Small Business Administration Loans

Josahn Oginga
Data Science II Final Project

- The Small Business Administration (SBA) is important in providing financial, educational, and other resources to support small businesses in the United States.
- SBA guarantees up to 85 percent of loans of $150,000 or less and up to 75 percent of loans above $150,000.
- This project uses SBA lending data from 1994-2014 and machine learning models to predict loan repayment, default, amount charged off, disbursed, and number of jobs created.

Research Questions

- What is the optimal amount of loan to approve for a business?
- What is the predictor of the size of the loan charged-off principal upon default?
- What are the predictors of the number of jobs created by SBA guaranteed loans?
- What are the predictors of a loan's repayment status?
- What is the predictor of a loan's likelihood to be in default?

Literature Review

- Li, Mickel & Taylor (2018) developed logistic regression framework for loan approval decisions
- Chehab & Xiao (2024) found positive correlation between county social capital and SBA loan approvals, with unemployment, population, income, and rural-urban status as key factors
- Glennon & Nigro (2005) used hazard models to show loan maturity, economic conditions, and firm-specific factors significantly predict default probabilities

National SBA Dataset

- Original dataset had 889,164 observations with 27 variables
- After pre-processing, there were 572,333 observations with 48 variables

Methodology

Data Cleaning

- Dropping rows with missing values in non-numeric variables, binary encoding for categorical variables, imputation, and dropping columns that are not part of key variables.

All the the research questions were predicted using Linear Regression, Random Forest, XGBoost, and SVM models.

Continuous Variables

Gross approved loan:

Predicting the optimal amount of loan to be approved by bank.

Charged-off principal amount:

Predicting the amount that is charged off the principal in the event of default.

Number of Jobs Created

Predicting the impact of SBA loan issuance in job creation

Binary Variables

Repayment Status

Predicting the loan repayment status and identifying important features.

Default prediction

Predicting the likelihood of loan default and the drivers.

Approach

Similarity:

- LinearSVC was used because it is less computationally intensive, however, the trade off is that it does not capture non-linear relationships
- Only 2% of the data was used on SVM model

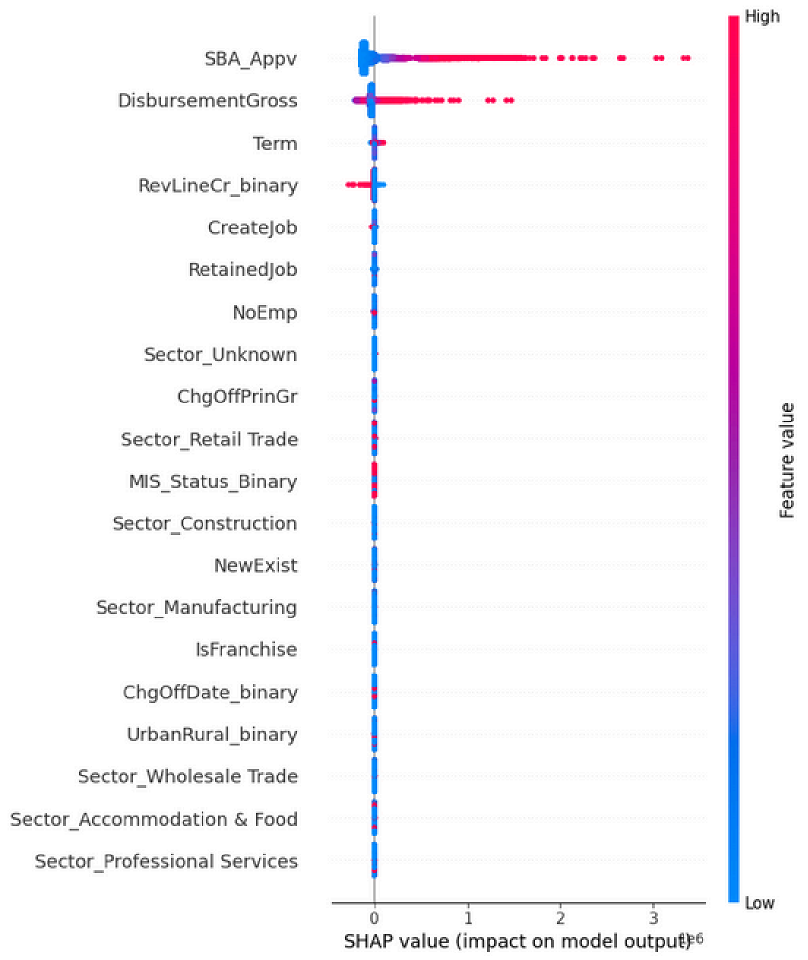
Difference

Continuous variable used regression models while binary models used classification models.

