2023 Business Analytics Competition
CSUS Center for Business Analytics

"Should This Loan be Approved or Denied?"
Predictive Modeling Using the SBA National Data

The U.S. Small Business Administration (SBA) was founded in 1953 on the principle of promoting and assisting small enterprises in the U.S. credit market (SBA Overview and History, US Small Business Administration (2015). Small businesses have been a primary source of job creation in the United States; therefore, fostering small business formation and growth has social benefits by creating job opportunities and reducing unemployment. One way SBA assists these small business enterprises is through a loan guarantee program which is designed to encourage banks to grant loans to small businesses. SBA acts much like an insurance provider to reduce the risk for a bank by taking on some of the risk through guaranteeing a portion of the loan. In the case that a loan goes into default, SBA then covers the amount they guaranteed.

There have been many success stories of start-ups receiving SBA loan guarantees such as FedEx and Apple Computer. However, there have also been stories of small businesses and/or start-ups that have defaulted on their SBA-guaranteed loans. The rate of default on these loans has been a source of controversy for decades. Conservative economists believe that credit markets perform efficiently without government participation. Supporters of SBA-guaranteed loans argue that the social benefits of job creation by those small businesses receiving government-guaranteed loans far outweigh the costs incurred from defaulted loans.

Since SBA loans only guarantee a portion of the entire loan balance, banks will incur some losses if a small business defaults on its SBA-guaranteed loan. Therefore, banks are still faced with a difficult choice as to whether they should grant such a loan because of the high risk of default. One way to inform their decision-making is through analyzing relevant historical data such as the SBA National Data with the following Data Dictionary (see Li, Mickel, and Taylor (2018) for background information about the data and analysis):

Data Dictionary

NAME: National SBA

TYPE: Census

SIZE: 899,164 observations, 27 variables

SOURCE: United States Small Business Administration

STORY BEHIND THE DATA: This data set is from the U.S. Small Business Administration (SBA) and provides historical data from 1987 through 2014, containing 27 variables and 899,164 observations. Each observation represents a loan that was guaranteed to some degree by the SBA. Included is a variable [MIS_Status] which indicates if the loan was paid in full or defaulted/charged off.

VARIABLE DESCRIPTIONS:

The data reside in a comma-separated values (csv) file. A header line contains the name of the variables.

Variable Name	**Data Type**	**Description of variable**
LoanNr_ChkDgt	Text	Identifier – Primary Key
Name	Text	Borrower Name
City	Text	Borrower City
State	Text	Borrower State
Zip	Text	Borrower Zip Code
Bank	Text	Bank Name
BankState	Text	Bank State
NAICS	Text	North American Industry Classification System code
ApprovalDate	Date/Time	Date SBA Commitment Issued
ApprovalFY	Text	Fiscal Year of Commitment
Term	Number	Loan term in months
NoEmp	Number	Number of Business Employees
NewExist	Text	1 = Existing Business, 2 = New Business
CreateJob	Number	Number of jobs created
RetainedJob	Number	Number of jobs retained
FranchiseCode	Text	Franchise Code 00000 or 00001 = No Franchise
UrbanRural	Text	1= Urban, 2= Rural, 0 = Undefined
RevLineCr	Text	Revolving Line of Credit: Y = Yes
LowDoc	Text	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	Date/Time	The date when a loan is declared to be in default
DisbursementDate	Date/Time	Disbursement Date
DisbursementGross	Currency	Amount Disbursed
BalanceGross	Currency	Gross amount outstanding
MIS_Status	Text	Loan Status, “CHGOFF” (charged off or defaulted) or “P I F” (paid in full)
ChgOffPrinGr	Currency	Charged-off Amount
GrAppv	Currency	Gross Amount of Loan Approved by Bank
SBA_Appv	Currency	SBA’s Guaranteed Amount of Approved Loan

The CSUS Center for Business Analytics would like to acknowledge the contribution of Professors Amy Mickel and Stanley Taylor to the publication (Li, Mickel, and Taylor (2018)) of this data set for public use.

Complete the following:

1. Data Exploration and Preprocessing
How is the outcome variable MIS_Status distributed? Identify predictors that may help predict MIS_Status using descriptive statistics and visualization.

2. Divide the data into training and validation partitions. Choose appropriate predictors and develop classification models using the following methods to classify the loan applications as "higher risk" or "lower risk" for loan approval: kNN, Classification trees (single tree, bagging, boosting, and random forest), the logit model (including Lasso, Ridge, and ElasticNet), neural networks, and discriminant analysis. The costs of incorrectly classifying a loan application as lower risk outweigh the benefits of correctly classifying a loan application as lower risk by a factor of 5. The average net profit table was derived from the average net profit per loan based on the variable DisbursementGross (Amount Disbursed) as shown in the table below:

Average Net Profit (U.S. dollars)

	Actual	
Predicted (decision)	**Paid in full**	**Default**
Paid in full (accept)	5% of DisbursementGross	-5 times 5% of DisbursementGross
Default (reject)	0	0

Incorporate this cost and net profit information in your modeling. Where appropriate, normalize predictors and select hyperparameters using cross validation. Where appropriate, justify the algorithm/solver, e.g., IRLS (Iterative Reweighted Least Squares), SGD (Stochastic Gradient Descent), SAGA (Stochastic Average Gradient Descent), used in the optimization problem for these methods. Where appropriate, ensure the solver you choose converges by adjusting parameters. For neural networks, justify the parameters chosen, including the size of the hidden layer(s), the activation function, the solver, and the learning rate. Select an appropriate cut-off probability for each classification method incorporating the cost and net profit information and justify your selection. Discuss all implementation details. Choose one model from each method and report the confusion matrix and the cost/gain matrix for the validation data. Which method produces the highest net profit?

3. Use the estimated probabilities (propensities) from your chosen model as a basis for selecting the least risky loan application first, followed by more risky loan applications. Create a vector containing the net profit for each loan application in the validation set. Use this vector to create gains and lift charts for the validation set that incorporates the net profit.
a. How far into the validation data should you go to get maximum net profit?

b. If this model is used to score to future loan applicants, what “probability of success” cut-off should be used in extending credit?

Instructions

1. You have from 12 am, November 6, 2023, to 11:59 pm, December 10, 2023, to complete this project and submit it.
2. You may work in a group of up to 3 students.
3. You must use open-source tools Python and/or R.
4. Save your code, output, results, and explanation into one file and then email it to cbaanalytics@csus.edu by 11:59 pm, December 10, 2023.

Each member of the winning teams will receive an iPhone 15 as the prize.