1. Dataset

https://data.sba.gov/en/dataset/7-a-504-foia/resource/d67d3ccb-2002-4134-a288-481b51cd3479

The dataset is sourced from the US Small Business Administration's official FOIA records. It contains comprehensive data on hundreds of thousands of small business loans from 2020 through early 2026. Unlike static credit risk datasets, this includes real historical interest rates, the SBA government guarantee portion, and actual outcomes (Paid in Full vs. Charged Off) during recent high-inflation and interest rate hike cycles.

Customer Profile
Variables(X) used to train the Machine Learning models to predict the Probability of Default (PD)

Variable Name	Project Role	Business Rationale
naicscode	Core Feature	Industry Risk: The 2-digit sector code is vital. The survival rate of a restaurant versus a medical clinic differs significantly during economic shifts.
businessage	Key Feature	Maturity: Distinguishes between startups and established firms. Startups generally have a much higher "infant mortality rate."
businesstype	Model Feature	Legal Structure: Corporations, partnerships, and sole proprietorships have different risk profiles and legal liabilities.
borrstate / borrcity	Geographic Feature	Regional Economy: Used to analyze risks associated with specific locations (e.g., high-cost-of-living areas vs. industrial zones).
jobssupported	Dashboard Visual	Social Impact: While not a risk predictor, showing "Jobs Supported" on the dashboard highlights the bank's ESG (Social) contribution.

Loan Details
Inputs for the financial formulas to calculate profitability and risk costs.

Variable Name	Project Role	Business Rationale
initialinterestrate	Profit Core	Revenue (i): This is the bank’s primary income stream and the main variable we are looking to optimize.

grossapproval	Total Exposure	Loan Principal (EAD): The total amount lent, which sets the upper limit for potential financial loss.
sbaguaranteedapproval	Risk Mitigation	Government Guarantee: Acts as the bank's "safety net." The guaranteed portion is not counted toward the bank's actual loss.
terminmonths	Risk Factor	Time Risk: Longer loan terms increase the likelihood of the borrower facing an economic recession.
collateralind	Loss Calculation	Collateral: Directly impacts the Loss Given Default (LGD). Secured loans allow the bank to recover value more easily.
fixedorvariableinterestind	Profit Analysis	Rate Risk: Variable rates behave differently in high-interest environments, affecting the bank's net margin.

Targets
The "Ground Truth"(Y) for model training (Labeling) and financial back-testing.

Variable Name	**Project Role**	**Business Rationale**
loanstatus	Target Label (Y)	Model Goal: We map CHGOFF to 1 (Risk) and Paid in Full to 0 (Safe). This is the "answer" the model learns to predict.
grosschargeoffamount	Cost Back-test	Actual Loss: The real dollar amount lost during a default. Used to calculate objective LGD parameters.
approvalfiscalyear	Trend Analysis	Temporal Dimension: Allows us to compare pre-pandemic, pandemic, and high-interest rate cycles to control for macro-economic noise.

2. Proposal

This project develops a machine learning–based loan decision system that combines credit risk prediction with interest rate optimisation to maximise expected profit. By leveraging recent SBA 7(a) data, the system trains machine learning models to predict the Probability of Default (PD) while incorporating a government guarantee as a risk-mitigation tool. The project moves beyond pure classification to Profitability Optimization, using real historical interest rates to validate a pricing strategy that balances revenue growth with credit risk. The final output is a Decision Support System, a dashboard that identifies the profit-maximising

interest rate for various industry segments (NAICS) and business types while accounting for market-driven price elasticity.

Notes from Rachel:
-Focus on which features we want to use first for predicting
-If we want to do prediction, we need to decide on pricing strategy (i.e. what features we want to look at to predict interest rates e.g. loan amount, business type)
-Look at descriptions of all variables and what they mean, have a set amount of variables (shouldn't have a ton), correlation heat map to see what's more closely correlated with interest rates. If continuous variable (it is hard to determine), categorize interest into high and low if it is hard to interpret.

3. Expected Outcome

Outcome 1: Customer Risk Profiling & Predictive Dashboard

Objective: To identify high-risk segments and quantify the key drivers of loan defaults using machine learning.

- Risk Segmentation (Tiering): Categorizes borrowers into distinct Risk Tiers (e.g., Tier 1: Low, Tier 2: Medium, Tier 3: High) based on their predicted Probability of Default (PD).
- Industry Risk Heatmap: Visualizes default concentration across various NAICS sectors, identifying which industries are most vulnerable under current economic conditions (2024–2026).
- Feature Importance Analysis: Highlights the most influential variables—such as Business Age, Project Location, and Collateral Status—that determine a borrower’s creditworthiness.
- Target Audience: Credit Risk Officers and Underwriters.

Outcome 2: Strategic Profitability & Pricing Simulator

Objective: To determine the optimal risk-adjusted interest rate that maximizes bank profitability while maintaining market competitiveness.

- Dynamic Profit Curves: Illustrates how Net Profit fluctuates as interest rates change, helping managers find the "Sweet Spot" where revenue outweighs the Expected Loss (EL).
- SBA Guarantee Impact Analysis: Specifically quantifies how the Government Guarantee Percentage reduces the bank's Net LGD (Loss Given Default), allowing for more aggressive pricing.
- "What-If" Strategy Simulation: Allows users to simulate macroeconomic shifts (e.g., interest rate hikes) to see how the total portfolio profit responds in real-time.
- Target Audience: Business Development Managers and Strategy Analysts.

Project Value Proposition

By integrating these two dashboards, the project addresses the trade-off between maximizing interest margins and mitigating customer churn, providing a data-driven answer to the "Optimal Pricing" challenge in a competitive lending market.

