Devyn Hughes, Zinnia Kamal, Randy Garcia

6/18/2024

IST 718

Prof. Jillian Lando

IST 718 Final Project

Prioritizing Proft with SBA 7A Loans

**Introduction**

The banking industry is one of the largest industries in the world. This term is just an umbrella term for all that can be done. This area of the economy is one of the biggest powerhouses for development and growth within a nation. In the US alone, the finance industry amasses majority of the wealth in the United States alone. According to the FDIC, the United Sates estimated revenue in 2023 was 256.9 billion dollars (FDIC,2024). Within this massive, and monumental business, various forms and arrangements of data are created. From simple daily transactions on a day-to-day basis to insurmountable expenses on a national or even global scale, this is all information that can be reviewed, evaluated, and utilized.

For this project, a small, minute portion of data relative to the banking industry was collected. From Kaggle, our team downloaded a data set that involves SBA 7a Loans. An 7a loan is the most common loan granted by the SBA, and is usually used for working capital, debt refinancing, etc. The SBA is the Small Business Administration in which banks and credit unions work with them to promote a loan. For example, if a small business was seeking to improve their own capital or grow their business, the institution can seek out the assistance of the Small Business Administration. With the data acquired about these given loans, we will be evaluating some business questions to try and find ways to improve the outcome of SBA loans.

**The Data**

The SBA 7A Loan dataset contains three CSV files that present all publicly available data from 1991 though year-end of 2022. This includes their approval amounts, terms, statuses, and other related financial details. There are a total of 39 different variables with more than 1 million rows. Below are the columns listed for the project.

* AsOfDate = Loan Status as of Date
* Program = Loan Program
* BorrName = Borrower Business Name
* BorrStreet = Borrower Street Address
* BorrCity = Borrower City
* BorrState = Borrower State
* BorrZip = Borrower Zip Code
* BankName = Name of Bank
* BankFDICNumber = FDIC Number of Funding
* BankNCUANumber = National Credit Union Assoc. Number
* BankStreet = Bank Address
* BankCity = Bank City
* BankState = Bank State
* BankZip = Bank Zip Code
* GrossApproval = Total Amount of Loan Approved
* SBAGuaranteed = Amount guaranteed by the SBA
* Approval Date = Approval Date of Loan
* ApproveFiscalYear = SBA Fiscal Year of Approved Loan
* FirstDisbursement = Date of Loan First Disbursement
* DeliveryMethod = Delivery Method
* SubPGMDesc = Subsidized Loan Data
* InitialinterestRate = Initial Interest Rate of Loan
* TermInMonths = Loan Tenor, Months
* NAICSCode = NAISCS Code of Borrowe
* NAICSDescription = Description of Borrower via NIACS Code
* FranchiseCode = Franchise Code of Borrower
* FranchiseName= Name of Franchise
* ProjectCounty = County of Business
* ProjectState = State of Project
* SBADistrictOffice = Responsible SBA Office for Loan
* CongressionalDist = Congressional District of Laon
* BusinessType = Type of Business
* BusinessAge = Age of Business
* LoanStatus = Ending Status of the Loan
* PaidInFullDate = Pay Off Date of Loan
* ChargeOffDAte = Date of Charge Off
* GrossChargeOff = Amount Charged Off
* RevolverStatus = Credit Revolver or Not
* JobsSupported = Estimated Number of Jobs Created

**Business Question/Project Pathway**

As expected of loans, the best result of lending a person or entity a loan is to profit from the loan. This is done by granting a certain amount of money that begins to accrue interest over time. The worst possible result for a loan is for the account to default or be charged off. Charging off the account means that the remaining amount of the loan is concluded to not be paid in full and is a loss of money in the overall scheme of money. Therefore, how can the SBA of the loaning entity find a way to balance profitability and risk mitigation in SBA loan allocations? By utilizing the SBA 7A loan data, the goal is to understand and identify key predictors of repayment reliability. This will empower a institution with actionable insights to refine lending strategies, minimize defaults, and enhance portfolio performance to ensure sustained profitability and prudent risk management.

