Decoding Small Business Administration (SBA) 7(a) Loan Charge Off Risk: A Predictive Model Odyssey

by Timalyn Eugenia Franklin-Bullock

B.S. in Math, May 1993, Spelman College
B.S. in Industrial and System Engineering, June 1995, Georgia Institute of Technology
MBA in Finance, June 2001, New York University Stern School of Business

A Praxis submitted to
The faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

May 19, 2024

Praxis directed by

Amir Etemadi
Associate Professor of Engineering and Applied Science

The School of Engineering and Applied Science of The George Washington University certifies that Timalyn Eugenia Franklin-Bullock has passed the Final Examination for the degree of Doctor of Engineering as of April 2, 2024 . This is the final and approved form of the Praxis.

Decoding Small Business Administration (SBA) 7(a) Loan Charge Off Risk: A Predictive Model Odyssey

Timalyn Eugenia Franklin-Bullock

Praxis Research Committee:

Amir Etemadi, Associate Professor of Engineering and Applied Science, Praxis Director

Michael Jones, Professorial Lecturer of Engineering and Applied Science, Committee Chair

Jonathan Bierce, Professorial Lecturer of Engineering and Applied Science, Committee Member

Abstract of Praxis

Machine learning (ML) and probability have been utilized by organizations such as the International Monetary Fund (IMF) to predict loan defaults for small and midsize enterprise (SME) loan defaults. Predominantly ML loan default models are based on loan application data and credit worthiness features. The goal of this praxis is to model US Small Business Administration (SBA) 7(a) loan charge off prediction utilizing loan application and economic features. The features are modeled with logistics regression, XG Boost, Decision Tree, Random Forest, and K-Nearest Neighbor ML methods. The model with the best precision for predicting if a loan will be paid in full or charged off is selected for a reusable model.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AML	Anti-Money Laundering
BSA	Bank Secrecy Act
CBSA	Core-Based Statistical Areas
CFPB	Consumer Finance Protection Bureau
CHGOFF	Charge Off
EIDL	Economic Injury Disaster Loan
FDIC	Federal Deposit Insurance Company
FINTECH	A blend of finance and technology
GAO	General Accounting Office
KYB	Know Your Borrower
KYC	Know Your Customer
LTSMNN	Long Term Short Term Neural Network
ML	Machine Learning
NAICS	North American Industry Classification System
OIG	Office of Inspector General
PIF	Paid in Full
PPP	Paycheck Protection Program
PPP	Proportion of Positive Predictions
SBA	Small Business Administration

SEC	Security and Exchange Commission
SVM	Support Vector Machine
U.S.	United States

Chapter 1—Introduction

1.1 Background

In 2019, the Consumer Financial Protection Bureau (CFPB) estimated the small and midsize (SME) lending market to total $1.4 trillion (Factsheet: Required Rulemaking on Small Business, 2023). The trillion dollar lending market was estimated to have more than $3.3 billion loan defaults as of 2019 (Patel et al., 2023). Financial institutions have been interested in predicting and managing SME business loan default risk since the 1970s(Ciampi et al., 2021). Most loan default prediction tools are tailored to the commercial market and do not address the intricacies of the US Small Business Administration (SBA) lending policies.

SBA offers small businesses government guaranteed, microloans, and direct loans to small businesses which account for 98.4% of registered US businesses (Brighi et al., 2019). Disaster loans are underwritten and managed by the federal government for victims of natural disasters. Since Disaster loans are not limited to businesses, public Disaster loan information is limited for analysis. Microloans target underserved markets, and the loan application data is not publicly available. SBA does publish data about loan guaranty programs.

SBA provides access to capital to small businesses with the 7(a) and 504 loan guaranty programs (*Lender and Development Company Loan Programs (50 10 6)*). SBA Guaranty loans are originated and serviced by an approved lender. Guaranty loan programs are designed for small business that are not able to obtain “Credit Elsewhere” (Temkin & Theodos, 2008). Lending partners are incentivized to originate and service

SBA guaranty loans to foster economic growth within the small business community. SBA guarantees lenders that up to 75% of the loan amount will be paid by SBA in the event of a default leading to a loan charge off (Siegel, 2014). The guaranteed portion of loans are funded each year as a cohort which is identified as a liability in the SBA financial report.

Defaults result from a business not paying a loan for more than 90 days during the servicing phase of the loan process. SBA works with lenders to recover the funds before sending the loans to Treasury for collection. Charge offs are the unrecovered funds from defaults. Lenders can only recover the guaranteed portion of a charged off loan if the lender proves that the loan met the subprogram code underwriting and SEC requirements. If a defaulted loan does not meet requirements, SBA is not required to honor the guaranty. Defaulted loans are not forgiven and will be pursued in collection by the US government using Treasury debt collection processes and/or the lender.

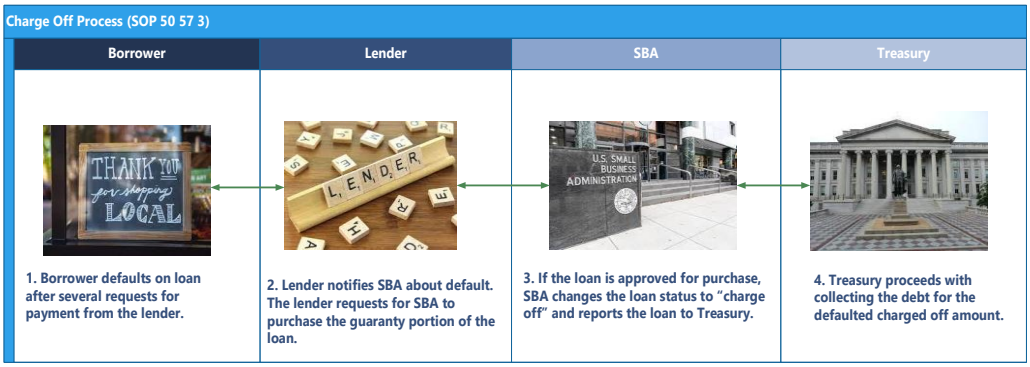


Figure 1. Charge Off Process

1.2 Research Motivation

In March of 2020, the United States (US) government declared the Coronavirus Disease 2019 (COVID-19) a national pandemic. The pandemic declaration negatively impacted the economy with businesses shutting down(Fairlie & Fossen, 2022b). To

counter the economic impact of the shutdown, the government offered Paycheck Protection Program (PPP) and COVID-EIDL loans(PUBLIC LAW 116–136—MAR. 27, 2020, 2020) to assist small business owners pay their employees. PPP loans are guaranteed 100% by SBA shifting the shared 75/25% SBA/Lender risk("Summary of Senate Omnibus CARES Act," 2020).

The pandemic increased awareness of SBA loan fraud risk (Shear, 2021). To supply oversight of the pandemic funds, 20 Inspector Generals formed the Pandemic Response Accountability Committee (PRAC) in 2020 ("Semiannual Report to Congress.," 2020). In 2023, the PRAC identified $5.4 billion in potentially fraudulent PPP and COVID-Economic Injury Disaster Loan (EIDL) (K. Singh, 2023). Loan defaults occur for legitimate reasons such as insolvency as well as due to fraud (Eweoya et al., 2019). Taxpayers pay the guaranty fee for defaulted loans whether due to fraud or business failure. The research motivation is to predict 7(a) guaranty loan charge off risk to support SBA and lender decision making. This research utilizes the public 7(a) loan data to explore a risk prediction model. Since the SBA publicly reports charged off loans this praxis will use the term charge off for SBA as being synonymous with loan defaults.

1.3 Problem Statement

Small business origination processes do not adequately estimate the risk of loan default as indicated by the issuance of over $16.4 billion in government backed loans that were charged off between fiscal years 2014 and 2023 (the SBA fiscal year begins each October 1st) based on performance data as of 12/31/2023 (Small Business Administration Loan Program Performance, n.d.).

In 2023, SBA Inspector General Hannibal Ware identified $16.4 billion in PPP loans that defaulted or have not received forgiveness (Ware, 2023). The agency does not have the workforce required to perform debt collection on over a million loans.

1.4 Thesis Statement

A small business loan charge off prediction model reduces the guaranteed cost associated with the 7(a) loan program by reducing the number of charged off loans originated each year.

During the loan servicing process past due loans are tracked as required by the Debt Collection Improvement Act (DCIA). A machine learning (ML) model can predict charge off risk in the loan origination process with loan application data and economic factors to reduce the number of charged off loans. A reduction in the number of charged off loans reduces the guaranty approval amount that SBA pays lenders.

1.5 Research Objectives

To achieve the objective, the research identifies features and metrics that indicate default/charge off. The features are identified from public SBA 7(a) loan data and economic factors to achieve the research objectives that include:

- Utilize publicly available loan data (US Small Business Administration (SBA) 1991-2022 7(a) loan data) to identify features to potential loan charge offs.
- Develop a machine learning charge off prediction model to support SBA's 7(a) loan program.

1.6 Research Questions and Hypotheses

The research questions supporting the thesis and objectives are:

RQ1: What public loan application features are associated with the likelihood of charge offs?

RQ2: What are the important features, economic features, and machine learning model that predicts 7(a) loan charge offs to reduce guaranty fees?

In support of the research questions, this praxis will explore the hypotheses below.

H1: Loan application data features such as loan amount, interest rate, and term in months as well as the 7(a) loan subprogram code and delivery method are predictors of loan charge off risk.

- **H2:** A small business loan charge off prediction risk model that includes loan features and historical economic data predict the risk of charge off for 7(a) loans to reduce guaranty fees paid by SBA.

1.7 Scope of Research

The research data source is SBA 7(a) loans from 1991 to 2022 and includes loans from the great recession (2007-2009) as well as the pandemic (2020-2022). The loan data is publicly available on the SBA website in compliance with the Freedom of Information Act (FOIA). No additional due diligence or non-public data is used in the risk model for small business loan defaults/charge offs.

1.8 Research Limitations

Restricting the analysis to SBA 7(a) and PPP loans is a limitation of the research. The model may not predict loan risk for small business commercial or government non-SBA 7(a) loans. In the past ten years, online lending institutions known as FINTECHs leverage technology for loan decisioning and validation of the digital profile. An SBA 7(a) loan risk model may not be applicable to FINTECH lending.

Internally, SBA tracks several loan statuses that are not publicly reported. Loans can go into "liquidation" prior to charge off. The public status of "Exempt" includes loans in good standing as well as loans with delinquent payments that are in liquidation. The industry definition of default is not applicable to this research. This research aligns with SBA's usage of the loan status "Charge off."

Congress recommended SBA implement a comprehensive loan risk management program in 2007 (Baum & Hsueh, 2008). This model may not address risk associated with SBA Microloans, 504, SBIC, and Disaster loan programs. In addition, the model does not differentiate charge off due to fraud.

1.9 Organization of Praxis

Chapter two is a literature review of loan default models and terms which will be leveraged for defining the SBA charge off prediction model. Chapter three details the data sources, data cleansing, testing methodology, and risk model detail. The ML results, approaches, and calibration are reviewed in Chapter four. Chapter five summarizes the benefits of the research to the engineering management community.

Chapter 2—Literature Review

2.1 Introduction

The literature review is compartmentalized into five sections to provide a foundation for a 7(a) loan charge off prediction model. Section 2.2 introduces the SBA 7(a) loan types. Section 2.3 reviews ML definitions and algorithms that will be used for predicting charge offs. Section 2.4 identifies small and medium enterprise (SME) default features. This section includes research correlating economic conditions with SME default risk. Section 2.5 is a compilation of ML loan default models evaluated by the International Monetary Fund (IMF), Asian Development Bank (ADB) and European Union. The risk model considerations for supporting loan decisioning is reviewed in section 2.6

2.2 Previous SBA research

The 7(a) Guaranty loan program includes traditional 7(a) and PPP loans. Since traditional 7(a) loans have different data requirements than PPP loans, SBA distinguishes between traditional 7(a) Guaranty loans and PPP loans for external reporting. This praxis will evaluate traditional 7(a) loans.

2.2.1 7(a) SBA Loan Types

The 7(a) Guaranty loan program is umbrella for loans that are further categorized by subprogram and delivery method. The 7(a) current include Standard 7(a), 7(a) Small, SBA Express, Export Express, Export Working Capital, and International Trade(Lake, 2019). Each loan type lists the processing delivery method to originate the loan. The current delivery methods include Preferred Lender Delegated, Non-Delegated, there are subprogram codes such as Preferred Lender Program (PLP), non-PLP, Community Express, and Payroll Protection Program (PPP), and Cap Lines.

There are specific underwriting and loan guaranty requirements for every subprogram and processing/delivery method. In 2022 a George Washington University dissertation conducted research on SBA loan defaults. The analysis assumes all 7(a) loan programs are underwritten with the same requirements (Liu, 2022). The default analysis does not navigate the complexity of the SBA underwriting model. Previous research identifies the current economy as an input to small business loan risk (Jeong, 2023). Hypotheses one and two will include features for subprogram code and processing method to account for differing requirements.

SBA's guaranty purchase process includes reviewing that a lender complies with the lending agreement, lender of last resort requirement, and subprogram specific underwriting policy associated to the cohort year. Loan purchases are denied for not following the requirements. The charged off loans that SBA publishes are loans that followed the subprogram rules and the guarantees were honored. SBA lenders are at risk of originating loans that have not been properly vetted and a FICO score is not the only requirement(*Lender and Development Company Loan Programs (50 10 6). (n.d.).*, n.d.).

Previous SBA loan research targets either loan defaults or reviewing the impact of SBA loans on the economy. Research coupling the economic risks on SBA loan defaults has been limited. Models that review SBA loan defaults falsely assume that every SBA loan guaranty is honored. SBA guaranteed loans are considered loans of last resort for small businesses that may not have the history or credit worthiness to meet the requirements of a traditional business loan (Small Business Loans: Additional Actions Needed to Improve Compliance with the Credit Elsewhere Requirement, 2018).

Economic research related to SBA loans targets the impact on economic factors such as increasing employment and increasing credit to underserved markets. In 2023 a study was published comparing the impact of SBA loans on various communities (Jeong, 2023). Using Generalized Methods of Moments (GMM), the study compared the change in the employment rate to low- and high-income communities that received loans. The study results identified a positive increase in the employment rate for loans in underserved markets. In addition, the study identified that SBA loans are not homogeneous, and their impact will vary depending on the loan program.

2.2.2 PPP SBA Loan Types

The were multiple rounds of PPP loans. The first round is characterized as benefitting larger businesses(Henry, 2020). With the focus on increasing access to capital for underserved and smaller business, there was a second round of PPP loans(Henry, 2020). Legislation allowed borrowers to apply for a second draw loan in a third round of funding (Fairlie & Fossen, 2022). The loans are tracked with the processing method indicating first draw loans as PPP and second draw loans as PPS.

PPP loan policy provides guidelines for loans to be forgiven if the business submitted information substantiating that the funds were used to pay employees. Unlike 7(a) loans, the loan amount charged off is not publicly available(Gradisher & Tassell-Getman, 2020). The PPP charge off total is not available in the public data.

2.3 Machine Learning Definitions and Metrics

Identifying ML classification features includes comparing predicted data to the actual data to identify true positives, true negatives, false positives, and false negatives (Ohsaki et al., 2017). The classifications create a confusion matrix and metrics to

measure the model's accuracy and precision. Tables 1 and 2 summarize the definitions and metrics used throughout the literature review (James et al., n.d.).

Table 1. Definitions

Metric	Definition
True Positive (TP)	A correct positive class prediction (X is X)
True Negative (TN)	A correct negative class prediction (X is NOT Y)
False Positive (FP)	An incorrect positive class prediction/Type 1 Error
False Negative (FN)	An incorrect negative class prediction/Type 2 Error

Table 2 .Metric Definitions

Metric	Formula	Ideally
False Negative Rate	FN/(TP+FN)	Low
False Positive Rate (FPR)	1-Specifity=(FP)/(TN+FP)	Low
Recall/TPR/Sensitivity	TP/ (TP +FN)	High
Precision	TP/(TP+FP)	High
Specificity/True Negative	TN/(TN+FP)	High
F1/Harmonic Mean	(2*Precision*Recall)/ (Precision + Recall)	High
Geometric means (G-mean)	Sqrt (Specificity*Sensitivity)	High
Accuracy/ACC	(TP+TN)/(TP+TN+FP+FN)	High
Area Under the Curve (AUC)	(1+Recall-FPR)/2	High
Errors	(FP+FN)/(TP+TN+FP+FN)	Low
Brier Score/Average Square Error	The Brier Score is between 0 and 1.0.	Low
Kolmogorov–Smirnov test	(Number of elements in sample<=t)/n	High
Proportion of Positive Predictions (PPP)	PPP(TPR = y)	Low
R-squared	1-(sum of squared difference between predicted and actual value/sum of squared difference between actual values and mean of dependent variable)	1
Root Mean Squared Error (RMSE)	Square root ((average of (sum of actual minus predicted values)) squared)	0
Mean Squared Error	Average of ((sum of actual minus predicted values)) squared)	0
Mean Absolute Deviation	Average of (Absolute value (sum of actual minus predicted values)))	0
Mean Absolute Percent Error	Average of sum of Absolute actual minus predicted values)	0

2.4 Loan Default Prediction Features

For this analysis, a defaulted borrower is likely to not pay the loan due to insolvency or fraud. The model will determine the likelihood of default based on features in the loan application as well as economic predictors. Preferably, the model should account for varying degrees of distress although the loan may not be considered bad (Lin et al., 2012).

2.4.1 Loan Application Features

Small business research identifies default predictors including business type, financial/credit risk, business location, and loan terms. Understanding the features of a good loan supports the risk default review process (Altman & Sabato, 2008). The traditional underwriting small business process includes: collecting data from a borrower reviewing business and personal documents such as credit history (rating, collections, charge-offs, foreclosures, loan history, bankruptcy, reorganization, federal loan or tax delinquencies), criminal history, lawsuit history, and child support arrears (Kumar & Motwani, 1999).Business type features include whether a business is a sole proprietor, LLC, or Corp. Small business loan application data provides additional information to identify loan default risk. Experian associated firms with 1-4 employees with a higher default risk than firms with 5 or more employees (Experian, 2021). The same study documented sole proprietors with higher default risk than firms organized as Limited Liability Corporation (LLC), Limited Liability Partnership (LLP), and/or corporations. The dependence on a small number of resources to maintain a business correlates to risk that should be considered during the loan decisioning process. The franchise also is a feature that influences default. In the 2008 economic crisis the lending community noted

the elevated risk of default for franchises; the SBA identified franchises as most likely to default on loans (Gibson, 2009). Another element of business type is the business age. There is a correlation between the firm/business age and the likelihood of loan denied due to risk of insolvency (Cassar et al., 2015). The correlation indicates that firms older than 15 years are more likely to not be denied a loan. This praxis does not include denials but will examine the correlation between loan defaults and business/firm age.

Financial risk is measured by reviewing credit and fraud risk features. Credit risk assumes the applicant is a legitimate business with the intent of repaying the loan. Research confirms the positive correlation between financial data and small business credit default risk (Kirschenmann & Norden, 2012). Credit worthiness is an additional feature that predicts small business default risk determined by historical factors such as debt to credit ratio, collateral, and credit score. (VanSomeren & Tarver, n.d.). This research will not directly include credit worthiness due to the publicly available limitations of individual credit information.

Fraud risk assumes the applicant is not a legitimate business and/or does not have the intent to repay the loan. Credit and fraud risk features can overlap limiting identification of the root cause of defaults to a primary factor. Non-traditional indicators of credit risk include owning multiple businesses. An owner with multiple businesses in multiple states is a strong indicator of fraud (McGlasson, 2010). A mismatch of a small business owner's city and state address with the business city and state address is an indicator that the loan application requires additional review (McGlasson, 2010). A business that does not have a location or the location is a PO Box is another indicator to perform additional due

diligence (McGlasson, 2010). The SBA public data does not include the guarantor address to compare with the business address.

According to Experian (King, 2023), FICO (Cox, 2022), and industry experts (Evans, 2023), there are three types of fraud: first, second, and third party fraud. Sixty-two percent of fraud is first party fraud with second- and third-party fraud being equally likely (Zaki, 2023). First party fraud is a legitimate, credit approved customer who does not plan to repay the lender. With first party fraud the legitimate person provides false documentation and/or banking information. First party fraud is mitigated with identity proofing and validating the legitimacy of loan documentation. Random Forest and XG Boost models identify first party fraud with features such as comparing the IP address with a loan application address (G. Yedukondalu et al., 2021).

Second party fraud occurs when a legitimate customer allows another entity to use their information to defraud a lender (Evans, 2023). Second party fraud is more common for money laundering and difficult to detect as a government lender. The borrower in second party fraud usually transfers money to another account. The borrower's lender can use the Patriot Act to report business transactions that are potentially money laundering.

Lack of authentication controls produces an opportunity for third-party fraud. The proliferation of fraud is increased by the difficulty with authenticating small businesses that are synthetic identity fraud victims. Synthetic identity fraud transpires when a digital identity is created from stolen legitimate data that is not traced to the original data (Rudegeair & Andriotis, 2018). Third- party fraud caused by synthetic fraud was heightened during the pandemic (Zhu et al., 2021).

McKinsey noted historical data can validate a person's identity to combat synthetic fraud (Richardson & Waldron, 2019). McKinsey compared 15,000 consumer accounts with nine data sources to evaluate the risk profile. The profiles are assigned a score based on depth and consistency of information related to 150 features. McKinsey's model is an example of a method that can be implemented by lenders. The article did not detail accuracy metrics.

Location features include rural vs. urban as well as the proximity to the originating lender. SBA rural borrowers and lenders in the same town have a lower default rate than loans originated by lenders and small businesses in the same urban city (DeYoung, 2015). In 2006, Federal Deposit Insurance Corporation (FDIC) identified lender borrower distance as a feature of default risk (FDIC, 2006).

A peer to peer (P2P) model was created to review loan defaults using loan approval data (Turiel & Aste, 2020). The analysis identified the loan term, FICO score, debt to income ratio, and loan amount to be the top four predictors of default for P2P loans. SBA does not publicly share the FICO score and debt to income ratio of loan borrowers due to privacy legislation. The SBA public data does include the loan term and loan amount which will be considered in the analysis.

The loan charge off prediction model will leverage SBA data that is consistently available to the public. The publicly available business information includes borrower address, interest rate, and term in months. The business type information that is consistently available includes the North American Industry Classification System (NAICS) codes and whether the business is an LLC or sole proprietor.

2.4.2 Economic Predictors of Loan Defaults

A businesses growth within an economy is the normal measure for loan default. Business growth is measured by reviewing financial data such as collateral and assets as well as non-financial assets such as the character of the business owner (Zhao & Lin, 2023). Technology SME defaults may even be predicted based on the number of patents and/or technology (Ciampi et al., 2021). Relying on the company's growth to predict loan defaults should be expanded to also include a review of the national economy that the business operates in at time of loan origination.

Emerging economies provide a perspective on the impact of the national economy to SME defaults. A study conducted with data from the Central Bank of Mexico reviews macroeconomic factors predicting loan default. Based on the study, economic conditions at the time of origination impact the business surviving. Specifically, businesses are sensitive to inflation which is a factor that should be considered when originating loans (Batiz-Zuk et al., 2022).

