Josahn Oginga

Machine Learning Project Using SBA Loan Data

May 5th, 2025

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 are the predictors of the number of jobs created by SBA guaranteed loans?*

iv. *What are the predictors 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 for 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 each 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, linear regression is used as a base model. Random Forest, XGBoost, and linear SVM are used.

For the binary target variables, logistic regression is used as the base model. Then Random Forest, XGBoost, and SVM classification models are deployed. In both continuous and binary variables, SVM was not trained on the whole dataset. This is because SVM scale poorly with large datasets, therefore, a subsample of 2 percent of the data is used which reduces the computational cost and training time while still capturing representative patterns.

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 what influences 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 572,000 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) is used to explain how each feature influences an individual prediction in the best performing model in each questionc.

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.

After testing for correlation between ChgOffDate_binary and ChgOffPrinGr and MIS_Status and finding a high correlation, I dropped them from the features list.

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.

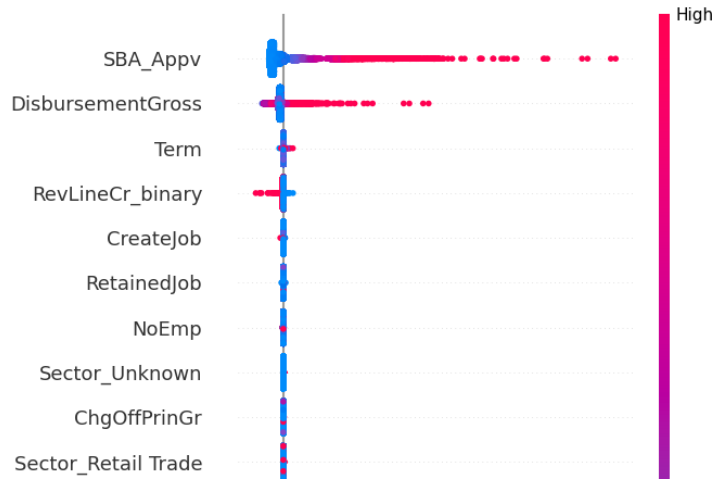
As shown in Table 2 below, the Linear Regression model exhibited strong predictive power, with a test R^2 of 0.9785, indicating that the model captured a substantial proportion of variance in the target variable. The Random Forest model outperformed all others, achieving a test R^2 of 0.9972 and the 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.

Table 2: Model performance comparison for GrAppv

	Model	Train MSE	Train R^2	Val MSE	Val R^2	Test MSE	Test R^2
0	Linear Regression	1.989391e+09	0.9783	1.794591e+09	0.9802	1.981585e+09	0.9785
1	Random Forest	1.294227e+08	0.9986	2.468923e+08	0.9973	2.594256e+08	0.9972
2	XGBoost	1.075210e+09	0.9883	1.834541e+09	0.9798	1.923854e+09	0.9791
3	Linear SVM (10%)	2.102963e+10	0.7705	2.093001e+10	0.7696	2.120981e+10	0.7695

In terms of feature importance using SHAP (SHapley Additive exPlanations) analysis, I chose to use the model with the best performance which in this case was Random Forest. The top ten most important features in predicting the size of approved loan is shown in Figure 1 below:

Figure 1: Feature importance using Random Forest model results



From the SHAP plot, the leading features influencing the gross approved loans is the size of SBA loan approved, the amount disbursed, loan term, and revolving line of credit. The first two features can be argued as having a correlation with the target variable as the size of SBA loan approved is directly correlated with the final amount that is approved based on SBA guideline (85% for loans up to USD 150,000 and 75% for loans above USD 150,000). Also, the amount disbursed is directly dependent on how much is approved (the disbursed amount is the approved amount minus any fees).

Loan term has a positive impact on the amount of loan approved while revolving line of credit has a high negative impact on predicting gross approved loan. In terms of industry sectors, the Unknown and Retail Trade sector are drivers of the size of approved loans.

b) Charged-Off Principal Prediction

Similar to gross approved loan prediction, the same four models were employed to predict the amount of principal charged off. The Linear Regression baseline did not do well in predicting the charge off principal. It only explained 26.39% of the data. SVM has the poorest performance with

a negative r-squared. The poor performance of SVM on this data might be because I used LinearSVC which is computationally less expensive, but the actual relationship between charge-off principal and the independent features is non-linear.

