Executive Summary: Predictive Modeling with SBA National Data

By: Brandon Chang, Darien Lin, & Dylan Ton

Tables of Contents

1 Background

Bank Loan Scenarios

Banks generate revenue by issuing loans to businesses and earning interest on repaid loans. However, lending decisions carry financial risks. The outcomes of issuing loans can be categorized into three scenarios:

1. **Loan Repaid**:
    - The bank earns a profit equal to 5% of the loan amount.
2. **Loan Not Issued**:
    - The bank neither gains nor loses money since no loan is disbursed.
3. **Loan Defaulted**:
    - The bank incurs a loss equal to 25% of the loan amount issued.

Dataset Overview

The dataset consists of SBA data about small businesses that either repaid their loans in full or defaulted. This data exists because small businesses frequently submit loan requests to banks through the SBA program. We can use this information to assess whether banks should approve or deny these loan requests, aiming to minimize financial risk while supporting eligible businesses.

- **Dataset size:** 899,164 rows × 27 columns
- **Features:** LoanNr_ChkDgt, Name, City, State, Zip, Bank, Bank State, NAICS, Approval Date, ApprovalFY, Term, NoEmp, NewExist, CreateJob, RetainedJob, FranchiseCode, UrbanRural, RevLineCr, LowDoc, ChgOffDate, DisbursementDate, DisbursementGross, BalanceGross, MIS_Status, ChgOffPrinGr, GrAppv, SBA_Appv

2 Problem Statement

The goal is to develop a machine learning model that predicts whether a small business will fully repay a loan or default. This model aims to assist banks in making informed lending decisions, reducing financial risk, and maximizing profits by identifying businesses that are more likely to repay their loans.

3 Approach

To effectively address the problem and understand our stakeholders, we began with comprehensive background research on banks and the Small Business Administration (SBA). This step provided valuable context about the stakeholders' goals and the challenges they face. Next, we analyzed the profitability framework, ensuring a clear understanding of how profits are calculated for each type of prediction (e.g., repayment or default).

We conducted an in-depth review of the dataset, researching the meaning and significance of each feature to determine its utility in modeling. This allowed us to identify and exclude features irrelevant for future loan predictions. Simultaneously, we studied various machine learning (ML) models to decide the most appropriate techniques for this project.

After cleaning the dataset and engineering features, we optimized the variables to minimize overfitting and maximize model performance. The focus was on achieving high specificity, precision, and average profit. Specificity was crucial to avoid costly false positives (incorrectly predicting repayment and granting a loan that defaults), while precision ensured that loans predicted as likely to be repaid were accurate, driving significant profits. Balancing these priorities allowed the model to maximize profitability effectively.

We applied this approach across several ML models, including logistic regression, neural networks, discriminant analysis, k-nearest neighbors (KNN), bagging, boosting, and single decision trees (CART). Each model was evaluated using validation data and metrics such as ROC/AUC curves, gains charts, and lift charts. The goal was to identify the model that yielded the highest average profit per loan.

To further enhance the model’s performance, we explored additional business strategies and provided practical recommendations. Finally, we created a comprehensive video for stakeholders, demonstrating how to use the model and its practical applications in improving their decision-making and profitability.

4 Key Results

The table below summarizes the performance of various machine learning models, highlighting their average profit per loan and total profit. This analysis helps evaluate the effectiveness of each model in predicting loan repayment outcomes. The values in the table can be used to compare how each model performs in terms of profitability, assisting in selecting the best model for loan decision-making. Here's how to interpret it:

Profit/Loss Analysis for Each Model

Model	Average Profit ($)	Total Profit ($)
Bagging (Random Forest)	$7,452.13	$1,951,532,847.00 (from 261,876 instances)
Single Tree	$7406.72	$1,293,094,098.05 (from 261,876 instances)
Boosting	$7123.60	$2,487,332,690.10 (from 261,876 instances)
Neural Network (MLP)	$6,416.44	$1,119,969,849.60 (from 174,547 instances)
Logistic Regression	$5,005.65	$1,310,858,367.60 (from 261,876 instances)
Discriminant (Quadratic)	$4,503.11	$786,170,263.75 (from 174,584 instances)
KNN	$4015.04	$80,300,795.0 (from 20,000 instances)