Inflation directly impacts the business cost of goods and services which may make a business less viable (Webster 2023). The inflationary pressures also result in higher small business loan interest rates which directly impacts loan defaults (Aeppel 2023). In addition, SBA loans have fixed and variable rates that make them susceptible to defaults (Angell, 2023).

Gross Domestic Product (GDP) is an indicator of a borrower's future ability to pay (Piffer, 2018). Small businesses require demand for their products to generate revenue. GDP is considered an indicator of product demand directly influencing a small

business' solvency (Kaya, 2022). GDP decreases also result in a contraction of the lending markets which impacts the availability of capital to small businesses (P. Singh, 2023).

GDP and unemployment rate are predictors of household consumption predicts the success of small business which predicts the likelihood of loan default (Caselli et al., 2008). In 2005 a small business default model was created using proprietary bank data(Agarwal et al., 2008). The model included owner risks, firm risks, loan contract, and macroeconomic risks. Businesses with higher local unemployment rates have a higher risk for loan default.

The research will compare SBA loans by year to national inflation, GDP, and unemployment rates to address national economic conditions related to hypothesis two. The influence of the national economy may improve risk predictions for charge offs/defaults.

2.5 Loan Default Machine Learning Analysis

According to the Oxford Dictionary Machine Learning (ML) is "the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data." SME loan defaults have been modeled but there is not a leading model that is accepted by the industry. This research review includes studies conducted International Monetary Fund (IMF), Asian Development Bank (ADB), and European Union (EU) for Italian SMEs. The literature includes a Chinese SME Credit Worthiness model, Historical Random Forest Model (HRF) for a European lender and

insurance company, and a Random Forest Model. The loan default literature review also briefly considers default models for non-SME loans.

2.5.1 SME Loan Default Models & Outcomes

The IMF conducted a study on SME risk prediction using China's MyBank data (Huang et al., 2020). Traditionally, China predicts SME defaults with a Scorecard that includes the SME's credit history. The provided data set includes bank credit information and information from MyBank's majority owner known as Ant Group*. The analysis compares the traditional technique with ML supervised technique known as Random Forest (RF). RF is equivalent to multiple decision trees.

The RF analysis was also conducted using cross validation to reduce the likelihood of model bias. The IMF research cross validation nodes are split based on the Gini index. The Gini index calculates the probability that similar features are grouped together (Tyagi, 2020). The Gini index is from 0 to 1 with zero being the ideal number for purity.

Based on the analysis of the Mybank data, RF outperformed the traditional model. The IMF research notes that the traditional model relies on lagging data with credit history as the main variable while the RF model with Gini purity is more accurate with business age as the main variable.

Table 3. IMF MyBank SME Data

Model	AUC	Ninety-five percent Confidence Intervals	
Random Forest+bank credit history+Ant information	0.8681	0.8636	0.8725
Random forest + bank credit history	0.8372	0.8318	0.8427
Scorecard+bank credit history+Ant information	0.7767	0.7713	0.7820
Scorecard+ bank credit history	0.7397	0.7337	0.7456

Logistics Regression is an algorithm for classifying binary and linear data (Subasi, 2020). Like the IMF, the Asian Development Bank (ADB) conducted a study on Japan's SME default prediction model. Japan considers a loan in default after three months non-payment. Researchers evaluated a machine learning approach based on data from 2014-2018 to determine if Japan's Credit Risk Database (CRD) predicted SME defaults (Nguyen 2020). Each year the CRD models are recalibrated to predict loan default for 1-3 years with financial ratios.

The study compared the accuracy of the model using logistic regression, RF and Extreme Gradient Boost (XG Boost). XG Boost is a decision tree ensemble machine learning algorithm that requires a large data set (Bentéjac et al., 2019). The dataset was evaluated for accuracy and then retested. As indicated by the results below XGBoost outperformed logistic regression and RF for predicting defaults with transaction data. The study noted that RF and XG business are prone to overfitting.

Table 4. ADB Results

Dataset	Method	AUC
Testing (parameter tuning)	RF	0.7070
Testing (parameter tuning)	XGBoost	0.7329
Testing (parameter tuning)	Logistic	0.7113
Back test	RF	0.7482
Back test	XGBoost	0.7728
Back test	Logistic	0.7174

The European Union is assessing a SME default prediction model using accounting and financial data from Italian SMEs. The test compared prediction results for Logistic Regression, Probit, Binary Generalized Extreme Value Additive (BGEVA), XG Boost and Feedforward Neural Networks (FNN) using Orbis Bureau van Dijk (BvD) dataset(Crosato et al., 2023). Probit regression is like logistics regression but uses a different cumulative distribution function (CDF). BGEVA is a regression model used for rare binary events. FNN is a deep learning technique to fit non-linear events. Sensitivity, specificity, Type 1, and Type 2 errors, AUC, Brier's score (BS), harmonic mean, and Kolmogorov-Smirnov statistic (KS) are metrics that were used for comparison (Crosato et al., 2023). XG Boost performed the best on predicting Italian SME loan defaults with accounting data.

Table 5. Italian SME Results

Model	Sensitivity	Specificity	Type 1	Type 2	Harmonic Mean	AUC	BS	KS
FNN	0.694	0.829	0.171	0.306	0.391	0.827	0.187	0.501
XGBoost	0.821	0.719	0.281	0.179	0.383	0.843	0.146	0.552
BGEVA	0.752	0.727	0.273	0.248	0.331	0.819	0.178	0.481
LR	0.745	0.736	0.264	0.264	0.303	0.809	0.151	0.483
Probit	0.738	0.737	0.263	0.262	0.299	0.809	0.109	0.448

A Chinese lender worked with SMEs to test a fusion ensemble credit worthiness model (Gao et al., 2021). Boosting, bagging, stacking, soft/hard voting fusion are ensemble techniques useful for predicting default risk. Researchers created a soft voting fusion model that includes logistic regression, support vector machine (SVM), random forest (RF), XGBoost, and light gradient boosted machine (Light GBM) to identify small and medium business credit risk in China. To create the model, the AUC is generated for each classifier. The weighted average of the weighted AUCs creates the soft voting fusion model. The AUC for the fusion model is higher than each classifier's AUC. Based on AUC, the fusion model is feasible for predicting credit risk of small and midsize businesses in China.

In 2021, a research team used European SME lender and insurance data to compare a Historical Random Forest Model (HRF) to a traditional probit model (labeled PB in the analysis). The models compared static and dynamic credit scoring using Balance Sheet (BS) and US Securities and Exchange Commission (SEC) performance ratios as input variables. The Bergmann and Hommel p-value metric indicates that HRF is more reliable for predicting small business credit risk than the PB model (Bitetto et al., 2021).

Lenders forecast credit risk with statistical and logistic regression techniques over a one-year horizon. Researchers compared a Random Forest Survival Model to a logistics regression model predicting credit risk with small business financial ratios as the key input variables. The logistics regression AUC of .84 is superior to the Random Forest Survival AUC of .76. The researcher performed cross validation by comparing the confidence interval. Logistics regression proved to be the better model with cross validation (Fantazzini & Figini, 2009).

The Basel Committee on Banking Supervision (BCBS) was established in 1974 to develop international banking standards. BCBS identifies stress testing related to market, operational and credit risk to measure lender viability (Goodhart, 2011). Researchers identified neural networks, SVM, Random Forest, Classification and Regression Tree (CART) decision tree, Multivariate Adaptive Regression Splines (MARS) and Vector Auto regression to measure Basel credit risk (Leo et al., 2019). Besides the Basel model, credit default features for risk models include financial ratios like profitability and liquidity (Lin et al., 2012). Growth metrics (i.e., return on assets (ROA), market cap growth, sales growth etc.) are meaningful features for small businesses lacking financial data for default risk models (Lin et al., 2012).

2.5.2 Non-SME Loan Default Models Analysis & Outcomes

Non-SME models were reviewed to identify additional default prediction considerations. The reviewed non-SME models include a Norwegian model to identify ML money laundering classification. A P2P model that addresses borrower and loan characteristics provides additional considerations given the imbalance of data available

for adequately defaults vs. paid in full (PIF) loans. Finally, credit worthiness models are reviewed given their dominant role in loan underwriting decisions.

To detect money laundering, legitimate and illegitimate transactions were identified with a supervised learning method analyzing 2014-2016 data (Jullum et al., 2020). The Jullum binary classification model predicts if a transaction is legitimate or suspicious by reviewing transaction data. The model includes XGBoost and 10-Fold Cross Validation algorithms to predict risk on transactions and not risk on accounts.

A Norwegian Model was reviewed by the same research team to classify suspicious transaction activity in the alert, case, and/or reporting stage (Jullum et al., 2020). Stage A is a legitimate untagged transaction. Stage B includes transactions identified in the alert stage but considered legitimate. Stage C transactions are identified as suspicious in the alert and case stage but cleared as legitimate. Stage D transactions are flagged in the alert, case, and reporting stages. Stage D is a mix of legitimate and illegitimate transactions. The Brier Score, AUC, and PPP (invented by the researchers) measured the results for all stages as well as alerts. The results are measured for all types (A-D), transactions with no alerts (B and C), no normal transactions (legitimate B and C), all types of multi-response three, and all types of multi-response four. The AUC score for legitimate and illegitimate transactions is 0.907.

Table 6. Money Laundering Machine Learning Study Results

Evaluation metric (test data)	All types (binary) (includes A-D)	No non-reported alerts/cases (binary) (B and C cases)	No normal transactions (binary) (B and C illegitimate)	All types (Multiresponse 3)	All types (Multiresponse 4)
AUC (all)	0.907	0.852	0.872	0.91	0.91
AUC (only alerts)	0.822	0.706	0.819	0.826	0.825
Brier (all)	0.025	0.34	0.024	0.025	0.025
Brier (only alerts)	0.047	0.655	0.047	0.047	0.047
PPP (TPR = .95) (all)	0.315	0.373	0.452	0.305	0.33
PPP (TPR = .8) (all)	0.203	0.26	0.268	0.195	0.196

Lending Club data is used for P2P default risk model includes loan characteristics, borrower characteristics, and borrower assessment categories from loan application data (Chen et al., 2021). The data is imbalanced with a ratio of fully paid to defaulted loans of 3.5:1; under sampling and over sampling address the imbalance (Chen et al., 2021). Under sampling is conducted with Tomek Links to remove fully paid loans near defaulted loans in the model (Y.-R. Chen et al., 2021). Synthetic Minority Oversampling Technique (SMOTE) increased the number of loan defaults in the model. The imbalanced model identified false positives meaning loans likely to be Bad Standing were identified as Good Standing. SMOTE and under sampling will address the imbalance of SBA Good Standing and Bad Standing loans (Y.-R. Chen et al., 2021).

The Peer-to-Peer data is trained and tested with the execution of Random Forest, Neural Networks, and Logistics Regression algorithms. The three ML models are compared for accuracy, recall, F1, and Geometric mean (G-mean). Each model is compared against the original ratio, random under sampling, Tomek Links, random over sampling, SMOTE, borderline SMOTE, and SMOTE+Tomek Links. The SBA models will be analyzed with SMOTE and under sampling techniques.

With three features, Random Forest accuracy is high while neural networks recall, F1, and G-mean are high. Comparing loans amounts below $5,000, Random Forest accuracy is high. Logistics regression outperforms Random Forest and Neural networks when considering recall, F1, and G-mean metrics. Comparing loans amount below $30k, Random Forest accuracy and F1 are high, neural networks have the best recall metrics, and logistic regression has a better G-mean.

Table 7 summarizes the results for Random Forest, Neural Networks, and Logistic Regression with the original ratio, Random Under sampling, and SMOTE.

Table 7. Summary of Peer-to-Peer Metrics

Machine Learning Technique	Sampling Technique	Accuracy	Recall	F-1	G-Mean
Logistic Regression	Original Ratio	77.885	6.414	11.443	25.119
Logistic Regression	Random Under sampling	63.628	66.246	44.7773	64.519
Logistic Regression	SMOTE	63.946	65.714	44.814	64.566
Neural Networks	Original Ratio	77.738	0.177	0.354	4.213
Neural Networks	Random Under sampling	64.061	65.220	44.706	64.470
Neural Networks	SMOTE	37.927	91.932	39.758	47.188
Random Forest	Original Ratio	77.79	9.260	15.606	30.108
Random Forest	Random Under sampling	64.721	61.369	43.662	63.488
Random Forest	SMOTE	77.390	12.212	19.398	34.253
Random Forest	Under sampling for loans below $5k	62.941	58.368	37.526	61.128
Neural Networks	Under sampling for loans below $5k	60.840	65.276	38.865	62.475
Logistics Regression	Under sampling for loans below $5k	61.237	67.182	39.795	63.403
Random Forest	Under sampling with three features	63.931	60.881	42.923	62.812
Neural Networks	Under sampling with three features	63.559	66.463	44.830	64.567
Logistics Regression	Under sampling with three features	63.236	66.146	44.494	64.2247

Another P2P study modeled loan transactions data such as age, gender, education level, occupation, borrowing amount/gross approval, interest, and repayment term are modeled for predicting fraud (J. J. Xu et al., 2022). Random Forest, XG Boost, Deep Neural Networks (DNN), and long-term short-term memory neural network (LTSMNN) models are generated for three datasets -A, B, and C. Data set A includes categorical features, set B includes categorical and numerical features, and C includes categorical, numerical, and sequence/historical features. The results are compared with metrics—accuracy, recall, precision, f-score, and AUC. Set B outperformed set A for Random Forest, XG Boost, and Deep Neural Networks models. The metrics for Set C are generated with LTSMNN and compares to set B's DNN results for balanced and imbalanced data set. Sets B and C performed similarly for the balanced data set. Set C outperformed B in all metrics for the imbalanced data. The study implies that predictions from an imbalanced data set is improved with a sequence such as payment history and borrower data (J. J. Xu et al., 2022).

Credit worthiness was modeled by researchers comparing statistical analysis, machine learning, and deep learning models with German and Australian public data sets (Shi et al., 2022). The analysis includes AdaBoost, Artificial Neural Networks (ANNs), Conventional Neural Networks (CNN), Decision Tree (DT), Deep Belief Neural Networks (DBNs), Deep Multi-Layer Perception (DMLP), Extreme Gradient Boost (XG Boost), Extreme Learning Machine (ELM), Genetic Algorithm (GA), KNN, Long Short-Term Memory (LSTM), RF, Recurrent Neural Networks (RNNs), Restricted Boltzmann Machines (RBMs), Stochastic Gradient Boosting (SGB), and SVM models. For the German data, the AUC was high for bagging and RF was high for ACC. For the

Australian model ANN has the best AUC and ELM had the best AUC. The analysis indicates that deep learning is the most effective tool for identifying credit worthiness of a debtor. The analysis was inconclusive about which deep learning method was best to prove credit worthiness (Shi et al., 2022). A summary of the credit models is in Table 8.

Table 8. Credit Worthiness Models

Unique Datasets	Source of Model	Model with Highest Accuracy Score for Dataset	Accuracy
Balanced FICO	(C. Chen et al., 2022)	Two-Layer Additive Risk Model	.7404
Credit Default Swap	(Luo et al., 2017)	DBN	1
Seventy-six small businesses from a bank in Italy	(Angelini et al., 2008)	Feedforward networks	.87
NYU's Solomon Center database	(Barboza et al., 2017)	Boosting	.8631
A leading European P2P Platform Bendor	(Byanjankar et al., 2015)	Neural Networks	.7438
An Original credit scoring dataset of a firm providing credit loans in Singapore	(Leong, 2016)	Neural Networks	.84
Lending club and Kaggle, German credit	(Barboza et al., 2017)	Balasso based RF	.8975
A dataset belonging to National Bank of Canada	(Marceau et al., 2019)	XGBoost	.7822 (AUC)
A real-world P2P China data platform	(Wang et al., 2019)	AM LSTM	.669 (AUC)
Lending club	(Fan & Yang, 2018)	A denoising-autoencoder-based neural network model	.875
A real-world dataset from a Chinese Company	(B. Zhu et al., 2018)	Relief-CNN	.6989 (AUC)
The credit data from CRBC, German dataset, Darden dataset	(Zhang et al., 2017)	FV-SMO	.849 (AUC)
The database from Taiwan	(R.-Z. Xu & He, 2020)	DBN	.9604
Bloomberg and Compustat	(Golbayani et al., 2020)	CNN2d	.9013
Mortgage loan data from a commercial bank	(Galindo & Tamayo, 2000)	CART	.917

2.6 Risk Model Considerations

Risk is “a potential loss, disaster, or other undesirable event measured with probabilities assigned to losses of various magnitudes” (Hubbard, 2020). The information from models provides decision makers with a method to calculate expected loss and quantifiable impact. When determining if a borrower is worthy of credit, the obligor/lender can consider the borrower solvent, liquid, and likely to repay the loan (Bouteille & Coogan-Pushner, 2022).

Risk for banks is measured as expected loss. To calculate the expected loss (EL), the lender assesses the probability of default (PD) within the first year, loss given default (LGD), and exposure at default (EAD) (Orlando & Pelosi, 2020). In the analysis of SBA loans, less than 1% (1720/211665) of 7(a) loans were charged off in the first year. The analysis for SBA will consider loss to include loans charged off at any time during the life of the loan.

Any model utilized used to deny loans based on fraud or credit risk must comply with Federal laws related to The Equal Credit Opportunity Act (ECOA). ECOA requires fair lending and prohibits discrimination. The selected risk model must be reviewed for possible bias which may cause noncompliance with Fair Lending Rules (Brotcke, 2022). As the SBA does not publish the protected group information for each loan, this bias assessment is beyond the scope of the research.

Cost sensitive learning is another technique to ensure that all errors are not treated equal (Elkan, 2001). The Elkan model improves on predicting defaults but less accurately predicts payment in full loans. The Elkan Model is defined on the reasonable condition that the cost of a false positive (FP) is more than the cost of a true negative (TN) and a

false negative (FN) cost more than a true positive (TP). In this example, defaults are represented as negative and paid in full is represented by positive. With the condition that FP>TN and FN>TP, it is more cost effective to accurately predict defaults than paid in full loans. A financial organization is likely to prefer improved prediction of defaults even if it is at reduced accurate prediction of loans paid in full. The misclassification of fraudulent and valid loans does not have an equal cost. The SBA analysis design considers the condition that the cost of FP (default predicted to be paid in full) costs more than TN (default) and FN (loan paid in full predicted to default) costs more than a TP (loan paid in full).

An effective risk model is continuously monitored and calibrated to account for new risks (Hubbard, 2020). The model outcomes should also be routinely reviewed for false positives and false negatives to refine the accuracy (X. Zhu et al., 2021). Calibration techniques include upper and lower bound, Repetition & Feedback, Klein's premortem, Equivalent Bets, and Avoid Anchoring (Hubbard 2020). The SBA model will be calibrated with repetition and feedback.

2.7 Summary and Conclusions

Previous SBA loan default research does not differentiate between the various loan programs. The research identifies risk at the higher level of 7(a) loans not accounting for unique underwriting requirements within each program under 7(a) loans. Loan defaults can be due to legitimate reasons such as insolvency as well as illegitimate activity including first, second, and third-party fraud. SBA loans have an additional complexity with unique underwriting requirements by cohort year and subprogram/delivery method. Previous SBA loan research oversimplifies SBA loan default events. Hypothesis one

highlights the importance of the loan application coupled with 7(a) loan program/delivery data to predict charge offs. This research will determine if the program and processing method impact the risk of default.

With respect to the economy and SBA loans, previous research identified the impact that SBA loans have on the economy. Research has not focused on the impact of the economy on SBA loan defaults. This praxis will define a model that includes economic data to identify defaults that SBA and lenders can consider for 7(a) loan decision making.

Hypothesis two includes a prediction model that accounts for program features and economic data to predict loan charge offs. The GDP, inflation, and unemployment will be evaluated for correlation as well as uncertainty of prediction given two independent events. Bayes theorem and the multiplication rule will be the basis for selecting the best economic factor. The loan default prediction model leverages previous research with the goal of identifying a model with high AUC, precision, and accuracy.

For hypothesis two, probability techniques will provide insight into GDP, inflation, and unemployment trends improving 7(a) loan charge off predictions. Economic trends provide an additional data point to refine charge offs predictions.

The literature identifies ML default prediction techniques with AUCs of .70 to .9. The features from hypothesis one and economic factors will culminate in a Decision tree, Extreme Gradient Boost (XG Boost), K-nearest neighbor (KNN), logistics regression and random forest (RF) models to support lenders and SBA decision making. The literature provides SME and non-SME models that support improving SBA loan default predictions. The features, algorithms, and risk default research breadth provides

foundational analysis that supports this praxis contributing a loan application and economic SBA 7(a) loan default prediction model to the SBA lending industry.

ML models have been developed to predict SME loan defaults based on borrower, loan, and credit worthiness factors. Loan default models have also been developed for non-SME loans. The research validates that statistics and machine learning provide default predictions with AUCs of .7 and above. In addition, various ML techniques provide data that supports decision making. All the models require fine tuning to address variance and bias with either boosting, bagging, oversampling, and/or under sampling. This praxis utilizes SMOTE and under sampling to address the imbalance of paid in full and charge off 7(a) loans.

Chapter 3—Methodology

3.1 Introduction

This praxis model predicts the likelihood of a US SBA 7(a) loan default in origination. The model identifies prediction features, predicts loan default from loan application data, and predicts risk given economic factors. The methodology for developing the model is defined below.

1. Collect and cleanse 7(a) and PPP loans as well as economic data.
2. Identify 7(a) default trends associated with loan application and economic data.
3. Develop and test prediction model based on hypotheses.

Section 3.2 identifies the data sources and data cleansing considerations for all data evaluated in this praxis. Section 3.3 is a summary of charge off trends that were instrumental in selecting the methodology for hypotheses 1 and 2. Sections 3.4 and 3.5 include the methodologies for evaluating hypotheses 1 and 2, respectively.

3.2 Data Collection & Cleansing

SBA 7(a) loans 1991-2022, PPP loans 2020-2022, and economic data are the foundation of the risk prediction model. The data was reviewed for anomalies and completeness. The 7(a) and PPP Data Dictionaries are in the Appendix.

3.2.1 7(a) Loan Data

The SBA files on the website as of September 30, 2023 are the source of the 7(a) loan data (*7(a) & 504 FOIA*, 2023). There are approximately 1.7 million SBA loans in .csv format organized by year with files in 10-year segments. SBA publishes a data dictionary for the 7(a)-loan file. Tables 9-10 summarizes the loan transaction totals.

Table 9. Summary of Loan Data

Year	1991-1999	2000-2009	2010-2019	2020-2022	Total
Total Observations	337,043	690,347	545,751	141,836	1,714,977
CHGOFF	35,322	145,929	29,598	816	211,665
PIF	256,038	455,539	342,177	18,505	1,072,259
Total PIF+CHGOFF	291,360	601,468	371,775	19,321	1,283,924

3.2.1 PPP Loan Data

The SBA files on the website as of September 30, 2023 are the source of the PPP loan data (*PPP Dataset*, 2023). There are approximately 11.4 million SBA loans in .csv format organized by year with files delineated in 10-year segments. The analysis is based on PIF and charged off loans. The data set includes 1.16M charged off loans and 10M PIF loans.