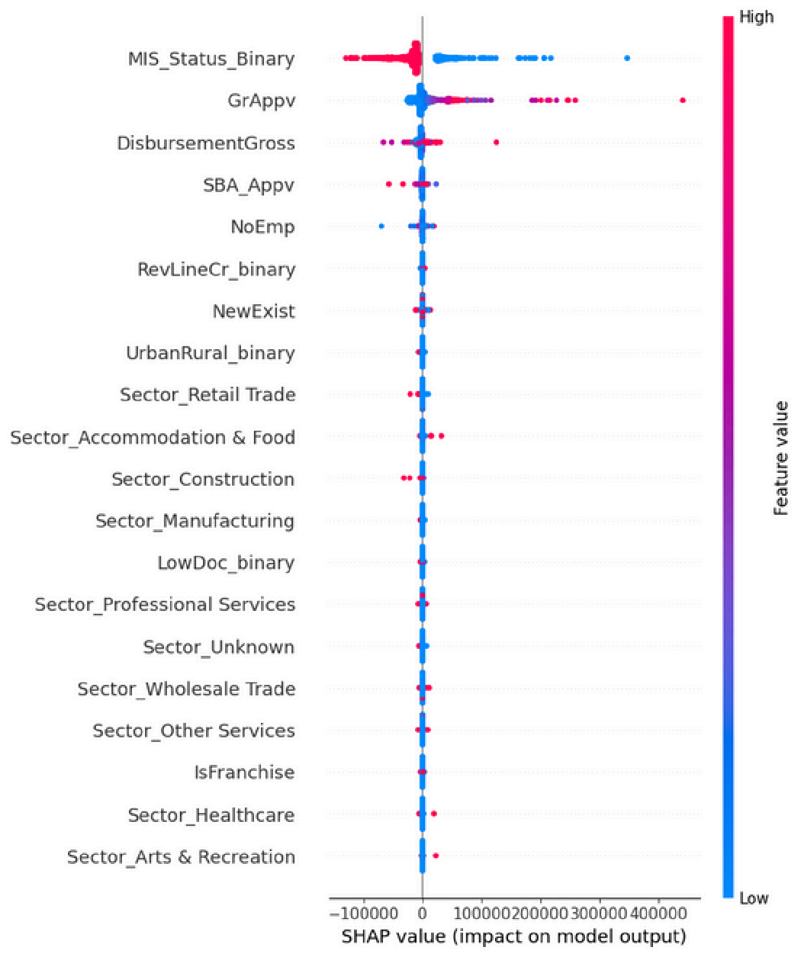
Findings

Continuous Target Features

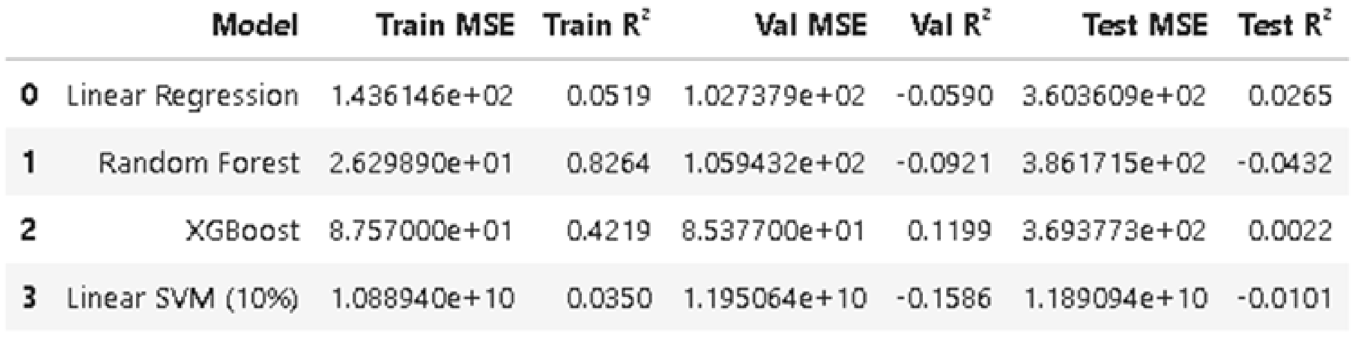
Gross Approved Loan



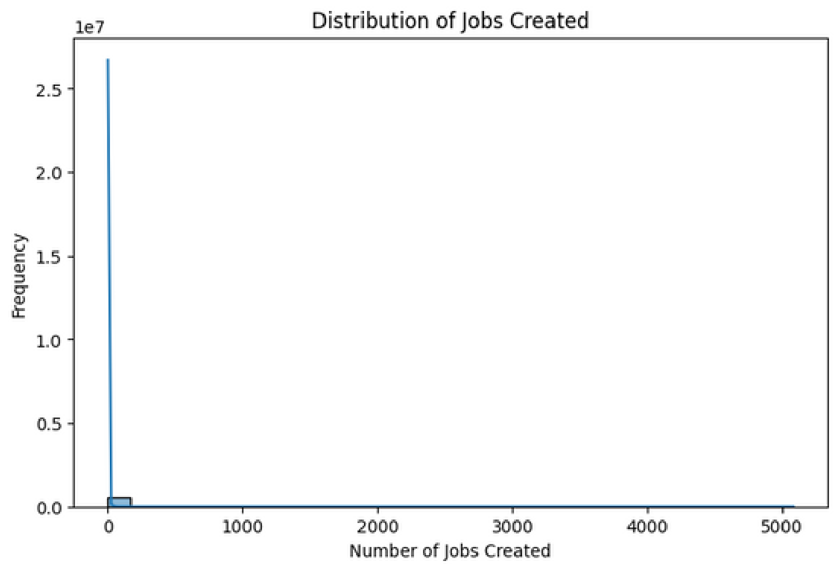
Charge-Off Principal

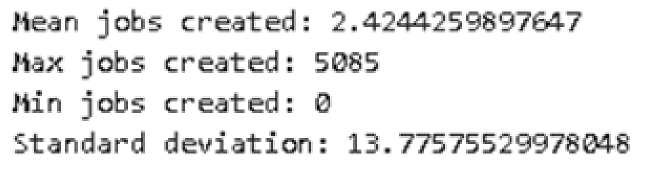


Number of Jobs Created



	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





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

Results:

- Random Forest performed the best with explaining 99% variation in dataset
- SVM had the smallest explanation of 76%
- The most important features in predicting approved loan is the SBA approved loan size, the amount disbursed, and the length of the loan.

Results:

- Random forest explained 81% of the variation in the dataset
- SVM had poorest performance with a negative r-squared
- The most important features is the repayment status of the loan, the amount approved, and the amount disbursed.

Results:

