THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

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

Timalyn Franklin-Bullock
DN#: 230428
Date: April 2, 2024
Time: 9-10 PM EST

Praxis Committee:

- Dr. Amir Etemadi
- Dr. Jonathan Bierce
- Dr. Michael Jones

Overview

- SBA Loan Guaranty Programs/Defaults/Charge Offs
- Research Motivation
- Literature Review
- Scope
- Methodology
- Results
- Discussion
- Conclusions/Contributions
- Future Research Directions

SBA Loan Guaranty Programs

The United States Small Business Administration (SBA) provides access to capital to small businesses with the 7(a) and 504 loan guaranty programs. 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 Loan Programs (Continued)

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/Charge Offs

SBA 7(a) loan guaranty defaults result from a business not paying a loan for more than 90 days. SBA loan defaults may result in the loan being charged off. Taxpayers pay the guaranty fee for charged off defaulted loans whether due to fraud or business failure. Charged off loans are not forgiven and will be pursued in collection by the US government using Treasury debt collection processes to offset expenses incurred by taxpayers.

The research motivation is to use 7(a) loan data to enhance the loan origination process with a predictive charge off risk model to minimize financial losses to taxpayers and lenders.

Charge Off Process (SOP 50 57 3)

Borrower	Lender	SBA	Treasury
			
1. Borrower defaults on loan after several requests for payment from the lender.	2. Lender notifies SBA about default. The lender requests for SBA to purchase the guaranty portion of the loan.	3. If the loan is approved for purchase, SBA changes the loan status to "charge off" and reports the loan to Treasury.	4. Treasury proceeds with collecting the debt for the defaulted charged off amount.

Research Motivation

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). Loan defaults occur for legitimate reasons such as insolvency as well as due to fraud (Eweoya et al., 2019).

The Federal government's growing concern about fraud defaults was elevated with the appointment of 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).

Problem & Thesis Statements

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.).

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.*

Research Objectives, Questions, &Hypotheses

Research Objectives	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.
Research Questions	**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?
Research Hypotheses	**H1:** Loan application data features such as loan amount, interest rate, and term in months as well as the 7a 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.

Literature Review

The literature review highlights:

- Previous SBA research to unveil insights and identify improvements
- Loan default prediction features
- SME loan default models
- Risk Model Considerations

Key Concept: Explore diverse small business lending default models to understand the intricacies and implications for effective risk prediction in the SME lending sector.

Literature Review

Highlight	Paper	Contributions	Gaps
Previous Research	Jeong 2023	The research identifies the impact of small business loans on the economy.	The model doesn't evaluate the impact of the economy on SBA defaults.
Loan Default Features	Turiel & Aste, 2020	Peer to peer (P2P) lending model research that identified the features of loan term, FICO score, debt to income ratio, and loan amount to be the top four predictors of default.	The model doesn't consider additional features including borrower and lender zip code, interest rate, and economic features.

Literature Review (Continued)

Highlight	Paper	Contributions	Gaps
SME Loan Default Models	Huang 2020	IMF developed small and medium enterprise (SME) default prediction model using China MyBank data. The analysis compared using a scorecard prediction model with a random forest model that considers features such as business age. Random forest outperformed the scorecard model.	The model didn't compare random forest against other ML models to determine if RF is the best model overall.
SME Loan Default Models	Nguyen 2020	Asian Development Bank (ADB) predicted loan defaults with financial ratios using logistics regression, XG Boost, and RF. XG Boost outperformed logistics regression and RF.	The analysis relies on financial data over 1-3 years and not loan application data.

