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 the 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 the 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 optimal amount of loan to approve for a business?*

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

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

iv. *What is the predictor of a loan's repayment status?*

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

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 to 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 48 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 the approved loan
IsFranchise	Boolean	1 = Franchise, 0 = otherwise
LowDoc_binary	Boolean	1 = low doc program participant, 0 = otherwise
RevLineCr_binary	Boolean	1 = has revolving line of credit, 0 = otherwise
Sector_Accommodation & Food	Boolean	1 = Business in Accommodation & Food Services
Sector_Admin & Waste Mgmt	Boolean	1 = Business in Administrative & Waste Management
Sector_Agriculture	Boolean	1 = Business in Agriculture
Sector_Arts & Recreation	Boolean	1 = Business in Arts, Entertainment, and Recreation
Sector_Construction	Boolean	1 = Business in Construction
Sector_Education	Boolean	1 = Business in Education
Sector_Finance	Boolean	1 = Business in Finance and Insurance
Sector_Healthcare	Boolean	1 = Business in Healthcare and Social Assistance
Sector_Information	Boolean	1 = Business in Information sector
Sector_Management	Boolean	1 = Business in Management of Companies and Enterprises
Sector_Manufacturing	Boolean	1 = Business in Manufacturing
Sector_Mining	Boolean	1 = Business in Mining, Quarrying, and Oil and Gas Extraction
Sector_Other Services	Boolean	1 = Business in Other Services (except Public Admin)
Sector_Professional Services	Boolean	1 = Business in Professional, Scientific, and Technical Services
Sector_Public Admin	Boolean	1 = Business in Public Administration
Sector_Real Estate	Boolean	1 = Business in Real Estate and Rental and Leasing
Sector_Retail Trade	Boolean	1 = Business in Retail Trade
Sector_Transportation	Boolean	1 = Business in Transportation and Warehousing
Sector_Unknown	Boolean	1 = Sector not identified
Sector_Utilities	Boolean	1 = Business in Utilities
Sector_Wholesale Trade	Boolean	1 = Business in Wholesale Trade

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'
7. Binarize Low_doc, RevLineCr, and IsFranchise where 1 is true and 0 otherwise
8. Hot encode sector according to the North American Industry Classification System (NAICS) code.

Methodology

The target variables are Gross Approved loan, Charge Off Principal, Job Creation, Loan Repayment Status, and Loan Default. The approach will be two-fold. The first approach will create predictive models for continuous variables (*disbursement gross*, *charge-off principal*, and *job creation*). The second approach will develop predictive models for binary variables (*loan repayment status* and *loan default*). For continuous target variables, Random Forest and XGBoost models will be deployed, with the possibility of using Linear SVM based on the performance of these two models. As for binary target variables, logistic regression, Random Forest, XGBoost, and SVM models are used.

A) Continuous variables

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, a large number of jobs created, and high loan default, which could inform policy decisions and considerations. This is how they are set up:

i) Predicting the gross approved loan

To identify the optimal amount of loan that ought to be approved, the gross approved loan (GrAppv) is used as the target variable and the following as independent variables: loan term, number of employees, new or existing business status, jobs created, jobs retained, gross disbursements, charge off principal, SBA approved loan amount, binary of the charge off date, urban or rural status, MIS status, franchise status, low doc status, revolving line of credit status and all the sector dummies.

ii) Predicting the amount of charged-off principal

To predict the amount of loans charged off the principial, the gross charge-off principal (ChgOffPrinGR) is used as the target variable and loan term, number of employees, new or existing business status, jobs created, jobs retained, gross disbursements, gross approved loan, SBA approved loan amount, binary of the charge off date, urban or rural status, MIS status, franchise status, low doc status, revolving line of credit status and all the sector dummies as the independent variables.

iii) Impact on job creation prediction

To predict the impact of SBA loans on job creation, the number of jobs created (CreateJob) will be the target variable with loan term, number of employees, new or existing business status, charge off principal, jobs retained, gross disbursements, gross approved loan, SBA approved loan amount, binary of the charge off date, urban or rural status, MIS status, franchise status, low doc status, revolving line of credit status and all the sector dummies as the independent variables.

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.

B) **Binary variable**

i) Loan repayment status

Logistic regression is used as the baseline model to predict loan repayment status (MIS_Status_Binary) with the key independent variables are loan term, number of employees, new or existing business status, charge off principal, jobs retained, gross disbursements, gross approved loan, SBA approved loan amount, binary of the charge off date, urban or rural status, jobs created, franchise status, low doc status, revolving line of credit status and all the sector dummies as the independent variables.

Here is the logistic regression formula for loan repayment status:

$$Logit\left(P(MIS\_Status\_Binary = 1)\right) = \beta 0 + \beta_1 * Term + \cdots + \beta_{47} * \text{Sector\_Wholesale Trade}$$

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

ii) Loan Default prediction

For loan default, the charge-off date (binary) is used as a proxy that the loan was defaulted and had to be charged off, 1 = default, 0= otherwise. The independent variables are loan term, number of employees, new or existing business status, charge off principal, jobs retained, gross disbursements, gross approved loan, SBA approved loan amount, MIS status, urban or rural status,

jobs created, franchise status, low doc status, revolving line of credit status and all the sector dummies as the independent variables.

