Joel Cabrera, Kouame Oura, Yifu Zhang Professor Guo

Data Mining (16:958:588) May 9, 2021

Credit Applications: Evaluating Decisions Using Classification Methods - Final Report

Abstract

Credit serves as one the main ways of consumers to obtain goods and services readily, conditional on them being able to repay those loans. However, credit lenders may be uncertain as to whether they should provide credit, as some people may default on their loans. Thus, we investigate this phenomenon based on whether someone was approved of receiving credit using six statistical models: logistic regression, classification trees, k-nearest neighbors (KNN), random forests, generalized additive models (GAMs), and boosted trees. Based on our analysis of the "credit" dataset from the UCI Machine Learning Repository, we find that random forests best predict whether someone is approved of receiving credit. Such a finding would perhaps make banks minimize the risk related to lending and expedite the process.

Keywords: credit, regression, trees, KNN, random forests, GAM, boosting, approved

I. Introduction

Credit plays an important role in our lives. It is one of the main ways consumers can obtain goods and services readily. Indeed, we may eventually need to apply for credit in order to obtain an education, a car, housing, or even starting a business. However, credit lenders tend to be unsure of whether to provide credit, as some people default on loans and thus hurt their businesses **[1]**. Evaluating credit applications manually by issuer staff is a time-consuming process, as well. It would be inefficient for both the issuer and applicants. Thus, automating such processes is quite important.

The goal of this project is to develop a model to make the most accurate decisions possible for credit approval. If we succeed, then we can ensure that credit approval can perhaps be automated and lessen credit lenders' fears.

II. Data & Description

The dataset that will be used for our data analysis is the "Credit Approval Data Set" (referred to as "credit") from the UCI Machine Learning Repository **[2]**. The data is formatted as a DATA file. It contains 690 observations (rows) and 16 variables (columns). Of these 16 variables, 6 are continuous and 10 are categorical. All but 1 of these variables will be used as

predictors - with the remaining one, “approved” (0 = denied credit, 1 = approved for credit), being the response variable. Note that the original data contain missing values and anonymized information. Both of these problems are accounted for in the upcoming sections. All analyses of these data were conducted using R.

III. Quantitative Methodologies

Given the properties of the credit dataset, we will conduct some data cleaning, exploratory data analysis (EDA), and data splitting to prepare the data for analysis. After this, we then analyze the data using six statistical models to process and analyze said dataset. These six statistical methods are: logistic regression, classification trees, k-nearest neighbors (KNN), random forests, generalized additive models (GAMs), and boosted trees. The performance of all six models will be evaluated via cross-validation – that is, to choose the optimal tuning parameters for their respective models. The model that will be chosen to be the best among them will be based on determining which model’s prediction accuracy for “approved” is the highest. We will also obtain their AUC values, in order to see how well a model predicts.

Data Cleaning

The first step was to replace the character variables with their corresponding numeric variables, then replace missing values represented by “?” by NA. We then defined the “education” variable as an ordered variable in R.

All the values can now be converted to numeric values and using missForest() function, missing values are imputed. Also, we replaced the anonymized names and values with their true names based on the provided data documentation. The names of these variables are below:

```
> names(data) # now has meaningful name
 [1] "male"            "age"             "debt"            "married"
 [5] "bank_customer"   "education"       "ethnicity"       "years_employed"
 [9] "prior_default"   "employed"        "credit_score"    "drivers_license"
[13] "citizenship"     "zip_code"        "income"          "approved"
```

EDA

In terms of checking and exploring the data, we focused mostly on bar charts and histograms. The bar charts of the categorical variables indicate that the data is balanced - particularly for “approved”. The histograms of the continuous variables show that the variables are skewed to the right. This is not a major concern, however, as the data is still balanced.

Splitting Data

After cleaning the data and conducting some EDA, we do an 80-20 split of the data into training and test sets. Both sets have dimensions of 552 x 16 and 138 x 16, respectively.. The "approved" variable is also balanced in each of them, allowing for more accurate results in our upcoming models.

IV. Results & Discussion

Logistic Regression

Before fitting the logistic regression model, we should check the linearity among the continuous variables, the assumption is that all the continuous variables are linearly independent. According to the scatter matrix in R, there is no obvious linear relationship among the continuous variables, which satisfied this assumption.

The response variable "approved" is binary. Thus, we will model this using three logistic regressions so that we can estimate the probabilities for "approved". We will use the training data to estimate the models' regression coefficients and evaluate the models via cross-validation.