Literature Review (Continued)

Highlight	Paper	Contributions	Gaps
SME Loan Default Models	Crosato 2023	Accounting data predicts loan defaults for Italian SMEs. XG Boost was the best model for predicting loan defaults.	The analysis does not include a model inclusive for all businesses including start ups lacking accounting ratios.
SME Loan Default Models	Batiz-Zuk 2022	Economic conditions are used to predict loan SME defaults in Mexico.	The research focuses on emerging economies.
Risk Model Considerations	Orlando & Pelosi, 2021	Defines a method for calculating default risk.	N/A

Scope

- SBA charged off and paid in full (PIF) 7a loans from 1991-2022 reported as of 9/30/2023 are analyzed to gain insight into the historical patterns of charge offs. For additional context, charged off and paid in full (PIF) PPP loans were reviewed to validate the features.
- Data Sources:
  - SBA 7a and PPP loans
  - Microtrends.com Economic Data
- Limitations:
  - May not be applicable to commercial/ non-SBA SME loan programs
  - Default is synonymous with charge off for this analysis which is not aligned with industry best practices
  - Fraud is not included in this analysis

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

File	Transaction Count	PPP CHGOFF Trans	PPP PIF Transcation	PPS CHGOFF Transactic	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**

H1 Methodology

- H1: Loan features are identified using 7a and PPP loan data. The metrics for the two groups are compared to identify group that best predicts charge offs.
  - 7a Group 1 includes continuous features (Borrower Zip, Bank Zip, Gross Approval, SBA Guaranteed Approval, Initial Interest Rate, Term in Months, Jobs Supported, Revolver Status, Business Age, Business Type, Approval Fiscal Year, Subprogram Code, and Delivery Method. 7a Group 2 includes Group 1 with Subprogram Code and Delivery Method excluded.
  - PPP Group 1 includes 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.

H1 7(a) 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 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
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

H1 PPP 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

H2 Methodology

- H2: 7a data is processed and cleansed to generate15 models: Logistic Regression, Logistic Regression SMOTE, Logistics Regression Under Sampling, KNN, KNN SMOTE, KNN Under Sampling, XGBoost, XG Boost SMOTE, XG Boost Under Sampling, Decision Tree, Decision Tree SMOTE, Decision Tree Under Sampling, Random Forest, Random Forest Smote, and Random Forest Under Sampling.
  - Shapley values are reviewed to refine the features.
  - The selected model is identified based on the metrics and cost savings.

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

H2 Methodology: Savings Results

- 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*

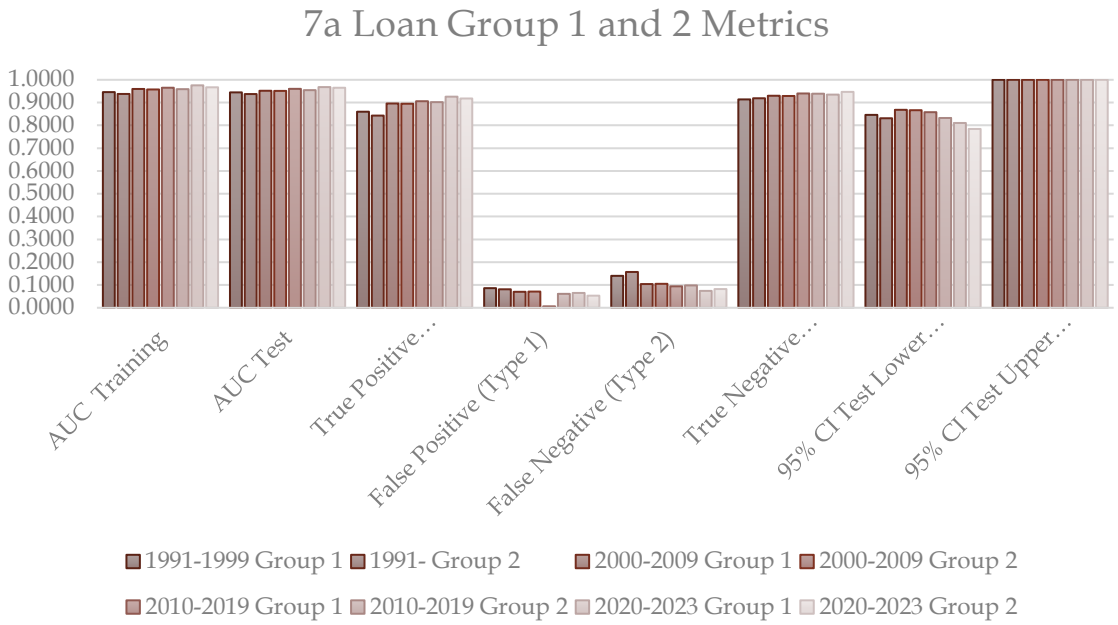
  *Servicing Fee Per Year = .01*$152250=$1523*

