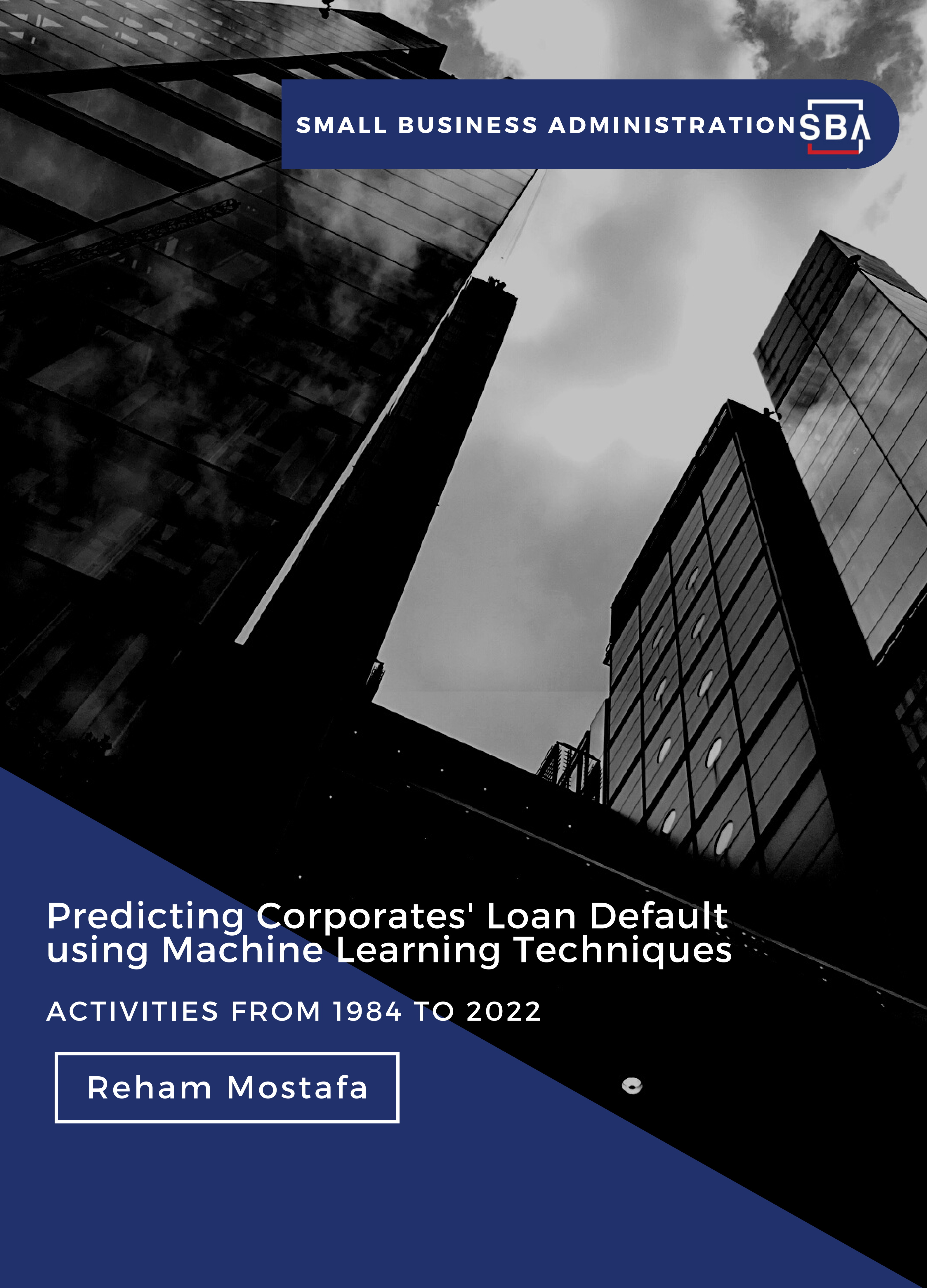


SMALL BUSINESS ADMINISTRATION

Predicting Corporates' Loan Default using Machine Learning Techniques

ACTIVITIES FROM 1984 TO 2022

Reham Mostafa

Introduction

According to Dias' analysis of 201 journal publications about business failure, the most well-defined causes are bankruptcy, business closure, ownership change, and failure to satisfy expectations. Forecasting, particularly before a crisis, may shield the company from its effects, making it possibly more advantageous for the stakeholders than forecasting bankruptcy. One of the most alarming signs that a business is on the verge of collapse is payment default. (Kohv & Lukason, 2021).