Table 10. PPP Loan Data

File	Transaction Count	PPP CHGOFF Trans	PPP PIF Transcation	PPS CHGOFF Transactio	PPS PIF Transc	Total
A	968,525	11,863	655,036	4,016	282,517	953,432
B	900,000	32,079	605,528	8,294	233,778	879,679
C	900,000	39,745	622,192	8,822	203,841	874,600
D	900,000	52,352	593,347	12,475	212,144	870,318
E	900,000	48,396	622,083	11,431	184,601	866,511
F	900,000	50,755	10,311	612,056	193,581	866,703
G	900,000	36,421	623,074	8,062	205,298	872,855
H	900,000	29,395	662,960	6,744	183,000	882,099
I	900,000	31,447	597,612	7,717	245,125	881,901
J	900,000	40,402	624,912	9,218	202,347	876,879
K	900,000	29,631	633,499	6,585	208,899	878,614
L	900,000	38,204	631,503	8,934	197,407	876,048
M	599,774	15,796	429,682	3,712	140,804	589,994
Total	**11,468,299**	**456,486**	**7,311,739**	**708,066**	**2,693,342**	**11,169,633**

3.2.3 Economic Data

The economic data for hypotheses two is from macrotrends.net. The macrotrends data includes unemployment, GDP, and inflation data reported on 12/31 of each year from 1991-2022. The economic data is in the appendix.

3.3 Charge Off Trends

This praxis includes analyzing the loans labeled as “PIF” (paid in full) and CHGOFF (charged off). The Great Recession (2007-2009) is considered the worst downturn since the Great Depression. It is estimated that 400,000 small businesses went out of business in the first quarter of 2009 (Mount, 2009). Figure 4 demonstrates that there were approximately 44k loans in default in cohort 2007 versus 10k in cohort 2009.The pattern indicates an increase in defaults for loans originated in a recession (2007). The economic circumstances during the time of origination reflects increased risk during an economic crisis (Carroll, 2010).

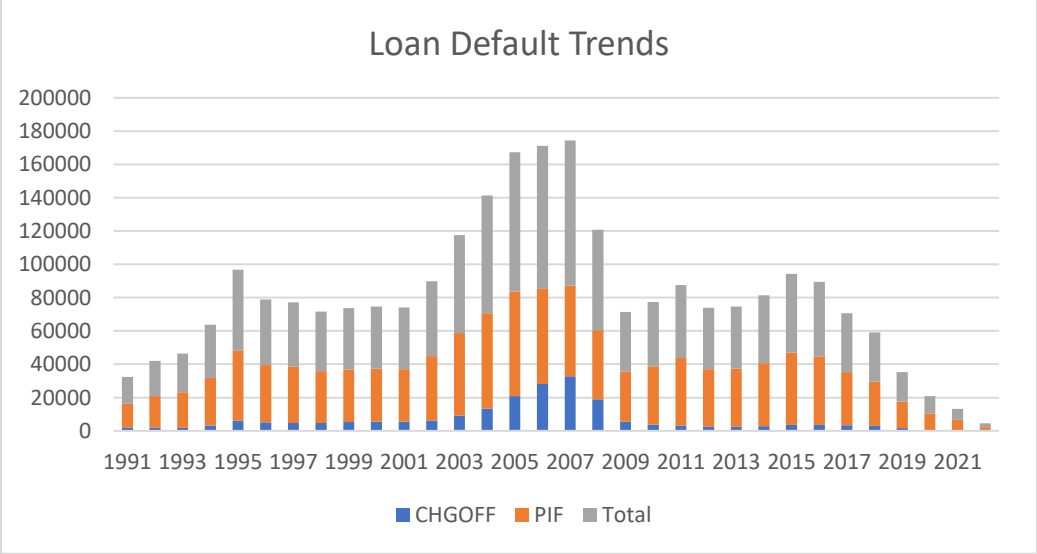


Figure 2. 7(a) PIF vs. Charged Off Loans

The aftermath of the Great Recession is characterized by decreased lending with lenders skeptical of small business credit worthiness (Hackney, 2023). To improve small business access to capital, SBA allowed non-traditional lenders to offer PPP loans (Simon & Rudegeair, 2023). Pandemic loans are perceived to be the largest fraud in US

history (Dilanian & Strickler, 2022). There are less charge off loans during the pandemic than during the Great Recession due to PPP forgiveness. The PPP loan legislation allowed loans to be eligible for forgiveness which may also result in a lower number of defaults for pandemic loans than the Great Recession.

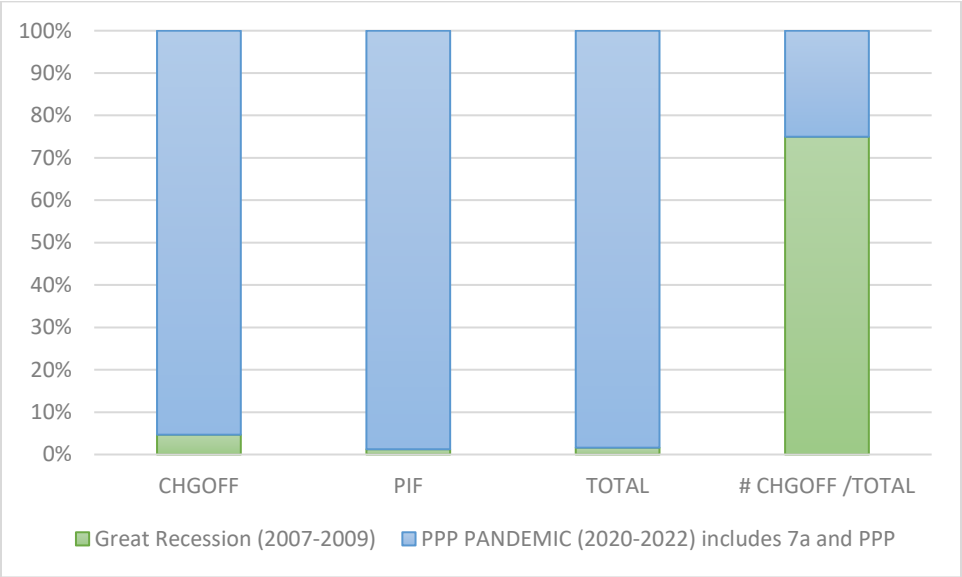


Figure 3. Comparison of Great Recession and PPP Charged Off Loans

3.4. Hypothesis 1 Methodology

Hypothesis 1: Loan application data features such as loan amount, interest rate, and term in months as well as the 7(a) loan subprogram code and delivery method are predictors of loan charge off risk. Hypothesis one is evaluated by identifying the features that predict charge offs. Using Minitab's Classification and Regression and Tree (CART) classification methodology, a decision tree is created to predict features. The methodology uses Gini node splitting within one standard error of minimum classification cost. The model is designed with 10-fold cross validation with a testing to training ratio of 70/30. The Gini index is from 0 to 1 with zero being the ideal number for purity. Pi is the probability of a feature classified as significant.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Equation 1. Minitab CART Gini Index

For the review of 7(a) loans (this does not include PPP loans), two groups are evaluated to identify features. Group 1 includes continuous features (Borrower Zip Code, Bank Zip Code, Gross Approval Amount, SBA Guaranteed Approval, I Amount, Initial Interest Rate, Term in Months, Jobs Supported, Revolver Status (revolver is equivalent to a line of credit loan), Business Age, Business Type, Approval Fiscal Year, Subprogram Code, and Delivery Method). Group 1 does not assume that all 7(a) loans are homogenous and accounts for the variations of the underwriting criteria within the 7(a) loan portfolio.

Group 2 does not include subprogram and delivery method. Group 2 is aligned with previous studies that treat all 7(a) loans equally when considering default. The year of origination/Approval Fiscal Year is a categorical feature for Group 1 and 2 to refine the default to account for the underwriting variations within a given year. For each period, the AUC training and testing are noted as indicators of the overall success of the features.

Additional feature nuances to consider when creating the model are listed below.

- The feature business age is continuous for 1991-2009 and categorical for 2010-2023. The reporting criteria changed from 2009-2010.
- 2000-2009 PLCP and Pollution are missing from test for 7(a) loans. Premier Certified Lenders Program (PLCP) and Pollution were pilot programs with minimal data.

- 2020-2023 subprogram description Small Contractors Section 7(a)(9) are missing from test. The subprogram has minimal data as of the writing of this praxis.

The data available for PPP loans is different from the data available for 7(a) loans.

For comparison purposes, PPP Group 1 and 2 were created with PPP files labelled as A, C, F, K, and M. Group 1 includes continuous features (Jobs Reported, Term, and Initial Approval Amount) and categorical features (Borrower Zip, Originating Lender Location ID, Business Type, Business Age, and Processing Method). Group 2 is equivalent to Group 1 without the processing method. Group 2 includes continuous features (Jobs Reported, Term, and Initial Approval Amount) and categorical features (Borrower Zip, Originating Lender Location ID, Business Type, and Business Age). Note, Minitab removes Originating Lender Location ID and Borrower Zip because there are more than 1024 distinct levels.

3.4.1 Hypothesis 1 Testing

The hypothesis is tested with the 7(a) and PPP SBA data files using Minitab CART to prove that Group 1 performs better than Group based on the True Positive, True Negative, Lower Bound, and Upper Bound.

Table 11. 7(a) Hypothesis 1 Testing Criteria

Test Data	AUC Test	True Positive (Sensitivity)	False Positive (Type 1)	False Negative (Type 2)	True Negative (Specificity)	95% CI Lower Bound	95% CI Upper Bound
1991-1999	Group 1 AUC >=Group 2 AUC	Group 1>=Group 2	Group 1<=Group 2	Group 1<=Group 2	Group 1>=Group 2	Group 1>=Group 2	Group 1>=Group 2
2000-2009	Group 1 AUC >=Group 2 AUC	Group 1>=Group 2	Group 1<=Group 2	Group 1<=Group 2	Group 1>=Group 2	Group 1>=Group 2	Group 1>=Group 2
2010-2019	Group 1 AUC	Group 1>=Group 2	Group 1<=Group 2	Group 1<=Group 2	Group 1>=Group 2	Group 1>=Group 2	Group 1>=Group 2

Test Data	AUC Test	True Positive (Sensitivity)	False Positive (Type 1)	False Negative (Type 2)	True Negative (Specificity)	95% CI Lower Bound	95% CI Upper Bound
	>=Group 2 AUC						
2020-2023	Group 1 AUC >=Group 2 AUC	Group 1>=Group 2	Group 1<=Group 2	Group 1<=Group 2	Group 1>=Group 2	Group 1>=Group 2	Group 1>=Group 2

Table 12. PPP Hypothesis 1 Testing Criteria

Test Data	AUC Test	True Positive (Sensitivity)	False Positive (Type 1)	False Negative (Type 2)	True Negative (Specificity)	95% CI Lower Bound	95% CI Upper Bound
File A	Group 1 AUC >=Group 2 AUC	Group 1>=Group 2	Group 1<=Group 2	Group 1<=Group 2	Group 1>=Group 2	Group 1>=Group 2	Group 1>=Group 2
File C	Group 1 AUC >=Group 2 AUC	Group 1>=Group 2	Group 1<=Group 2	Group 1<=Group 2	Group 1>=Group 2	Group 1>=Group 2	Group 1>=Group 2
File F	Group 1 AUC >=Group 2 AUC	Group 1>=Group 2	Group 1<=Group 2	Group 1<=Group 2	Group 1>=Group 2	Group 1>=Group 2	Group 1>=Group 2
File K	Group 1 AUC >=Group 2 AUC	Group 1>=Group 2	Group 1<=Group 2	Group 1<=Group 2	Group 1>=Group 2	Group 1>=Group 2	Group 1>=Group 2
File M	Group 1 AUC >=Group 2 AUC	Group 1>=Group 2	Group 1<=Group 2	Group 1<=Group 2	Group 1>=Group 2	Group 1>=Group 2	Group 1>=Group 2

3.5. Hypothesis 2 Methodology

3.5.1 Hypothesis 2 Features

Hypothesis 2: A small business loan charge off prediction risk model that includes loan features and historical economic data predict the risk of charge off for 7(a) loans to reduce guaranty fees paid by SBA. The features identified in hypothesis one are integrated with economic factors to create prediction risk model. Inflation, GDP, and unemployment data are evaluated for consideration in the loan prediction risk model. The economic data is compared to the corresponding cohort's paid in full and charged off/

defaulted loans to identify correlations using Pearson and Spearman. The r values for the economic data compared to the charge off rates indicate that Based on the results, there are not linear or monotonic relationships between the defaults and economic data.

Table 13. Pearson and Spearman Results

Type of Correlation	Data Compared	r value	p
Pearson	CHGOFF Percent, Inflation Rate	0.082	0.657
Pearson	CHGOFF Percent, GDP	-0.041	0.823
Pearson	CHGOFF Percent, Unemployment	-0.248	0.17
Spearman	CHGOFF Percent, Inflation Rate	0.221	0.225
Spearman	CHGOFF Percent, GDP	0.124	0.497
Spearman	CHGOFF Percent, Unemployment	-0.231	0.204

Economic evaluation for hypothesis two evaluation includes probability theorems related to independent events. The probability of two events is reviewed by first understanding if there is no relationship, a linear relationship, or a nonlinear relationship. Linear correlation strength is measured with the Pearson correlation coefficient r and p values (Xiao et al., 2016). The r coefficient ranges from -1 (negative linear relationship) to 1 (positive linear relationship) for two continuous variables. The p value determines the strength of the correlation using the t test with n equaling the sample size (Jaadi, 2019). If the p value is less than or equal to .05, there is a strong relationship between the variables (Jaadi, 2019).

Equation 2. Pearson r Coefficient

$$r = r_{xy} = (cov(x, y))/(S_x * S_y)$$

Nonlinear correlation is calculated using Spearman's rank coefficient, ρ . Like the Pearson coefficient determines positive correlation if the value is closer to one, negative correlation if the value is closer to -1, and no correlation when the value is approximately zero. For the Spearman rank di is the difference between two ranks in n observations

Equation 3. Spearman's Rank

$$\rho = 1 - (6\sum d_i^2\ )/(n(n^2\ -1))$$

Bayes Theorem identifies the probability of one event occurring depending on the probability of another event occurring (Devore & Berk, 2012).

Equation 4. Bayes Theorem

P(A/B) = P (A ⋂ B)/P(B)=(P(A) * P(B/A))/P(B).

The multiplication rule for statistically independent events is the product of each event (Hardle et al., 2015).

Equation 5. Multiplication Rule for Independent Events

P (A ⋂ B)= P(A)*P(B)

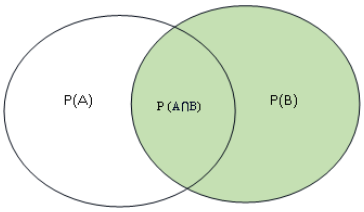


Figure 4. Example of Independent Events Intersection

Additional analysis is performed to refine the understanding of the correlation between economic data and defaults. The 5-, 3-, 2-, and 1-year running averages of the default and economic data are calculated. The probability of charge off is calculated based on the probability of charge off given the probability of GDP, inflation, or

unemployment. Using Bayes Theorem and Multiplication Rule for Independent Events, the probability of defaults is calculated.

The 5-year unemployment data provides the best predictor of SBA 7(a) loan defaults. For most years, the 5-year unemployment is an accurate predictor within +-5%. As expected, the 5-year unemployment data is not a great predictor when there are major unemployment fluctuations. The years prior to the Great Recession predicted a default of 20% and the actual default for the 2007 cohort is 37%. The unemployment data predicted a default rate of 20% in 2012 after the recovery of the recession and the actual default is 6.6% for the 2012 cohort.

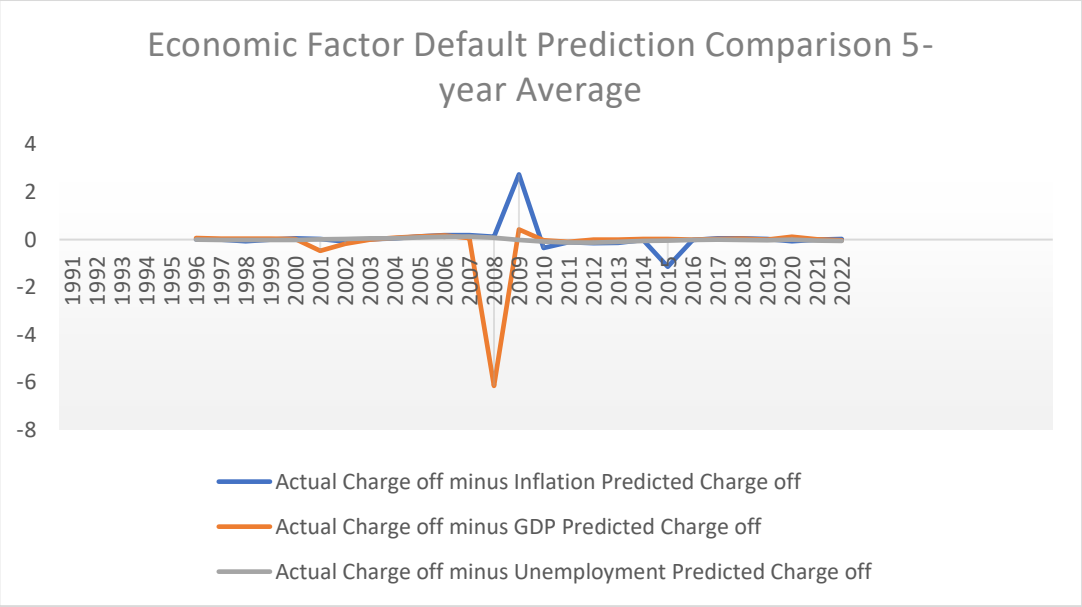


Figure 5. 5 Year Economic Default Prediction vs. 7(a) Loan Defaults

As a sensitivity analysis review, the 3-year average provides improved results. The variation for the unemployment factor prediction was -10 to 9.53%. The 2007 cohort with the 3-year average is predicted to be 28% but is 37%. The 2012 cohort is predicted to be 12.3% and the cohort had 6.6% defaults.

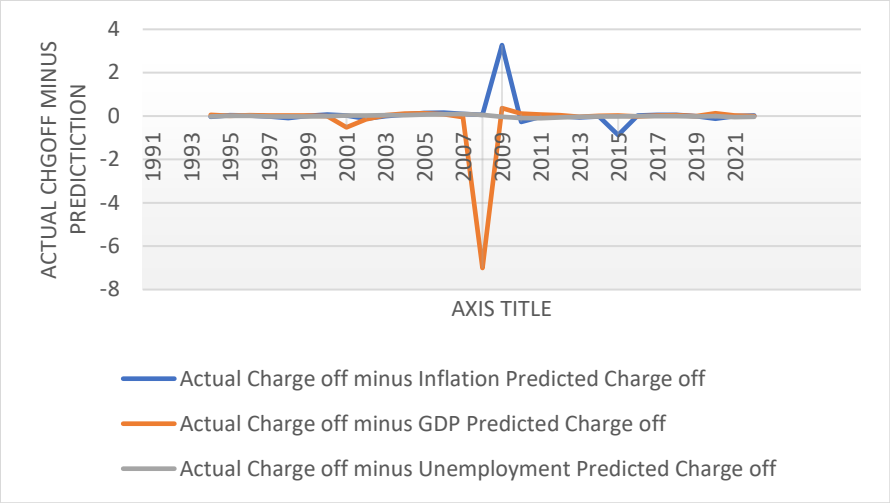


Figure 6. 3-Year Prediction

The unemployment factor outperforms GDP and Inflation factors for the 2-year average. The variation for the unemployment factor prediction -8.6 to 7.7%. The 2007 cohort with the 2-year average is predicted to be 28% but is 30.4%. The 2012 cohort is predicted to be 9.7% and the cohort had 6.6% defaults.

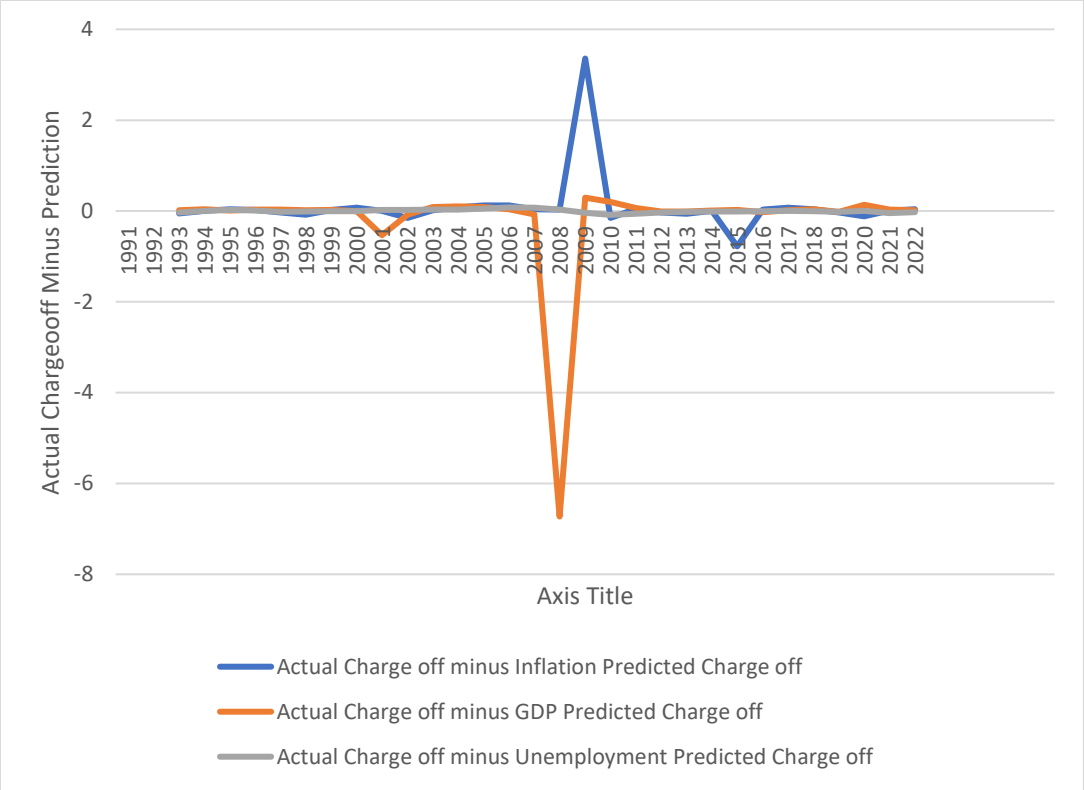


Figure 7. 2-Year Prediction

Overall, the previous year's unemployment factor is the best predictor. The variation for the unemployment factor prediction was -5 to 5.4%. The 2007 cohort with the 2-year average is predicted to be 33% but is 30.4%. The 2012 cohort is predicted to be 8% and the cohort had 6.6% defaults.