**Preliminary Data Exploration**

For our preliminary data exploration, reviewing the data was a crucial step in determining our business questions. First, we determined the type of each column, identifying categorical columns vs. numerical columns. Also, we calculated the number of Non-Null Values to start cleaning the data for further processing. Using Python, we were able to identify the identity of each of the columns. The data that was collected was not fully complete as there were many null values shown in Figure 1, below. The result of the summary determines that there will need to be some cleaning of the data to make sure there are proper evaluations of the data.

A screenshot of a computer

Description automatically generated

**Figure 1**

After analyzing our data on a preliminary level, we were able to begin making some changes for a proper assessment. First, we removed any column that had more than 40% of the total length as NA’s. This helped create a more streamlined data set to then create categorical and numerical sets to assist with feature creature. Through our findings, we did have some rows that had “MISSINGBANKID”. For better accuracy, we removed those rows from the data. Here is an example of the code utilized for **Cleaning and Categorizing the Data.**

A screenshot of a computer code

Description automatically generated

**Figure 2**

**Chi Squared Test**

The analysis of the associations between BorrState and GrossChargeOffCategory, as well as BankName and GrossChargeOffCategory, reveals significant insights into the distribution patterns of gross charge-offs across states and banks.

First, the Chi-Square test for the relationship between BorrState and GrossChargeOffCategory produced a Chi-Square statistic of 4034.32 with 336 degrees of freedom, accompanied by a p-value of 0.0. This extremely low p-value, well below any common alpha level such as 0.05, indicates a highly significant relationship between the two variables. The data clearly show that the distribution of GrossChargeOffCategories is not uniform across different states. The expected frequencies, which represent the distribution that would be expected if there were no association between BorrState and GrossChargeOffCategory, deviate significantly from the observed frequencies. These deviations contribute to the Chi-Square statistic and confirm the dependency between state and charge-off category. The contingency table, which details the observed counts of GrossChargeOffCategories for each state, further supports this finding by revealing substantial variations from the expected values.

Similarly, the Chi-Square test for the association between BankName and GrossChargeOffCategory resulted in an even larger Chi-Square statistic of 74635.91 with 14958 degrees of freedom, and a p-value of 0.0. This also underscores a very strong relationship between these variables. The expected frequencies, illustrating how GrossChargeOffCategories would be distributed across banks if there were no association, differ markedly from the observed frequencies. These differences contribute to the high Chi-Square value, indicating that the distribution of charge-off categories is significantly different across banks.

In both tests, the p-values being effectively zero provide overwhelming evidence to reject the null hypothesis of no association between the variables. This confirms that there are indeed significant associations: both geographic (state-level) and institutional (bank-level) factors play crucial roles in the distribution of gross charge-offs.

In conclusion, the Chi-Square tests reveal that gross charge-off categories are not uniformly distributed either across different states or among different banks. The association between BorrState and GrossChargeOffCategory suggests that state-specific factors might influence charge-offs, while the relationship between BankName and GrossChargeOffCategory implies that institutional policies or customer demographics linked to specific banks are significant determinants. These findings highlight the importance of considering both geographic and institutional contexts when analyzing and managing gross charge-offs.

The Chi-Square test for the association between BusinessType and GrossChargeOffCategory resulted in a Chi-Square statistic of 6872.66 with 12 degrees of freedom and a p-value of 0.0. This highly significant result indicates a strong relationship between these two variables, suggesting that the distribution of GrossChargeOffCategories varies significantly across different business types. The expected frequencies, which represent the expected counts if there were no association between BusinessType and GrossChargeOffCategory, show substantial deviations from the observed frequencies. For instance, corporations have far higher actual counts in the "Low" and "Very Low" categories than expected, while individuals and partnerships show lower actual counts in higher charge-off categories compared to the expected counts. These discrepancies contribute to the high Chi-Square statistic. The contingency table further highlights these differences, with corporations exhibiting significantly higher counts in most categories, especially "Low" and "Very Low," compared to individuals and partnerships. This analysis underscores the impact of business type on the distribution of charge-off categories, indicating that corporations, individuals, and partnerships experience different patterns of charge-offs.

**Data Analysis and Models**

In this section, we delve into the detailed steps involved in the data analysis and model development for predicting SBA 7(a) loan repayment outcomes. We employed two machine learning models: a Decision Tree Classifier and Logistic Regression. Both models were evaluated for their effectiveness in classifying loans as either fully paid or charged off.

