Should this Loan be Approved or Denied?

Siddharth Singhal - Data Architect
Anukriti Yadav - Business Analyst
Sebastian Salazar - Project Manager
Jawad Toufaili - Data Analyst
Larbi Farihi - Strategist

https://github.com/McGill-MMA-EnterpriseAnalytics/Loan-Approve-Deny

TABLE OF CONTENTS

CONTEXT

DATASET U.S. Small Business Administration 1982-2013

USE CASE Help SBA make better data-driven decisions when granting loans to small businesses

GOAL Assess risk factors for borrowers and build a model to decide whether an SBA loan should be approved

$13,147,241,525

Lost by **SBA** as of 2005

6.06% Loss Rate

DATA DICTIONARY

of loans : 899,164
Timeframe : 30 years

SME specific : 11 features
Bank specific : 6 features
Loan specific : 11 features

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, N = No
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 charged off = CHGOFF, Paid in full = PIF
ChgOffPrinGr	Currency	Charged-off amount
GrAppv	Currency	Gross amount of loan approved by bank
SBA_Appv	Currency	SBA's guaranteed amount of approved loan

STEPS TAKEN

01

DATA EXPLORATION

02

DATA TRANSFORMATION & PREPROCESSING

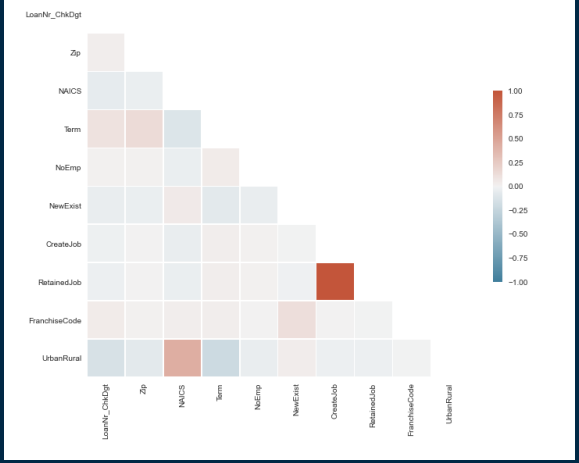