The first model was obtained by taking all of the 15 variables we have. After fitting the model, based on p-values some variables were dropped for simplification, we obtained a second model - having only 8 predictors.. In the same way, we also obtained a third model - having only 4 predictors.

With the three candidate logit models, we then evaluated their performance on the test set. Via cross-validation. In R, we received the following result: The three models correctly predicted 84.78%, 80.43% and 84.06% of the test observations, respectively. We can see that among the three candidates, the third model is the simplest one, having a prediction accuracy close to that of the first model which has more variables. Therefore the third model is retained as the best model with accuracy 84.06% and AUC value 0.849. The model contains four predictors: married, prior default, citizenship and income.

Classification Tree

Fitting a classification tree in R, we find that 83.33% of the test observations for "approved" are correctly classified. However, to find a more accurate classification rate, we have to prune the optimal number of nodes for the tree. Performing cross-validation on the tree in R, we find that 2 nodes are the optimal number. Pruning the tree based on this, we ultimately find that 84.06% of the test observations for "approved" are correctly classified. This suggests that

pruning the tree based on 5 nodes results in a higher classification rate by 0.73% points. Its associated AUC value is 0.847, as well.

KNN

KNN is a non-parametric method. Its basis principle is to find the k neighbors of the new observation to classify then assign it to the class with most observations. In R, we used the knn() function. It takes as input the training and testing data of predictor variables as matrix, and the response variable of the training set as vector. We also standardized the predictors prior to data analysis, so for both sets of predictors, the mean and the standard deviation of the predictors in the training data have been used.

Cross-validation was conducted for different values of k, and each one has its own prediction accuracy rate. We find that the optimal k value is 8 - with a corresponding prediction accuracy of 87.68%. Its associated AUC value is 0.864, as well.

Random Forests

We now fit a random forest to the "credit" data. We first obtain the minimum out-of-bag-error (OOB) to select the optimal terminal node size and number of parameters (variables). Running our algorithm in R, we find that the optimal node size and number of parameters are 1 and 4, respectively. Fitting and evaluating our random forest model using the test set, we find that the model's prediction accuracy is 89.13%. Its associated AUC value is 0.884, as well.

GAM

Since "approved" is a binary response variable, it would make sense to also fit a logistic GAM model to the data. For our analysis, we fit five GAMs. The first GAM contains a degree of freedom (DF) of 1 for the smoothing terms (e.g. age, debt, years employed, and income). The second to fifth GAMs individually contain a DF of 2 for each smoothing term. Evaluating their performances on the test set, we find that they have prediction accuracies of 84.78%, 84.06%, 83.33%, 84.78%, and 85.51%, respectively. Thus, the fifth model - with "income' having DF of 2 - is the best GAM. The AUC associated with this model is 0.847, as well.

Boosting

The shrinkage parameter, depth of trees, and number of trees have been optimized to find the optimal boosted model using 5-folds cross-validation. Then, the optimal cutoff for the

probabilities has been calculated to improve prediction accuracy. Finally, using the testing data set, the accuracy of 88.41% has been found and the AUC is 0.882.

V. Summary of Results

Now that we were able to obtain the best models for each statistical method, we summarize their prediction accuracy rates and AUC values in the table below. This will be useful in making our final conclusions about our models.

Model	Accuracy %	AUC
Logistic Regression	84.06	0.849
Classification Tree	84.06	0.847
KNN	87.68	0.864
Random Forest	89.13	0.884
GAM	85.51	0.847
Boosting	88.41	0.882

VI. Conclusion

Considering all six models used to predict "approved" using all other 15 variables, we find that the most optimal logistic regression, classification tree, KNN, random forest, GAM, and boosting models have classification accuracy rates of 84.06%, 84.06%, 87.68%, 89.13%, 85.51%, and 88.41%, respectively. Additionally, their corresponding AUC values are 0.849, 0.847, 0.864, 0.884, 0.847, and 0.882, also respectively. Thus, it would be reasonable to believe that our random forest model has the best predictive performance of the six models we used. Such a finding would perhaps make banks minimize the risks related to lending and expedite the lending process.

References

[1] Arnold, Chris. "Millions Of Americans Skip Payments As Tidal Wave Of Defaults And Evictions Looms." *NPR*, NPR, 3 June 2020, www.npr.org/2020/06/03/867856602/millions-of-americans-skipping-payments-as-tidal-wave-of-defaults-and-evictions-.

[2] *UCI Machine Learning Repository: Credit Approval Data Set*, archive.ics.uci.edu/ml/datasets/Credit+Approval.