Based on Basel II-which is used for risk assessment, in the banking sector, reduced-form regression-based default forecasting techniques have gained popularity. These techniques entail multivariate regression models that forecast a firm's credit quality using information about the firm's economic and financial fundamentals. A credit score is an output. Banks and investors use the probability of default (PD) forecasts to assess debtors before deciding whether or not to lend to them. (Moscatelli et al., 2020).

Not only are large corporates the clients of banks, but also small and medium enterprises (SMEs) are special clients who deal with banks for loans. Because lending to SMEs is riskier than lending to large corporations, techniques and risk management tools have been developed by their characteristics (Kohv & Lukason, 2021), to make wiser credit decisions, to fairly compensate the risk in the expected returns, or to prevent financing unhealthy enterprises. The interest rates are to be shifted in the range between 4.25% - 4.50% (Federal Reserve Board). In the first place, it is essential to evaluate a firm's likelihood of failure. Nowadays, financial ratios from historical accounting data are employed as predictor variables in machine-learning approaches to forecasting loan failure.

METHODOLOGY

Data Source: Small Business Administration (SBA) dataset is from Kaggle. The number of observations is 899,164 and 26 features and one target. The target is determining if the firm will pay in full or default. it is a binary classification problem. The analysis will be conducted after the data-cleaning process.

For implementation, Kaggle Notebook for (Artificial Neural Network), Visual Studio Notebook (Python 3.9.7 for pre-processing, visualization, and building machine learning models). The researchers used Kaggle Notebook, Visual Studio Notebook, Numpy, Pandas, Matplotlib, Seaborn, Date-Time, Plotly, SHAP and Power Bi (for some visualizations).

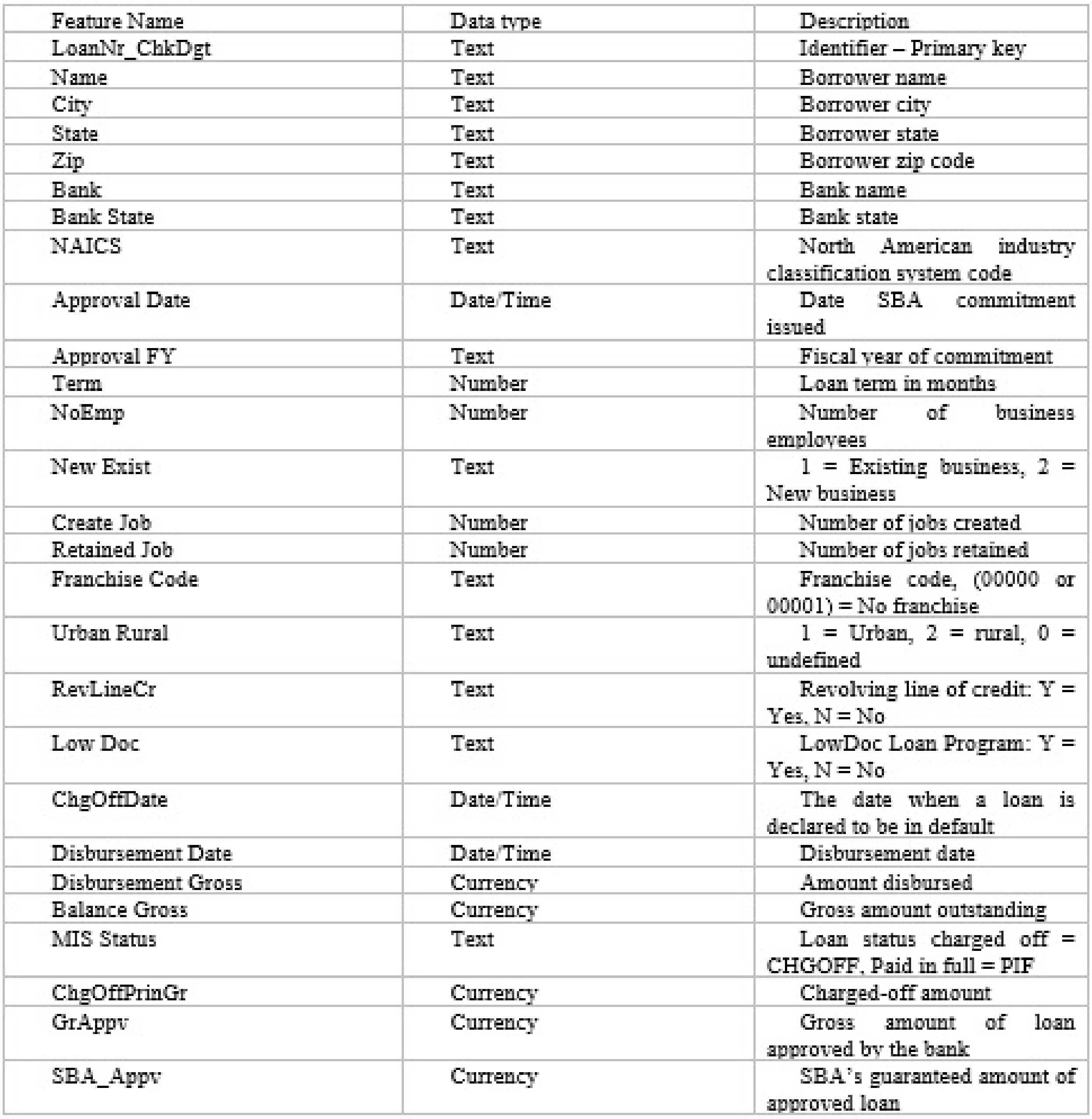
EDA process: plotting data distribution to know if there are outliers, which distribution data follows, and if the target is balanced or not, if there are typos and strange symbols. The data cleaning process contains: removing null values, removing typos, and dropping unnecessary features..

