A Multi-Faceted Approach to Credit Risk Modeling

Math 358
Fall 2019

Austin Jones, August Horkan, Zach Kennedy, Brandon Shelton

James Madison University

4/15/2020

Contents

1 Background

1.1 Abstract

Our modeling process began with exploratory data analysis and visualization to gain familiarity with the features. We then construct three loan default predictive models, utilizing logistic regression, decision tree, and QDA/LDA modeling structures. The logistic regression model was trained on an 80-20 split and targeted at a binary variable derived from loan status.

1.2 Introduction

With fortunes on the line, the analysis of credit risk is a crucial focus of financial services firms. It is vital for institutions to manage loan defaults, as these represent losses to their lending revenue. Credit risk models are also critical to Dodd-Frank Stress Tests and more generally to the Federal Reserve's Comprehensive Capital Analysis and Review. As financial institutions continue to amass data at an outstanding rate, there is now an opportunity to leverage this big data for accurate loan default prediction.

1.3 Related Works

Credit Risk modeling has a backbone of research from industry and academia. Loan default modeling in particular has been approached a number of ways. From logistic regressions to decision trees, there is not just one way to build a loan default model. While this research is robust, most publicly available research has been conducted on much smaller samples an only takes one modeling approach. Our research will be different in that we have thirteen years of data totaling over a million observations, and we're hoping to compare models from at least two approaches, possibly logistic regression and decision tree. We could also develop K-Nearest Neighbors and Support Vector Machine Models.

1.4 Data

The data used for this analysis is Lending Club Loan Data, sourced from Kaggle. This is a complete data frame of all loans issued from 2007 to the first quarter of 2020 (or 2007-2015). There over a million observations and 75 variables. We will particularly be interested in loan status as our variable of interest. Some features we are interested in are loan amount, grade of employment, annual income, term, education, gender, and age. We cleaned the data set by selecting variables that may have an impact, which meant that we removed variables that were either repetitive, almost certainly unimportant, or that had too many missing variables to clean. We then filtered out the missing values that were shared by too many variables to clean without creating averages of the existing averages, especially with so few missing values anyways, which left no missing values.

1.5 Methodology

The classification tree is made using a by choosing the variable with the smallest gini index, which is

$$Gini = -\sum_{i=1}^{n} p_i^2 \tag{1.1}$$

Where i is the classes, n is the number of classes, and p is the probability of that class happening.

The formula repeats itself to go over all trees, and then prunes itself if the cp of adding the branch is not significantly different compared to not adding it, according to the given cp level.

A Linear Discriminant Algorithm works by calculating the separability between classes, in class variance, and then applies Fisher's criterion. The LDA then uses Bayes theorem to calculate the probability to which class each observation belongs to. The purpose of this is to find the linear combinations of the variables that gives the best separation between groups in the dependent variable.

The Linear Discriminant model was trained on a 70-30 split, selecting the variable "loan_status". An LDA fit was applied on the training data using the MASS package, and histograms were plotted for the first and second discriminant functions. Accuracy assessments were also made using predictions from the test set.

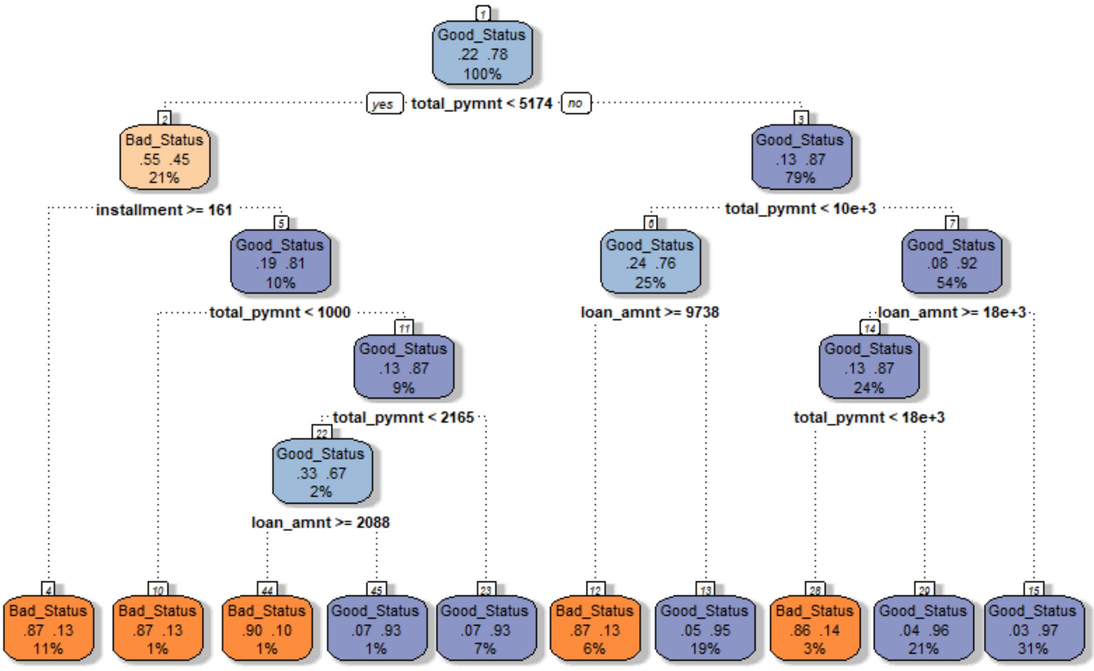
The logistic regression model was trained on an 80-20 split and targeted at a binary variable derived from loan status. Defaults and loans in collection are 1's, all else are 0's. For linear regression, we initially selected a subset of around 20 variables to investigate as predictors. We then split the data, we then found our best logistic regression model using the caret package and compared it to our training data. We evaluated the predictions, create a confusion matrix, investigate the ROC Curve, and analyze sensitivity, specificity, and accuracy. This logistic regression model provides interpretable coefficients and overall ease of explanation. That said, our model could suffer from omitted variable bias, given we did not consider macroeconomic indicators.

