grid search because the bank may value both precision and recall. While recall may be valued greater (because the bank wants to ensure it does not offer credit to individuals who will default), it is important to use a balanced metric to ensure that both precision and recall are maximized.

Once the parameters are tuned, the model will be run on the testing set and we will assess its accuracy. Because multicollinearity can also be an issue in random forest models, we will use the same set of variables that was used in the logistic regression model.

The optimal values for the parameters in the random forest model are shown in Table 13.

Table 13: Optimal Values of Tuned Hyperparameters.

Parameter	Value
Number of trees	250
Predictor subset size	10
Leaf/bag size	12,000

After fitting the random forest model using the tuned parameters, we can analyze its performance based on a confusion matrix and several performance metrics, these are shown in Table 14 and 15, respectively.

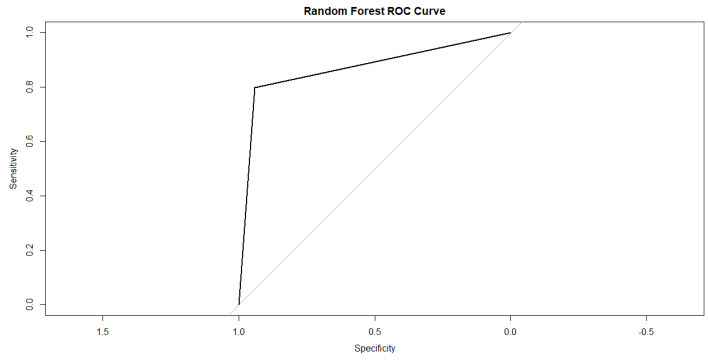
Table 14: Confusion Matrix of Test Labels

		Observed	
		0	**1**
Predicted	**0**	4,569	282
	1	30	119

Table 15: Prediction Metrics of Test Data

Metric	Value
Precision	0.7987
Recall	0.2967
Accuracy	0.9376
F1 Score	0.4327

Figure 9: Random Forest ROC Curve



Our random forest model has a precision, which indicates it does not over guess positive results. However, the recall rate is very low, which indicates it poorly captures all of the true positive cases.

The random forest model also allows us to see which predictors are most important in the predictions of the model, the feature importance in terms of the mean decrease in Gini Index are shown in Table 16.

Table 16: Feature Importance for Random Forest Model

Predictor	Mean Decrease in Gini Index
tot_credit_debt	140.44
avg_card_debt	240.45
avg_cred_miss	35.68
credit_good_age	110.87
card_age	125.79
credit_past_due_amount	303.24
inq_12_month_num	64.64
card_inq_24_month_num	67.63
card_open_36_month_num	16.52
auto_open_36_month_num	12.32
uti_50plus_pct	134.70
uti_max_credit_line	134.88
uti_card_50plus_pct	164.64
ind_acc_XYZ	14.74
rep_income	111.60
AL	12.79
FL	11.31
SC	12.80
LA	12.36
MS	11.46
NC	11.53

5 Conclusion

The comparison of performance metrics of the logistic model and the random forest model are shown below.

Table 17: Logistic Regression Results

Metric	Value
Precision	0.2192
Recall	0.7107
Accuracy	0.7738
F1 Score	0.3351

Table 18: Random Forest Results

Metric	Value
Precision	0.7987
Recall	0.2967
Accuracy	0.9376
F1 Score	0.4327

According to the performance metrics, the random forest model performs significantly better than the logistic regression model in most aspects. There is a clear difference between the precision and recall of the two models. The logistic regression has a high recall and low precision, while the random forest has a low recall and high precision. This means that the logistic model is good at capturing true cases of default, at the risk of over predicting for default cases. Because the random forest is more selective about predicting for default cases, it has a better overall prediction ability but poorly captures the true cases of default.

From the bank's perspective, approving a loan that will default is much more harmful than denying a loan that will not default. Because the logistic regression model has a significantly higher recall, we value its classifications greater than the random forest model's classifications. Thus, we will propose our logistic regression model to bank XYZ to use for future decisions on providing credit.

