SACRAMENTO STATE

CENTER FOR BUSINESS ANALYTICS

2024 CSU Systemwide Business Analytics Competition
CSUS Center for Business Analytics

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

Show us you can apply your business analytics skills to a real-world scenario and win an iPad Air!!

Submission Deadline: November 30, 2024 at 11:59pm

Winners Announced on December 6, 2024 at California State University, Sacramento

Register to enter the competition

Email cbaanalytics@csus.edu if you have any questions.

SACRAMENTO STATE

CENTER FOR BUSINESS ANALYTICS

2024 CSU Systemwide Business Analytics Competition

Li, Mickel, and Taylor (2018) published a data set from the U.S. Small Business Administration (SBA) to teach students making business decisions through statistical modeling. This data set has been used by others for machine learning projects on Kaggle and GitHub, thesis research, and course assignment. This data set will be used in this competition.

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 Apple Computer and FedEx. 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 probability of default. One way to inform their decision-making is through predicting this probability of default 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):

SACRAMENTO STATE

CENTER FOR BUSINESS ANALYTICS

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 resides 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

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

The decision to grant such a loan or not based on the predicted Probability of Default (PD) involves solving the same problems needed in adopting the Basel Framework developed by the Basel Committee on Banking Supervision to set international regulatory standards for bank capital adequacy, liquidity requirements, and stress testing (see stress tests by the Federal Reserve under Section 165(i)(2) of the Dodd-Frank Wall Street Reform and Consumer Protection Act). PD is a critical component in the Expected Loss (EL) approach in the Basel Framework for estimating potential losses from credit risk over one year:

EL amount = PD x EAD x LGD

where EAD is the Exposure At Default (the face value of the loan) and LGD is the Loss Given Default (expressed as a percentage of the loan). The expected loss helps banks and regulators determine the amount of capital needed to be set aside to cover potential losses, especially under stressful conditions, to ensure financial stability of the banking system. More specifically, banks are required to calculate Risk-Weighted Assets (RWA), a measure of the total assets of a bank adjusted for the risk of these assets. Loans to small businesses in this SBA National Data belong to one asset class of the bank's portfolio since they generate interest income for the bank. However, small businesses have a higher risk of default than large companies and higher PD on these loans to small businesses are expected. The portion of the loan not guaranteed by SBA is part of LGD for the bank. Higher PD corresponds to higher risk

of this asset class, and hence higher RWA. To determine the minimum amount of capital needed to be set aside to cover potential losses, i.e., the capital requirement, banks calculate RWA using the estimated PD, LGD, and EAD from their internal models. To ensure the banks' internal models do not underestimate PD leading to insufficient capital reserve under stressful conditions, the Basel framework imposes a minimum threshold for PD estimates used by banks in their internal models. This minimum threshold is called the input floor to PD, which was proposed to be raised from three basis points (0.03%) to five basis points (0.05%) in 2023 as part of the implementation of the Capital Requirements Regulation (CRR3) amendment, part of the Basel III Endgame. Deutsche Bundesbank provided an argument for raising this input floor:

"*the smaller a parameter value – the PD of a low- default portfolio, for example – the greater the number of observations needed to validate that parameter value to a statistically significant degree. However, since observations are often scarce in practice, it is not possible to sufficiently validate small parameter values, hence the risk of underestimating the risks involved. One impact of the use of input floors, though, is that the resulting increase in parameter values will primarily affect risks previously deemed to be minor. As a result, there is the danger that institutions will tend to take on greater risks which promise superior returns but will lead to similar capital requirements on account of the input floor.*"

However, many U.S. Senators have opposed these proposed changes in the Basel III Endgame, believing they would significantly increase capital requirements for U.S. banks, which could harm the U.S. economy. They also believe the changes are unnecessary because U.S. banks are already well-capitalized. The outcome of the 2024 United States Presidential election on November 5, 2024 will likely determine whether the U.S. will implement the proposals in the Basel III Endgame (see "Bank capital levels, deregulation plans at stake in election outcome").

In this competition, you are asked to develop these internal models for the banks to predict PD by applying several classification methods to the SBA National Data. In addition, while the new PD input floor of 0.05% from the Basel III Endgame may be lower than historical default rates from this SBA National Data, an appropriate cut-off probability (threshold) for classification methods needs to be determined to make the business decision as to whether to grant the loan to maximize the profit for the bank. What should it be?

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 (implement all the following methods for comparison) to classify the loan applications as "higher risk" or "lower risk" for loan approval:

- *k*NN
- Classification trees (single tree, bagging, boosting, and random forest)
- Logit model (including Lasso, Ridge, and ElasticNet)
- Neural networks
- Discriminant analysis

The cost of incorrectly classifying a loan application as lower risk outweighs 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 (grant the loan)	5% of DisbursementGross	-5 times 5% of DisbursementGross
Default (deny the loan)	0	0

Incorporate this cost and net profit information in your modeling. Include and discuss all implementation details where appropriate:
- normalize predictors.
- select hyperparameters using cross-validation.
- justify the algorithm/solver used in the optimization problem for these methods, e.g., IRLS (Iterative Reweighted Least Squares), SGD (Stochastic Gradient Descent), SAGA (Stochastic Average Gradient Descent), and ensure the solver you choose converges by adjusting parameters.
- for neural networks, justify the parameters chosen, including the size and number of the hidden layer(s), the activation function, the solver, and the learning rate.

- discuss the appropriate accuracy measures such as the accuracy, sensitivity (recall), specificity, ROC Curve, precision, F1-score, the cost/gain matrix, gains and lift charts incorporating costs and benefits, etc. used for these classification methods.
- select an appropriate cut-off probability (threshold) for each classification method incorporating the cost and net profit information and justify your selection.

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 granting the loan and extending credit?

Instructions

1. You have from 12 am, October 28, 2024, to 11:59 pm, November 30, 2024, to complete this project.

2. You must use open-source tools Python and/or R only.

3. Include an executive summary, a write-up of the procedures and results, Python/R code, its output and explanation into ipynb/Rmd files, and then email these files to **cbaanalytics@csus.edu** by 11:59 pm, November 30, 2024.

4. Each member of the winning teams identified as having produced exceptional work on December 6, 2024, at California State University, Sacramento will receive an iPad Air.