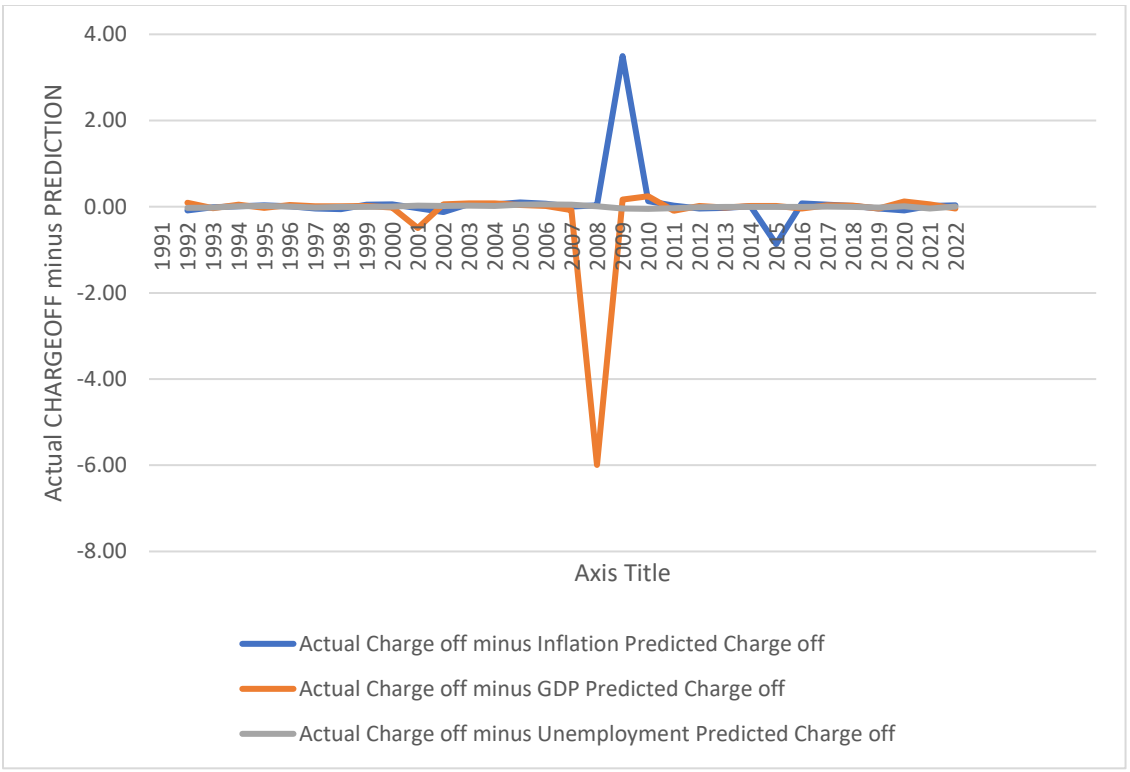


Figure 8. 1 Year Prediction

Unemployment data is a predictor of SBA 7(a) loan defaults. The predictor provides the government information on the expected liability and government's guaranty risk for a cohort. Ideally, the previous year's unemployment rate is the best predictor but may not be available for the planning cycle. The Federal government budget year is from October to September. The annual unemployment rate is reported in January. For example, the 2024 fiscal year budget started in October 2023 and was planned in October 2021. To have the prediction data available for the budget year, the 2020 unemployment data reported in January 2021 will be the most recent data available to support the 2024 fiscal year. The three average unemployment rates are the most recent data when planning for risk.

A model is developed with the three-year running average and the prediction features. Logistic Regression, KNN, XGBoost, Decision Tree, and Random Forest models are generated for 2010-2019 data to evaluate accuracy as a default metric. For all models 20% of the data is set aside for testing the model. Oversampling with SMOTE and under sampling is also generated for each model to compare accuracy, precision, harmonic mean, and geometric mean. The table below identifies the model equations and minimum target metrics.

Table 14. Model Equations

Algorithm	Formula
Logistic Regression	$\sigma(x) = 1/1 + \exp((-x))$
KNN Euclidean	$\sqrt{\left(\sum_{i=1}^{k} (x_i - y_i)\right)^2}$
XGBoost	$L(\phi) = \sum_i (\hat{y_i}, y_i) + \sum_k \Omega(f_k) \, where \, \Omega(f) = \gamma T + \frac{1}{2}\lambda \, \|\|w\|$
Decision Tree	IG(Dp)=I(Dp)−(Nleft/Np)I(Dleft)−(Nright/NpI)(Dright) Where I is entropy, Gini Index, or classification error. Entropy=−∑jpjlog2pj Gini=1−∑jp2j ClassificationError=1−maxpj
Random Forest	Uses decision tree formula.

Table 15. Target Metrics

Metric	Target
Accuracy	Accuracy should be more than 80%, The highest value identified is 77.8 .(Y.-R. Chen et al., 2021)
AUC	AUC should be greater than .5.
Geometric Mean	Geometric mean should be greater than .5.
Harmonic Mean	Harmonic mean should be greater than .5.
Precision	Precision should be more than .5.
Recall	Recall should be more than .5.
Specificity	Specificity should be more than .5.

To prepare the data for ML, the business age, business type, subprogram code, delivery method, and loan status are converted to numerical values.

- Business Age conversion is defined as zero if the field is blank, there is a change of ownership, the business is a startup, and/or the business is less than two years old. The Business Age is one for all remaining application categories.
- Business Type is zero for sole proprietor and one for all remaining business types.
- The subprogram code is converted to a numerical value from 1-17.
- The Delivery Method is converted to a numerical value 1-15.
- Loan Status is converted to zero for charge off and one for paid in full.

3.5.2 Hypothesis 2 Risk

The savings associated with implementing the model is based on the risk and expected loss. The expected loss for SBA will be measured by calculating the average cost of charge off times the number of charge offs that the model is expected to prevent. The cost of charge off is calculated by dividing the total cost of 7(a) charge offs from 1991-2022 ($19.1 billion) divided by the total number of charge off transactions from 1991-2022 (211665). The average cost of a charge off is approximately $90k.

The total number of transactions per year (40000) are based on the average total of PIF and charge off transactions from 1991-2022. On average 16 percent of the loans are charged off per year. The number of charge offs that are expected to be prevented by using the model is calculated based on the model's precision value for charge offs. The precision percentage will be multiplied by the total number of expected charge off transactions. The expected loss from charge offs equals 6400*Charge Off Precision*90000. A model that predicts charge offs saves taxpayers from the expected loss.

The cost of declining a potentially good loan is calculated using the formula devised by the FDIC's Federal Examiners (FDIC, 2011). Lenders earn a premium the first year which totals the loan approval amount * SBA guaranteed percent*ten percent. The annual service fee totals one percent of the guaranteed portion of the loan. The calculated average total of the loan term is 113 months or 9.4 years. The average gross approval amount of a paid in full loan is calculated by the paid in full gross approvals 1991-2022 ($218 billion) divided by the total number of paid in full transactions 1991-2022 (1072259) for a rounded down value of $203000 per loan. The average loan guaranteed percent is 75%.

*Loan Portion Guaranteed: $203000*75%=$152250*

*Average Premium Per Paid in Full Loan: .1*$152250=$15225*

*Servicing Fee Per Year = .01*$152250=$1523*

*Average Servicing Fee per Paid in Full loan=9*1523=$13707*

Average Premium and Servicing Fee Per Paid in Full Loan=$28932

The model should account for the loans that it marks as denials that would have resulted in paid in full loans. Based on 84 percent of the loans being paid in full, there are a total of 33600 loans per year that the model should classify as potentially paid in full. The one minus the paid in full loan precision rate equals the rate of loans that are denied that may have resulted in a $28932 benefit to a lender. The annual impact of denying loans that would have been paid in full totals (1-precision rate)*33600*28932. The total model savings is calculated from the charge off savings minus the paid in full denials.

3.5.3 Hypothesis 2 Testing

Testing hypothesis two includes reviewing fifteen models: Logistics Regression, Logistics Regression SMOTE, Logistics Regression Under Sampling, XG Boost, XG Boost SMOTE, XG Boost Under Sampling, KNN, KNN SMOTE, KNN Under Sampling, Decision Tree, Decision Tree SMOTE, Decision Tree Under Sampling, Random Forest, Random Forest SMOTE, and Random Forest Under Sampling. Fifteen models are generated utilizing the 2010-2019 7(a) loan data. The selected model will meet the minimum requirements for AUC, accuracy, precision, and specificity and have the highest annual expected Model Savings. The selected model is tested with the 7(a) general 2020-2022 loan data.

Table 16. Hypothesis 2 Criteria

Algorithm	AUC	Accuracy	Precision	Specificity	Annual Expected Charge Off Savings	Annual Expected Cost of Denied PIF Loans	Annual Expected Model Savings
Model	>=.8	>=.5	>=.5	>=.5	6400*90000*Charge Off Precision Rate	(1-Paid in Full Precision)*33600*28932	Annual Expected Charge Off Savings Annual Expected Cost of Denied PIF Loans

3.6 Summary

Classifying and modeling 7(a) loan charge offs with features and macroeconomic data can reduce the cost of risk to the SBA. Using 7(a) loan data and Minitab, the features for classifying charge offs and paid in full loans is defined to test hypothesis one. The results from hypothesis one along with macro-economic factors model charge offs and paid in full loans using Logistics Regression, XG Boost, RF, Decision Tree, and KNN. The selected model tested for hypothesis two risk savings is calculated based on the charge off precision.

Chapter 4—Results

4.1 Introduction

Chapter 4 includes results of the hypotheses explored in Chapter 3. Minitab, Excel, and Python ML libraries including Pandas, NumPy, Imblearn, Sklearn, and Matplotlib are used to evaluate the four hypotheses. Default predictors provide insight that may support innovative loan program policies to address weaknesses that lead to default.

H1: Loan application data features such as loan amount, interest rate, and term in months as well as the 7(a) loan subprogram code and delivery method are predictors of loan charge off risk.

H2: A small business loan charge off prediction risk model that includes loan features and historical economic data predict the risk of charge off for 7(a) loans to reduce guaranty fees paid by SBA.

4.2 Hypothesis 1 Results

Hypothesis 1: Loan application data features such as loan amount, interest rate, and term in months as well as the 7(a) loan subprogram code and delivery method are predictors of loan charge off risk. There are 13 and 11 predictors for Group 1 and 2, respectively. Subprogram is consistently one of the top features for predicting charge offs and defaults. It is noted that subprograms are less important in the 2020-2022 data. There are less traditional 2020-2022 7(a) loans since most small businesses applied for PPP loans during that period.

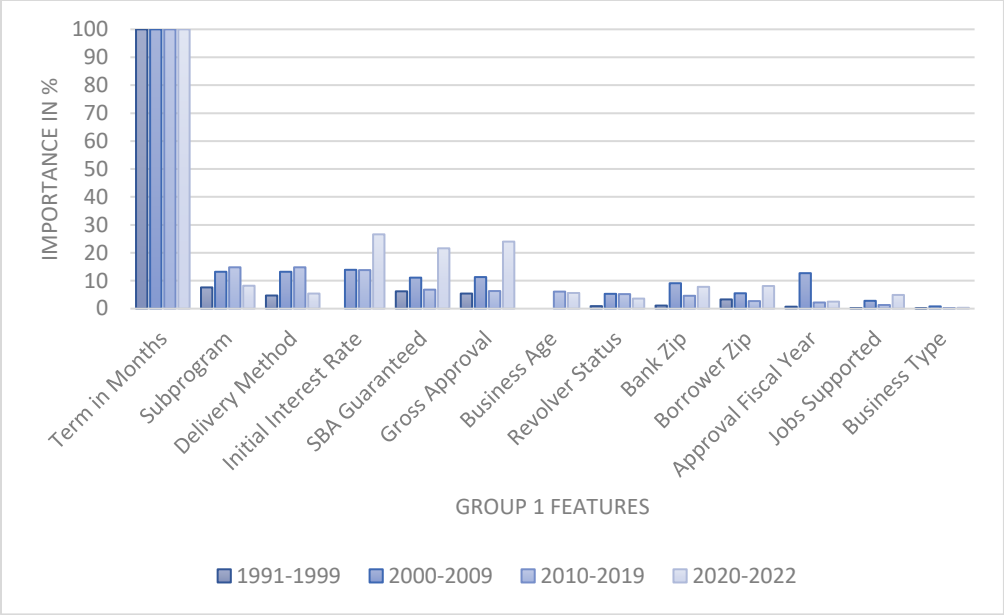


Figure 9. Group 1 7(a) Loans Features

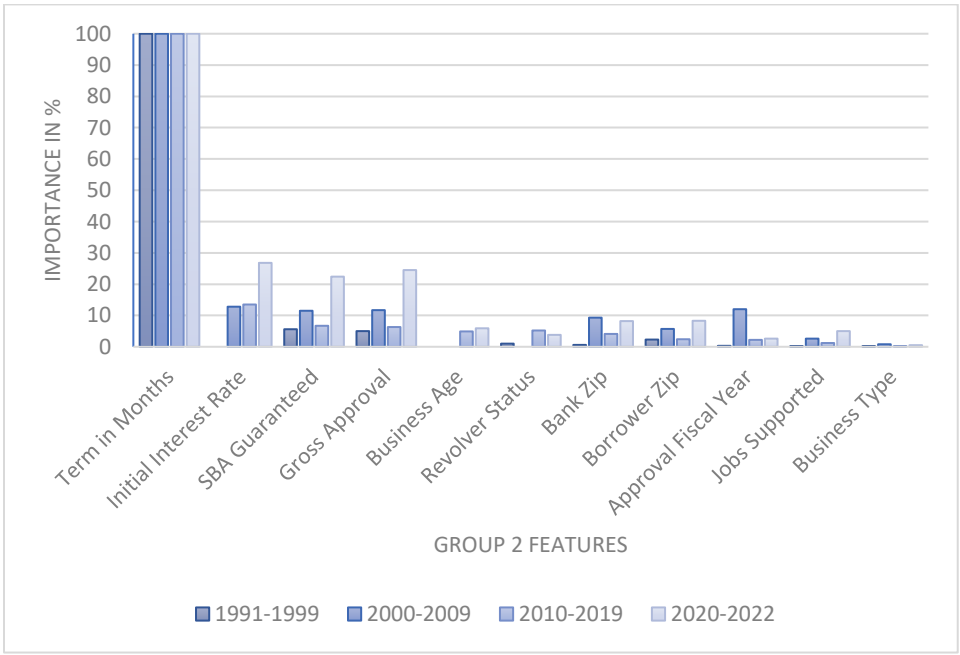


Figure 10. Group 2 7(a) Loans Features

The AUC, True Positive, False Positive, False Negative, True Negative, Upper Bound, and Lower Bound were compared for Group 1 and 2. The 7(a) 1991-1999 loan data Group 1 performed as expected apart from the True Negative (Specificity result).

Group 2 specificity is .9190 and Group 1 specificity is .9140. The 7(a) 2000-2009 and 2010-2019 results performed as expected. The 7(a) 2020-2023 data performed as expected for all metrics except True Negative/Specificity. For 2020-2023, Group 2 specificity is .947 and Group 1 is .935.

Group 1 features produce a more accurate 95% confidence interval with improved lower bound results. Group 1's lower bound is .17-2.62% more accurate than Group 2's lower bound. From 1991 to 2022, there have been a total of $19B in 7(a) charge off loans. Reducing the number of charge off/defaulted loans .17-2.62% per year over 31 years can result in an annual savings of $1-16M per year.

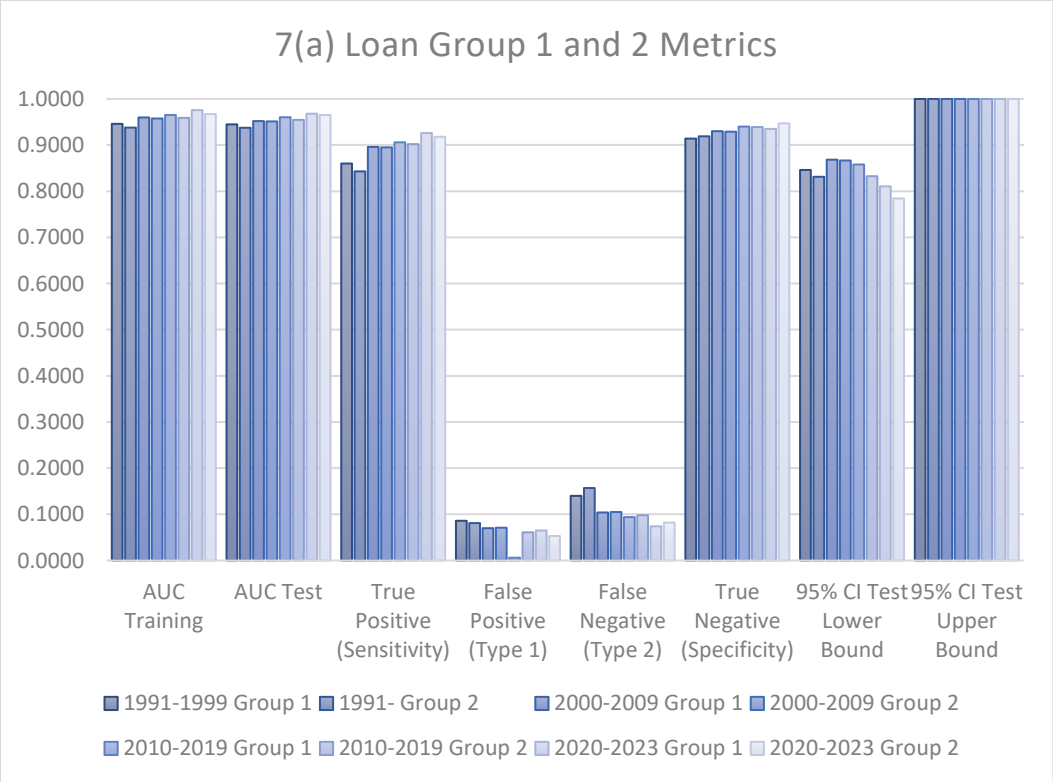


Figure 11. 1991-2023 7(a) Loan Group 1 and 2 Performance

The processing method for PPP indicates if the loan is a first or second draw. The requirements are the same for the first and second draw. The results indicate that processing method is a feature when included but does not improve the AUC, True Positive, False Positive, False Negative, True Negative, and/or Lower Bound results. Group 1 and 2 for PPP resulted in the same outcomes.

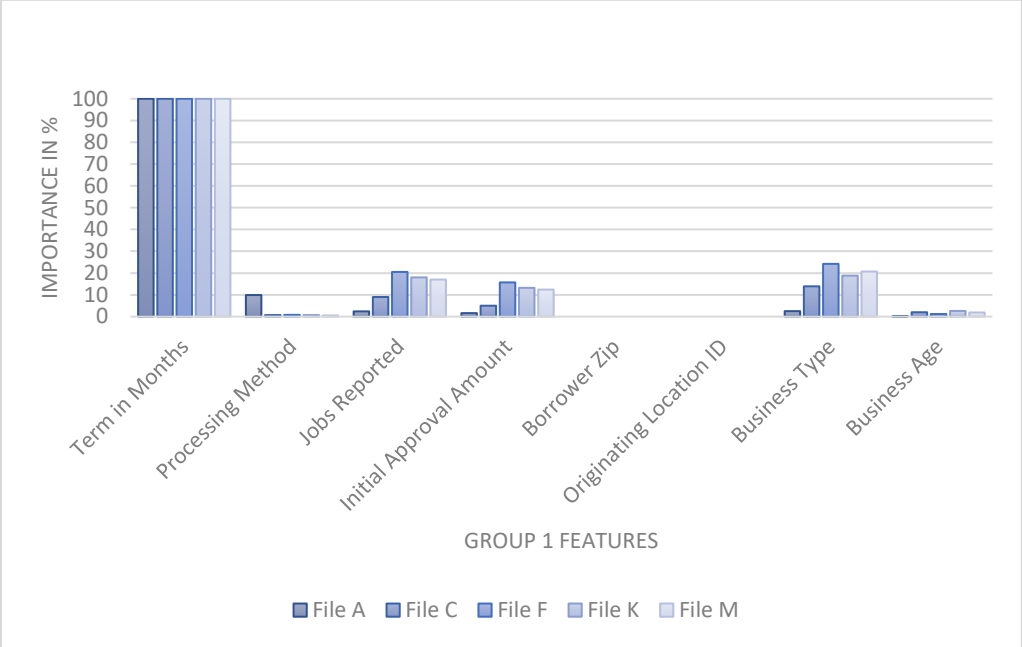


Figure 12. PPP Group 1 Features

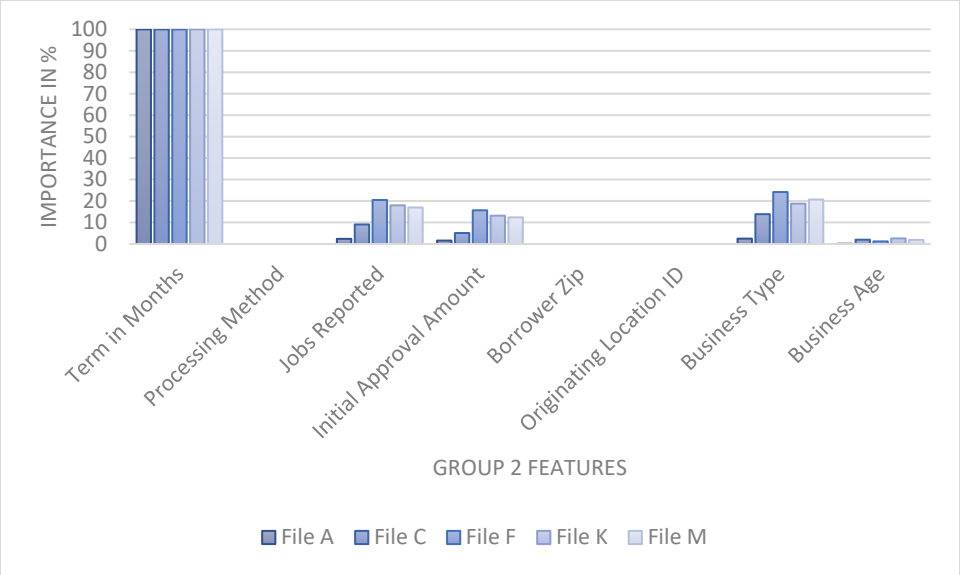


Figure 13. PPP Group 2 Features

A model inclusive of the subprogram code for 7(a) loans improves default vs. PIF prediction results. The improvement is minor but given the volume of loans the minor improvement reduces taxpayer cost of paying for defaults. Differentiating between the first and second draw for PPP loans does not result in improved prediction results due to both draws having the same terms and requirements.

The analysis validates hypothesis one: Loan application data features such as loan amount, interest rate, and term in months as well as the 7(a) loan subprogram code and delivery method are predictors of loan charge off risk. The minimal improvement of the prediction when including subprogram code and delivery method may be considered survivorship bias. Additional research should be conducted before definitively stating that delivery method and subprogram code improve prediction of charge off loans.