Highest Avg. Profit

Lowest Avg. Profit

The table provides insights into the performance of various machine learning models based on their average profit per loan and total profit. Here's how to interpret it:

- If you use a particular model, the average profit indicates how much money the bank can expect to make per loan (on average) based on all the model's predictions.
- For example, if you grant loans to 100,000 businesses:
    - Using the Bagging (Random Forest) model with an average profit of $7,452.13, the total profit would be approximately $745,213,000.
    - Using the Neural Network (MLP) model with an average profit of $6,416.44, the total profit would be approximately $641,644,000.
    - With KNN, the average profit is only $4,015.04, leading to a much lower total profit of $401,504,000.

From this analysis, Bagging (Random Forest) is the most profitable model, with an average profit of $7,452.13 per loan, accurately predicting repayments and minimizing defaults.

Given these results, the Bagging (Random Forest) model is the best choice to maximize the bank's profitability when granting loans. By prioritizing this model, the bank can confidently make more accurate decisions and achieve the highest possible returns.

Training Accuracy Metrics for Each Model

Model	Accuracy	Precision	Recall	F-1 Score	Specificity
Bagging (Random Forest)	90.66%	97.92%	90.66%	94.15%	90.68%
Single Tree	90.57%	97.85%	90.62%	94.09%	90.16%
Boosting	89%	97%	89%	93%	88.82%
Neural Network	87.31%	96.72%	87.66%	91.97%	85.62%
Logistic Regression	70.66%	94.87%	68.20%	79.39%	82.54%
Discriminant (Quadratic)	64.48%	81.44%	64.48%	68.85%	74.65%
KNN	88%	90%	96%	93%	48.41%

Based on the metrics, Bagging (Random Forest) is the most reliable model for loan prediction and should be prioritized. It performs consistently across all key metrics, ensuring effective risk management and profitability. Here's how each metric translates to business terms for loans:

- **Specificity**: Identifies loans likely to default. High specificity minimizes losses by reducing approvals for risky loans.
- **Precision**: Ensures that approved loans are highly likely to be repaid. High precision reduces costly false positives.
- **Recall**: Captures loans that will be successfully repaid. High recall helps maximize profits by approving viable loan opportunities.
- **F1-Score**: Balances precision and recall. A high F1-score ensures the model can approve profitable loans while avoiding defaults effectively.
- **Accuracy**: Reflects overall correctness of predictions. High accuracy means the model performs well across both loan repayment and default predictions.

The Bagging (Random Forest) model delivers the best performance in all these areas, striking the optimal balance for loan approval decisions. Single Tree is a close second and could be an alternative for simpler implementation. However, models like KNN and Discriminant should be avoided due to poor specificity and accuracy, which could lead to higher financial losses.

Validation Accuracy Metrics for Each Model

Model	Accuracy	Precision	Recall	F-1 Score	Specificity
Bagging (Random Forest)	90.54%	97.93%	90.51%	94.08%	90.69%
Single Tree	90.42%	97.81%	90.47%	94.00%	90.16%
Boosting	89.00%	97.00%	89.00%	93.00%	88.27%
Neural Network	86.92%	96.41%	87.51%	91.74%	84.04%
Logistic Regression	70.76%	95.00%	68.34%	79.50%	82.51%
Discriminant (Quadratic)	64.60%	81.56%	64.60%	68.96%	74.94%
KNN	87%	89%	95%	92%	44.07%

The validation accuracy metrics reflect how well the models perform on unseen data. These results provide a realistic assessment of the models' ability to generalize beyond the training set. The validation set metrics are very close to the training set metrics for most models, suggesting that overfitting is not a major issue for most models.