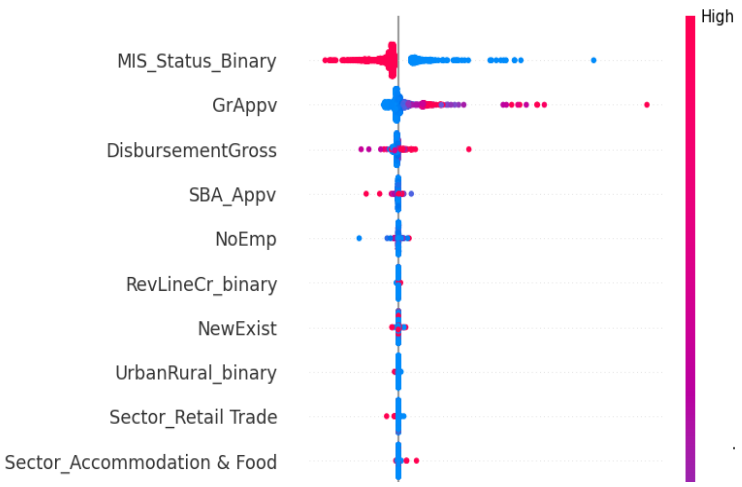
XGBoost, implemented as a regressor, exhibited strong in-sample performance and a good generalization of 81.25%. Out of all the four models, Random Forest Regressor achieved the best predictive performance with just 0.24% better than XGBoost. Table 3 below show summary results of each model across training, validation, and test datasets.

Table 3: Model performance comparison for ChgOffPrinGr

	Model	Train MSE	Train R^2	Val MSE	Val R^2	Test MSE	Test R^2
0	Linear Regression	4.222555e+09	0.2544	3.914277e+09	0.2850	4.280890e+09	0.2639
1	Random Forest	1.312258e+04	0.9696	3.566790e+04	0.7676	3.281153e+04	0.8149
2	XGBoost	2.385226e+08	0.9579	1.185832e+09	0.7834	1.090246e+09	0.8125
3	Linear SVM (10%)	1.088940e+10	-0.7378	1.195064e+10	-1.0756	1.189094e+10	-1.1637

SHAP analysis was then conducted using random forest results. Figure 2 below shows the top ten most important features in predicting charge-off principal.

Figure 2: Feature importance using Random Forest model results



There is a clear high impact of the charge-off amount and the status of the loans (charged off or not). Beyond that, the size of approved loan have the highest impact on charge off amount. In terms of sectors of the applicant, retail trade and construction sectors have a negative effect on charge-off amount (less likely to be charged off) versus accommodation & food sector that had a positive effect on charge off amount (more likely to be charged off). As a lender, this can provide sector specific information to approve or decline a loan based on sectoral likelihood of being charged off (risk of default).

c) Impact of Loans on Job Creation

As in the other continuous variable questions, the four models were utilized to assess the impact of loans on job creation. As shown in Table 4 below, all models show poor generalization to the test set, with low or negative R^2 values. Random Forest and XGBoost perform a little better on training data but overfit, as indicated by a sharp drop in validation and test R^2. Linear Regression underfits across all sets, while the Linear SVM performs poorly overall with extremely high errors and negligible explanatory power.

Table 4: Model performance comparison for CreateJob

	Model	Train MSE	Train R^2	Val MSE	Val R^2	Test MSE	Test R^2
0	Linear Regression	1.436146e+02	0.0519	1.027379e+02	-0.0590	3.603609e+02	0.0265
1	Random Forest	2.629890e+01	0.8264	1.059432e+02	-0.0921	3.861715e+02	-0.0432
2	XGBoost	8.757000e+01	0.4219	8.537700e+01	0.1199	3.693773e+02	0.0022
3	Linear SVM (10%)	1.088940e+10	0.0350	1.195064e+10	-0.1586	1.189094e+10	-0.0101

Due to the poor performance, I was curious to dig more and find out what might have been the reason. At first, I suspected data leakage, but then after making sure I did the right split, eliminated similarity in naming train, validation, and test sets, I ruled it out.