4.3 Hypothesis 2 Results

Hypothesis 2: A small business loan charge off prediction risk model that includes loan features and historical economic data predict the risk of charge off for 7(a) loans to reduce guaranty fees paid by SBA. The features identified in section 3 are modeled using Logistic Regression, XG Boost, KNN, Decision Tree, and RF. Each algorithm is generated using the 2010 to 2019 7(a) loan data with an 80/20 training/test split. Each algorithm is generated with the data, SMOTE, and under sampling. The accuracy, precision, AUC, Geometric Mean, Harmonic Mean, Recall, and Specificity are compared for each model.

4.3.1 Economic Features

The 3-year running average of inflation, GDP, and unemployment are generated in comparison with the Group 1 Features for Logistics Regression, XG Boost, and Decision Tree. Due to the processing time required for RF and KNN, the shap summaries are not generated for this research. The logistics regression model notable features based on the shap summaries include Term in Months, SBA Guaranteed Approval, and Gross Approval. The economic features are the least important in the logistics regression model. Within the three economic features, the unemployment rate is the most important economic factor in the model with inflation having the lease influence on predicting the loan status.

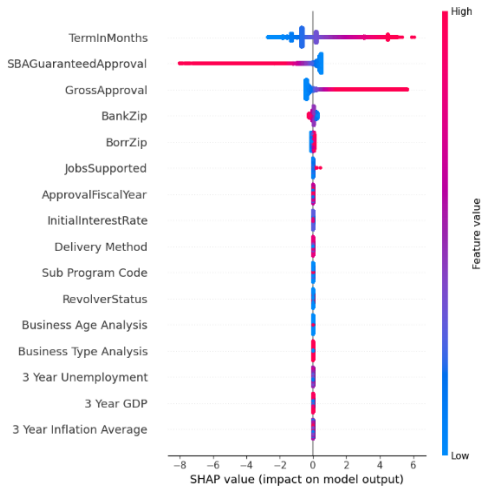


Figure 14. Logistics Regression SHAP Summary

The XGBoost model's top three features include Term in Months, Initial Interest Rate, and Bank Zip. Unemployment is twelve out of 16th on the feature rankings and is the highest economic feature. GDP is the lowest economic feature ranking 15th out of 16th.

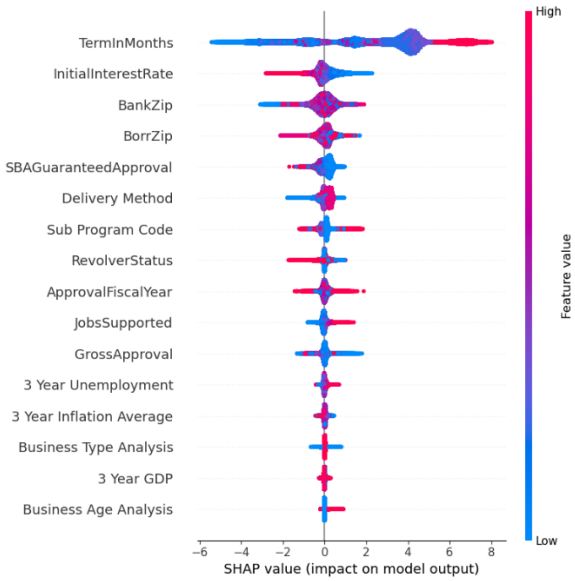


Figure 15. XG Boost SHAP Summary

Decision Tree model is most influenced by Term in Months, Revolver Status, and SBA Guaranteed Approval. Unemployment is the most important economic factor, ranking twelve out of sixteen. GDP is the lowest ranking economic factor ranking 14th out of 16th.

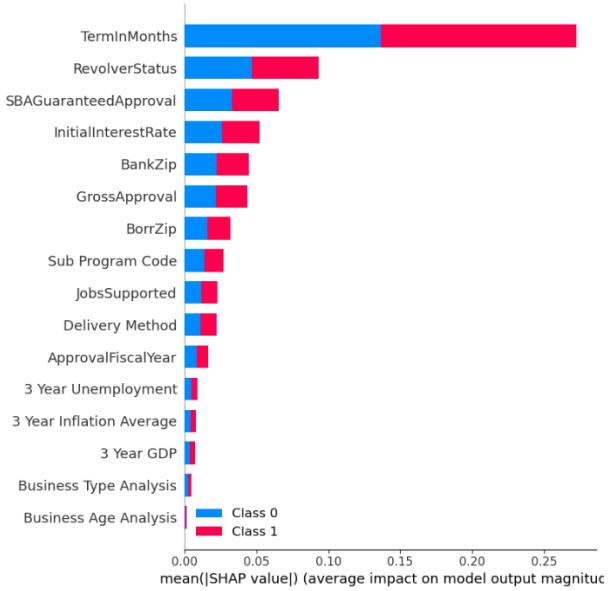


Figure 16. Decision Tree SHAP Summary

4.3.2 Models

Based on the Shapley values Group 1 features Jobs Supported, Revolver Status, Business Age, and Business Type are removed from the final model. Although previous research indicates that the number of jobs supported by a small business indicates the likelihood of the business repaying a loan, SBA 7(a) loans paid in full behavior does not align with the research. Revolver status indicates whether a loan is a line of credit or not. The prediction of charge off is not highly correlated to whether a loan is a revolver.

Previous research indicates that more mature businesses have lower defaults. The Shapley summaries consistently indicate that business age is not a significant influencer on a business paying a loan off. SBA loan underwriting and lender of last resort criteria

may reduce the correlation of business age with paying a loan in full. Like business age, business type is not a differentiator for SBA loans. A loan likely to be paid in full is not influenced by whether a business is set up as a sole proprietor, LLC, or LLP.

The final 7(a) model features include Borrower Zip Code, Bank Zip Code, Gross Approval Amount, SBA Guaranteed Approval Amount, Initial Interest Rate, Term in Months, Approval Fiscal Year, Sub Program Code, Delivery Method, and 3 Year Unemployment. The selected features will be modeled fifteen times. The models include Logistics Regression, Logistics Regression SMOTE, Logistics Regression Under Sampling, KNN, KNN SMOTE, and KNN Under Sampling, XGBoost, XGBoost SMOTE, XG Boost Under Sampling, Decision Tree, Decision Tree SMOTE, Decision Tree Under Sampling , Random Forest, Random Forest SMOTE, and Random Forest Under Sampling.

4.3.2.1 Logistics Regression

The logistics regression model has an accuracy and precision of .92. The low specificity (.02) and AUC (.51) indicate that the model is not ideal for predicting charge offs. Logistics SMOTE improves the AUC to .74 and precision .92 but the accuracy reduces from the original model to .73. Logistics under sampling AUC remains .74 and the precision reduces to .76. Despite logistics regression under sampling having the lower metrics, the model has the best charge off and paid precision resulting in an overall savings of $181M per year with predicted charge offs.

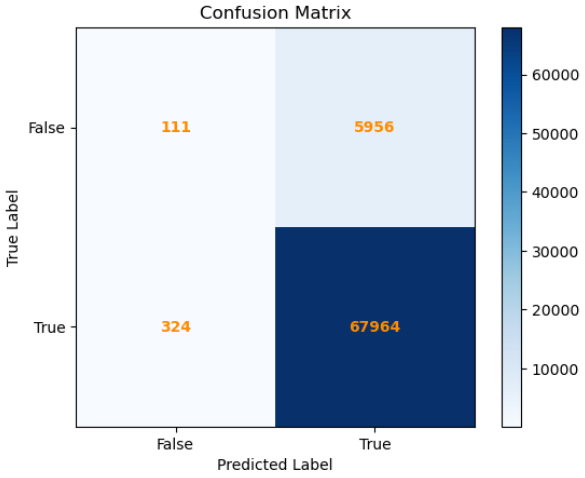


Figure 17. Logistics Regression Confusion Matrix

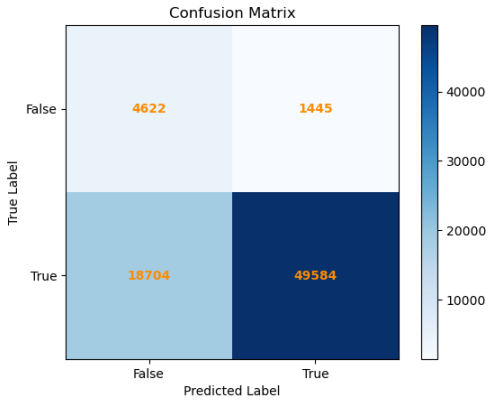


Figure 18. Logistics Regression SMOTE Confusion Matrix

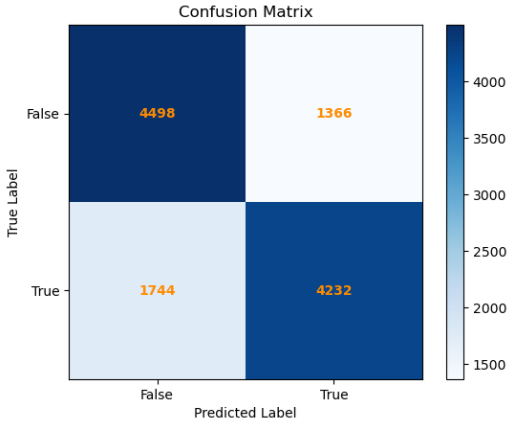


Figure 19. Logistics Regression Under Sampling Confusion Matrix

Table 17. Logistics Regression Cost Model

Model	Charge Off Precision	Paid in Full Precision	1-Paid in Full Precision	Annual Expected Charge Off Savings	Annual Expected Cost of Denied PIF Loans	Annual Expected Model Savings
Logistics Regression	0.26	0.92	0.08	$149,760,000	$77,769,216	$71,990,784
Logistics Regression SMOTE	0.2	0.97	0.03	$115,200,000	$29,163,456	$86,036,544
Logistics Regression Under Sampling	0.72	0.76	0.24	$414,720,000	$233,307,648	$181,412,352

4.3.2.2 XG Boost

The XG Boost model reflects acceptable AUC (.84), accuracy (.96), precision (.97), and specificity (.71). XG Boost SMOTE also indicates strong metrics with AUC (.89), accuracy (.96), precision (.98), and specificity (.8). XG Boost under sampling metrics improve with AUC (.93), accuracy (.92), precision (.99), and specificity (.94). XG Boost has the best charge off precision metric compares to XG Boost SMOTE and under sampling with an expected annual savings of $437M.

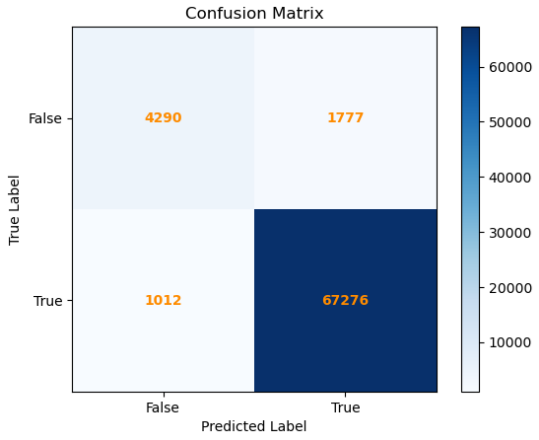


Figure 20. XG Boost Confusion Matrix

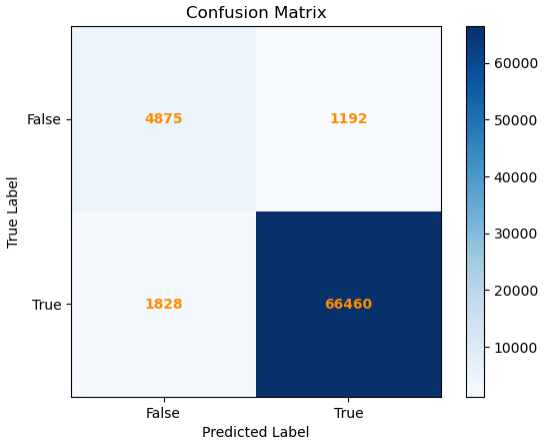


Figure 21. XG Boost SMOTE Confusion Matrix

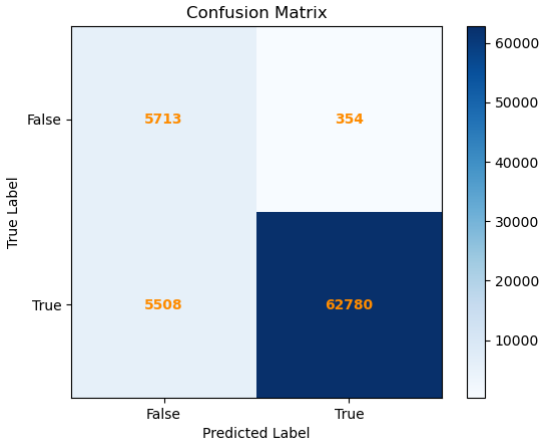


Figure 22. XG Boost Under Sampling Confusion Matrix

Table 18. XG Boost Cost Savings

Model	Charge Off Precision	Paid in Full Precision	1-Paid in Full Precision	Annual Expected Charge Off Savings	Annual Expected Cost of Denied PIF Loans	Annual Expected Model Savings
XGBoost	0.81	0.97	0.03	$466,560,000	$29,163,456	$437,396,544
XGBoost SMOTE	0.73	0.98	0.02	$420,480,000	$19,442,304	$401,037,696
XGBoost Under Sampling	0.51	0.99	0.01	$293,760,000	$9,721,152	$284,038,848

4.3.2.3 KNN

KNN has a low AUC (.57) and specificity (.15) make it a less than ideal model for predicting charge offs. KNN does score well for accuracy (.91) and precision (.93). The accuracy (.85) and precision (.88) decrease for KNN SMOTE compared to KNN. KNN SMOTE's AUC (.85) and specificity (.89) improved compared to KNN. KNN under sampling metrics are not satisfactory with low AUC (.61), accuracy(.61), precision(.61), and specificity (.61) metrics. KNN SMOTE has a higher savings than KNN and KNN under sampling with a total savings of $361M per year.

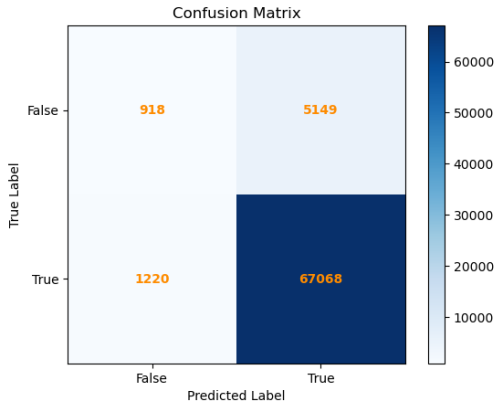


Figure 23. KNN Confusion Matrix

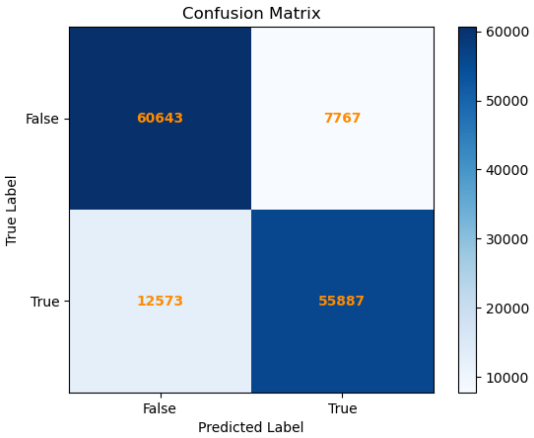


Figure 24. KNN SMOTE Confusion Matrix

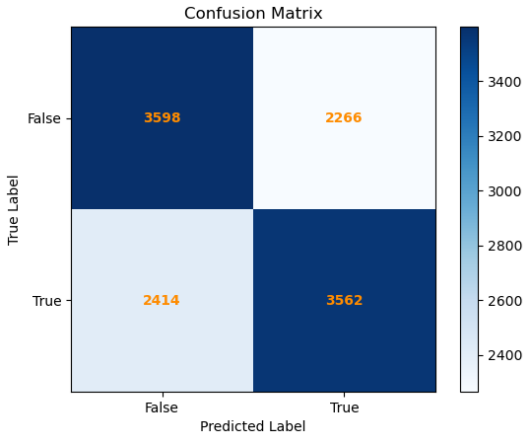


Figure 25. KNN Under Sampling Confusion Matrix

Table 19. KNN Cost Savings

Model	Charge Off Precision	Paid in Full Precision	1-Paid in Full Precision	Annual Expected Charge Off Savings	Annual Expected Cost of Denied PIF Loans	Annual Expected Model Savings
KNN	0.43	0.93	0.07	$247,680,000	$68,048,064	$179,631,936
KNN SMOTE	0.83	0.88	0.12	$478,080,000	$116,653,824	$361,426,176
KNN Under Sampling	0.6	0.61	0.39	$345,600,000	$379,124,928	-$33,524,928

4.3.2.4 Decision Tree

Decision Tree accuracy (.94), precision (.97), and AUC (.82) are adequate for the model. Specificity (.66) is low for the Decision Tree model. Decision Tree SMOTE metrics are like Decision Tree with accuracy (.93), precision (.97), AUC (.82), and specificity (.69). Decision Tree under sampling accuracy (.89), precision (.89), and AUC (.89) are lower and specificity (.89) is higher than for Decision Tree and Decision Tree SMOTE. The Decision Tree Under Sampling charge off precision (.89) and paid in full precision (.89) has the best cost savings at $406M per year.

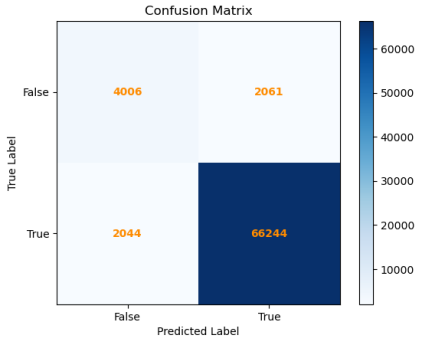


Figure 26. Decision Tree Confusion Matrix

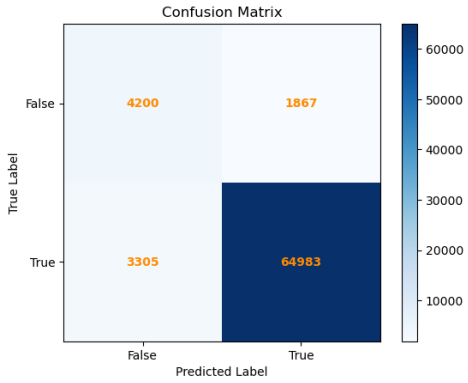


Figure 27. Decision Tree SMOTE Confusion Matrix

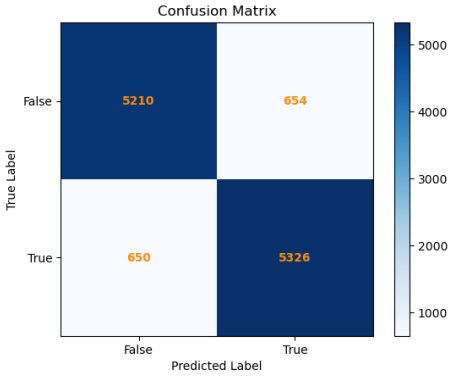


Figure 28. Decision Tree Under Sampling Confusion Matrix

Table 20. Decision Tree Cost Savings

Model	Charge Off Precision	Paid in Full Precision	1-Paid in Full Precision	Annual Expected Charge Off Savings	Annual Expected Cost of Denied PIF Loans	Annual Expected Model Savings
Decision Tree	0.66	0.97	0.03	$380,160,000	$29,163,456	$350,996,544
Decision Tree SMOTE	0.56	0.97	0.03	$322,560,000	$29,163,456	$293,396,544
Decision Tree Under Sampling	0.89	0.89	0.11	$512,640,000	$106,932,672	$405,707,328

4.3.2.5 Random Forest

Random Forest accuracy (.96), precision(.97), and AUC (.83) metrics are ideal. Specificity (.67) is low. Random Forest SMOTE accuracy (.96), precision (.98), AUC (.86), and specificity (.75) are acceptable. Random Forest under sampling accuracy (.92), precision (.93), AUC(.92), and specificity (.93) are the ideal performers. Random Forest under sampling has the best cost savings for a total of $456M per year.

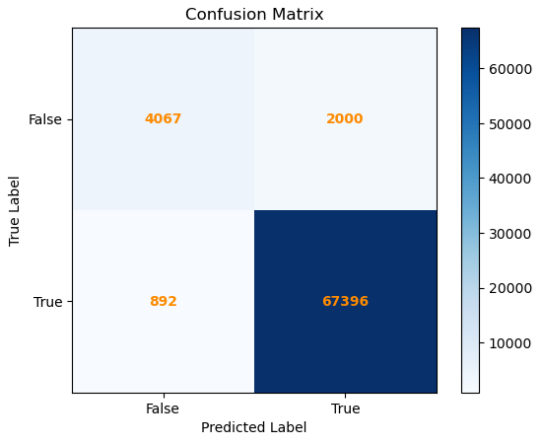


Figure 29. Random Forest Confusion Matrix

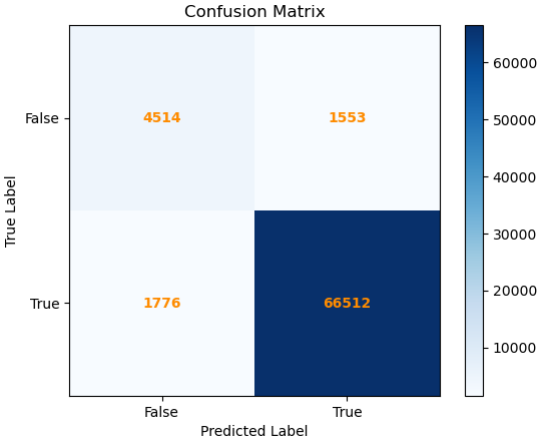


Figure 30. Random Forest SMOTE Confusion Matrix

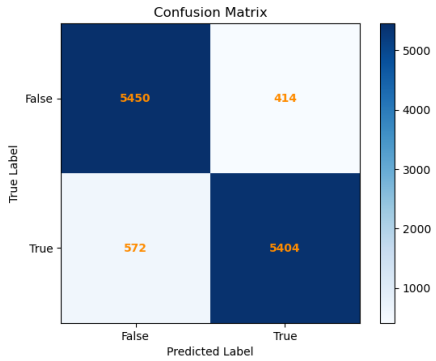


Figure 31. Random Forest Under Sampling Confusion Matrix

Table 21. Random Forest Cost Savings

Model	Charge Off Precision	Paid in Full Precision	1-Paid in Full Precision	Annual Expected Charge Off Savings	Annual Expected Cost of Denied PIF Loans	Annual Expected Model Savings
Random Forest	0.82	0.97	0.03	$472,320,000	$29,163,456	$443,156,544
Random Forest SMOTE	0.72	0.98	0.02	$414,720,000	$19,442,304	$395,277,696
Random Forest Under Sampling	0.91	0.93	0.07	$524,160,000	$68,048,064	$456,111,936

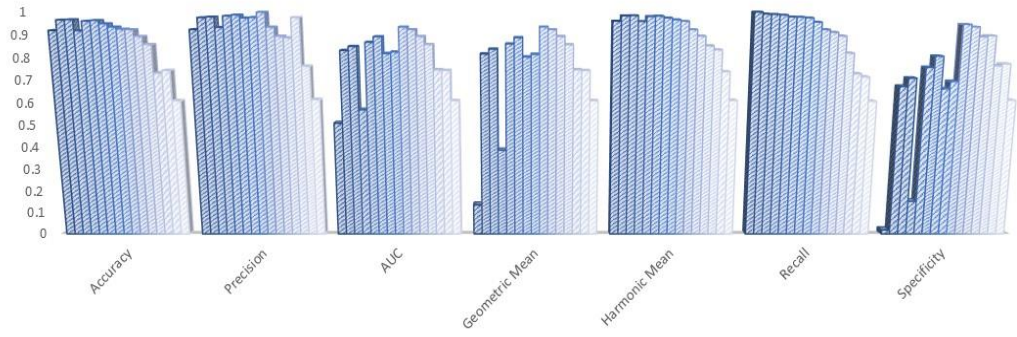
4.3.3 Performance of Models

Accuracy indicates the model's capability to classify loans as paid in full or charge off. XG Boost's accuracy score of .962 indicates that it is the best model for predicting the classification of loans as charge off or paid in full. Accuracy indicates the number of times that the classification is correct. Since the cost savings model also includes incorrect predictions, accuracy is not the only metric for selecting the best model.