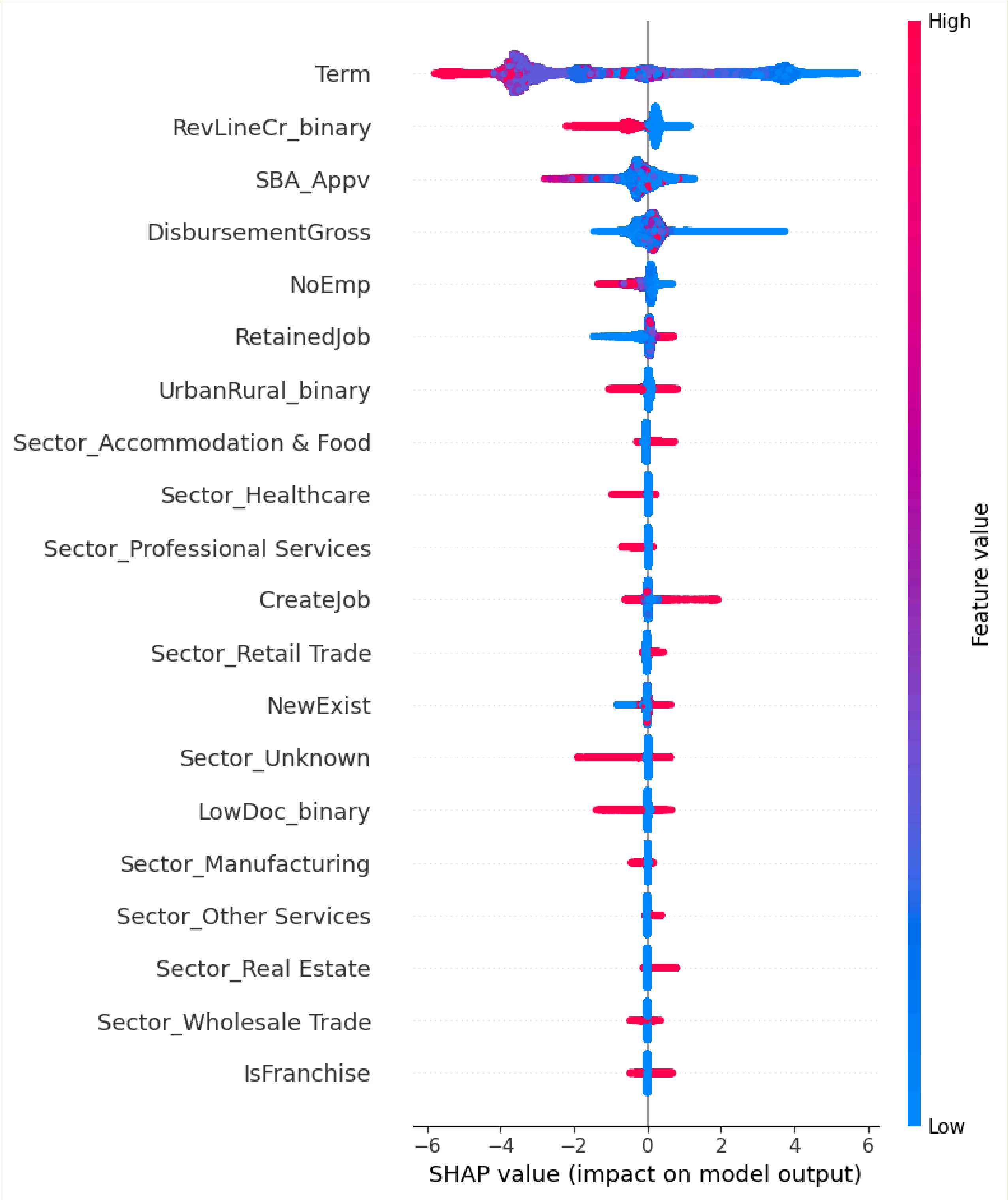
- Linear Regression and XGBoost had low but positive r-squared (0.03 and 0.002).
- Random Forest and SVM reported negative r-squared, evidence for poor performance
- Poor performance was because, CreateJob has a skewed distribution with extreme outliers - the standard deviation is over five times the mean.
- No SHAP analysis was done

The Best Performing Model for the first two questions was Random Forest. The poorest performance recorded was in SVM.