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

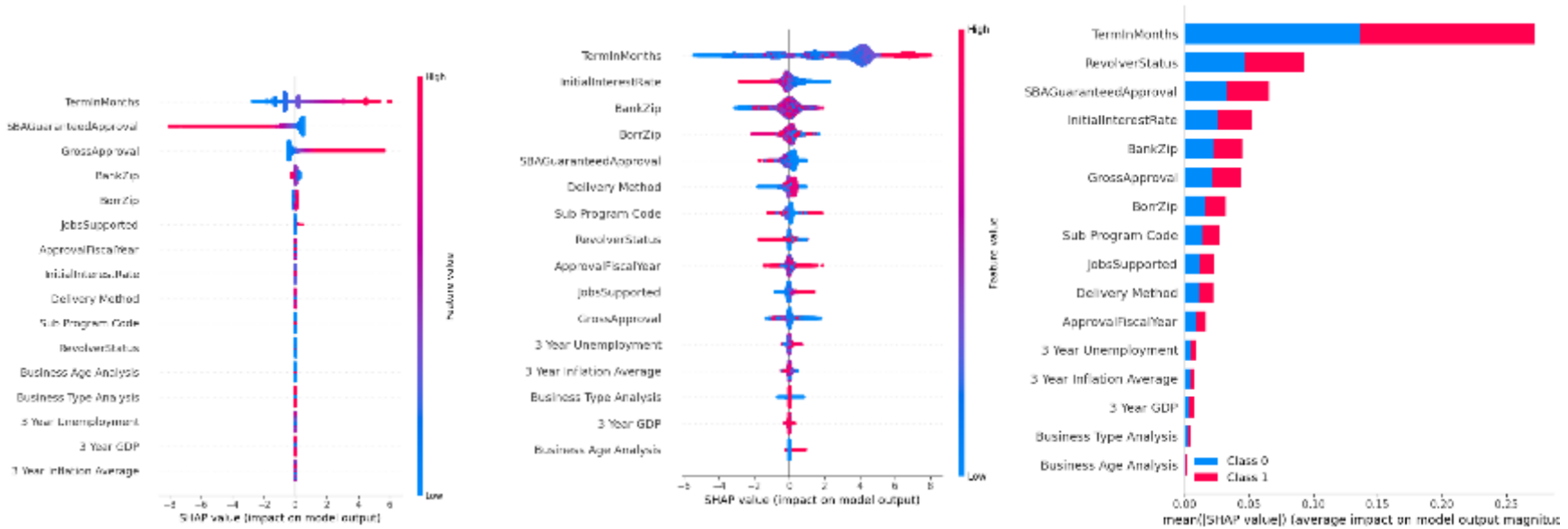
  Average Premium and Servicing Fee Per Paid in Full Loan=$28932

Results: H1

7a loans, 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 7a 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.

