Data Science II

Stage I

Background

The Small Business Administration (SBA) plays an important role in providing financial, education, 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.

Understanding the determinants of SBA loan approvals can yield insights transferable to development finance in emerging markets. By examining factors beyond traditional credit risk assessments and SBA policy terms, 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 research question: *What are the determinants of loan approval beyond standard credit risk factors and SBA terms and conditions?*

I aim to identify key predictive variables influencing loan decisions by applying machine learning models such as random forests and gradient boosting (XGBoost). This approach enhances predictive accuracy and contributes to a more nuanced understanding of how data-driven insights can support equitable economic growth through informed policymaking.

Literature Reviews

Li, Mickel, & Taylor provide a framework for deciding loan approval using logistic regression (2018). Chehab and Xiao (2024) uses 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 also 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 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. It was uploaded to Kaggle by Toktogaraev in 2020. The dataset has 899,164 observations and 27 variables, all of which are described in Table 1 below.

Table 1: Description of the variables in the dataset.

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

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.

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