Precision determines the best model for predicting the target group. In this model, paid in full is the target group. XG Boost under sampling with a precision of .99 is the best model for identifying loans that will be paid in full. The precision metric does not account for false positives.

AUC indicates the overall performance of a model by addressing false positives and true positives. XG Boost under sampling is the best overall model with an AUC of

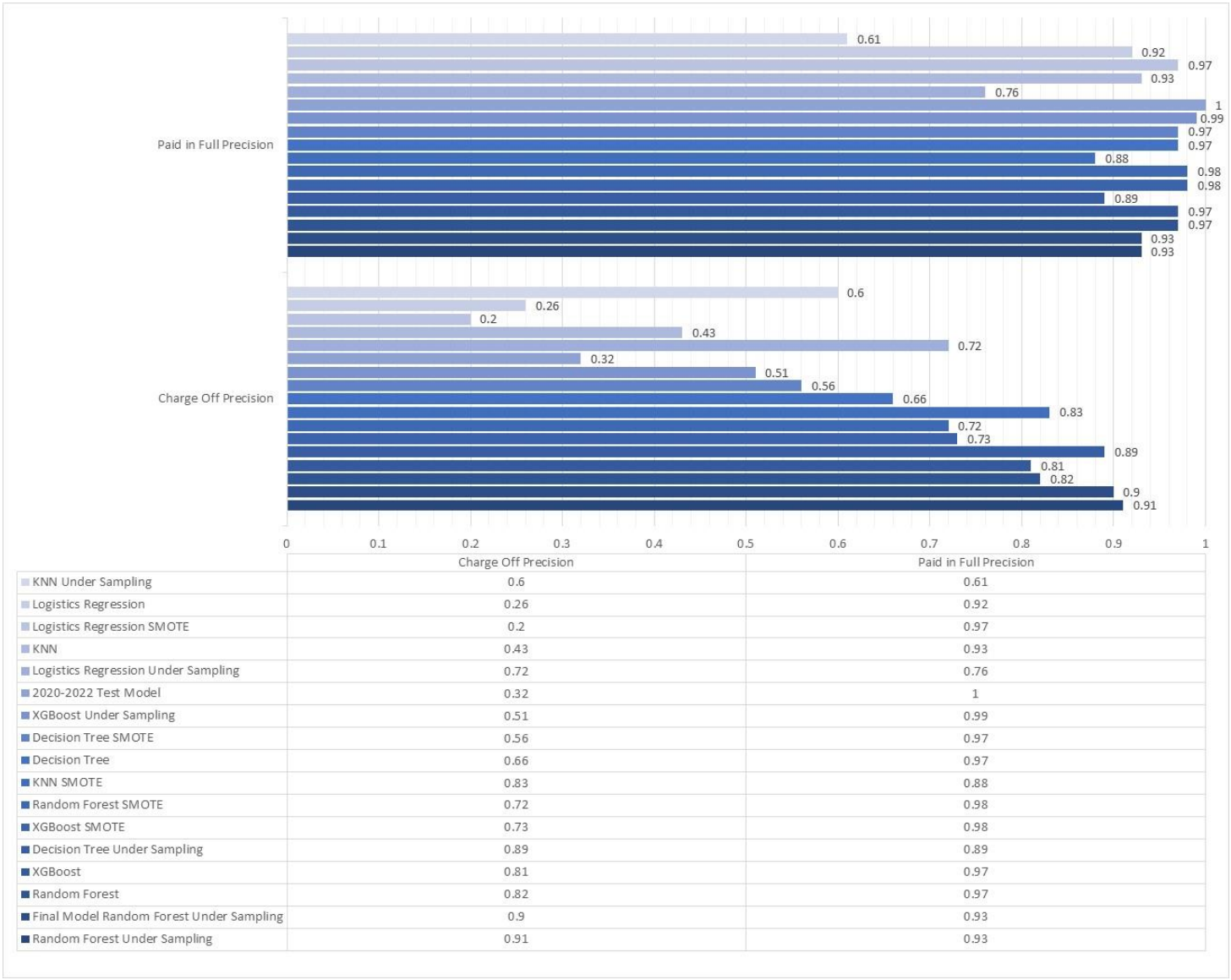
.93. Recall denotes the model's performance for finding all the target objects. Logistics regression with a recall of .99 is the best for finding all the paid in full objects. Specificity identifies the model best for identifying true negatives. XG Boost under sampling with a specificity of .94 is the best model for identifying true charge offs.



	Accuracy	Precision	AUC	Geometric Mean	Harmonic Mean	Recall	Specificity
Logistics Regression	0.915540313	0.919426407	0.506775543	0.134940328	0.955839334	0.995255389	0.018295698
Random Forest	0.961038262	0.971204981	0.828756302	0.81354046	0.97895769	0.986835169	0.670677435
XGBoost	0.962490754	0.974266143	0.846142209	0.834640651	0.979692881	0.985180412	0.707104005
KNN	0.914343353	0.928700998	0.566722428	0.385495954	0.954670652	0.982134489	0.151310368
Random Forest SMOTE	0.956169726	0.978077076	0.864176963	0.857156127	0.976089333	0.974109653	0.754244272
XGBoost SMOTE	0.959384036	0.982380417	0.88837915	0.884317632	0.977784317	0.973231022	0.803527279
Decision Tree	0.944791877	0.969826513	0.815180669	0.80033084	0.969947215	0.970067948	0.66029339
Decision Tree SMOTE	0.930441799	0.972071803	0.821935847	0.811643527	0.961728011	0.951602038	0.692269656
XGBoost Under Sampling	0.921161993	0.994392879	0.930496585	0.930429718	0.955395596	0.919341612	0.941651558
Random Forest Under Sampling	0.919087838	0.930360206	0.919197253	0.919124475	0.91885482	0.907630522	0.930763984
Decision Tree Under Sampling	0.889864865	0.890635452	0.889851813	0.889850743	0.890933423	0.891231593	0.888472033
KNN SMOTE	0.854665011	0.883888204	0.854678875	0.853835752	0.848980398	0.816725095	0.892632656
Logistics Regression SMOTE	0.729016206	0.971682769	0.743963746	0.743749276	0.831130518	0.726101218	0.761826273
Logistics Regression Under Sampling	0.740878378	0.759364966	0.741150824	0.740590995	0.735106199	0.712349398	0.769952251
KNN Under Sampling	0.606672297	0.611722853	0.606695427	0.606690499	0.607963633	0.604250335	0.609140518

Figure 32. Model Metrics

After comparing the metrics of each model, the cost savings associated with each model is reviewed for model selection. To calculate savings from predicting charge offs, the models' performance for precision with predicting charge off and paid in full loans is evaluated. Random Forest Under Sampling is the best model for predicting charge offs with a precision rate of .91. XG Boost under sampling is the best model for predicting paid in full with a precision rate of .99.



	Charge Off Precision	Paid in Full Precision
KNN Under Sampling	0.6	0.61
Logistics Regression	0.26	0.92
Logistics Regression SMOTE	0.2	0.97
KNN	0.43	0.93
Logistics Regression Under Sampling	0.72	0.76
2020-2022 Test Model	0.32	1
XGBoost Under Sampling	0.51	0.99
Decision Tree SMOTE	0.56	0.97
Decision Tree	0.66	0.97
KNN SMOTE	0.83	0.88
Random Forest SMOTE	0.72	0.98
XGBoost SMOTE	0.73	0.98
Decision Tree Under Sampling	0.89	0.89
XGBoost	0.81	0.97
Random Forest	0.82	0.97
Final Model Random Forest Under Sampling	0.9	0.93
Random Forest Under Sampling	0.91	0.93

Figure 33. Paid in Full & Charge Off Precision Calculations

Evaluating the model with savings associated with correctly predicting charge offs and minimal cost associated with denying potentially viable paid in full loans determines the preferred model. Random Forest under sampling is the best performer financially with expected savings of $450M per year and Random Forest is a close second with an expected savings of $443M per year. Although XG Boost under sampling has a high specificity score, precision of identifying charge offs results in the model not a top five performer for cost savings.

Table 22. Top Five Performing Models Based on Cost Savings

Model	Charge Off Precision	Paid in Full Precision	1-Paid in Full Precision	Annual Expected Charge Off Savings	Annual Expected Cost of Denied PIF Loans	Annual Expected Model Savings
Random Forest Under Sampling	0.91	0.93	0.07	$524,160,000	$68,048,064	$456,111,936
Random Forest	0.82	0.97	0.03	$472,320,000	$29,163,456	$443,156,544
XGBoost	0.81	0.97	0.03	$466,560,000	$29,163,456	$437,396,544
Decision Tree Under Sampling	0.89	0.89	0.11	$512,640,000	$106,932,672	$405,707,328
XGBoost SMOTE	0.73	0.98	0.02	$420,480,000	$19,442,304	$401,037,696

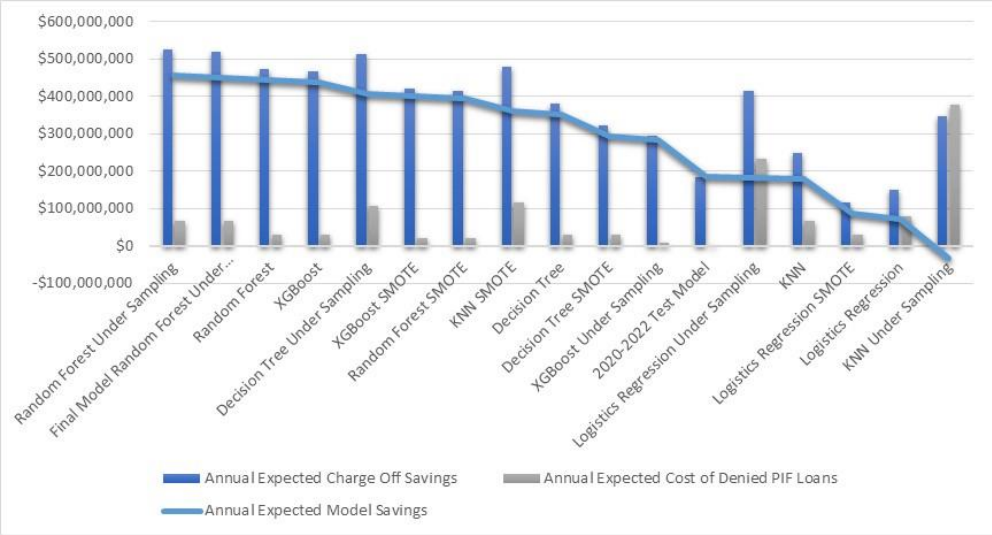


Figure 34. Model Cost Savings

4.3.4 Selected Model

The final Random Forest Under sampling model is saved as a python pickle file generated from the 2010-2019 7(a) data. The final model saves $450M per year which is less than the expected $456M per year. Note when comparing the actuals the total is $286M per year (actuals $405M with the FNs costing $119M). The False Negative loans that the model predicted from 2010-2019 had an estimated value of $38,604 for the premium and servicing fee compared to the cost model estimate of $28,932.

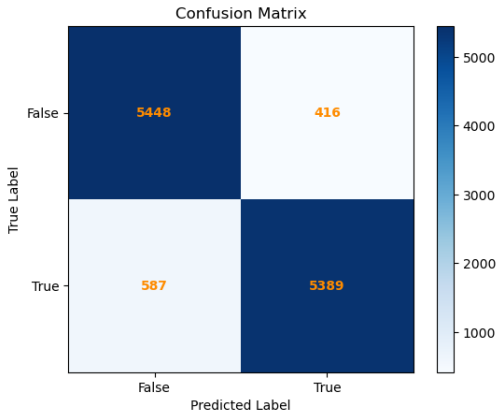


Figure 35. Selected Model Confusion Matrix

Table 23. Selected Model Savings

Model	Charge Off Precision	Paid in Full Precision	1-Paid in Full Precision	Annual Expected Charge Off Savings	Annual Expected Cost of Denied PIF Loans	Annual Expected Model Savings
Random Forest Under Sampling	0.91	0.93	0.07	$524,160,000	$68,048,064	$456,111,936
Final Model Random Forest Under	0.9	0.93	0.07	$518,400,000	$68,048,064	$450,351,936
2010-2019 Actuals Random Forest	0.9	0.93	0.07	$404,722,451	$119,167,319	$285,555,132

The reusable model requires loan application and unemployment formatted data. To format the data, convert the Delivery Method and Subprogram Description to numerical values. The data format template includes an identifier/borrower name so that the results can be tracked. The final model also includes the borrower zip, bank zip, Gross Approval, SBA Guaranteed Approval, Approval Fiscal Year, Delivery Method, Subprogram Code, Initial Interest Rate, Term in Months, and 3 Year Unemployment

Rate. The code creates an output file that includes the input items as well as a prediction one for paid in full and zero for charge off. The model also calculates the probability of being paid in full.

Table 24. Model File Format

BorrName	BorrZip	BankZip	GrossApproval	SBAGuaranteedApproval	ApprovalFiscalYear	DeliveryMethod	subpgmdesc	InitialInterestRate	TermInMonths	3 Year Unemployment	Sub Program Code	Delivery Method
Sample Business	99208	59101	142000	71000	2016	SBA EXPRES	FA$TRK (Small Loan Express)	4.75	84	0.062733333	6	14
Information Needed to Identify Loan	Borrower Zip Code	Bank Zip Code	Gross Approval	SBA Guaranteed Approval	Approval Fiscal Year/Current Fiscsal Year	Required to calculate Delivery Method	Required to Calculate Sub Program Code	Initial Inteest	Term in Months	3 year unemployment	Numeric Version of Subprogram Code	Numeric Portion of Delivery Method

Table 25. Sample Output File

BorrName	BorrZip	BankZip	GrossApproval	SBAGuaranteedApproval	ApprovalFiscalYear	DeliveryMethod	subpgmdesc	InitialInterestRate	TermInMonths	3 Year Unemployment	Sub Program Code	Delivery Method	Loan Status	Predictions	Probabilities
A	99208	59101	142000	71000	2016	SBA EXPRES	FA$TRK (Small Loan Express)	4.75	84	0.062733333	6	14	1	1	0.97
B	85204	57104	2498000	1873500	2016	PLP	Guaranty	5.55	300	0.062733333	7	12	1	1	0.99
C	28779	57104	5000	2500	2016	SBA EXPRES	FA$TRK (Small Loan Express)	8	84	0.062733333	6	14	1	1	1
D	59901	53066	3383100	2537325	2016	OTH 7A	Guaranty	5.5	300	0.062733333	7	10	1	1	0.95
E	89131	57104	5000	2500	2016	SBA EXPRES	FA$TRK (Small Loan Express)	10	84	0.062733333	6	14	1	1	1
F	83815	99216	1000000	750000	2016	OTH 7A	Standard Asset Based	4.75	120	0.062733333	17	10	1	1	0.91

The model is tested with the 2020-2022 7(a) loan data. The AUC totals .93 but the charge off precision is .32. The paid in full precision of one does not have an expected loss so the model would save $184M per year from 2020-2022 for a total of $552M over three years. The Random Forest Under Sampled Model is generated with the 2020-2022 data. The model precision of .32 for charge off and 1 for paid in full results in a savings of $184M per year 2020 – 2022. The actual identified loans total $65.5M or an average of $21.8M.

The loan charge off amount and volume from 2020-2022 is less than expected. On average SBA loans take 4.7 years to default(Voigt & Weiner Campbell, 2017). The model may be accurate as loans default over time.

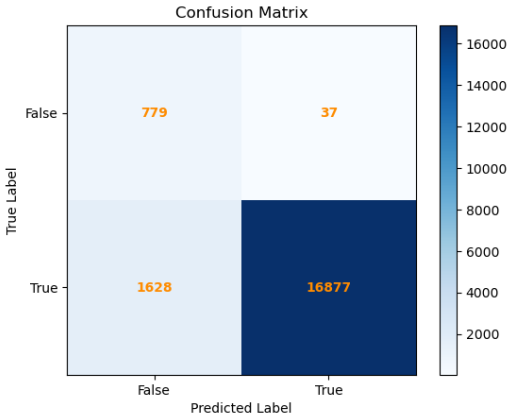


Figure 36. 2020-2023 Confusion Matrix

Table 26. Test Model Savings

Model	Charge Off Precision	Paid in Full Precision	1-Paid in Full Precision	Annual Expected Charge Off Savings	Annual Expected Cost of Denied PIF Loans	Annual Expected Model Savings
2020-2022 Test Model	0.32	1	0	$184,320,000	$0	$184,320,000

Chapter 5—Discussion and Conclusions

5.1 Discussion

Hypothesis one validates that not all 7(a) loan subprograms and delivery methods should be evaluated as equal. There are underwriting policies that reduce the risk of some subprograms and delivery methods. The research validates that future research should consider the differences in the loan subprograms and methods.

Hypothesis two provided additional information for the model features as the Shapley values were reviewed. The analysis for hypothesis two also validated that economic factors such as unemployment may improve charge off prediction. Hypothesis two contributes a reusable model to the SBA lending community to predict loan charge offs and lower risk.

5.2 Conclusions

Predicting loan charge offs provides a tool to support SBA lenders and policy makers to identify risks that may be mitigated with additional underwriting or fee structures. Machine learning predicts charge off loans using loan application, loan processing/delivery, and unemployment data. Logistics Regression, XG Boost, KNN, Decision Tree, and Random Forest predict charge offs. The borrower and lender zip codes highlight the importance of considering local factors when modeling features. Random forest is selected as the best model which is like the results with the IMF study.

The models are compared for savings associated with predicting charge off as well as the cost of identifying false negatives. Except for KNN Under sampling, the models save $71-456M per year. KNN Under Sampling has a cost of $33M per year because 39% of the time it incorrectly identifies loans that would be paid in full as charge offs.

Currently, SBA does not predict loan charge offs prior to origination. SBA, like most lending organizations, relies on credit worthiness data to make loan decisions. The loan prediction charge off model is a tool for reducing small business lending risk and improving SBA's fiduciary duty to taxpayers. The model can be used to explore innovative loan policies such as a dynamic loan guaranty program. A dynamic loan guaranty program would attract more lenders to SBA by increasing the amount guaranteed by SBA for loans likely to be paid in full.

5.3 Contributions to Body of Knowledge

This praxis provides a 7(a)-loan charge off model to reduce SBA lending risk and improve cost savings. SBA currently relies on credit scores and public data to originate loans. ML provides a model to predict 7(a) loan charge off based on loan application, program/processing method, and cohort year. Praxis validates loan application data features such as loan amount, interest rate, and term in months as well as the 7(a) loan subprogram code and delivery method are predictors of loan charge off risk.

Previous research identifies the impact of 7(a) loans on the economy. This praxis provides insight on the impact of economic data on loan default predictions. A small business loan charge off prediction risk model that includes loan features and historical

economic data predict the risk of charge off for 7(a) loans to reduce guaranty fees paid by SBA. This praxis contributes a ML reusable model based on loan application and economic features to predict loan charge offs.

The analysis identifies a method to calculate cost savings and expected loss for SBA's implementation of charge off prediction. The model's precision with predicting charge off and paid in full loans generates an expected savings and loss. Expected savings is derived from the model correctly predicting charge offs. Expected loss is calculated from the precision in predicting paid in full loans. The selected model has expected savings of up to $450M per year. SBA does not currently use a charge off prediction model for decisioning. This research enhances data for decision making.

5.4 Recommendations for Future Research

The research limitations may be addressed in future research.

1. Future research may address second and third-party default risk factors. ML money laundering and synthetic fraud detection model improves online lending default prediction.
2. A model that differentiates between credit and fraud risk would improve the policies for loan decisioning.
3. The risk model can be enhanced with reducing features that target specific groups that have historically been impacted by unfair lending practices.
4. Develop a model based on local economic indicators.

References

7(a) & 504 FOIA. (2023). [dataset]. https://data.sba.gov/en/dataset/7-a-504-foia

Agarwal, S., Chomsisengphet, S., & Liu, C. (2008). *1—Determinants of small business default* (pp. 1–12). Academic Press. https://doi.org/10.1016/B978-075068158-2.50004-4

Angelini, E., di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, *48*(4), 733–755. https://doi.org/10.1016/j.qref.2007.04.001

Angell, M. (2023). Exclusive: New Senate Report Shows What May Befall Small Businesses If the U.S. Defaults. *Inc*. https://www.inc.com/melissa-angell/exclusive-new-senate-report-shows-what-may-befall-small-businesses-if-us-defaults.h

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Batiz-Zuk, E., López-Gallo, F., Mohamed, A., & Sánchez-Cajal, F. (2022). Determinants of loan survival rates for small and medium-sized enterprises: Evidence from an emerging economy. *International Journal of Finance and Economics*, *27*(4), 4741–4755. https://doi.org/10.1002/ijfe.2397

Baum, N., & Hsueh, W. (2008). *SBA lender oversight: Preventing loan fraud and improving regulation of lenders: Hearing before the Committee on Small Business and Entrepreneurship, United States Senate, One Hundred Tenth Congress, first session, November 13, 2007*. U.S. G.P.O.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). *A Comparative Analysis of XGBoost*.

Bitetto, A., Cerchiello, P., Filomeni, S., Tanda, A., & Tarantino, B. (2021). Machine Learning and Credit Risk: Empirical Evidence from SMEs. *IDEAS Working Paper Series from RePEc*. ProQuest Central. http://proxygw.wrlc.org/login?url=https://www.proquest.com/working-papers/machine-learning-credit-risk-empirical-evidence/docview/2587465818/se-2

Bouteille, S., & Coogan-Pushner, D. (2022). *The handbook of credit risk management: Originating, assessing, and managing credit exposures* (Second edition.). John Wiley & Sons, Incorporated.