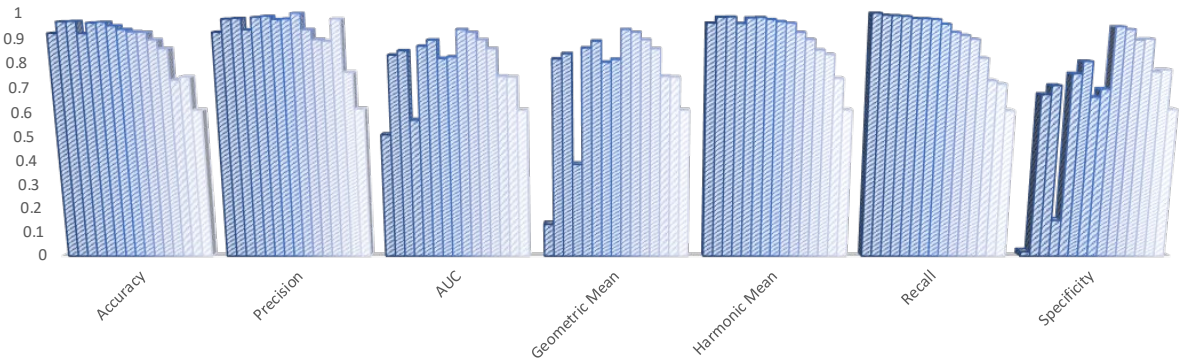


Results: H2 Economic Features



The SHAP value summaries are generated for Logistics Regression, XG Boost, and Decision Tree. Based on the summaries, unemployment is the most consistent economic feature that predicts loan status. 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. (Business age and business type were not influencers.)