The results of our logistic regression model show that having an account with XYZ Bank decreases the odds of defaulting (Table 7), meaning that our model shows slight favorable consideration towards applicants who already have an account open with XYZ Bank.

Given the interpretability of the estimated coefficients, you could explain to each applicant specifically which aspects of their profile led to their rejection. For example, an applicant with a large amount of credit that is past due, would be far more likely to be rejected for the loan.

6 Appendix

Table 19: VIF Scores of all remaining predictors.

Variable	VIF Score
tot_credit_debt	2.144
avg_card_debt	1.863
avg_cred_miss	1.280
credit_good_age	2.160
card_age	2.169
credit_past_due_amount	1.049
inq_12_month_num	4.227
card_inq_24_month_num	4.218
card_open_36_month_num	1.012
auto_open_36_month_num	1.013
uti_50plus_pct	1.782
uti_max_credit_line	1.782
uti_card_50plus_pct	2.105
ind_acc_XYZ	1.005
rep_income	1.002
AL	1.824
FL	1.780
SC	1.825
LA	1.823
MS	1.834
NC	1.813

Figure 10: Average Card Debt vs Total Credit Debt

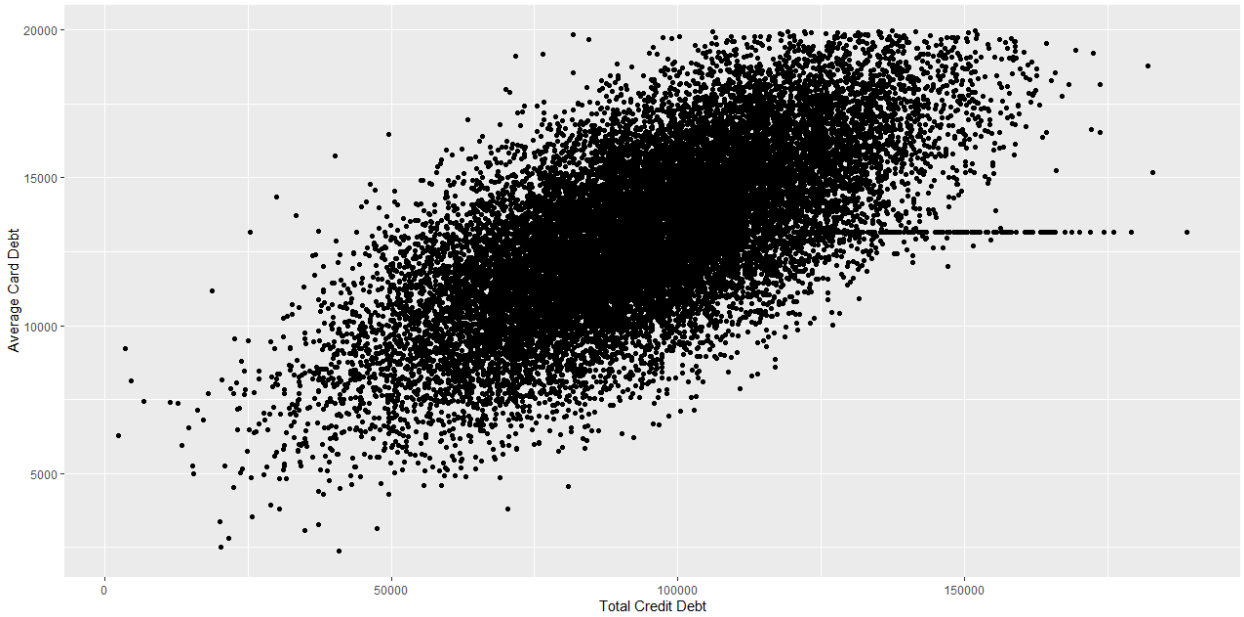


Figure 11: Age of First Credit Product vs Age of First Credit Card