Brighi, P., Lucarelli, C., & Venturelli, V. (2019). Predictive Strength of Lending Technologies in Funding SMEs. *Journal of Small Business Management*, *57*(4), 1350–1377. https://doi.org/10.1111/jsbm.12444

Brotcke, L. (2022). Time to Assess Bias in Machine Learning Models for Credit Decisions. *Journal of Risk and Financial Management*, *15*(4). https://doi.org/10.3390/jrfm15040165

Byanjankar, A., Heikkilä, M., & Mezei, J. (2015). *Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach*. https://doi.org/10.1109/SSCI.2015.109

Carroll, P. (2010). Rethinking Underwriting. *The RMA Journal*, *93*(4), 28.

Caselli, S., Gatti, S., & Querci, F. (2008). The Sensitivity of the Loss Given Default Rate to Systematic Risk: New Empirical Evidence on Bank Loans. *Journal of*

Financial Services Research, *34*(1), 1–34. https://doi.org/10.1007/s10693-008-0033-8

Cassar, G., Ittner, C. D., & Cavalluzzo, K. S. (2015). Alternative information sources and information asymmetry reduction: Evidence from small business debt. *Journal of Accounting and Economics*, *59*(2), 242–263. https://doi.org/10.1016/j.jacceco.2014.08.003

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2022). A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. *Decision Support Systems*, *152*, 113647. https://doi.org/10.1016/j.dss.2021.113647

Chen, Y.-R., Leu, J.-S., Huang, S.-A., Wang, J.-T., & Takada, J.-I. (2021). Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets. *IEEE Access*, *9*(No.), 73103–73109. https://doi.org/10.1109/ACCESS.2021.3079701

Ciampi, F., Giannozzi, A., Marzi, G., & Altman, E. I. (2021). Rethinking SME default prediction: A systematic literature review and future perspectives. *Scientometrics*, *126*(3), 2141–2188. https://doi.org/10.1007/s11192-020-03856-0

Cox, M. (2022, July 25). *What Is First-Party Fraud? From banks to telcos to debt collection agencies, what looks like unrecoverable bad debt may in fact be first-party fraud*. https://www.fico.com/blogs/what-first-party-fraud

Crosato, L., Liberati, C., & Repetto, M. (2023). Lost in a black-box? Interpretable machine learning for assessing Italian SMEs default. *Applied Stochastic Models in Business and Industry*, *39*(6), 829–846. https://doi.org/10.1002/asmb.2803

Devore, J., & Berk, K. (2012). *Modern Mathematical Statistics with Application Second Edition* (Second). Springer.

DeYoung, R. (2015). Personal Touch Makes Big Difference in Small-Business Loans; New research shows loans are less likely to end in default when borrower and lender have personal relationship. *The Wall Street Journal. Eastern Edition*.

Dilanian, K., & Strickler, L. (2022, March 28). 'Biggest Fraud In a Generation': The Looting of the COVID Relief Plan Known as PPP, Retrieved from "Biggest fraud in a generation": The looting of the Covid relief program known as PPP. *NBC News*. https://www.nbcnews.com/politics/justice-department/biggest-fraud-generation-looting-covid-relief-program-known-ppp-n1279664

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01).*, *IJCAI 01*.

Evans, P. (2023, March 29). A Complete Guide to Second and Third Party Fraud. *Feature Space*. https://www.featurespace.com/newsroom/a-complete-guide-to-second-and-third-party-fraud/

Eweoya, I. O., Adebiyi, A. A., Azeta, A. A., & Amosu, O. (2019). Fraud prediction in loan default using support vector machine. *Journal of Physics. Conference Series*, *1299*(1), 12039. https://doi.org/10.1088/1742-6596/1299/1/012039

Experian. (2021, April). *SME's with four or less employees are most at risk for payment default* [Experian]. https://www.experian.nl/over-experian/2021/04/22/smes-with-four-or-less-employees-are-most-at-risk-for-payment-default/

Factsheet: Required Rulemaking on Small Business. (2023). Consumer Financial Protection Bureau. https://files.consumerfinance.gov/f/documents/cfpb_small-business-lending-rule-fact-sheet_2023-03.pdf

Fairlie, R., & Fossen, F. M. (2022a). The 2021 Paycheck Protection Program Reboot: Loan Disbursement to Employer and Nonemployer Businesses in Minority Communities. *AEA Papers and Proceedings*, *112*, 287–291. https://doi.org/10.1257/pandp.20221028

Fairlie, R., & Fossen, F. M. (2022b). The early impacts of the COVID-19 pandemic on business sales. *Small Business Economics*, *58*(4), 1853–1864. https://doi.org/10.1007/s11187-021-00479-4

Fan, Q., & Yang, J. (2018). A Denoising Autoencoder Approach for Credit Risk Analysis. In *ICCAI 2018: Proceedings of the 2018 International Conference on Computing and Artificial Intelligence* (p. 65). https://doi.org/10.1145/3194452.3194456

Fantazzini, D., & Figini, S. (2009). Random Survival Forests Models for SME Credit Risk Measurement. *Methodology and Computing in Applied Probability*, *11*(1), 29–45. ProQuest Central. https://doi.org/10.1007/s11009-008-9078-2

FDIC. (2006, March). *FDIC Center for Financial Research Working Paper No. 2006-04, Borrower-Lender Distance, Credit Scoring, and the Performance of Small Business Loans* [Working Paper]. https://www.fdic.gov/analysis/cfr/working-papers/2006/2006-04.pdf

FDIC. (2011). *From the Examiner's Desk: SBA Lending: Insights for Lenders and Examiners*.

https://www.fdic.gov/regulations/examinations/supervisory/insights/sisum11/sisummer11-article2.pdf

G. Yedukondalu, K. Thrilokya, T. M. Reddy, & K. S. Vasavi. (2021). Antifraud Model For Internet Loan Using Machine Learning. *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1534–1537. https://doi.org/10.1109/ICECA52323.2021.9675968

Galindo, J., & Tamayo, P. (2000). Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics*, *15*, 107–143. https://doi.org/10.1023/A:1008699112516

Gao, G., Wang, H., & Gao, P. (2021). Establishing a credit risk evaluation system for smes using the soft voting fusion model. *Risks (Basel)*, *9*(11), 202. https://doi.org/10.3390/risks9110202

Gibson, R. (2009). Defaults by Franchisees Soar as the Recession Deepens: List of Small Business Administration-Backed Bad Loans at 500 Brands Increased 52% in Most Recent Fiscal Year. *Wall Street Journal (1923-)*, 1. ProQuest Historical Newspapers: The Wall Street Journal.

Golbayani, P., Wang, D., & Florescu, I. (2020). Application of Deep Neural Networks to assess corporate Credit Rating. *IDEAS Working Paper Series from RePEc*.

Goodhart, C. (2011). *The Basel Committee on Banking Supervision: A History of the Early Years, 1974–1997*.

Gradisher, S. M., & Tassell-Getman, T. (2020). Paycheck Protection Program: Piecing Together the Loan-Forgiveness Phase. *Journal of Financial Service Professionals*, *74*(5), 71.

Hackney, J. (2023). Small Business Lending in Financial Crises: The Role of Government-Guaranteed Loans*. *Review of Finance*, *27*(1), 247–287. https://doi.org/10.1093/rof/rfac002

Hardle, W. K., Klinke, S., & Ronz, B. (2015). *Introduction To Statistics, Using MM*Stat Elements*. Springer.

Henry, J. (2020). Second-round PPP loans are moving slowly, here's why. *Central Penn Business Journal*.

Huang, Yiping., Zhang, Longmei., Li, Zhenhua., Qiu, Han., Sun, Tao., & Wang, Xue. (2020). *Fintech Credit Risk Assessment for SMEs: Evidence from China*. International Monetary Fund.

Hubbard, D. W. (2020). *The failure of risk management: Why it's broken and how to fix it* (Second edition.). Wiley.

Jaadi, Z. (2019). Everything you need to know about interpreting correlations. *Towards Data Science*. https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). *An Introduction to Statistical Learning with Applications in R* (Second). Springer. https://www.statlearning.com/

Jeong, J. (2023). Do government guaranteed small business loans increase employment? Evidence from us counties, 2010-2016. *Journal of Policy Studies*, *38*(2), 11–21.

Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., & Lorentzen, J. (2020). Detecting money laundering transactions with machine learning. *Journal of Money*

Laundering Control, *23*(1), 173–186. https://doi.org/10.1108/JMLC-07-2019-0055

Kaya, O. (2022). Determinants and Consequences of SME Insolvency Risk During the Pandemic. *Economic Modeling*, *115*. https://www-sciencedirect-com.proxygw.wrlc.org/science/article/pii/S0264999322002048?via%3Dihub

King. (2023, January). The different types of fraud and how they're changing. *Fraud Prevention*. https://www.experian.co.uk/blogs/latest-thinking/fraud-prevention/what-is-first-second-and-third-party-fraud/

Kirschenmann, K., & Norden, L. (2012). The Relationship between Borrower Risk and Loan Maturity in Small Business Lending. *Journal of Business Finance & Accounting*, *39*(5–6), 730–757. https://doi.org/10.1111/j.1468-5957.2012.02285.x

Kumar, A., & Motwani, J. (1999). Reengineering the lending procedure for small businesses: A case study. *Work Study*, *48*(1), 6–12. ProQuest Central.

Lake, R. (2019, October 2). How Does an SBA 7(a) Loan Work? *US News*. https://money.usnews.com/loans/small-business-loans/articles/how-does-an-sba-7a-loan-work

Lender and Development Company Loan Programs (50 10 6). (N.d.). (n.d.). https://www.sba.gov/document/sop-50-10-lender-development-company-loan-programs

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks (Basel)*, *7*(1), 29. https://doi.org/10.3390/risks7010029

Leong, C. K. (2016). Credit Risk Scoring with Bayesian Network Models. *Computational Economics*, *47*(3), 423–446. https://doi.org/10.1007/s10614-015-9505-8

Lin, S.-M., Ansell, J., & Andreeva, G. (2012). Predicting default of a small business using different definitions of financial distress. *The Journal of the Operational Research Society*, *63*(4), 539–548. https://doi.org/10.1057/jors.2011.65

Liu, Y. (2022). *Small Business Administration Guaranteed Loan Default Detection Model with Optimized Boosting Methods*.

Luo, C., Wu, D., & Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, *65*, 465–470. https://doi.org/10.1016/j.engappai.2016.12.002

Marceau, L., Qiu, L., Vandewiele, N., & Charton, E. (2019). *A comparison of Deep Learning performances with others machine learning algorithms on credit scoring unbalanced data*.

Mount, I. (2009). And 7 Businesses That Did Not Survive. *New York Times*.

Ohsaki, M., Peng Wang, Matsuda, K., Katagiri, S., Watanabe, H., & Ralescu, A. (2017). Confusion-Matrix-Based Kernel Logistic Regression for Imbalanced Data Classification. *IEEE Transactions on Knowledge and Data Engineering*, *29*(9), 1806–1819. https://doi.org/10.1109/TKDE.2017.2682249

Orlando, G., & Pelosi, R. (2020). Non-Performing loans for Italian companies: When time matters. An empirical research on estimating probability to default and loss given default. *International Journal of Financial Studies*, *8*(4), 1–22. https://doi.org/10.3390/ijfs8040068

Patel, P. C., Mostaghel, R., & Oghazi, P. (2023). Local individualism culture and lower third-party guarantee loan defaults: Evidence from SBA loans in the US. *Applied Economics*, *55*(59), 7033–7047. https://doi.org/10.1080/00036846.2023.2206993

Piffer, M. (2018). Monetary Policy and Defaults in the United States. *International Journal of Central Banking*. https://www.ijcb.org/journal/ijcb18q3a8.pdf

PPP Dataset. (2023). [dataset]. https://data.sba.gov/dataset/ppp-foia/resource/b6528428-fbd9-4ca6-ae08-9e3416f8ee7f?inner_span=True

PUBLIC LAW 116–136—MAR. 27, 2020 (2020). https://www.congress.gov/116/plaws/publ136/PLAW-116publ136.pdf

Richardson, B., & Waldron, D. (2019). Fighting back against synthetic identity fraud. *McKinsey on Risk*, *7*, 1–6.

Rudegeair, P., & Andriotis, A. (2018). The New ID Theft: Millions of Credit Applicants Who Don't Exist; Synthetic-identity fraud is one of the fastest growing forms of identity theft—And the hardest to spot and combat. *WSJ Pro. Cyber Security*.

SBA 7a Dictionary. (2023, September 30). https://data.sba.gov/en/dataset/7-a-504-foia/resource/6898b986-a895-47b4-bb7e-c6b286b23a7b

SBA PPP Data Dictionary. (2023, September 30). https://data.sba.gov/dataset/ppp-foia/resource/5158aae1-066d-4d01-a226-e44ecc9bdda7?inner_span=True

Semiannual report to Congress. (2020). *Semiannual Report to Congress.*

Shear, W. B. (2021). *Small Business Administration: Actions needed to improve COVID-19 loans internal controls and reduce their susceptibility to fraud: Testimony before the Committee on Small Business and Entrepreneurship, U.S. Senate*. United States Government Accountability Office.

Shi, S., Tse, T., Luo, W., D'Adonna, S., & Pau, G. (2022). Machine learning-driven credit risk: A systemic review. *Neural Network Computing & Application, 34,* 14327–14339.

Siegel, L. (2014). Pitfalls and Windfalls of SBA 7(a) lending. *The RMA Journal, 97*(3), 54.

Simon, R., & Rudegeair, P. (2023). SBA overhaul will make it easier for fintechs and other nonbank lenders to issue loans, prompting worries about defaults. *WSJ*. https://www.wsj.com/articles/small-business-lending-is-about-to-change-with-simpler-requirements-8578a895?mod=business_lead_pos4

Singh, K. (2023, January 30). U.S. Watchdog Identifies $5.4 Billion in Potentially Fraudulent COVID-19 Loans. *Reuters, Thomson*. https://www.reuters.com/world/us/us-watchdog-identifies-54-billion-potentially-fraudulent-covid-19-loans-2023-01-30/

Singh, P. (2023, March 15). Goldman Sachs Cuts GDP Forecast Because of Stress on Small Banks, which are Key to U.S. Economy. *CNBC*. https://www.cnbc.com/2023/03/15/goldman-sachs-cuts-gdp-forecast-because-of-stress-on-small-banks.html

Small Business Administration Loan Program Performance. (n.d.). https://www.sba.gov/document/report-small-business-administration-loan-program-performance

Small Business Loans: Additional Actions Needed to Improve Compliance with the Credit Elsewhere Requirement. (2018). In *Policy File*. US Government Accountability Office.

Subasi, A. (2020). Chapter 3—Machine learning techniques. In A. Subasi (Ed.), *Practical Machine Learning for Data Analysis Using Python* (pp. 91–202). Academic Press. https://doi.org/10.1016/B978-0-12-821379-7.00003-5

Summary of Senate Omnibus CARES Act. (2020). *The RMA Journal*, *102*(8), 42–45.

Temkin, K., & Theodos, B. (2008). An Analysis of the Factors Lenders Use to Ensure Their SBA Borrowers Meet the Credit Elsewhere Requirement. In *Policy File*. Urban Institute.

Turiel, J. D., & Aste, T. (2020). Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society Open Science*, *7*(6), 191649–191649. https://doi.org/10.1098/rsos.191649

Tyagi, N. (2020, March 24). Understanding the Gini Index and Information Gain in Decision TreesTyagi. *Medium.Com*. https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8

VanSomeren, L., & Tarver, J. (n.d.). What is Credit Worthiness? *Forbes Advisor*. https://www.forbes.com/advisor/credit-score/what-is-creditworthiness/

Voigt, K., & Weiner Campbell, C. (2017, October 3). 1 in 6 Small Business Administration Loans Fail, Study Finds. *Nerd Wallet*. https://www.nerdwallet.com/article/small-business/study-1-in-6-sba-small-business-administration-loans-fail

Wang, C., Han, D., Liu, Q., & Luo, S. (2019). A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM. *IEEE Access*, *7*, 2161–2168. https://doi.org/10.1109/ACCESS.2018.2887138

Ware, H. (2023). *Top Management and Performance Challenges Facing the Small Business Administration in Fiscal Year 2024* (24–01).

Xiao, C., Ye, J., Esteves, R. M., & Rong, C. (2016). Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation*, *28*(14), 3866–3878. https://doi.org/10.1002/cpe.3745

Xu, J. J., Chen, D., Chau, M., Li, L., & Zheng, H. (2022). PEER-TO-PEER LOAN FRAUD DETECTION: CONSTRUCTING FEATURES FROM TRANSACTION DATA. *MIS Quarterly*, *46*(3), 1777–1792. https://doi.org/10.25300/MISQ/2022/16103

Xu, R.-Z., & He, M.-K. (2020). *Application of Deep Learning Neural Network in Online Supply Chain Financial Credit Risk Assessment* (p. 232). https://doi.org/10.1109/CIBDA50819.2020.00058

Zaki, A. (2023, January 11). 70% Of Financial Institutions Lost over $500k to Fraud in 2022: Weekly Stat. *CFO Magazine*. https://www.cfo.com/technology/cyber-security-technology/2023/01/cybersecurity-fraud-lending-breach-recovery/

Zhang, Q., Wang, J., Lu, A., Wang, S., & Ma, J. (2017). An Improved SMO Algorithm for Financial Credit Risk Assessment–Evidence from China's banking. *Neurocomputing*, *272*. https://doi.org/10.1016/j.neucom.2017.07.002

Zhao, Y., & Lin, D. (2023). Prediction of Micro- and Small-Sized Enterprise Default Risk Based on a Logistic Model: Evidence from a Bank of China. *Sustainability (Basel, Switzerland)*, *15*(5), 4097. https://doi.org/10.3390/su15054097

Zhu, B., Yang, W., Wang, H., & Yuan, Y. (2018). A hybrid deep learning model for consumer credit scoring. *ICAIBD*, 205–208. https://doi.org/10.1109/ICAIBD.2018.8396195

Zhu, X., Ao, X., Qin, Z., Chang, Y., He, Q., & Liu, Y. (2021). Intelligent financial fraud detection practices in post-pandemic era. *The Innovation*, *2*(4). https://doi.org/10.1016/j.xinn.2021.100176

Appendix A-Microtrends Data

Table 27. Unemployment Microtrends Data

Date	Unemployment Rate (%)	Annual Change
12/31/1991	6.8	
12/31/1992	7.5	0.7
12/31/1993	6.9	-0.6
12/31/1994	6.12	-0.78
12/31/1995	5.65	-0.47
12/31/1996	5.45	-0.2
12/31/1997	5	-0.45
12/31/1998	4.51	-0.49
12/31/1999	4.22	-0.29
12/31/2000	3.99	-0.23
12/31/2001	4.73	0.74
12/31/2002	5.78	1.05
12/31/2003	5.99	0.21
12/31/2004	5.53	-0.46
12/31/2005	5.08	-0.45
12/31/2006	4.62	-0.46
12/31/2007	4.62	0
12/31/2008	5.78	1.16
12/31/2009	9.25	3.47
12/31/2010	9.63	0.38
12/31/2011	8.95	-0.68
12/31/2012	8.07	-0.88
12/31/2013	7.37	-0.7
12/31/2014	6.17	-1.2
12/31/2015	5.28	-0.89
12/31/2016	4.87	-0.41
12/31/2017	4.36	-0.51
12/31/2018	3.9	-0.46
12/31/2019	3.67	-0.23
12/31/2020	8.05	4.38
12/31/2021	5.35	-2.7
12/31/2022	3.611	-1.74

Table 28. GDP Microtrends Data

U.S. GDP - Historical Data			
Year	GDP	Per Capita	Growth
1991	$6,158.13B	$24,342	-0.11%
1992	$6,520.33B	$25,419	3.52%
1993	$6,858.56B	$26,387	2.75%
1994	$7,287.24B	$27,695	4.03%
1995	$7,639.75B	$28,691	2.68%
1996	$8,073.12B	$29,968	3.77%
1997	$8,577.55B	$31,459	4.45%
1998	$9,062.82B	$32,854	4.48%
1999	$9,631.17B	$34,515	4.79%
2000	$10,250.95B	$36,330	4.08%
2001	$10,581.93B	$37,134	0.95%
2002	$10,929.11B	$37,998	1.70%
2003	$11,456.44B	$39,490	2.80%
2004	$12,217.19B	$41,725	3.85%
2005	$13,039.20B	$44,123	3.48%
2006	$13,815.59B	$46,302	2.78%
2007	$14,474.23B	$48,050	2.01%
2008	$14,769.86B	$48,570	0.12%
2009	$14,478.06B	$47,195	-2.60%
2010	$15,048.96B	$48,651	2.71%
2011	$15,599.73B	$50,066	1.55%

2012	$16,253.97B	$51,784	2.28%
2013	$16,843.19B	$53,291	1.84%
2014	$17,550.68B	$55,124	2.29%
2015	$18,206.02B	$56,763	2.71%
2016	$18,695.11B	$57,867	1.67%
2017	$19,477.34B	$59,908	2.24%
2018	$20,533.06B	$62,823	2.95%
2019	$21,380.98B	$65,120	2.29%
2020	$21,060.47B	$63,529	-2.77%
2021	$23,315.08B	$70,219	5.95%
2022	$25,462.70B	$76,399	2.06%

Table 29. USA Inflation Data

Year	Inflation Rate (%)	Annual Change
1991	4.24%	-1.16%
1992	3.03%	-1.21%
1993	2.95%	-0.08%
1994	2.61%	-0.34%
1995	2.81%	0.20%
1996	2.93%	0.13%
1997	2.34%	-0.59%
1998	1.55%	-0.79%
1999	2.19%	0.64%
2000	3.38%	1.19%
2001	2.83%	-0.55%
2002	1.59%	-1.24%
2003	2.27%	0.68%
2004	2.68%	0.41%
2005	3.39%	0.72%
2006	3.23%	-0.17%

2007	2.85%	-0.37%
2008	3.84%	0.99%
2009	-0.36%	-4.19%
2010	1.64%	2.00%
2011	3.16%	1.52%
2012	2.07%	-1.09%
2013	1.46%	-0.60%
2014	1.62%	0.16%
2015	0.12%	-1.50%
2016	1.26%	1.14%
2017	2.13%	0.87%
2018	2.44%	0.31%
2019	1.81%	-0.63%
2020	1.23%	-0.58%
2021	4.70%	3.46%
2022	8.00%	3.30%

Appendix B-SBA Data

Table 30.SBA Performance Data as of 12/31/2022

Table 5 - Charge Off Amount by Program

	Fiscal Year									
Program	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
Guaranteed Business										
7(a) Regular	$806,928,976	$1,474,525,906	$1,430,439,372	$689,553,988	$472,374,120	$641,019,903	$364,825,722	$373,015,281	$456,134,205	$95,371,700
504 Regular	$402,437,587	$317,485,370	$217,585,054	$101,690,600	$104,180,106	$79,260,787	$78,316,762	$34,460,536	$95,375,087	$11,107,834
SBIC Debentures	$70,077,734	$44,492,463	$4,773,821	$11,652,153	$37,148,146	$24,039,000	$878,421	$71,589,681	$2,172,541	$0
SBIC Participating Securities	$89,400,101	$47,547,090	$57,574,744	$63,888,169	$83,254,811	$46,988,629	$36,924,459	$6,959,450	$1,384,128	$0
ARC 506	$3,676,757	$2,020,121	$668,762	$448,391	$718,386	$349,767	$197,000	$21,180	$34,688	$0
Dealer Floor Plan	$0	$0	$366,413	$0	$0	$0	$0	$0	$0	$0
504 First Lien	$0	$0	$0	$6,555,142	$-623,845	$477,708	$420,387	$0	$1,683,273	$0
504 Refi	$698,905	$1,192,058	$6,653,255	$9,248,168	$5,973,270	$2,808,099	$7,151,546	$1,479,847	$4,726,803	$3,073,987
PPP	N/A	N/A	N/A	N/A	N/A	N/A	$0	$0	$4,823,195,167	$2,509,111,225
All Other	$6,492,648	$19,492,216	$23,306,730	$5,093,978	$5,683,814	$24,060,060	$17,386,933	$-1,800,000	$1,151,603	$0
Subtotal	**$1,379,712,708**	**$1,906,755,223**	**$1,741,368,151**	**$888,130,591**	**$708,708,808**	**$819,003,953**	**$506,101,229**	**$485,725,975**	**$5,385,857,494**	**$2,618,664,747**
Direct Business										
Microloan Direct	$356,710	$151,549	$732,698	$452,653	$188,494	$21,604	$326,549	$0	$31,446	$17,654
All Other	$971,715	$0	$0	$0	$341,157	$0	$101,137	$0	$0	$0
Subtotal	**$1,328,425**	**$151,549**	**$732,698**	**$452,653**	**$529,651**	**$21,604**	**$427,685**	**$0**	**$31,446**	**$17,654**
Disaster										
Disaster	$204,728,937	$102,428,523	$87,747,318	$85,987,230	$98,191,070	$190,022,038	$129,320,597	$18,405,594	$180,342,594	$70,643,945
COVID EIDL	N/A	N/A	N/A	N/A	N/A	N/A	$0	$21,530,939	$198,192,908	$110,577,723

This table displays the total charge off amount by program as of the end of each fiscal year. Since data are not available through the end of the most recent fiscal year, the data displayed in 2023 are as of 12/31/2022.
Charge off amount is defined as the total dollar amount of principal and interest outstanding at the time that the loan is charged off.
Loans are charged off if SBA determines no additional principal and interest from the borrower will be recovered via the agency.
For guaranteed loans, the charge off amounts reflect the SBA guarantied portion and exclude the non-guarantied portion of the loan.
Charge off amounts for a given fiscal year may be adjusted due to data updates.
Guaranteed Business, Direct Business, and Disaster loan programs include all loans that are subject to the Credit Reform Act of 1990; which are loans SBA approved on or after 10/01/1991.
The 7(a) and 504 loans in the DELTA and STAR programs are included in the 'All Other' category, not the '7(a) Regular' and '504 Regular' categories, of guarantied business programs.

