**NAME:** SBA Case

**TYPE:** Census

**SIZE:** 2,102 observations, 35 variables

**ARTICLE TITLE:** “Should This Loan Be Approved or Denied?”: A Large Data Set with Class Assignment Guidelines

**DESCRIPTIVE ABSTRACT: “**SBA Case” is a subset of the file “National SBA.” “SBA Case” is the data set used for the assignment described in the article.

**SOURCE:** United States Small Business Administration

**STORY BEHIND THE DATA:**  The case study in the article focuses on loans pertaining to the *Real Estate and Rental and Leasing* industry in California. The relevant data is extracted from the National SBA file to create this file which has 2,102 observations and 35 variables. This file is utilized in the case study. The SAS code to create this file is shown below:

data SBAnational;

  infile 'C:\temp\SBAnational.csv' DLM=',' FIRSTOBS=2 DSD MISSOVER;

  input

        LoanNr\_ChkDgt    : ?? BEST10.

        Name             : $CHAR30.

        City             : $CHAR30.

        State            : $CHAR2.

        Zip              : ?? BEST5.

        Bank             : $CHAR30.

        BankState        : $CHAR2.

        NAICS            : ?? BEST6.

        ApprovalDate     : ?? DATE9.

        ApprovalFY       : $CHAR5.

        Term             : ?? BEST3.

        NoEmp            : ?? BEST4.

        NewExist         : ?? BEST1.

        CreateJob        : ?? BEST4.

        RetainedJob      : ?? BEST4.

        FranchiseCode    : ?? BEST5.

        UrbanRural       : ?? BEST1.

        RevLineCr        : $CHAR1.

        LowDoc           : $CHAR1.

        ChgOffDate       : ?? DATE9.

        DisbursementDate : ?? DATE9.

        DisbursementGross : ?? DOLLAR15.

        BalanceGross     : ?? DOLLAR12.

        MIS\_Status       : $CHAR6.

        ChgOffPrinGr     : ?? DOLLAR14.

        GrAppv           : ?? DOLLAR14.

        SBA\_Appv         : ?? DOLLAR14. ;

if MIS\_Status='CHGOFF' then y=1; else y=0;  /\* If the loan is charged off, the response y is set to be 1. \*/

    if NewExist='2' then New=1; else New=0;  /\* x1=1 if the business is less than 2 years old; x1=0 if the business is more than 2 years old \*/

    if Term ge 240 then RealEstate=1; else RealEstate=0; /\*x2=1 if loans secured by real estate \*/

run;

data one; set SBAnational;

if DisbursementDate <='31DEC2010'd; /\* Exclude loans disbursed after 2010; see Section 3.3; \*/

  NAICS\_2=INT(NAICS/10000); /\* Extract first two digits of NAICS code \*/

run;

data two; set one;

  if State eq 'CA' AND NAICS\_2 eq 53; /\* The case study only uses "SBA Case" data \*/

  Portion=SBA\_Appv/GrAppv;

  proc freq; tables MIS\_Status;  /\* This frequency table shows 32.64% of the loans were charged off or defaulted. \*/

run;

data ca53;

  set two;

  PROC SURVEYSELECT OUTALL OUT=dataca53 METHOD=SRS

  SAMPSIZE=1051 SEED=18467; \*training and test;

  RUN;

data SBAcase;

  set dataca53;

   Recession=0;

   y1=y;

   daysterm=Term\*30;

   xx=DisbursementDate+daysterm;

  if Selected=0 then y1=.;

     Default=y1;

  if xx  ge '1DEC2007'd AND xx le '30JUN2009'd then Recession=1;

run;

data casedata; set SBAcase (drop=y y1 NAICS\_2);

  proc export data=casedata outfile='C:\temp\SBAcase.csv' DBMS=CSV REPLACE; /\*Create the CA dataset.  The dates are in SAS format. \*/

run;

**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 |
| ChgOffPrinGr | Currency | Charged-off Amount |
| GrAppv | Currency | Gross Amount of Loan Approved by Bank |
| SBA\_Appv | Currency | SBA’s Guaranteed Amount of Approved Loan |
| New | Number | =1 if NewExist=2 (New Business), =0 if NewExist=1 (Existing Business) |
| Portion | Number | Proportion of Gross Amount Guaranteed by SBA |
| RealEstate | Number | =1 if loan is backed by real estate, =0 otherwise |
| Recession | Number | =1 if loan is active during Great Recession, =0 otherwise |
| Selected | Number | =1 if the data are selected as training data to build model for assignment, =0 if the data are selected as testing data to validate model |
| Default | Number | =1 if MIS\_Status=CHGOFF, =0 if MIS\_Status=P I F |
| daysterm | Number | Extra variables generated when creating “Recession” in Section 4.1.6 |
| xx | Number | Extra variables generated when creating “Recession” in Section 4.1.6 |

**PEDAGOGICAL NOTES:**  The authors have used the data set to illustrate how logistic regression can be used to classify a loan application as a “lower risk” (approve) or “higher risk” (deny).

**SUBMITTED BY:**

Name: Min Li, Amy Mickel and Stanley Taylor

Affiliation: California State University, Sacramento

Address: College of Business Administration, CSUS, Sacramento, CA 95819-6088, USA

Email: [limin@csus.edu](mailto:limin@csus.edu) [mickela@csus.edu](mailto:mickela@csus.edu) sataylor@csus.edu