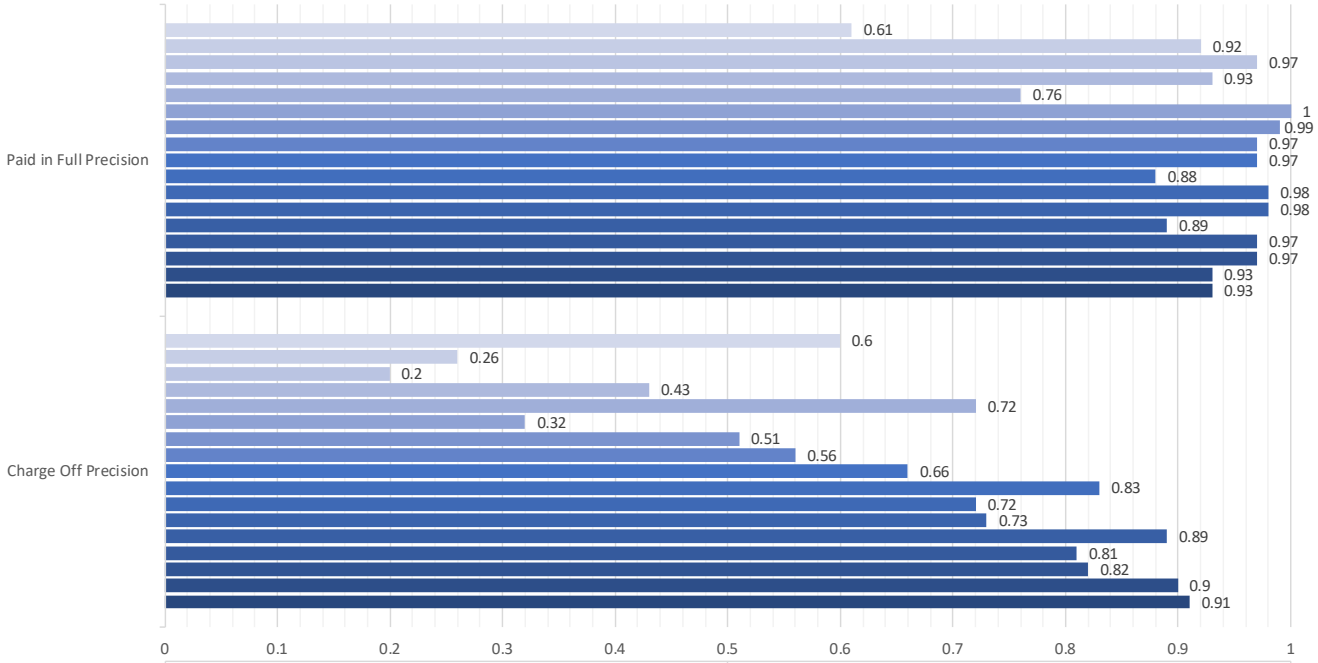
Results: Model Metric Results



	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

- XG Boost has an accuracy of .96
- XG Boost under sampling with a precision of .99 is the best model for predicting paid in full.
- XG Boost under sampling is the best overall model with an AUC of .93.
- XG Boost under sampling has a geometric mean of .93.
- XG Boost has a harmonic mean of .98.
- Logistics regression with a recall of .99 is the best for finding all the paid in full objects.
- XG Boost under sampling with a specificity of .94 is the best model for predicting charge offs.
- Based on the metrics, XG Boost under sampling is the selected model for predicting charge offs.

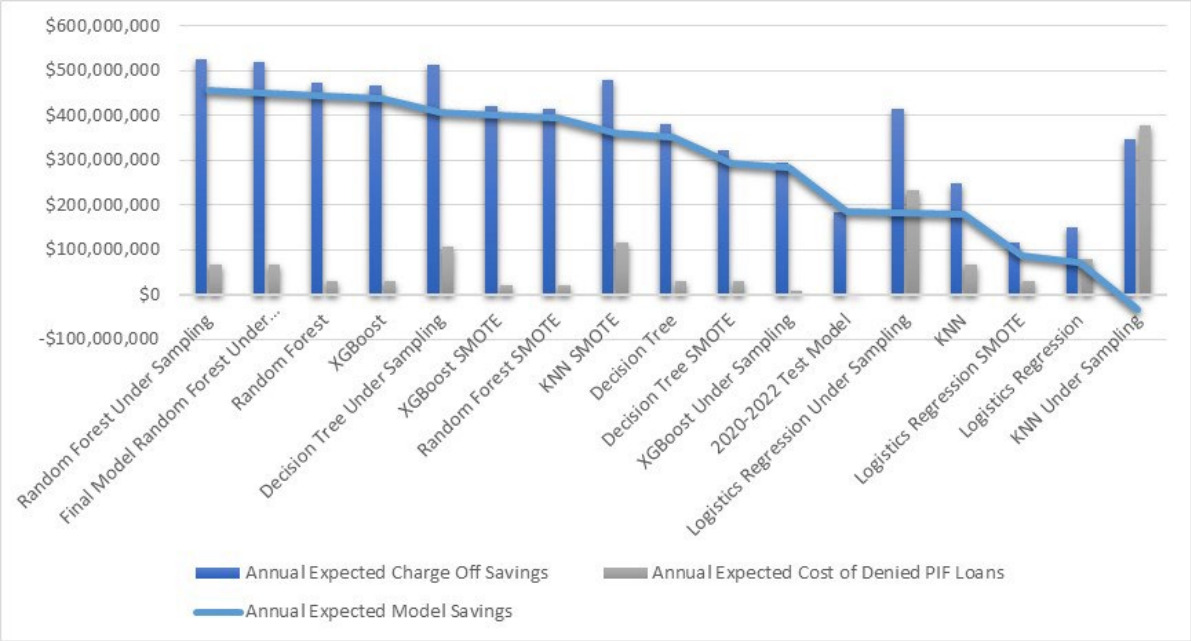
Results: Precision Results



	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

- 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.

Results: Savings Results



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 likely 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.

Random Forest Under Sampling is the best performer financially with expected savings of $456M per year. The file used to create the pickle file has an average savings pf $450M per year. Note when comparing the actuals the total is $286M per year (actuals $405M with the FNs costing $119M).