Here is the logistic regression formula for loan default:

$$Logit\left(P(ChgOffDate\_binary = 1)\right) = \beta 0 + \beta_1 * Term + \cdots \beta_{47} * Sector\_Wholesale\_Trade$$

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

Results

a) Gross Approved Loan Prediction

Four regression models, linear regression, random forest, XGBoost, and support vector regression (SVR), were evaluated to predict the gross approved loan amount (GrAppv). Model performance was assessed using Mean Squared Error (MSE) and the coefficient of determination (R^2) across training, validation, and test datasets.

The Linear Regression model exhibited strong predictive power, with a test R^2 of 0.9785, indicating that the model captured a substantial proportion of variance in the target variable. The Random Forest model outperformed all others, achieving a test R^2 of 0.9972 and the lowest MSE, highlighting its superior capacity to model complex, non-linear relationships. XGBoost demonstrated competitive performance (test $R^2 = 0.9791$), closely aligning with Linear Regression but slightly underperforming relative to Random Forest. In contrast, the SVR model yielded a test R^2 of 0.7695, indicating limited effectiveness despite the application of dimensionality reduction via principal component analysis. These findings suggest that ensemble-based methods, particularly Random Forest, provide the most accurate and robust predictions for gross approved loan amounts.

b) Charged-Off Principal Prediction

The same four models were employed to predict the amount of principal charged off: Linear Regression, Support Vector Machine (SVM), XGBoost, and Random Forest. The Linear Regression baseline performed well across all datasets, achieving a test MSE of 0.0090 and R^2 of 0.9535. The SVM model demonstrated slightly stronger generalization, with a test MSE of 0.0087 and R^2 of 0.9549, marginally outperforming Linear Regression on the test set.

XGBoost, implemented as a regressor, exhibited strong in-sample performance (MSE = 238,522,649.42; $R^2 = 0.9579$), but lower generalization on test data ($R^2 = 0.8125$), indicating the presence of some overfitting and comparatively higher error magnitudes. In contrast, the Random Forest Regressor achieved exceptional predictive performance across all datasets, with near-perfect test results (Test RMSE = 0.0000; $R^2 = 1.0000$), suggesting a highly effective fit to the data.

However, the unusually perfect test performance warrants further scrutiny for potential data leakage or overfitting, despite the model's consistency across training, validation, and test sets.

Overall, Random Forest emerged as the top-performing model in terms of raw predictive accuracy, while both Linear Regression and SVM offered strong and stable generalization. XGBoost, though powerful in training, demonstrated slightly reduced test accuracy and greater prediction error in this context.

c) Impact of Loans on Job Creation

Random Forest, XGBoost, and Linear Regression were utilized to assess the impact of loans on job creation. The Random Forest model showed relatively high explanatory power on training data ($R^2 = 0.7754$); however, its performance deteriorated sharply on validation ($R^2 = -0.0636$) and test sets ($R^2 = -0.4830$), indicating severe overfitting. XGBoost demonstrated marginally better generalization (validation $R^2 = 0.0866$; test $R^2 = -0.0025$) but still failed to produce meaningful predictions. Linear Regression also performed poorly, with a test R^2 of 0.0535. Collectively, the models failed to achieve robust out-of-sample accuracy, suggesting that the current features lack predictive power for job creation, and that additional data or alternative modeling strategies may be necessary.

d) Loan Repayment Status Classification

Four models were implemented to classify loan repayment status: Logistic Regression, Random Forest, XGBoost, and SVM. The Logistic Regression model delivered strong and consistent performance across all datasets, achieving a test accuracy of 0.9913, with MSE = 0.0087 and $R^2 = 0.9535$.

The Random Forest model attained excellent training performance (MSE = 0.0017; $R^2 = 0.9905$) and maintained equivalent test metrics to the baseline, suggesting strong generalization. XGBoost achieved near-perfect training accuracy ($R^2 = 0.9986$; accuracy = 0.9998) but showed minor overfitting, with a test R^2 of 0.9442 and accuracy of 0.9895. The SVM model, despite feature standardization, failed to improve upon Logistic Regression, producing identical test results. All models demonstrated high classification accuracy (>98.9%), with XGBoost offering slightly better validation performance but marginally lower generalization on the test set compared to Logistic Regression and Random Forest.

e) Loan Default Prediction

Three classification models—logistic regression, random forest, and XGBoost—were assessed for predicting loan defaults. The Logistic Regression model achieved strong and stable performance across all datasets, with accuracy scores of 0.9925 (training), 0.9828 (validation), and 0.9913 (test).

The Random Forest model achieved perfect training results (accuracy, precision, recall, and F1 score all = 1.0000), with slight degradation on the test set (accuracy = 0.9913; F1 = 0.9831), reflecting minimal overfitting. The test confusion matrix indicated only one false negative. XGBoost demonstrated flawless performance on all datasets, with perfect scores across all

classification metrics and no test misclassifications, as confirmed by the confusion matrix. These results affirm that all models performed exceptionally in predicting loan defaults, with XGBoost offering optimal performance and perfect generalization to unseen data.

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

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