## **Decision Tree Classifier**

The Decision Tree Classifier is a supervised learning method used for classification and regression. It splits the dataset into subsets based on the value of input features, forming a tree structure where each node represents a decision point based on a feature, and each leaf node represents a class label (in this case, fully paid or charged off). The primary objective of using a Decision Tree Classifier in our analysis is to create a model that can segment the loans into different risk categories, allowing for a clear and interpretable set of rules for predicting loan repayment status.

The Decision Tree algorithm works by recursively partitioning the data into subsets that contain instances with similar values (homogeneous). Each split is made based on a feature that results in the most significant reduction in impurity, which can be measured using criteria such as Gini impurity or information gain. In our analysis, relevant features such as LogGrossApproval, SBAGuaranteedApproval, TermInMonths, and GrossChargeOffAmount were selected, as they significantly contribute to predicting the loan status.

The cleaned dataset was split into training and testing sets using an 80-20 split to ensure the model is trained on a substantial portion of the data while retaining a significant portion for evaluation. We trained the Decision Tree Classifier on the training set and evaluated its performance using a confusion matrix. The confusion matrix showed the number of true positives, true negatives, false positives, and false negatives, providing clear insights into the model's accuracy and misclassification rates.

The Decision Tree model effectively classified loans into fully paid and charged-off categories. The confusion matrix revealed a high number of correct classifications (true positives and true negatives) and a lower number of incorrect classifications (false positives and false negatives). Specifically, the model correctly predicted a large majority of the fully paid loans, with fewer misclassifications. This accuracy demonstrates the model's reliability in identifying loan repayment status. Additionally, the analysis of feature importance indicated that LogGrossApproval and SBAGuaranteedApproval were among the most influential factors in predicting loan repayment outcomes. This insight helps lenders focus on critical features when assessing loan applications.

## **Logistic Regression**

Logistic Regression is a statistical method used for binary classification problems, which models the probability of a binary outcome based on one or more predictor variables. It uses a logistic function to map predicted values to probabilities, making it particularly well-suited for scenarios where the outcome variable is dichotomous, such as predicting whether a loan will be fully paid or charged off. The primary objective of using Logistic Regression in our analysis is to create a probabilistic model that can accurately classify the loan repayment status and provide insights into the relationship between predictor variables and the likelihood of loan repayment.

The logistic function, or sigmoid function, is used to model the probability of the dependent variable. The function maps any real-valued number into a value between 0 and 1, which represents the probability of the event occurring. In our context, this event is the loan being fully paid. The coefficients of the logistic regression model represent the change in the log odds of the outcome for a one-unit change in the predictor variable.

For the Logistic Regression model, we selected relevant features like the Decision Tree model. The cleaned dataset was split into the same training and testing sets to ensure consistency in evaluation. After training the Logistic Regression model on the training set, we evaluated its performance using a confusion matrix and the Receiver Operating Characteristic (ROC) curve.

The confusion matrix indicated a minimal number of misclassifications, demonstrating the model's high performance. The model correctly predicted 11,837 loans as fully paid and incorrectly classified only 120 charged-off loans as fully paid. The ROC curve and the Area Under the Curve (AUC) were used to evaluate the model's discriminatory ability. The Logistic Regression model showed a high true positive rate and low false positive rate, with an AUC of 0.99, suggesting excellent discrimination between fully paid and charged-off loans. This high AUC value indicates that the model has a strong ability to distinguish between the two classes. Additionally, the coefficients of the Logistic Regression model were analyzed to understand the impact of each feature on the probability of a loan being fully paid. For instance, a positive coefficient indicates that an increase in the predictor variable is associated with a higher probability of the loan being fully paid, while a negative coefficient suggests the opposite.

The detailed analysis using Decision Tree Classifier and Logistic Regression models provided robust insights into the factors influencing SBA 7(a) loan repayment outcomes. Both models demonstrated strong predictive capabilities, with the Logistic Regression model excelling in classification accuracy. The Decision Tree model provided an interpretable set of rules for predicting loan repayment, while the Logistic Regression model offered a probabilistic framework that highlighted the impact of predictor variables on loan outcomes. These findings can significantly enhance lending strategies, allowing financial institutions to better assess and mitigate the risks associated with SBA 7(a) loans. Future research could explore additional variables and advanced machine learning techniques to further refine these predictive models, ultimately contributing to more effective and sustainable lending practices.

