Data Science Capstone Project July 2023

Credit Alchemy: Small Business Creditworthiness Demystified

Project Collaborators

Brian Njau

Grace Nyongesa

Josiah Okumu

Leo Kariuki

Lynne Mutwiri

Stella Kitur

GitHub Repository

TABLE OF CONTENTS

Business Understanding

Business Overview

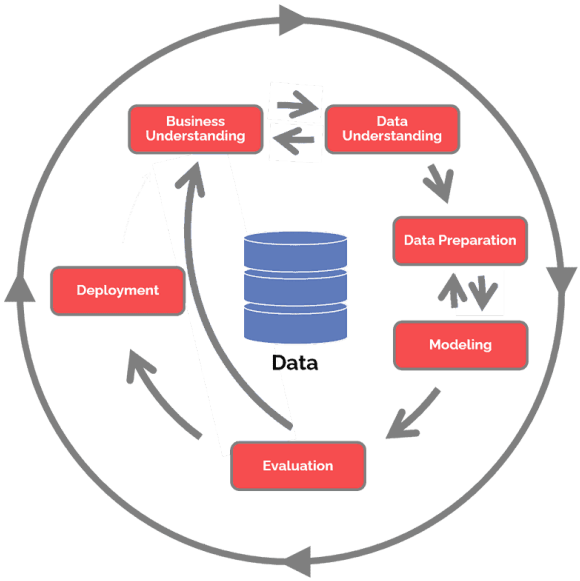
The Small Business Administration (SBA) is a crucial organization in the United States that supports and promotes small enterprises by providing access to credit. Through loan guarantees, the SBA encourages lenders to credit small businesses, which play a vital role in job creation and reducing unemployment. However, there have been instances of loan defaults, posing a challenge for the SBA in accurately assessing the creditworthiness of applicants and mitigating risks.

To address this, the SBA focuses on evaluating creditworthiness by considering factors such as *financial history, business plans, collateral, and projected cash flows*. They also strive to adapt to market dynamics, embracing innovative approaches and technology for informed lending decisions that minimize the likelihood of defaults. Overall, the SBA's commitment to supporting small businesses goes beyond loan guarantees, encompassing mentorship programs, training resources, government contracting opportunities, and disaster recovery assistance to foster an environment conducive to small business success.

Stakeholders: Small Businesses, Governments, and Lending institutions such as the SBA.

The Data Science Process that is adhered to in this analysis is the **CRISP-DM** Process, as illustrated in the image below.

Fig. 1 : CRISP-DM iterative process



Business Challenges

- ➔ Small businesses, particularly those in their early stages often face challenges in accessing credit due to limited financial resources and a lack of collateral. When starting a new business, entrepreneurs typically have minimal or no track record of financial performance, making it difficult for traditional lenders to assess their creditworthiness.
- ➔ Additionally, most banks want to see a business's growth projections before financing them. However, small businesses often grow slowly, which makes it hard for banks to service their loans.
- ➔ Understanding the creditworthiness of small businesses is crucial for lenders, policymakers, and investors. Assessing the viability of small businesses ensures that lenders can mitigate risks and make informed decisions when extending credit as well.

Project Overview

This project aims to develop a predictive model for SBA-guaranteed loan applications. It will analyze financial indicators and metrics to accurately assess the likelihood of loan approval, helping small business owners make informed decisions. Leveraging machine learning and statistical techniques, the model will streamline approvals and identify key financial factors. Ultimately, the project addresses credit access challenges for small businesses, providing a data-driven tool for assessing loan approval probabilities. By streamlining the process and fostering informed decision-making, it supports small business growth while mitigating risks for lenders and policymakers.

Business Problem Statement

By analyzing the financial indicators and metrics that influence loan approval decisions, our goal is to develop a reliable predictive model that accurately assesses whether an SBA-guaranteed loan application will be approved or rejected.

This project aims to address the challenge faced by small business owners in understanding their chances of securing funding and provide them with a tool to make informed decisions, while also streamlining the loan approval process for the SBA and identifying key financial factors that impact loan approvals.

Project Objectives

The main project objectives will be to:

1. Conduct a comprehensive analysis of the dataset from the U.S. Small Business Administration to identify patterns and trends for accurate loan approval predictions.
2. Develop a robust machine learning model that utilizes the identified patterns and trends to predict loan approval outcomes effectively.
3. Deploy and evaluate the developed machine learning model in the loan approval process of the U.S. Small Business Administration, continuously optimizing its predictive capabilities for informed lending decisions.

Data Understanding

Our data was sourced from the official SBA Open Data source. The data was then merged based on important features to be used. The data from kaggle had 899164 rows and 27 columns and the data from data.gov had 168211 rows and 40 columns.

Data Preparation included the following: Handling of missing values. Changing of data types to appropriate format. We checked for duplicates and dealt with them in the notebook. Feature engineering was also done to some columns that were important for modeling

Success Metrics

We settled on accuracy as our success metric and concluded the following that where:

- If the accuracy score is at least 85 % or higher the model would classify/filter the loan applications as either Approved or Rejected.

The accuracy score is preferred in a loan classification/filtering model as it provides a straightforward measure of how well the model correctly predicts loan approvals or rejections, allowing for a clear assessment of its overall performance and effectiveness in minimizing false positives and false negatives, which is crucial in this case.