The Bagging (Random Forest) model remains the most robust and reliable, offering a strong balance across all metrics. It generalizes well to unseen data, with high accuracy, precision, recall, and specificity. Models like Single Tree, Boosting, MLP, Logistic Regression, and Discriminant also perform well, making them good alternatives. However, KNN should be avoided due to its overfitting issues, particularly in specificity, which could lead to significant financial losses by incorrectly approving risky loans.

5 Recommended Model

The Bagging (Random Forest) model is the recommended choice for the bank because it consistently generates the highest profit compared to all other models. With its robust performance across all key metrics—accuracy, precision, recall, F1-score, and specificity—it effectively minimizes the risk of loan defaults while maximizing the number of correctly approved loans. This balance ensures that the bank achieves the highest profitability, making it the optimal model for decision-making in loan approvals.

6 Implementing the Model for Optimal Results

We have provided a video demonstration on how to use the model effectively. In summary, the process begins by uploading your data into a data cleaning notebook, where the dataset is preprocessed and prepared for model input. After cleaning and transforming the data, the model processes it to predict whether the bank should approve or reject a loan application. The final output includes both the decision—'Approve' or 'Reject'—as well as the probability associated with each decision, providing valuable insights for loan approval processes.

For a step-by-step guide, refer to the video: Loan Approval Prediction Model Demonstration

7 Bank Strategies

We have prepared a set of slides to guide you through the data analysis and business strategies for banks. These slides provide a comprehensive overview of how the model can be applied to real-world banking scenarios. They cover key aspects such as analyzing customer data, identifying trends, and making data-driven decisions to optimize loan approval processes. Additionally, the slides highlight how business strategies can be enhanced using the insights gained from the model's predictions.

For a detailed presentation, please refer to the slides: Loan Approval Business Strategy Deck

Additionally, you can explore loan payments trends and patterns through the following dashboard link: Loan Payments Dashboard

8 Conclusion

The results of this project clearly demonstrate how predictive modeling can revolutionize decision-making in loan approvals, directly increasing profitability for banks. Among all models evaluated, the Bagging (Random Forest) model is the most effective and tailored solution for our stakeholders, generating the highest average profit of $7,452.13 per loan, far exceeding other models like Neural Networks ($6,416.44 per loan) and KNN ($4,015.04 per loan). For every 100,000 loans approved using this model, the bank could earn approximately $745 million in total profit, highlighting its significant financial impact.

The model achieves this exceptional profitability by excelling in all key performance metrics:

- **Accuracy**: 90.54%, ensuring the model reliably predicts repayment or default outcomes.
- **Precision**: 97.93%, minimizing costly false positives by ensuring loans predicted as repaid are highly likely to succeed.
- **Recall**: 90.51%, capturing many loans that will be successfully repaid, maximizing profit opportunities.
- **Specificity**: 90.69%, effectively identifying and rejecting risky loans, preventing unnecessary losses.
- **F1-Score**: 94.08%, achieving a balanced measure of precision and recall optimizing decision-making.

These results emphasize the model's ability to align with the bank's primary goals—reducing losses from defaults while approving the most profitable loans. The Bagging model is specifically designed to cater to the needs of banks, addressing their unique challenges by leveraging accurate predictions to improve loan approval strategies.

For our stakeholders, this model is not just a tool—it is a tailored solution to increase profitability and reduce risk. By integrating this model into the Loan Approval Business Strategy Deck , stakeholders can expect a significant boost in overall financial performance. The Loan Approval Prediction Model Demonstration and Loan Payments Dashboard created as part of this project ensure ease of implementation, allowing stakeholders to apply the model seamlessly in real-world scenarios.

This project demonstrates the power of data-driven decision-making to empower banks, increase profits, and support small businesses effectively. By adopting the Bagging model, stakeholders are positioned to make smarter, more informed loan approvals, maximizing returns while minimizing risk. This model is built for you—to drive success, profitability, and long-term growth.