Final Model

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

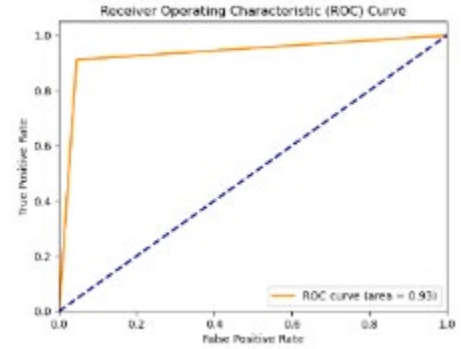
BorrName	BorrZip	BankZip	GrossApproval	SBAGuaranteedApproval	ApprovalFiscalYear	DeliveryMethod	subpgmdesc	InitialInterestRate	TermInMonths	3 Year Unemployment	Sub Program Code	Delivery Method	Loan Status	Predictions	Probabilities
A	99208	59101	142000	71000	2016	SBAEXPRES	FA$TRK(Small Loan Express)	4.75	84	0.062733333	6	14	1	1	0.97
B	85204	57104	2498000	1873500	2016	PLP	Guaranty	5.55	300	0.062733333	7	12	1	1	0.99
C	28779	57104	5000	2500	2016	SBAEXPRES	FA$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	SBAEXPRES	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

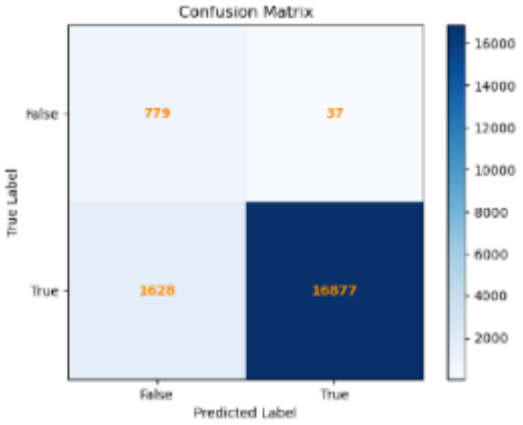
- The Random Forest Under Sampled Model is created to accept a file template and return a prediction and probability for each loan in the file.

Test 2020-2022 Loans

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. There are still current loans that may become charged off.





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	Actual Total Per Year
2020-2022 Test Model	0.32	1	0	$184,320,000	$0	$184,320,000	$21,842,078

Discussion

- **H1:**
  - Loan application data is a predictor of loan default risk.
  - Subprogram and delivery method improve accuracy. But there may be survivor bias.
  - The improvements are better than doing nothing.
- **H2:**
  - Unemployment data improves accuracy for each cohort.
  - Business age and business type were not key features for 7(a) charge off.
  - A ML model can improve charge off predictions saving up to $450M per year. (Based on actuals more likely up to $405M/year from 2010-2019 for a total of $4B.)

Conclusions

- Loan application data can predict defaults and hence charge offs.
- Subprogram and delivery method reduce Type 1 and 2 errors as well as improves the 95% lower bound.
- Borrower and lender zip codes are features for predicting 7(a) loan charge offs.
- Unemployment is a predictor for loan charge offs to balance the economic factors of each cohort.
- ML can reduce the cost of charge offs to the government.

Contributions

- Determined loan application, loan subprogram/processing method and economic features predict 7(a) loan charge offs
- Identified a charge off prediction model for SBA 7(a) loans with an expected savings of up to $450M per year.
  - The model identified $4B from 2010-2019.
  - The model identified $65M 2020-2022.
- Provided a savings model for selecting charge off prediction model.

Future Research Directions

- Refine the model to improve charge off precision and expected loss calculations.
- Future research may address second and third-party default risk factors. ML money laundering and synthetic fraud detection model improves online lending default prediction.
- A model that differentiates between credit and fraud risk would improve the policies for loan decisioning.
- The risk model can be enhanced with reducing features that target specific groups that have historically been impacted by unfair lending practices.
- Expand the model to consider local unemployment rates for predicting charge off.

Questions

?