**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. Toktogaraev uploaded the original data to Kaggle in 2020 with 899,164 observations and 27 variables. After preprocessing, the final clean data has 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 | Gross amount of loan approved by 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 following approach is used:

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 charge off date from a date into a binary: 0 – if there is no date, and 1 if there is a date, and store in ChgOffDate\_binary column.
4. Remove undefined location in UrbanRural column, and using the new information, create UrbanRural\_binary column, assign 0 to urban and 1 to rural.
5. Drop off columns that are not part of the key variables: 'LoanNr\_ChkDgt', 'ChgOffDate','UrbanRural', 'RevLineCr', 'LowDoc', 'BalanceGross'

**Methodology**

The target variable would be Disbursement Gross, Gross Amount approved by 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, urban rural binary, charge off date binary, disbursement date, and MIS status,

**Models**

1. **Predict loan amount (DisbursementGross) using:**
2. **Linear regression**

Setup a linear regression model as shown below:

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 there might be a non-linear relationship that will not be captured by this regression model.

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

1. **Gradient boosting and random forest**

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

Random forest can be robust and slow because the dataset is very large while XGBoost can be more efficient but require more hyperparameter tuning.

**Further analysis can be done to:**

1. Identify loan default prediction based on other features such as the business size, loan terms, and industry.
2. Time series analysis to see the loan distribution and change over time by industry or geographical location. This analysis can provide an insight into the loan patterns and whether specific economic factors or seasonal trends affect loan approval or size.

**Data**

**Important features**

**City, state, zip, bank, bankstate, NAICS, approval Date, ApprovalFY, Term, Number of Employeees, New**

**Methodology**

1. **Key variables**

**Location: State, Rural vs Urban**

**Business type: NAIC code**

**Trends in loans: disbursement date**

**Loan amount: disbursement gross**

**Franchise status**

**Employment: job created and retained jobs**

**Status of loan: MIS**

1. **Methods and modeling techniques**
2. **Justification for approach**
3. **Anticipated issues based on the approach**

**Model evaluation and validation e.g. cross-validation or performance** easures such as accuracy, precision, recall, ROC/AUC);

Supervised (decision trees and SVM) vs unsupervised learning (clustering)

**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>