03

MODELING

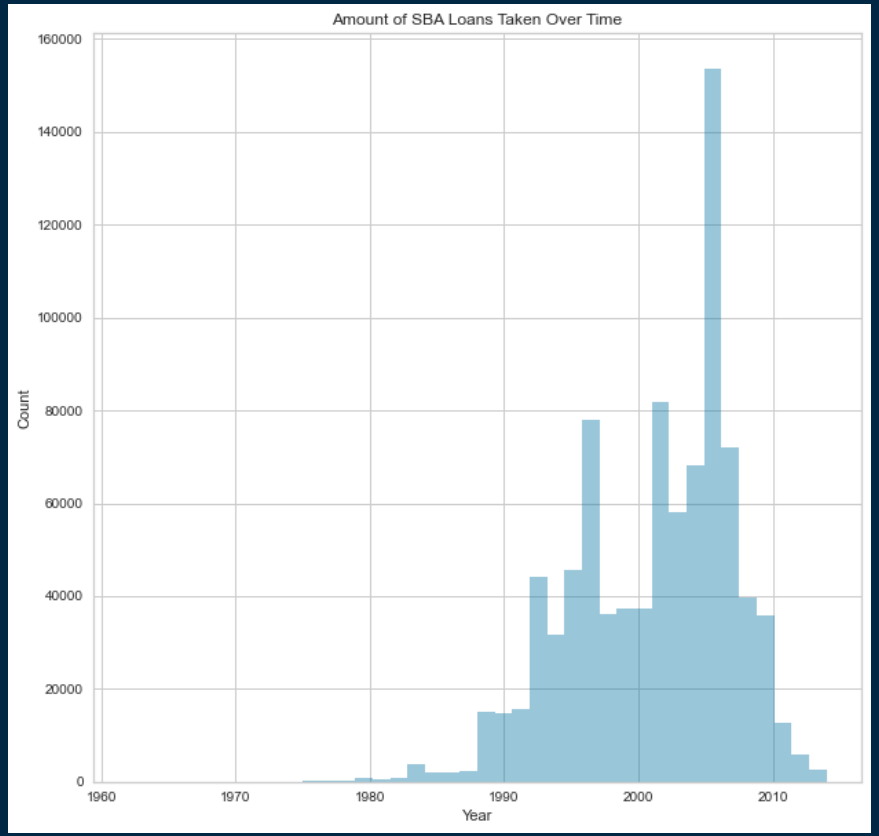
Data Exploration

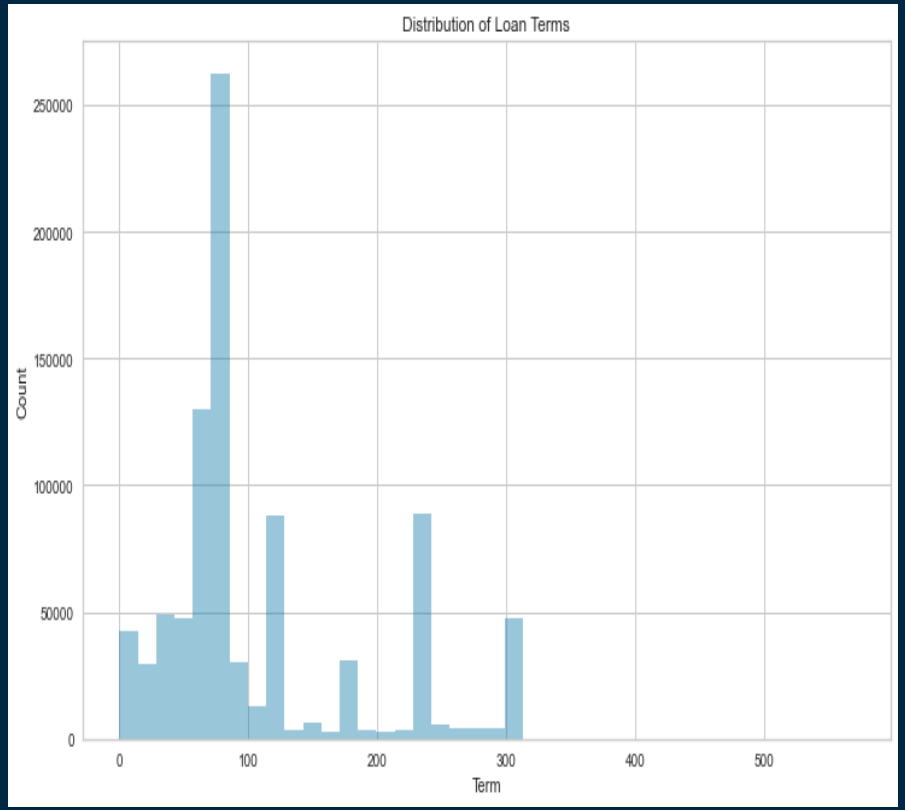
- Provide descriptive statistics
- Visualize the relationship between our variables
- Check for missing values
- Visualize missing values pattern
- Build a correlation matrix to make sure multicollinearity is not present

	Missing Values	% of Total Values
ChgOffDate	736465	81.9
RevLineCr	4528	0.5
LowDoc	2582	0.3
DisbursementDate	2368	0.3
MIS_Status	1997	0.2
BankState	1566	0.2
Bank	1559	0.2
NewExist	136	0.0
City	30	0.0
Name	14	0.0
State	14	0.0



DATA Visualization





Data Processing and Transformation

Perform Feature Selection using RFE-Logistic Regression

Impute missing values using the median value

Specify our target variable and predictors

Remove Outliers using Isolation Forest

Change the datatype of column Currencies to float

Dropped columns that:
Had no effect on the model
Are related to after the loan was signed

NewExist column contained values 1 or 2 so it was converted to binary

Convert NAICS code to relevant industry and then dummify

Converted categorical columns to binary using Ordinal Encoder
RevLineCr
LowDoc
Franchise