Table 5 - Charge Off Amount by Program

	Fiscal Year									
Program	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
Subtotal	**$204,728,937**	**$102,428,523**	**$87,747,318**	**$85,987,230**	**$98,191,070**	**$190,022,038**	**$129,320,597**	**$39,936,534**	**$378,535,503**	**$181,221,668**
Total	**$1,585,770,071**	**$2,009,335,295**	**$1,829,848,167**	**$974,570,474**	**$807,429,529**	**$1,009,047,594**	**$635,849,511**	**$525,662,508**	**$5,764,424,443**	**$2,799,904,069**

*Table 31. SBA 7(a) Loan Data Dictionary(*SBA 7(a) Dictionary, *2023)*

Field Name	Definition
AsOfDate	Date when the data was recorded
Program	Indicator of whether loan was approved under SBA's 7(a) or 504 loan program
BorrName	Borrower name
BorrStreet	Borrower street address
BorrCity	Borrower city
BorrState	Borrower state
BorrZip	Borrower zip code
BankName	Name of the bank that the loan is currently assigned to
BankFDICNumber	The Federal Depository Insurance Corporation certificate ID of the lender
BankNCUANumber	The National Credit Union Association charter number of the lender

BankStreet	Bank street address
BankCity	Bank city
BankState	Bank state
BankZip	Bank zip code
GrossApproval	Total loan amount
SBAGuaranteedApproval	Amount of SBA's loan guaranty
ApprovalDate	Date the loan was approved
ApprovalFiscalYear	Fiscal year the loan was approved
FirstDisbursementDate	Date of first loan disbursement (if available)
DeliveryMethod	Specific delivery method loan was approved under. See SOP 50 10 5 for definitions and rules for each delivery method. 7(a) Delivery Methods: • CA = Community Advantage • CLP = Certified Lenders Program • COMM EXPRS = Community Express (inactive) • DFP = Dealer Floor Plan (inactive) • DIRECT = Direct Loan (inactive) • EWCP = Export Working Capital Program • EXP CO GTY = Co-guaranty with Export-Import Bank (inactive) • EXPRES EXP = Export Express • GO LOANS = Gulf Opportunity Loan (inactive) • INTER TRDE = International Trade • OTH 7(a) = Other 7(a) Loan • PATRIOT EX = Patriot Express (inactive) • PLP = Preferred Lender Program • RLA = Rural Lender Advantage (inactive) • SBA EXPRES = SBA Express • SLA = Small Loan Advantage • USCAIP = US Community Adjustment and Investment Program • Y2K = Y2K Loan (inactive)
subpgmdesc	Subprogram description - specific subprogram loan was approved under. See SOP 50 10 5 for definitions and rules for each subprogram.
InitialInterestRate	Initial interest rate - total interest rate (base rate plus spread) at time loan was approved
TermInMonths	Length of loan term
NaicsCode	North American Industry Classification System (NAICS) code
NaicsDescription	North American Industry Classification System (NAICS) description
FranchiseCode	Franchise Code

FranchiseName	Franchise Name (if applicable)
ProjectCounty	County where project occurs
ProjectState	State where project occurs
SBADistrictOffice	SBA district office
CongressionalDistrict	Congressional district where project occurs
BusinessType	Borrower Business Type - Individual, Partnership, or Corporation
BusinessAge	SBA began collecting the following business age information in fiscal year 2018: • Change of Ownership • Existing or more than 2 years old • New Business or 2 years or less • Startup, Loan Funds will Open Business
LoanStatus	Current status of loan: • COMMIT = Undisbursed • PIF = Paid In Full • CHGOFF = Charged Off • CANCLD = Cancelled • EXEMPT = The status of loans that have been disbursed but have not been cancelled, paid in full, or charged off are exempt from disclosure under FOIA Exemption 4
PaidInFullDate	Date loan was paid in full (if applicable)
ChargeOffDate	Date SBA charged off loan (if applicable)
GrossChargeOffAmount	Total loan balance charged off (includes guaranteed and non-guaranteed portion of loan)
RevolverStatus	Indicator of whether a loan is a term loan or revolving line of credit (0=Term, 1=Revolver)
JobsSupported	Total Jobs Created + Jobs Retained as reported by lender on SBA Loan Application. SBA does not review, audit, or validate these numbers - they are simply self-reported, good faith estimates by the lender.
Soldsecmrtind	An indicator if the loan was sold on the secondary market. This is a static field once it is sold on the secondary market.Equals 'Y', if sold on the secondary market. Once it is 'Y' it will stay 'Y' for its entirety.

*Table 32. SBA PPP Loan Data Dictionary(*SBA PPP Data Dictionary, *2023)*

Field Name	Field Description
LoanNumber	Loan Number (unique identifier)
DateApproved	Loan Funded Date
SBAOfficeCode	SBA Origination Office Code
ProcessingMethod	Loan Delivery Method (PPP for first draw; PPS for second draw)
BorrowerName	Borrower Name
BorrowerAddress	Borrower Street Address
BorrowerCity	Borrower City
BorrowerState	Borrower State
BorrowerZip	Borrower Zip Code
LoanStatusDate	Loan Status Date - Loan Status Date is blank when the loan is disbursed but not Paid in Full or Charged Off
LoanStatus	Loan Status Description - Loan Status is replaced by 'Exemption 4' when the loan is disbursed but not Paid in Full or Charged Off
Term	Loan Maturity in Months
SBAGuarantyPercentage	SBA Guaranty Percentage
InitialApprovalAmount	Loan Approval Amount(at origination)
CurrentApprovalAmount	Loan Approval Amount (current)
UndisbursedAmount	Undisbursed Amount
FranchiseName	Franchise Name
ServicingLenderLocationID	Lender Location ID (unique identifier)
ServicingLenderName	Servicing Lender Name
ServicingLenderAddress	Servicing Lender Street Address
ServicingLenderCity	Servicing Lender City
ServicingLenderState	Servicing Lender State
ServicingLenderZip	Servicing Lender Zip Code
RuralUrbanIndicator	Rural or Urban Indicator (R/U)
HubzoneIndicator	Hubzone Indicator (Y/N)
LMIIndicator	LMI Indicator (Y/N)
BusinessAgeDescription	Business Age Description
ProjectCity	Project City
ProjectCountyName	Project County Name
ProjectState	Project State
ProjectZip	Project Zip Code
CD	Project Congressional District
JobsReported	Number of Employees
NAICSCode	NAICS 6-digit code

Race	Borrower Race Description
Ethnicity	Borrower Ethnicity Description
UTILITIES_PROCEED	Note: Proceed data is lender reported at origination. On the PPP application the proceeds fields were check boxes.
PAYROLL_PROCEED	
MORTGAGE_INTEREST_PROCEED	
RENT_PROCEED	
REFINANCE_EIDL_PROCEED	
HEALTH_CARE_PROCEED	
DEBT_INTEREST_PROCEED	
BusinessType	Business Type Description
OriginatingLenderLocationID	Originating Lender ID (unique identifier)
OriginatingLender	Originating Lender Name
OriginatingLenderCity	Originating Lender City
OriginatingLenderState	Originating Lender State
Gender	Gender Indicator
Veteran	Veteran Indicator
NonProfit	'Yes' if Business Type = Non-Profit Organization or Non-Profit Childcare Center or 501(c) Non-Profit
ForgivenessAmount	Forgiveness Amount
ForgivenessDate	Forgiveness Paid Date

Appendix C-ROC Curves

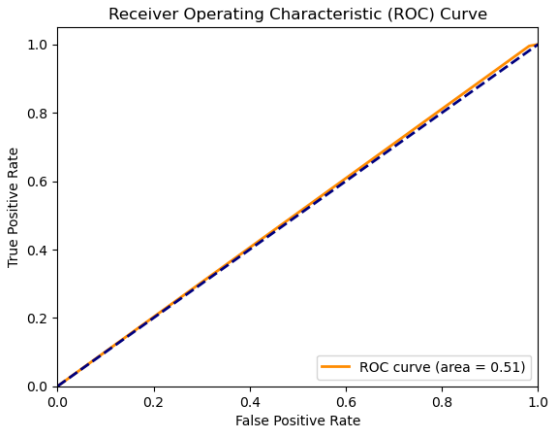


Figure 37. Logistics Regression ROC Curve

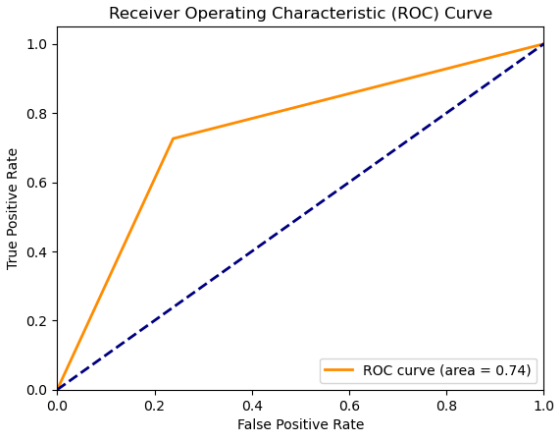


Figure 38. Logistics Regression SMOTE ROC Curve

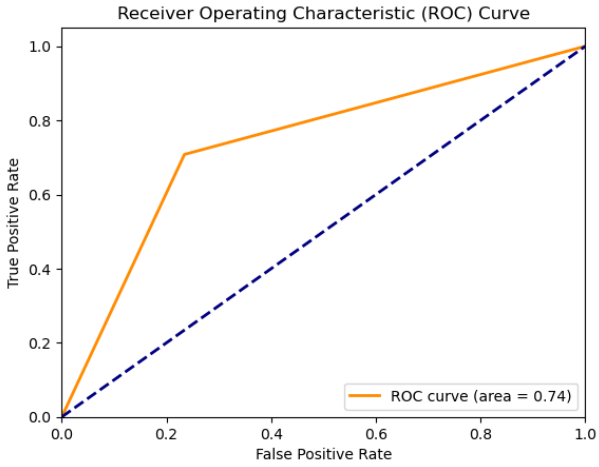


Figure 39. Logistics Regression Under Sampling ROC

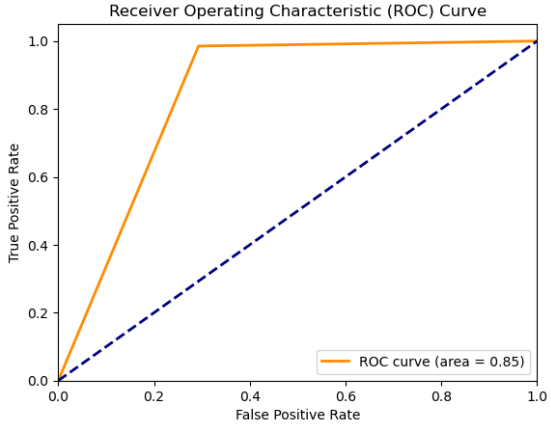


Figure 40. XG ROC Curve

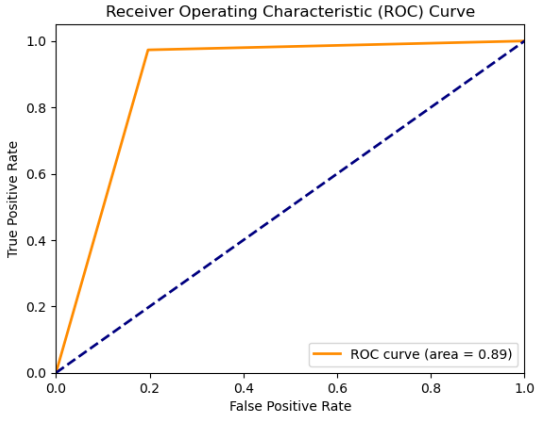


Figure 41. XG Boost SMOTE ROC Curve

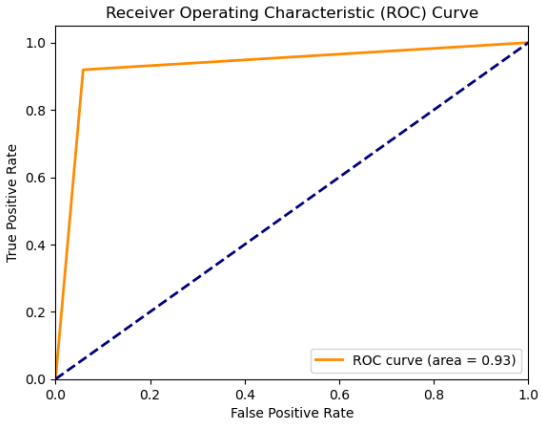


Figure 42. XG Boost Under Sampling ROC

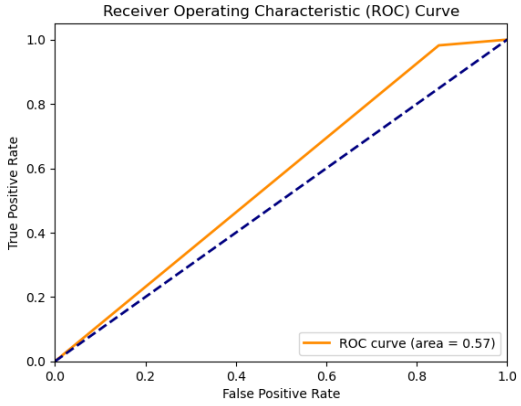


Figure 43. KNN ROC Curve

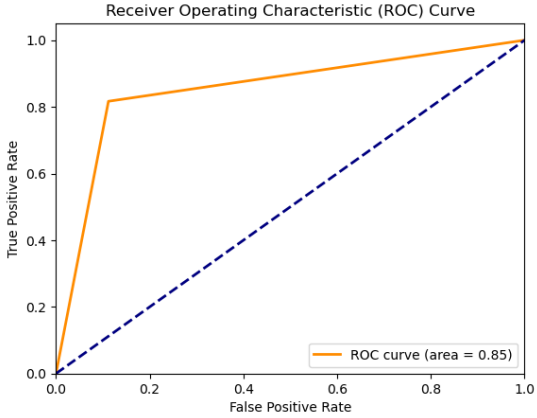


Figure 44. KNN SMOTE ROC Curve

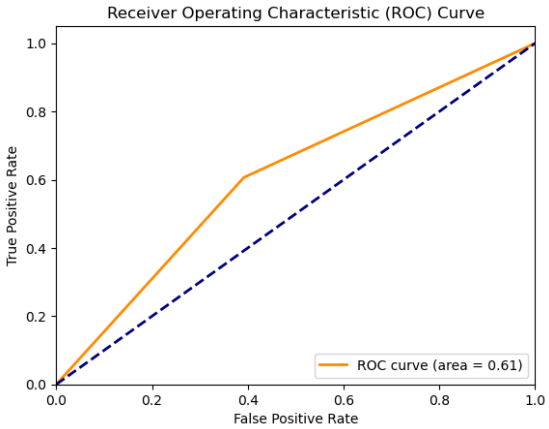


Figure 45. KNN Under Sampling ROC Curve

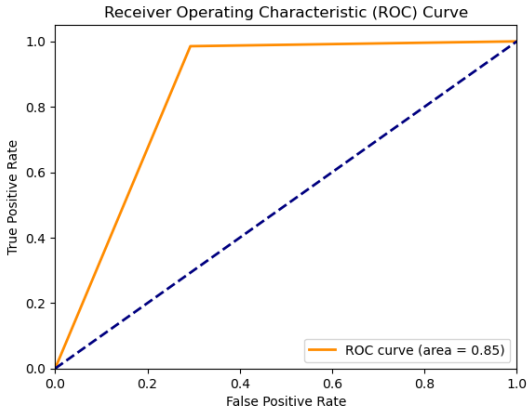


Figure 46. Decision Tree ROC Curve

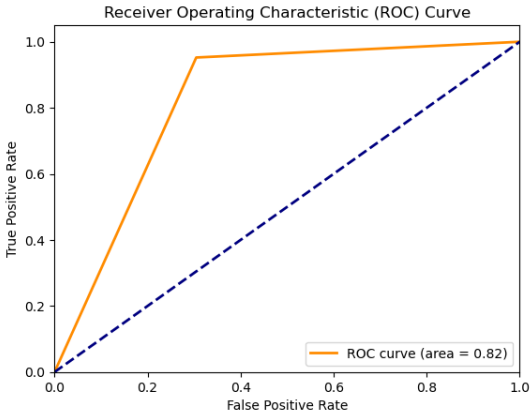


Figure 47. Decision Tree SMOTE ROC Curve

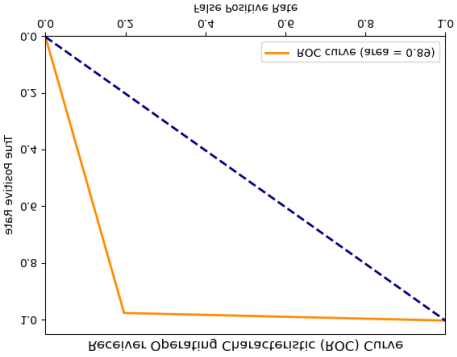


Figure 48. Decision Tree Under Sampling ROC Curve

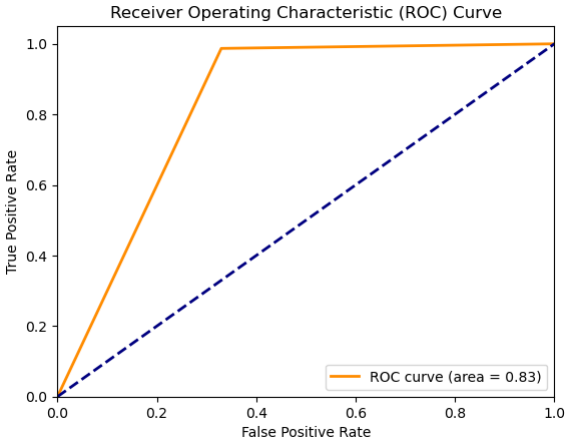


Figure 49. Random Forest ROC Curve

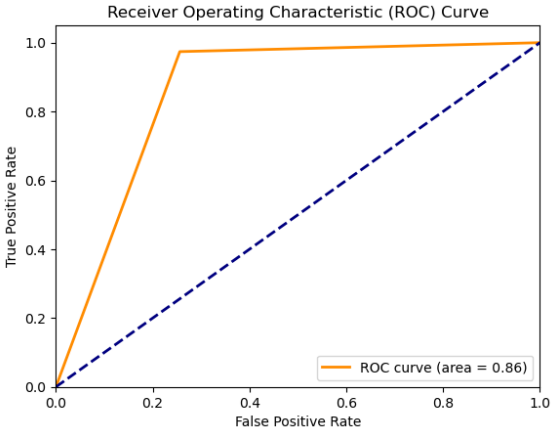


Figure 50. Random Forest SMOTE ROC Curve

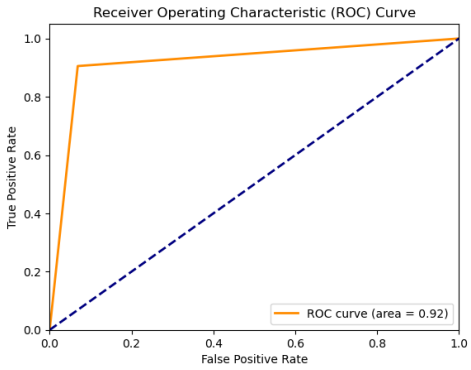


Figure 51. Random Forest Under Sampling ROC Curve

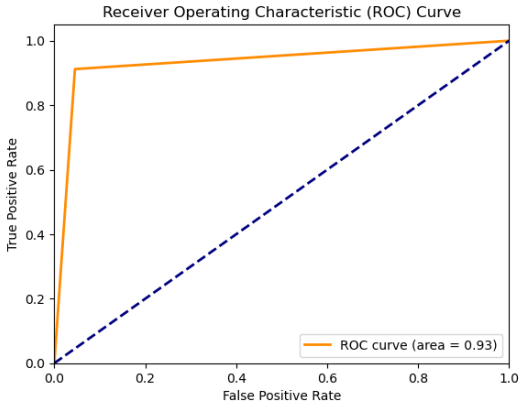


Figure 52. 2023 ROC Curve