Modeling

Six models were used and gave the following Accuracy as illustrated in the table below:

Model	Train Accuracy	Test Accuracy
Logistic Regression	78%	75%
Decision Tree	98%	89%
Random Forest	87%	88%
Support Vector Machine	77%	78%
Neural Network	89%	89%
XGBoost	**91%**	**90%**

Table 1: Model and the corresponding accuracy percentage values

Evaluation

After determining that the performance of the Baseline model did not meet our predefined metrics of success, we embarked on exploring alternative models in order to enhance the accuracy and efficacy of our loan classification/filtering system. Among the various models evaluated, XGBoost emerged as the best-performing model, exhibiting an exceptional performance with a remarkable test accuracy of up to 91%. Recognizing the significance of this substantial improvement, we decided to leverage XGBoost as our primary model for deploying the loan classification app, ensuring its effectiveness in accurately predicting loan approvals or rejections.

Deployment

The next step in the project was to create a web application that would help us to meet our project objectives effectively. This loan filtering app was deployed on Streamlit.

The input features were: Gross Approval, Industry, Loan Term, Number of Employees, NewExist, and UrbanRural to predict loan approval.

Upon doing a feature importance on the pre-trained model, we found out that the Loan Term and the Gross Approval amount are the two key features in determining whether a loan would pass the initial filtering stage.

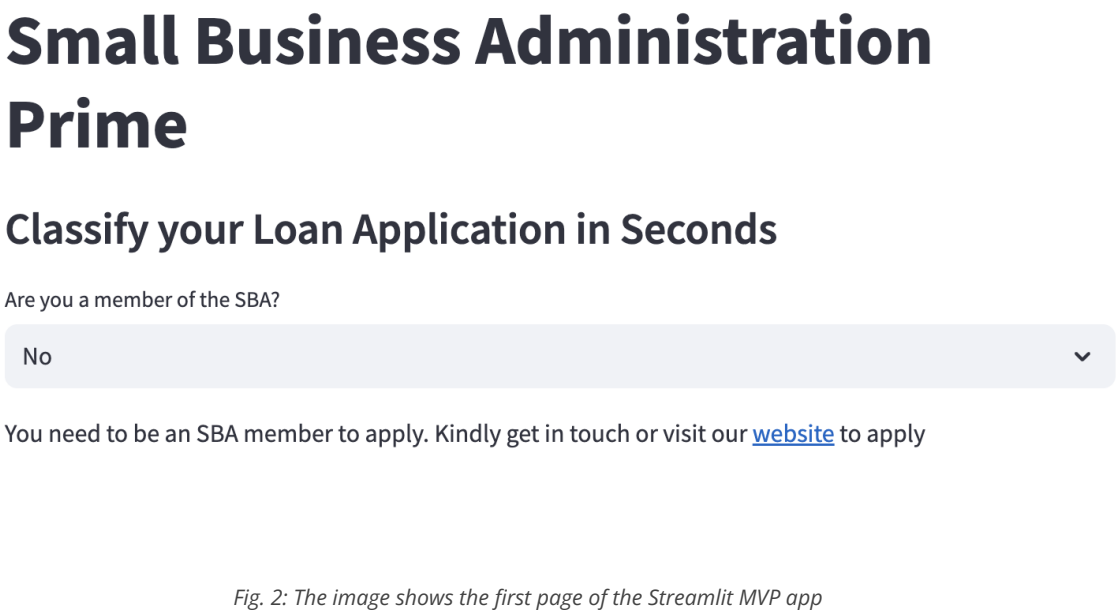


Fig. 2: The image shows the first page of the Streamlit MVP app

Recommendations

Based on the analysis that we have done, we came up with five main areas in which the SBA can dive deeper, these areas included:

1. **Loan Term**

   The SBA should consider advising small businesses on the appropriate loan amount to apply for based on the loan amount and term distribution. This can be achieved by identifying the most common loan sizes and advising small businesses to apply for loans within those ranges. Additionally, we recommend that the SBA identifies the most common loan terms and set loan term limits or eligibility criteria based on those terms.

2. **Size of the Business**

   We advise the SBA to look into the loan amount ranges and/or eligibility criteria for businesses of different sizes based on employee distribution. Through this, we recommend that they set loan amount limits that would be appropriate for different business sizes.

3. **Location**

   We would recommend the SBA advise the small businesses to seek loans that are appropriate for their location and the needs of their local community. Additionally, the SBA should ensure the distribution of loans is easily accessible to those in the Rural/Semi-urban regions, in order to encourage development

4. **Business Existence Period**

   Considering whether the business has been in existence for a long or short period of time, we recommend that the SBA advises the businesses seeking loans to apply for loans that are in alignment with their growth plans, additionally, the SBA can recommend and/or implement new training and mentorship programs to help the businesses achieve their targets.

5. **Industry**

   Finally, we would suggest that the SBA advises its members to apply for loans that are tailored to the specific needs of their industry.

Conclusion

In conclusion, the enhanced predictive capabilities of the SBA have revolutionized the loan application process, enabling automated filtration and empowering clients to identify crucial factors influencing their loan eligibility. This advancement not only streamlines the lending process but also fosters a supportive environment for entrepreneurs and small businesses. By encouraging more individuals to seek SBA support, this initiative contributes to reducing unemployment rates and fostering economic growth through increased business opportunities and job creation.