For feature extraction: generating new columns that are important for analysis and prediction such as SBA proportion, duration between approval date and disbursement date, and loans backed by real estate). One Hot Encoding for categorical features and Label Encoding for the target. For an imbalanced target, SMOTE will be used.

For prediction: several supervised ML models will be used to compare prediction MCC, F1 score, and AUC score for evaluating models (Random Forest Tree, Xgboost, AdaBoost, Catboost, Logistic Regression, Neural Network). Overfitting will be solved using hyperparameter tuning.

DATA TABLE

Table 1: Data Table



Feature Name	Data type	Description
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
Bank State	Text	Bank state
NAICS	Text	North American industry classification system code
Approval Date	Date/Time	Date SBA commitment issued
Approval FY	Text	Fiscal year of commitment
Term	Number	Loan term in months
NoEmp	Number	Number of business employees
New Exist	Text	1 = Existing business, 2 = New business
Create Job	Number	Number of jobs created
Retained Job	Number	Number of jobs retained
Franchise Code	Text	Franchise code, (00000 or 00001) = No franchise
Urban Rural	Text	1 = Urban, 2 = rural, 0 = undefined
RevLineCr	Text	Revolving line of credit: Y = Yes, N = No
Low Doc	Text	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	Date/Time	The date when a loan is declared to be in default
Disbursement Date	Date/Time	Disbursement date
Disbursement Gross	Currency	Amount disbursed
Balance Gross	Currency	Gross amount outstanding
MIS Status	Text	Loan status charged off = CHGOFF, Paid in full = PIF
ChgOffPrinGr	Currency	Charged-off amount
GrAppv	Currency	Gross amount of loan approved by the bank
SBA_Appv	Currency	SBA's guaranteed amount of approved loan

Data Pre-processing is the process of transforming raw data into meaningful data, this stage includes Exploratory Data Analysis, Data Cleaning, and Data Transformation such as changing data type, removing null values, drop unvaluable columns for prediction and analysis, and generating new columns.

1) To ensure that all values in columns (Name, City, State, Bank State, Bank) are upper case, upper() function is used.
2) Remove “$” from (disbursement gross, balance gross, Charged-off amount, gross approval, and SBA approval).
3) Change the data type (loan number identifier, Zip, NAICS, New Exist, Franchise code, urban and rural, revolving line of credit, low doc, MIS Status) columns to data type (string).
4) Change the data type (disbursement gross, balance gross, Charged-off amount, gross approval, SBA approval) to datatype (integer).
5) Drop columns (zip, approval fiscal year, charge-off date) because they are not useful features for analysis and prediction (charge-off date has 736.465 null values and approval fiscal year has inconsistent data).
6) Remove strange symbols from columns (Name, City, State, Bank, Bank State).
7) Drop the disbursement gross column because there is a gross approval column which is an alternative, that can do the same task.
8) The researchers dropped 810 observations whose term equals zero.
9) There are 6623 observations (number of employees) (NoEmp) equals zero and the RetainedJob column has 440175 observations equals zero, a new column generated (Total_jobs), which is the summation of NoEmp and RetainedJob, which decreases the number of observations of zero of both columns and filled them by the median.
10) Generating new column Industry based on NAICS column, portion SBA/Gross Approval that indicates the percentage of SBA contribution of loans, and duration between approval date and disbursement date (DisbusementDate – ApprovalDate).

