Data Science II

Stage III

Background

The Small Business Administration (SBA) is important in providing financial, educational, and other resources to support small businesses in the United States. A critical role of SBA is to act as a guarantor for a portion of loans issued to small businesses, thereby reducing risk to lenders and enhancing credit access for entrepreneurs. For most 7(a) loan programs, SBA guarantees up to 85 percent of loans of $150,000 or less and up to 75 percent of loans above $150,000. SBA Express loans, however, carry a 50% guarantee, while Export Express, Export Working Capital Program (EWCP), and International Trade loans benefit from a 90% guarantee (SBA, 2025; Glassman, n.d.).

My interest in this topic is rooted in a broader curiosity about how government policies influence economic development through financial instruments like development banks, grants, and subsidies, particularly in emerging markets. Previously, I conducted qualitative analyses of development banking in China and South Africa, focusing on policy frameworks and institutional structures. Even though I would have liked to explore the inquiry through quantitative methods, I faced challenges in accessing comprehensive data from developing economies. As a result, I turned my focus to the SBA, which has an extensive dataset that offers an opportunity to apply machine learning techniques to explore loan approval dynamics.

This study will predict the loan repayment status, loan default, the amount of loan charged-off principal, the number of jobs created, and the size of the disbursed loan. The findings can yield insights transferable to development finance in emerging markets, especially on the default risk and impact of access to credit on job creation. This research aims to uncover patterns that could inform strategies to enhance credit accessibility, reduce systemic biases, and optimize policy interventions. This study seeks to answer the following research questions:

i. *What is the predictor of a loan’s repayment status?*

ii. *What is the predictor of a loan’s likelihood to be in default?*

iii. *What is the predictor of the size of loan charged-off principal upon default?*

iv. *What is the predictor of the number of jobs created by SBA guaranteed loans?*

v. *What predicts the amount of loan issues to businesses?*

I aim to address these questions by applying machine learning models including logistic regression, random forests, gradient boosting (XGBoost), and SVM. These approaches enhance predictive accuracy and contribute a more nuanced understanding of how data-driven insights can support equitable economic growth through informed policymaking.

Literature Reviews

Li, Mickel, and Taylor provide a framework for deciding loan approval using logistic regression (2018). Chehab and Xiao (2024) use regression analysis to study the relationship between U.S. County social capital and aggregate SBA gross loan approvals, identifying a positive correlation. Their regression analysis highlights other influential factors, including unemployment levels, population, per-capita income, and rural-urban classification.

Additionally, some studies have explored what it takes to get approved for an SBA loan and the behavior of loan recipients. Further, Glassman examines SBA loan approval requirements, offering insights into the criteria influencing loan decisions (n.d.). Glennon & Nigro analyze the repayment behavior of small firms receiving SBA loans using a discrete-time hazard model (2005). Their findings indicate that loan maturity, economic conditions, and firm-specific factors significantly predict default probabilities.

Dataset

This project will use the National SBA dataset, which includes historical data from 1987 through 2014 from the U.S. Small Business Administration. Toktogaraev uploaded the original data to Kaggle in 2020 with 899,164 observations and 27 variables. After preprocessing, the final clean data is from 1994 to 2014, with 572,333 observations and 23 variables. Table 1 below describes all the key variables in the clean dataset.

Table 1: Description of the variables in the dataset.

Variable name	**Data type**	**Description of variable**
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_binary	Boolean	0 = Urban, 1 = Rural
ChgOffDate	Date/Time	The date when a loan is declared to be in default
ChgOffDate_binary	Boolean	0 = No default, 1 = default
DisbursementDate	Date/Time	Disbursement date
DisbursementGross	Numeric	Amount disbursed
MIS_Status_Binary	Text	Loan status charged off = 0, Paid in full = 1
ChgOffPrinGr	Numeric	Charged-off amount

GrAppv	Numeric	The gross amount of loans approved by the bank
SBA_Appv	Numeric	SBA’s guaranteed amount of approved loan

Even though this data is rich and can be instrumental in a machine-learning project, pre-processing will be needed to correct any missing values and ascertain the quality of the data before it is used for training and testing the machine-learning models.

In pre-processing the dataset, the following approach was taken:

1. Calculate the percentage of missing values per column to identify columns that need to be imputed or if the missing values are too low and won’t reduce the data’s richness and complexity.
2. Then drop rows that have missing values on key variables that cannot be imputed. In this case, empty rows in these columns were dropped: Name, City, State, DisbursementDate, and MIS status
3. Convert the charge-off date from a date into a binary: 0 – if there is no date, and 1 if there is a date and store it in the ChgOffDate_binary column.
4. Remove the undefined location in the UrbanRural column. Using the new information, create the UrbanRural_binary column and assign 0 to urban and 1 to rural.
5. Encode MIS_status into binary data where loan status charged off = 0, Paid in full = 1 and assign to a new column MIS_Status_Binary.
6. Drop off columns that are not part of the key variables: 'LoanNr_ChkDgt', 'ChgOffDate', 'UrbanRural', 'RevLineCr', 'LowDoc', ‘MIS_Status’, and 'BalanceGross'

Methodology

The target variable would be Disbursement Gross, Loan Repayment Status, Loan Default, Charge off principal, and job creation.