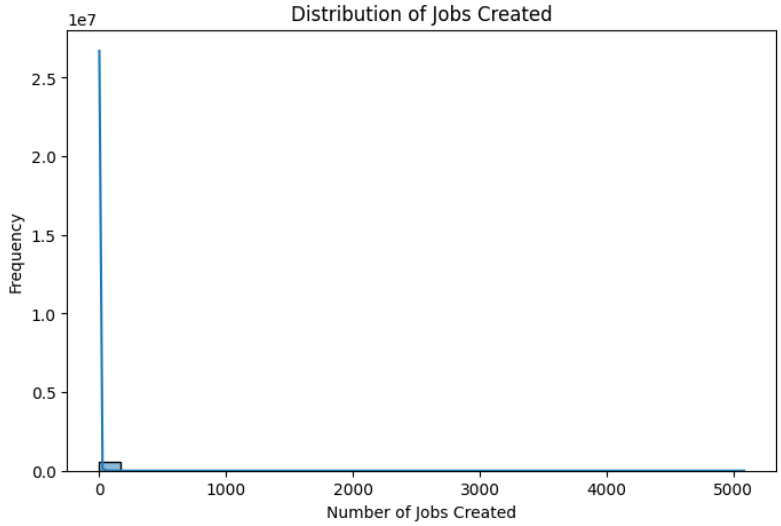
I then took a quick statistic of the target variable as shown in Figure 3 below.

Figure 3: Quick statistics of CreateJob

```
Mean jobs created: 2.4244259897647
Max jobs created: 5085
Min jobs created: 0
Standard deviation: 13.77575529978048
```

I then plotted the distribution of the data to visualize it and make it easier to identify any outliers as shown in Figure 4 below.

Figure 4: Distribution of CreateJob



Given my findings, I believe that the models performed poorly because the target variable, CreateJob, has a highly skewed distribution with extreme outliers — the standard deviation is over five times the mean. This imbalance distorts error metrics, in this case it is MSE, especially for models like Linear Regression and SVM, which are highly sensitive to outliers. Random Forest and XGBoost overfitted on these large values since proper transformation and regularization was not done, therefore limiting the tree model's ability to generalize on unseen data. I could not do a SHAP analysis in this case due to poor performance by all the four models.

d) Loan Repayment Status Classification

Similar to the previous questions, four models were implemented to classify loan repayment status. The Logistic Regression model delivered strong and consistent performance across all datasets, achieving a test accuracy of 80.39% and explaining 96.18 percent of the variation in the dataset.

The best performing model was XGBoost with an accuracyof 92.93% and an r-squared of 0.9617. Random Forest performed almost similarly to its tree-based counterpart, achieving a 90.73% accuracy and an r-squared of 0.9602. The poorest performer was SVM with a test accuracy of 79.68% but a negative r-squared.

Table 5: Model performance comparison for MIS_Status_Binary

	Model	Train MSE	Train R^2	Train Accuracy	Val MSE	Val R^2	Val Accuracy	Test MSE	Test R^2	Test Accuracy
0	Logistic Regression	0.0070	0.9611	0.8035	0.0071	0.9608	0.8042	0.0068	0.9618	0.8039
1	Random Forest	0.0016	0.9912	0.9936	0.0074	0.9587	0.9055	0.0071	0.9602	0.9073
2	XGBoost	0.0070	0.9613	0.9303	0.0071	0.9608	0.9284	0.0069	0.9617	0.9293
3	Linear SVM	0.1980	-0.1258	0.8020	0.2054	-0.1394	0.7946	0.2032	-0.1324	0.7968

Figure 5 below show SHAP analysis results identifying the top ten most important features in predicting loan repayment status.