11) Drop null values in the disbursement date column.

12) Generating a new column (Crisis) based on the disbursement date and term. From first December 2007 until last June 2009 was a global economic crisis.

13) Generating Loans Backed by Real Estate where the Term > 240 months.

14) Drop columns(my months, my date months, my date, Approval date, NAICS, SBA Approval, Disbursement Date).

15) Drop null values from the entire data frame.

16) Change data type (Crisis and Loans Backed by Real Estate) to data type string.

17) Generate a new feature (The same state or not) 1 = means the bank state and the firm's state are the same and 0 = is not the same state.

18) Generate new feature (Has Franchise Code?) 1= Yes, it has 0 = No, it has not.

19) Change data type of (The same state or not) and (Has Franchise Code?) columns to data type string.

20) Drop Franchise Code, as a new column generated depending on Franchise Code and Revolving line of credit has inconsistent data.

21) Drop inconsistent values (LowDoc) (0, C, S, A, R, 1 and nan)

22) Drop nan values from MIS Status and New Exist column.

23) Drop 0 value from New Exist as this column as its value should be 1=Existing Business or 2=New Business

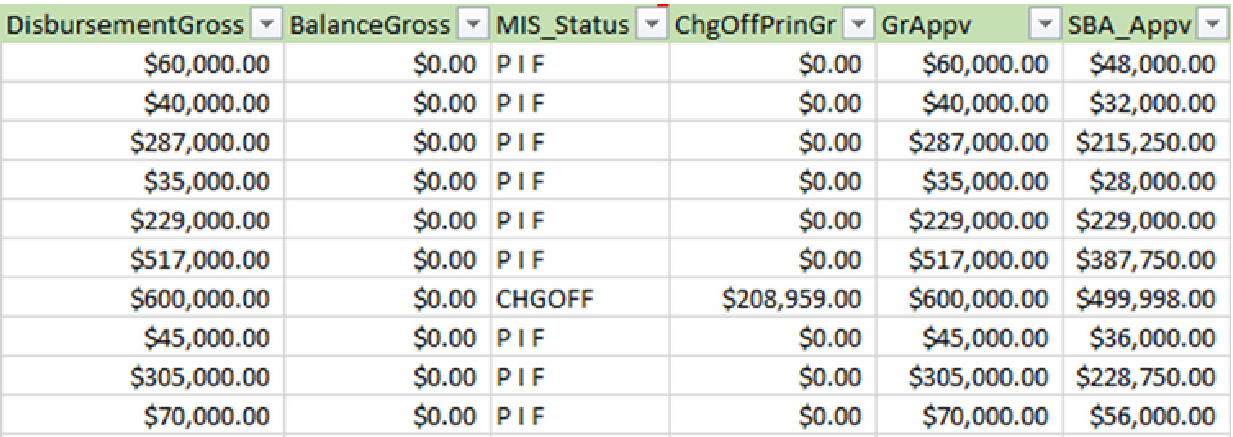
24) The researchers dropped 0 value from urban and rural as its values should be 1=Urban or 2=Rural

25) The researchers dropped columns(CreateJob, RetainedJob, NoEmp, RevLineCr)

26) The researchers dropped columns (Name, loan number identifier, City, Bank, Bank State) as They are not useful for prediction.

27) The researchers dropped the Charge-off amount column because it is the reason for data leakage.

PRE-PROCESSING



DisbursementGross	BalanceGross	MIS_Status	ChgOffPrinGr	GrAppv	SBA_Appv
$60,000.00	$0.00	P I F	$0.00	$60,000.00	$48,000.00
$40,000.00	$0.00	P I F	$0.00	$40,000.00	$32,000.00
$287,000.00	$0.00	P I F	$0.00	$287,000.00	$215,250.00
$35,000.00	$0.00	P I F	$0.00	$35,000.00	$28,000.00
$229,000.00	$0.00	P I F	$0.00	$229,000.00	$229,000.00
$517,000.00	$0.00	P I F	$0.00	$517,000.00	$387,750.00
$600,000.00	$0.00	CHGOFF	$208,959.00	$600,000.00	$499,998.00
$45,000.00	$0.00	P I F	$0.00	$45,000.00	$36,000.00
$305,000.00	$0.00	P I F	$0.00	$305,000.00	$228,750.00
$70,000.00	$0.00	P I F	$0.00	$70,000.00	$56,000.00

Figure 2: Balance Gross, Charged Off amount, Gross Approval, SBA Approval columns