The approach will be two-fold, the first approach will create a predictive model for binary variables (loan repayment status and loan default). The second approach will create a predictive model for continuous variables (disbursement gross, charge-off principal, and job creation). For binary target variables, logistic regression, random forest, XGBoost, and SVM models are used while for the continuous variables, random forest and XGBoost models would be deployed.

A) Binary variables

i) Loan repayment status

Logistic regression is used as the baseline model to predict loan repayment status with the key independent variables are term, disbursement gross, gross approved, SBA approved, number of employees, new or existing business, jobs created, franchise code, charge off principal, charge off date (status), approval financial year, urban-rural binary, and the disbursement year.

Here is the logistic regression formula for loan repayment status:

$$
\begin{aligned}
&Logit\left(P(MIS\_Status\_Binary = 1)\right) \\
&\qquad = \beta 0 + \beta_1 * Term + \beta_2 * DisbursementGross + \beta_3 * GrAppv + \beta_4 \\
&\qquad * SBA\_Appv + \beta_5 * NoEmp + \beta_6 * NewExist + \beta_7 * CreateJobs + B_8 \\
&\qquad * RetainedJob + \beta_9 * FranchiseCode + \beta_{10} * ChgOffPrinGr + \beta_{11} \\
&\qquad * ChgOffDate\_binary + \beta_{12} * ApprovalFY + \beta_{13} * DisbursementYear \\
&\qquad + \beta_{14} * UrbanRural\_binary
\end{aligned}
$$

Using the same independent features, Random Forest, XGBoost, and SVM are used to predict loan repayment status.

ii) Loan Default prediction

For loan default, charge off date (binary) is used as proxy that the loan was defaulted and had to be charged off, 1 = default, 0= otherwise. The independent variables in this model are number of employees, new or existing business status, disbursement gross, gross approved, SBA approved, and urban or rural status.

Here is the logistic regression formula for loan default:

$$
\begin{aligned}
&Logit\left(P(ChgOffDate\_binary = 1)\right) \\
&\qquad = \beta 0 + \beta_1 * UrbanRural\_binary + \beta_2 * DisbursementGross + \beta_3 \\
&\qquad * GrAppv + \beta_4 * SBA\_Appv + \beta_5 * NoEmp + \beta_6 * NewExist
\end{aligned}
$$

Random Forest, XGBoost, and SVM are used to further predict loan default using similar independent variables.

B) Continuous variable

i) Predicting the amount of charged-off principle

With gross charge-off principal (ChgOffPrinGR) as the target variable and number of employees, new or existing business status, disbursement gross, gross approved, SBA approved, and urban rural status as the independent variables. Given that this is a continuous variable, Random Forest, XGBoost, and SVM models are used in prediction.

ii) Impact on job creation prediction

The number of jobs created (CreateJob) will be the target variable with disbursement gross, term, new or existing business status, and urban or rural status as the independent variables. For prediction, Random Forest, XGBoost, and SVM models are deployed.

i) Disbursement Gross

To predict the gross amount disbursed to businesses, the loan term, number of employees, new or existing business status, franchise code, and rural or urban status will be used as independent variables. Similar to the other two continuous variables, Random Forest, XGBoost, and SVM will be used.

Using XGBoost, Random Forest, and SVM models will help identify factors (feature importance) that drive loan sizes, charge off principal, and impact on job creation. It is important to identify the factors associated with large loan amounts, large number of jobs created, and high loan default which could inform policy decisions and consideration.

A potential issue with Random Forest and SVM is that, while robust, they can be slow on large datasets – the processed dataset has more than 500,000 observations. XGBoost is more efficient but requires careful hyperparameter tuning to avoid overfitting. Additionally, redundant features may impact model performance.

Even though these models might be more challenging to interpret, SHAP (Shapley Additive Explanations) values can help explain how each feature influences an individual prediction in both models.

Bibliography

Li, M., Mickel, A., & Taylor, S. (2018) “Should This Loan be Approved or Denied?”: A Large Dataset with Class Assignment Guidelines, Journal of Statistics Education, 26:1, 55-66, DOI: 10.1080/10691898.2018.1434342

Toktogaraev, M. (2020). Should This Loan be Approved or Denied?” A dataset from the U.S. Small Business Administration (SBA). https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied?

Glennon, D., Nigro, P. An Analysis of SBA Loan Defaults by Maturity Structure. *J Finan Serv Res* 28, 77–111 (2005). https://doi.org/10.1007/s10693-005-4357-3

Chehab, A. & Xiao, Y. (2024). How does Social Capital Impact SBA Loan Approvals in US Counties? Editura ASE. https://www.ceeol.com/search/article-detail?id=1280617

Glassman, G. (n.d.). What does it take to get my loan approved? Burzenski & Company, P.C. East Haven, CT. https://www.cabidigitallibrary.org/doi/pdf/10.5555/20093018844

Glassman’s paper looks at the requirements to get a loan approved by SBA (n.d.). It also highlights that SBA’s guaranteed coverage is 85% for loans up to $150,000, and 75% for loans above $150,000 but less than $2 million.

SBA (2025). Terms, Conditions, and eligibility. https://www.sba.gov/partners/lenders/7a-loan-program/terms-conditions-eligibility