Data Science II

Stage I

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 identify determinants of **the loan amount (DisbursementGross), the gross amount approved by the bank (GrAppv), and SBA_guarantee (SBA_Appv). The findings 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 following research questions:

i. *What factors influence the amount of loans received by businesses?*

ii. *What factors influence the amount of loans approved by banks?*

iii. *What factors influence the amount of loan guaranteed by the SBA?*

I aim to identify key predictive variables influencing loan decisions by applying machine learning models such as linear regression, random forests, and gradient boosting (XGBoost). 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, the Gross Amount approved by the bank, and SBA's guaranteed amount of approved loan. The key variables are state, city, NAICS, approval date, approval financial year, loan term, number of employees, created jobs, retained job, franchise code, UrbanRural binary, charge-off date binary, disbursement date, and MIS status,

Models

1. **Predict loan amount (DisbursementGross), the gross amount approved by the bank (GrAppv), and SBA_guarantee (SBA_Appv).**
i.   **Linear regression**

Set up a linear regression model as follows:

$$Loan\ Amount\ (DisbursementGross) = \beta 0 + \beta 1 * Term + \beta 2 * NoEmp + \beta 3 * NewExist + \beta 4 * FranchiseCode + \beta 5 * UrbanRural\_binary$$

The linear regression will help identify if there is a linear relationship between the loan amount and key variables in the dataset. The downside of using a linear relationship is that this regression model might not capture a non-linear relationship.

A similar model will be run but with the Gross amount approved by the bank and SBA's guaranteed amount as the target variables:

$$Gross\ amount\ approved\ by\ bank\ (\text{GrAppv})$$
$$= \beta0 + \beta1 * Term + \beta2 * NoEmp + \beta3 * NewExist + \beta4$$
$$* FranchiseCode + \beta5 * UrbanRural\_binary$$

$$SBA\ guarantee\ (\text{SBA\_Appv})$$
$$= \beta0 + \beta1 * Term + \beta2 * NoEmp + \beta3 * NewExist + \beta4$$
$$* FranchiseCode + \beta5 * UrbanRural\_binary$$

ii.   Gradient boosting and random forest

XGBoost and Random Forest can be used to further understand what variables affect the three target features. In both models, the feature variables will be Term, NoEmp, NewExist, FranchiseCode, and UrbanRural_binary, and the target variable will be loan amount (DisbursementGross), gross amount approved by bank (GrAppv), and SBA_guarantee (SBA_Appv).

Using these models will help identify factors (feature importance) that drive loan sizes, or the characteristics of the loan approved. It is important to identify the factors associated with large loan amounts or more favorable loan conditions (loan terms).

A potential issue with Random Forest is that, while robust, it can be slow on large datasets. 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 than a linear regression model, SHAP (Shapley Additive Explanations) values can help explain how each feature influences an individual prediction in both models.

Further quantitative analysis can be done to:

a) Identify loan default prediction based on the feature variables such as the business size, loan terms, and industry.
b) Time series analysis can be conducted to see the loan distribution and change over time by industry or geographical location. This analysis can provide insight into the loan patterns and whether specific economic factors or seasonal trends affect loan approval or size.

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