2 Results and Conclusion

2.1 Decision Tree

The data set was further cleaned/changed by removing a couple of variables and changed emp_length to be numeric, as well as created dummy variables for grade instead of just having the grade itself. We also changed the response variable into a variable with either a good status or a bad status so we could have a simpler decision tree confusion matrix. Used an 80/20 split on the target variable loan status and then created a classification tree on the training data. We were looking for accuracy, compared with the no information accuracy.

	Reference	
Prediction	Bad_Status	Good_Status
Bad_Status	10217	1586
Good_Status	1854	41720

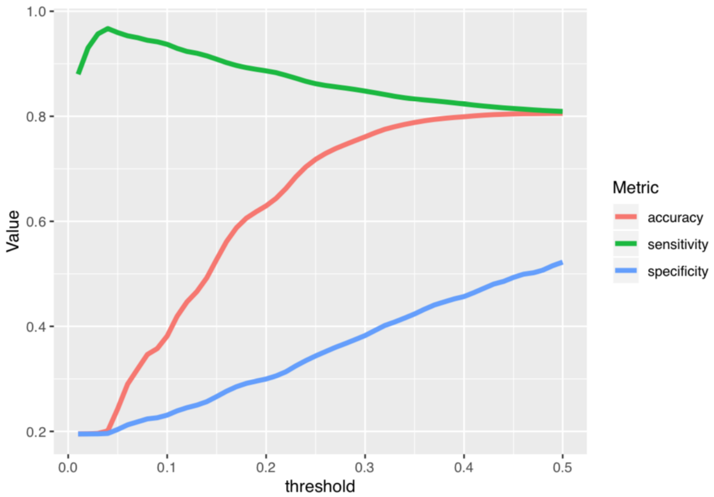


The accuracy was 93.79%, roughly 15% better than the no information rate. This model was fairly successful in predicted whether the loan_status was good or bad for the lenders.

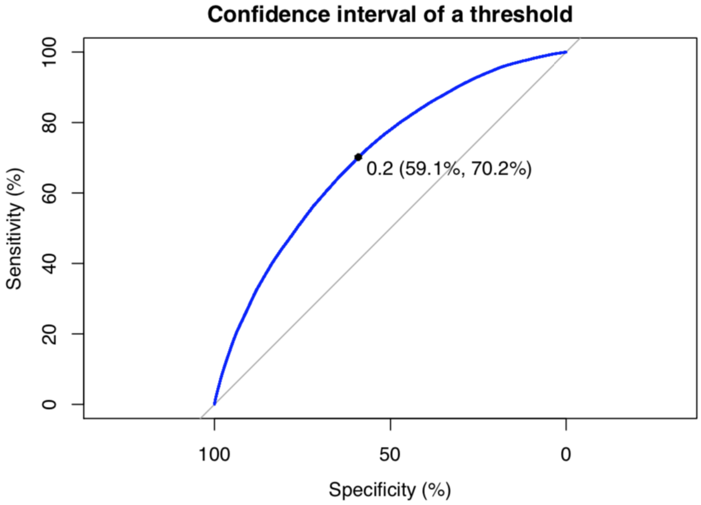
2.2 Logistic Regression

The logistic regression model was trained on an 80-20 split and targeted at a binary variable derived from loan status. Defaults and loans in collection are 1's, all else are 0's. We then fit a model to the training data and conduct a number of evaluation measures. We evaluated the predictions, create a confusion matrix, investigate the ROC Curve, and analyze sensitivity, specificity, and accuracy.