Findings

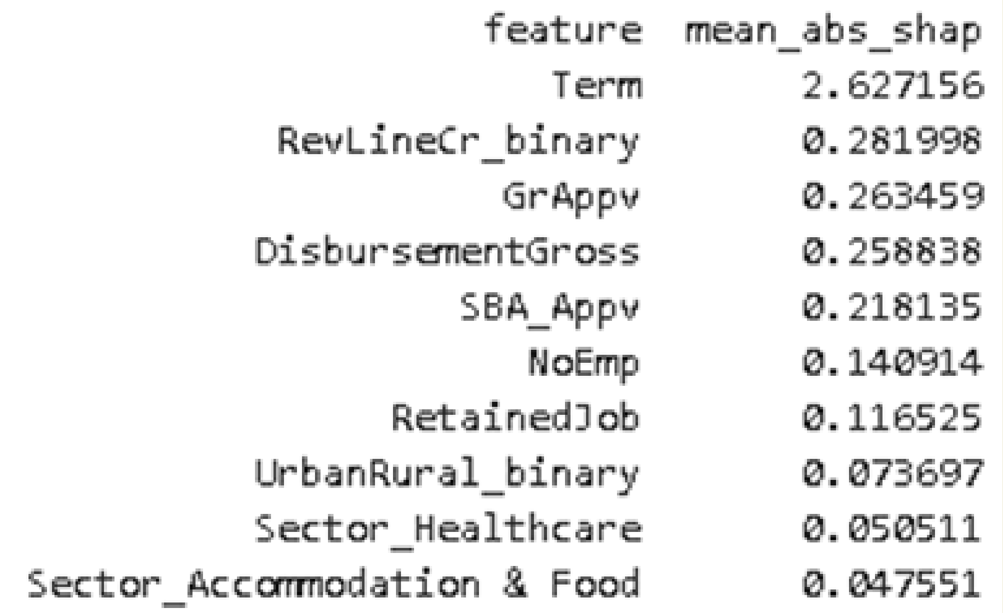
Binary Variable

Important predictors in loan defaults



Repayment Status Prediction

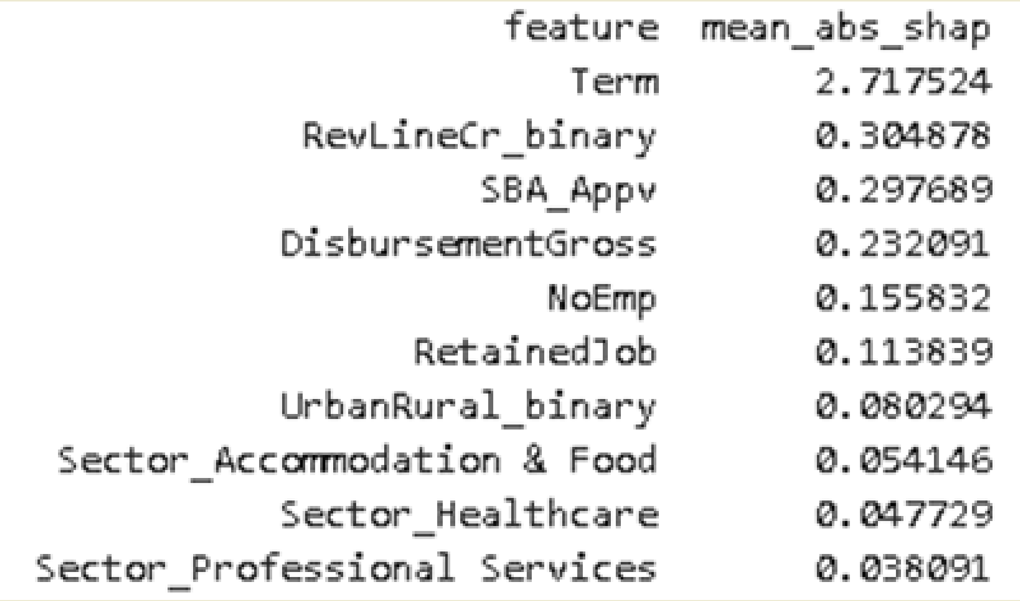
- XG Boost is the best model with a 93% accuracy, and explaining 96% variation in dataset
- The poorest performer is SVM at 79% accuracy and negative r-squared
- The most important features is the loan term, revolving line of credit, and gross amount approved.



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

Loan Default Prediction

- XG Boost is the best performing model with a 94% accuracy and explaining a 96% variation in dataset
- The poorest performer is SVM at 79% accuracy and a negative r-squared
- The most important feature is loan term, revolving line of credit, and size of SBA approved



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

Healthcare and accommodation & food are the sectors that affect loan repayment status the most across both research questions.

Conclusion

Key Findings & Lessons Learned

- Tree-based models (XGBoost and Random Forest) outperformed logistic regression and SVM across all research questions with lower computational cost.
- Loan maturity (term) is a critical default predictor in all SHAP analyses, supporting Glennon & Nigro findings (2005)
- Data understanding proved more crucial than model development

Future Research Directions

- Transform and regularize the CreateJob feature to improve tree-based model generalization
- Address extreme outliers that distort metrics and harm outlier-sensitive models such as linear regression and LinearSVC
- Increase computing resources for SVM to potentially improve its comparative performance
- Focus on balancing model complexity with interpretability for real-world loan decision applications