Figure 5: Feature importance using XGBoost Model results

feature	mean_abs_shap
Term	2.627156
RevLineCr_binary	0.281998
GrAppv	0.263459
DisbursementGross	0.258838
SBA_Appv	0.218135
NoEmp	0.140914
RetainedJob	0.116525
UrbanRural_binary	0.073697
Sector_Healthcare	0.050511
Sector_Accommodation & Food	0.047551

From the SHAP analysis, the length of the loan term emerged as the most influential factor affecting loan repayment status. Other key predictors included whether the loan was a revolving line of credit, the size of the approved loan, the SBA-approved loan amount, the disbursed amount, and the number of employees. Among sectoral features, the Healthcare sector and the Accommodation & Food sector stood out, with SHAP values of 0.051 and 0.047 respectively,

indicating that borrowers in these industries are more likely to be predicted as at risk of default compared to other sectors.

e) Loan Default Prediction

The base model, logistic regression had an accuracy of 80.71 percent and the best test r-squared of 96.18%. Similar to the loan repayment status classification, the best performing model in predicting default was XGBoost with 93.56% accuracy and explain 96.17% variation in the dataset. Random Forest was second place at an accuracy of 90.77% and r-squared of 96.02%. The poorest performer again was SVM with a negative r-squared and an accuracy of 79.87%. Table 6 shows a summary of the models' performance.

Table 6: Model performance comparison for ChgOffDate_Binary

	Model	Train MSE	Train R^2	Train Accuracy	Val MSE	Val R^2	Val Accuracy	Test MSE	Test R^2	Test Accuracy
0	Logistic Regression	0.0070	0.9611	0.8069	0.0071	0.9608	0.8070	0.0068	0.9618	0.8071
1	Random Forest	0.0016	0.9912	0.9920	0.0074	0.9587	0.9076	0.0071	0.9602	0.9077
2	XGBoost	0.0070	0.9613	0.9365	0.0071	0.9608	0.9347	0.0069	0.9617	0.9356
3	Linear SVM	0.1975	-0.1230	0.8052	0.2051	-0.1378	0.7974	0.2028	-0.1303	0.7987

Using the results of XGBoost, I conducted a SHAP analysis to get the top ten most important features as shown in Figure 6 below.

Figure 6: Feature importance using XGBoost Model Results

feature	mean_abs_shap
Term	2.717524
RevLineCr_binary	0.304878
SBA_Appv	0.297689
DisbursementGross	0.232091
NoEmp	0.155832
RetainedJob	0.113839
UrbanRural_binary	0.080294
Sector_Accommodation & Food	0.054146
Sector_Healthcare	0.047729
Sector_Professional Services	0.038091

In terms of predicting loan default, the most influential feature is loan terms. In terms of sectors, the accommodation & food and healthcare sectors were the top in predicting loan default followed by the healthcare sector. The results on these questions are almost similar to the repayment status prediction which also identified these two sectors as important predictive features. This can be informative for lenders to identify high-risk sectors and adjust their assessments.

Reflections

Similar to Li, Mickel, and Taylor, who used logistic regression for loan approval decisions, my analysis found that logistic regression performed relatively well in predicting loan repayment status and default with approximately 80% accuracy, outperforming SVM (2024). However, the tree-based models consistently demonstrated superior performance across all research questions.

My findings align with Glennon & Nigro's research, confirming that loan maturity (term) is a critical factor in predicting default and repayment status, as evidenced across all four SHAP analyses I conducted (2005). Additionally, I corroborated their conclusion that firm-specific characteristics—including number of employees, sector, and revolving line of credit availability—serve as important predictors of loan default status.

Both Random Forest and XGBoost models exhibited strong performance in regression and classification tasks. Their low computational cost makes them practical choices for these applications without significant trade-offs between accuracy and processing requirements. Throughout this project, I've learned that investing time in data familiarization and preprocessing is substantially more valuable than model development itself. My deep understanding of the dataset enabled me to identify why all four models underperformed when predicting job creation outcomes.

For future research on job creation prediction, I recommend proper transformation and regularization of the CreateJob feature to enhance tree-based models' ability to generalize to unseen data. Researchers should also address extreme outliers that distort error metrics and compromise the performance of outlier-sensitive models like linear regression and SVM. Finally, I suggest ensuring sufficient computing power to run SVM on complete test datasets, which would likely improve its comparative performance across all research questions.

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