MODEL CREATION

Data Split

Split the data into
Train Validation
Test Datasets

Test different Models

Gradient Boosting
Random Forest
AdaBoost
Logistic Regression
ANN

Choose the best model

ROC-AUC Score
Classification Report
Class Prediction Error
Confusion Matrix
PR- Curve
Learning Curve
Discrimintation
Threshold

Hyper-Parameter Tuning

Perform Grid Search

Different Models Used

Baseline Model

Accuracy score: 0.5

Artificial Neural Network

Accuracy score: 0.79

Gradient Boosting

Accuracy score: 0.82

Random Forest

Accuracy Score: 0.79

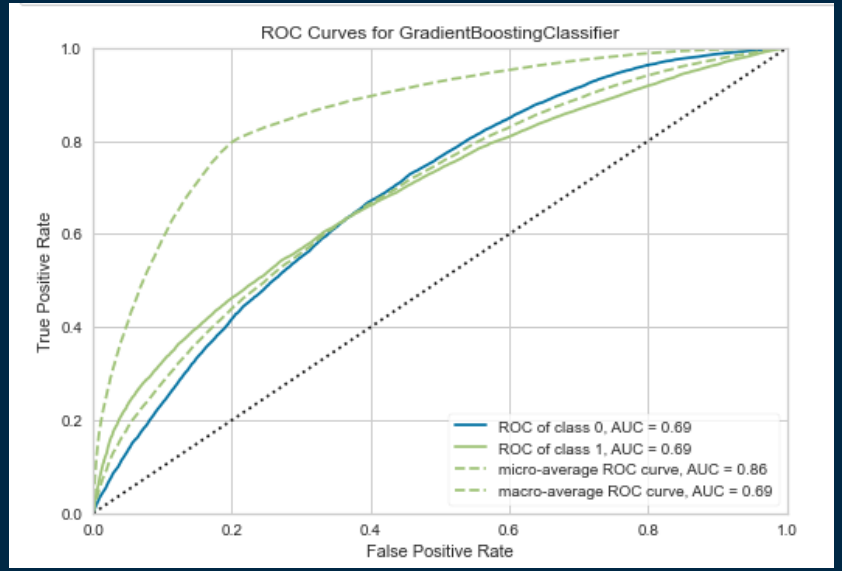
AdaBoost

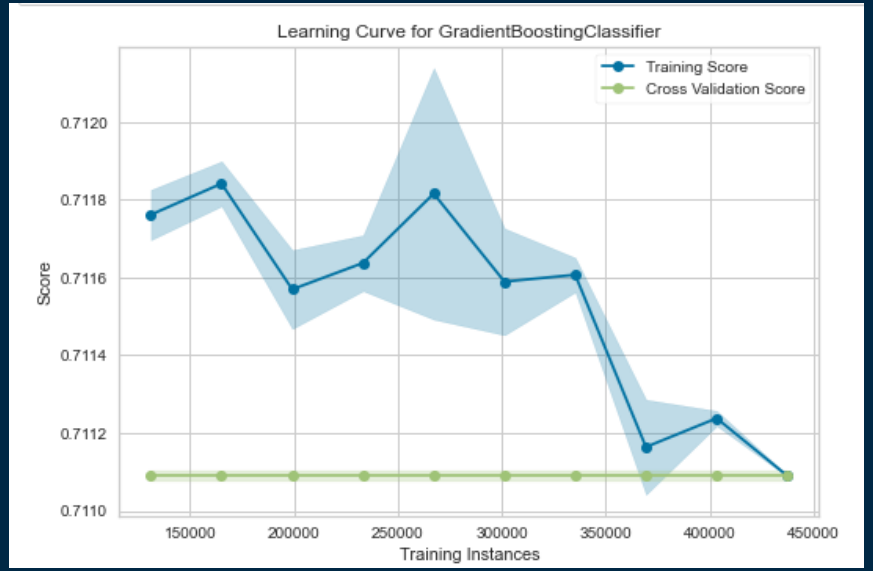
Accuracy Score: 0.79

Logistic Regression

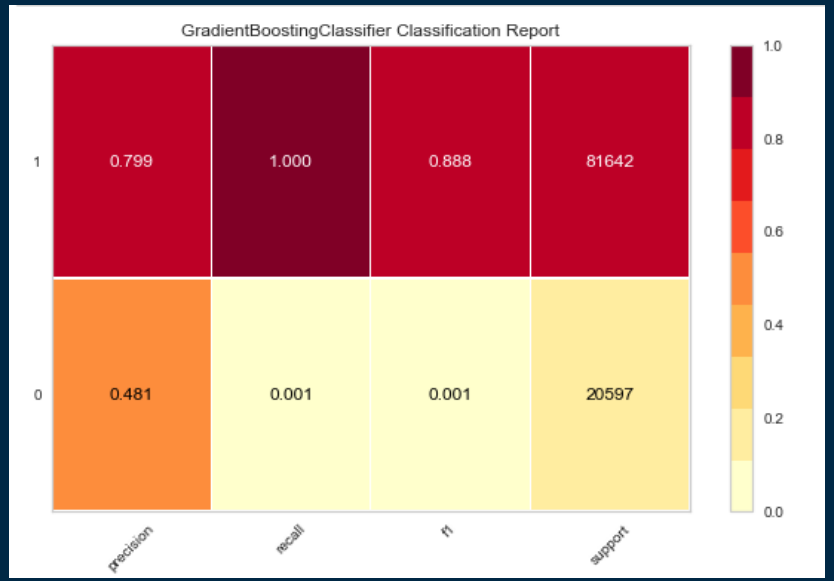
Accuracy Score: 0.8