**Apriori Test**

Apriori testing, or the Apriori algorithm, is a classical algorithm used in data mining for discovering frequent item sets and generating association rules. Some of the key concepts for this form of testing is to group interactions together based on the dataset with a frequency above a user defined threshold. This is also known as support. The other concepts are confidence and lift. Confidence is the measure of reliability from the associations derived from the data. The last concept for this testing is what is called lift. Lift is the likelihood of the item set of interaction between the grouping identified.

After reviewing the loan data, a question was conjured for analysis. Is there a way to try and use this method of unsupervised learning to see if there are any groups or associations based on certain variables? In the previous analysis, the target variable was where the account was paid in full or charged off. Using this Loan status as the target, the features were identified. Due to the sizing of the data, the apriori featureset consisted of “BorrState”, “GrossApproval”, “SBAGuaranteedApproval”, “SubPGMDesc”, “TermsInMonths”, “BusinessType”, and “LoanStatus”.

The last step after setting our parameters for the test, we had to make sure that we cleaned out any values that could be “NA” and set up dummy variables for processing. The purpose of this is to remove blank cells and create an ordinal value within any categorical cell. The Apriori test will only work when there are numerical values within the cell. Lastly, to check our data set, we checked for non-boolean values to assure the process with work.

To conclude our testing, lift was the determining factor for this assessment. The likelihood a loan for a small business seemed to be a better assessment than relying solely on confidence. From adjusting the parameters to lift and setting the consequent to Paid In Full, the calculation determined that the top 10 rules with the best likelihood mostly consisted of accounts that were on a individual basis within a certain amount of term and gross approval. Regarding future insight, this evaluation could delve deeper into identifying buckets that would help evaluate certain bins for better analysis.

**Conclusion**

The analysis of the SBA 7(a) Loan dataset has provided insightful and actionable conclusions for optimizing loan allocation strategies to balance profitability and risk mitigation. This study involved extensive data cleaning, exploratory data analysis, and the implementation of advanced machine learning models to identify key predictors of loan repayment reliability. Initially, significant missing values in the dataset necessitated thorough cleaning. Columns with more than 40% missing values were removed, and rows with critical missing identifiers were excluded, ensuring a robust dataset for accurate analysis.

Chi-Squared tests revealed significant associations between BorrState, BankName, BusinessType, and GrossChargeOffCategory. These tests indicated that charge-off rates varied significantly by state, bank, and business type, highlighting the importance of considering geographical and institutional factors in loan risk assessments. Predictive modeling using a Decision Tree Classifier and Logistic Regression provided further insights. The Decision Tree Classifier effectively segmented loans into risk categories, offering a clear set of rules for predicting loan repayment status, with key features such as LogGrossApproval and SBAGuaranteedApproval emerging as critical predictors. The Logistic Regression model, offering a probabilistic approach to predicting loan repayment, achieved high classification accuracy with an AUC of 0.99, and its coefficients provided insights into the impact of various features on loan outcomes.

Additionally, the Apriori algorithm was used to discover frequent item sets and generate association rules based on loan status and other features. While this unsupervised learning method highlighted potential groupings and interactions within the data, further refinement and additional testing are required for conclusive results. The findings from this project can significantly enhance lending strategies for financial institutions. By understanding state-specific and bank-specific charge-off patterns, institutions can tailor their lending strategies to mitigate regional risks. The decision tree and logistic regression models provide a framework for evaluating loan applications, focusing on critical predictors to minimize defaults, thereby improving risk assessment and leading to more sustainable and profitable lending portfolios.

Further research could explore additional variables and more advanced machine learning techniques to refine predictive models. Expanding the dataset to include more recent data and additional financial metrics could provide deeper insights into loan performance trends. In conclusion, this project demonstrates the power of data science in transforming raw financial data into actionable insights, ultimately driving more effective and sustainable lending practices.