4. Project Plan

Week	What we will do (Tasks)	Skills we will use
1-2	**Data Engineering & Decoding:** Filter for closed loans (Paid vs. Charged Off); Group 6-digit NAICS codes into 2-digit sectors.	SQL & Python (Data Cleaning)
3	**Predict Risk:** Build a classification model (XGBoost/Random Forest) to find the **Probability of Default (PD)**.	Machine Learning (Python)
4	**Financial Logic:** Integrate the **SBA Guarantee** into the profit formula to set risk-adjusted interest rates.	Financial Modeling (Math)
5	**Build the Dashboard:** Create a **Strategy Simulator** in Tableau to test "What-If" scenarios for interest rate hikes.	Tableau (Visualization)
6	**Final Pitch:** Present how the model-driven strategy outperforms historical bank pricing in terms of Net Profit.	Data Storytelling

5. Machine Learning Variable Mapping

Category	Key Columns from SBA Dataset	Role in the Project	Plain English Explanation
1. ML Inputs (Predicting PD)	naicscode, businessage, businesstype, projectstate, collateralind	Used to train the model to calculate the Probability of Default (PD).	These factors tell us how likely a business is to fail based on its industry, age, and location.
2. Financial Variables	grossapproval, initialinterestrate, terminmonths, loanstatus	Used to calculate Expected Loss (EL) and total interest income.	grossapproval is the total money lent. We use this to calculate the actual dollars at risk.
3. Strategic Dimensions	sbaguaranteedapproval, bankname, approvalfiscalyear	Used in Tableau for segment-based profit analysis.	These help us see how the government guarantee or different banks impact overall profitability.

6. Calculation of Financial Formulas

1. Group Customers into "Risk Tiers"

find the best rate for each **segment**

- **Tier 1 (Low Risk):** PD < 5\%
- **Tier 2 (Medium Risk):** PD between 5% to 15%
- **Tier 3 (High Risk):** PD > 15%

2. The Profitability Formula

In your Tableau dashboard, you will calculate the **Net Expected Profit** for every loan using this logic:

Net Profit = Interest Income - Expected Loss - Cost of Funds - Operating Expenses

- **Interest Income:** grossapproval x initialinterestrate
- **Expected Loss (EL):** PD x LGD x (grossapproval - sbaguaranteedapproval)
    - *Note: Risk is only calculated on the portion NOT guaranteed by the SBA.*
- **Cost of Funds (CoF):** grossapproval x Benchmark Rate (e.g., 4%)
- **Operating Expenses (OpEx):** Fixed fee per loan or 1% of the loan amount.

The Three Components of the Formula

A. Interest Income (The Revenue)

This is based on the actual historical rates found in the dataset.

Interest Income = grossapproval x initialinterestrate

- **Example:** For a **$100,000** loan (grossapproval) at a **10%** rate (initialinterestrate), the Income is **$10,000**.

B. Expected Loss (The "SBA-Adjusted" Risk Cost)

In a standard loan, the bank risks the whole amount. In an SBA loan, the government guarantees a huge chunk (usually 75%-85%). The bank only calculates risk on the **unguaranteed portion**.

Expected Loss (EL) = PD x LGD x (grossapproval - sbaguaranteedapproval)

- **Example:** If your ML model predicts a **5% PD**, and the SBA guarantees **$75,000**, the bank only risks the remaining **$25,000**.
- **Calculation:** $0.05 x 0.50 x ($25,000) = $625 (Instead of $2,500 without the guarantee!).

C. Operating & Funding Costs (The Business Reality)

To be realistic, we subtract the cost of borrowing the money and the cost of the staff/tech.

Total Costs = Cost of Funds (CoF) + Operating Exp (OpEx)

Variable	Symbol	Source	Suggested Value
Interest Rate	i	SBA Dataset / Simulation	Actual (e.g., 6.5% - 12%)

Loss Given Default	LGD	Calculated from `grosschargeoffamount`	~50% (Industry Standard)
Cost of Funds	CoF	Market Data (e.g., SOFR + Spread)	4% - 5%
Operating Exp.	OpEx	Internal Assumption	1%

How this helps you find the "Best Rate"

The **"Best Rate"** is the one where your Net Profit is highest.

- **If you set the rate too low (e.g., 3%):** Your Income ($30) won't even cover your Risk ($25) and Overhead ($20). You lose money.
- **If you set the rate too high (e.g., 25%):** Good customers will leave, and only the riskiest people will accept the loan, causing your **PD** (and thus your Expected Loss) to skyrocket.

Managerial Decision Support

Dashboard Insight	Managerial Question	Action Taken
Low Net Margin in High-Risk NAICS	Is our interest rate covering the risk for this industry?	Increase the risk premium for that specific industry code.
High Churn in Low-Risk Segments	Are our rates too high for the best businesses?	Lower the interest rate to remain competitive against other banks.

SBA Guarantee Impact	How much does the government guarantee help our bottom line?	Prioritize loans with higher guarantee percentages to lower capital requirements.

Finding the "Sweet Spot" (The Optimization)

To find the **Best Rate**, you look for the point where you maximize profit without losing your best customers. This is a balance of two forces:

- **The Risk Force:** If the rate is too low, you don't make enough money to cover the people who default.
- **The Competitive Force (Elasticity):** If the rate is too high, the "Good" customers (Low PD) will leave and go to a competitor like HSBC or Monzo.

Dashboard Insight	Managerial Question	Action Taken
High Expected Loss	Are we taking too much risk?	Tighten the credit score requirements.
Negative Net Profit	Are our loans too small or too cheap?	Increase the minimum loan amount or the base rate.
Segment Performance	Who is our "Best" customer?	Launch a loyalty program for that specific demographic.
OpEx Overload	Is our process too expensive?	Automate the approval process for low-value loans.