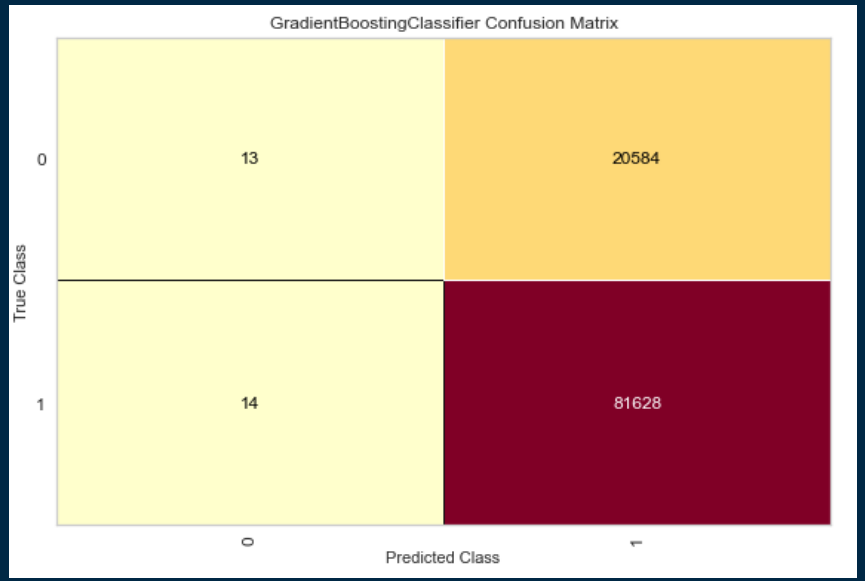
Choosing the best model: Gradient Boosting Results





Selected Model Performance





Imbalanced data problem will be addressed on final model

$12,141,676,859

Defaulted amount by **SBA**

$6,295,684,297

Predicted loss by **SBA** using our solution

Model in improvement

LIMITATIONS AND FUTURE WORK

Analysis does not consider Financial Statements of businesses

Model Robustness (USA specific)

Model does not incorporate the loan purpose

Potential defaulters not in the dataset (loan requests dataset)

Incorporate Financial Data in analysis

Analyse the impact of binning the loan term

Include personality, attitude and drive of business owners

Address existing outliers and skewness in some of the features

THANKS