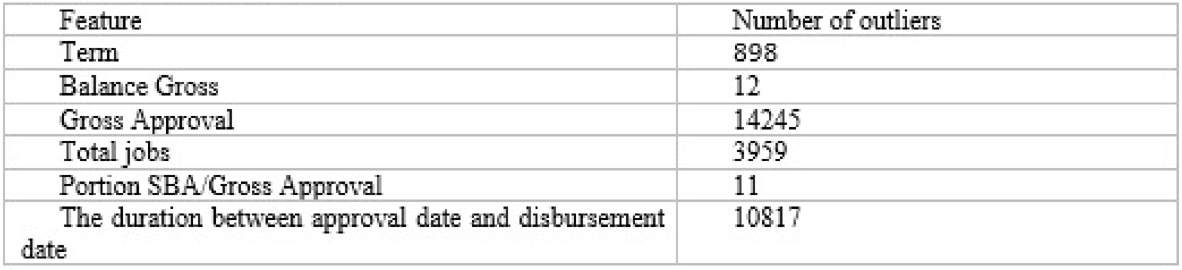
Table 2: List of cleaned features

Feature	Reason
Name, City, State, Bank State, Bank	Upper them
disbursement gross, balance gross, Charged- off amount, gross approval, and SBA approval	$
loan number identifier, Zip, NAICS, New Exist, Franchise code, urban and rural, revolving line of credit, low doc, MIS Status	Change data type from integer to string
Disbursement gross, Balance gross, Charged- off amount, Gross approval, SBA approval	Change data type from string to integer
Name, City, State, Bank State, Bank	Remove Strange Symbols
Crisis	Change data type from integer to string
Loans Backed by Real Estate	Change data type from integer to string
The same state or not	Change data type from integer to string
Has Franchise Code?	Change data type from integer to string

Table 3: Feature Engineering (Extraction) Table

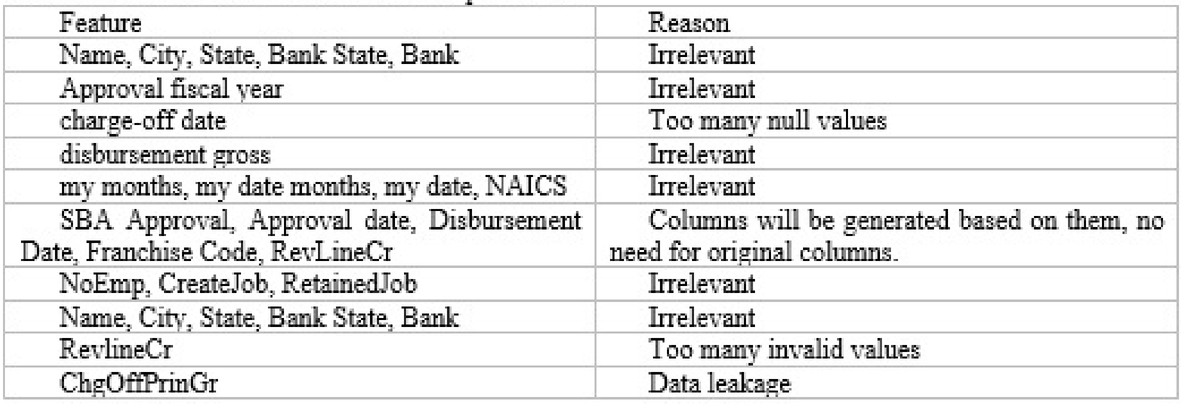
Feature Name	Description
SBA proportion	The percentage of SBA covers the loan.
The duration between approval date and disbursement date	-
Industry	The sector of each firm.
Loans Backed by Real Estate	1=Loans backed by Real Estate, 0=No
Total Jobs	-
Crisis	1=Y, 0=N
The same state or not	1=Y,0=N
Has Franchise code?	1=Y,0=N

Table 4: Z score (threshold=3)



Feature	Number of outliers
Term	898
Balance Gross	12
Gross Approval	14245
Total jobs	3959
Portion SBA/Gross Approval	11
The duration between approval date and disbursement date	10817

Table 5: List of Removed Features before prediction



Feature	Reason
Name, City, State, Bank State, Bank	Irrelevant
Approval fiscal year	Irrelevant
charge-off date	Too many null values
disbursement gross	Irrelevant
my months, my date months, my date, NAICS	Irrelevant
SBA Approval, Approval date, Disbursement Date, Franchise Code, RevLineCr	Columns will be generated based on them, no need for original columns.
NoEmp, CreateJob, RetainedJob	Irrelevant
Name, City, State, Bank State, Bank	Irrelevant
RevlineCr	Too many invalid values
ChgOffPrinGr	Data leakage