We plot the values for accuracy, sensitivity, and specificity:



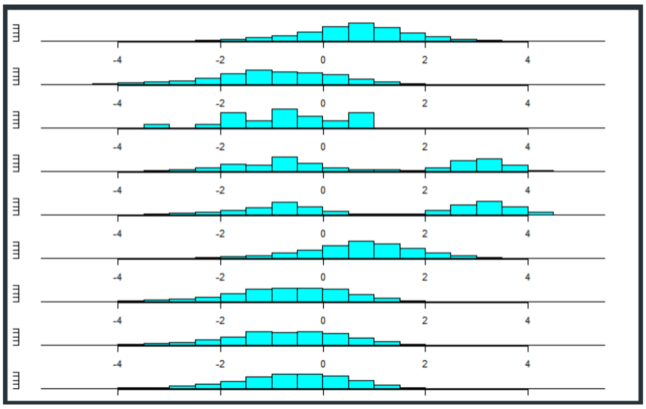
These values tell us that our model is very accurate at identifying the loan defaults, but too often predicts loans to default when in fact they do not.

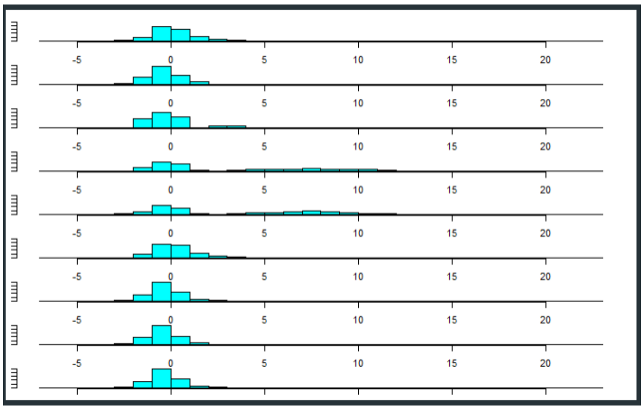


The area under the ROC curve is somewhat underwhelming at 0.7801. This does not inspire confidence in this model as a highly efficient predictor of loan defaults.

2.3 LDA

The percentage of separation accounted by LD1 is 73.05%, and the percentage of separation accounted by LD2 is 22.51%. The average compactness for each class is 15549.441 for “Charged Off”, 15887.445 for “Current”, 17089.773 for “Default”, 9668.715 for “Does not meet the credit policy. Status:Charged Off”, 8852.741 for “Does not meet the credit policy. Status:Fully Paid”, 8852.741 for “Fully Paid”, and 17763.482 for “In Grace Period”. The histograms for LD1 and LD2 respectively are:





2.4 Conclusion

The logistic regression model was clearly far from perfect. It overpredicted defaults and didn’t have the accuracy we look for in a predictive model. On the contrary, our decision tree model yielded great accuracy. The model was significant and effective in predicting loan defaults. The linear discriminant model performed decently well…

This analysis has found that a decision tree-based algorithm is the most efficient predictor of loan defaults. In practice, logistic regression models are nearly an industry standard, with few firms deviating from this norm. Further research could be targeted at the models in practice, and with the consideration of omitted variables. For example, there are countless macroeconomic indicators not taken into consideration. These models took only the data pertaining to each individual loan, and not the economic environment surrounding the lending markets.

3 Contributions

Austin Jones:
- Came up with original idea and found data set
- General data filtering (cleaning, feature engineering, visualization)
- Wrote up proposal
- Created a logistic regression, including model evaluation and ROC curve
- Wrote abstract, introduction, related works, conclusion

August Horkan:
- Came up with unused idea and data set
- General data filtering (cleaning, feature engineering)
- Created a Classification Tree, including confusion matrix
- Wrote up Data
- Put everything into latex for report and presentation.

Zachary Kennedy:
- Contributed by creating the LDA model (including histograms and predictions output), writing up the results of the LDA model, writing up the methodology of the LDA model, worked on unused models (k-NN and QDA), and helped give the final presentation

Brandon Shelton:
- Nothing

4 Bibliography

Dataset:
Kan, Wendy. "Lending Club Loan Data." Kaggle, 29 Apr. 2016, www.kaggle.com/wendykan/lending-club-loan-data.

R Packages:
dplyr
caret
readr
ggplot2
rsample
MASS
vip
ROCR
tree
rpart
ISLR
RColorBrewer
rattle