|  |  |
| --- | --- |
| Loan fulfillment analytics  Loan Default Dataset | Youssef maher elsebaey  YASSIN WAIL MASRY  ABDELRAHMAN AHMED KHEIR  YOUSSEF MOHAMED RADY  HASSAN MOHAMED SHERIF  HANNIA TAREK  ALI HASSAN ALI |

Table of contents:

1. Project Description
2. Problem Definition
3. Objectives
4. Dataset
5. Visualization
6. Preprocessing
7. Applied Hypothesis and Testing
8. Machine Learning Model
9. Discussion/Quantification for Relevant Project Findings
10. Project Description:

Our project aims to analyze loan data gathered to predict whether a loanee will be able to pay off the loan, or become a default (declared to not being able to pay). This will be done by analyzing data such as credit grade, delinquencies (late payments, even if one day), interest rate, loan amount, and salary. Visualizations facilitate understanding of the dataset. After cleaning our data and selecting the features to predict, the model will be designed to evaluate individual’s creditworthiness and forecast whether they are likely to default on their loans.

1. Problem Definition:

Banks and financial institutions are confronted with a difficult challenge posed by loan defaults, resulting in huge financial losses and adverse effects on a nation's economic state. Yearly, these institutions encounter significant losses due to loanees failing to repay their loans, slowing economic growth and stability. This project aims to address this issue by using data analysis.

1. Objective:

This project seeks to develop a predictive model capable of proactively identifying potential loan defaulters before extending credit. By leveraging insights gathered from data analysis, banks and financial institutions can actively lower risks and minimize financial losses. This approach will enable careful lending decisions, ensuring that credit is primarily extended to individuals and businesses most likely to fulfill their repayment obligations. Moreover, the project aspires to have broader economic implications by reducing instances of loan defaults.

1. Dataset:

The Loan Default prediction dataset consists of 35 columns, including the Loan Status (which is the label, default or not). It is split into train and test datasets. The training dataset of 67,463 rows, and the testing dataset of 28,913 rows. The columns are made of the following:

1. ID: unique ID of representatives (people who applied for or received loans).
2. Loan Amount: Loan amount applied.
3. Funded Amount: Loan amount funded.
4. Funded Amount Investor: loan amount approved by the investors.
5. Term: Term of loan (in months)
6. Batch Enrolled: Batch numbers to representatives.(A system used by lenders where loans are categorized)
7. Interest Rate: Interest rate (%) on loan
8. Grade: Grade by the bank
9. Sub Grade: Sub-grade by the bank.
10. Employment duration: Length of time a loanee has been employed at their current job.
11. Home ownership: Indicates whether the loanee owns or rents their primary residence.
12. Verification status: Status of the borrower's information and documents as confirmed or verified.
13. Payment plan: The agreed structure for repayment of the loan amount (monthly installments).
14. Loan title: The purpose or type of loan as described by the borrower (home improvement, debt consolidation).
15. Debit to Income: Ratio of representative's total monthly debt repayment divided by self-reported monthly income excluding mortgage.
16. Delinquency - two years: Number of 30+ days delinquency in past 2 - years
17. Inquiries- six months: total number of inquiries in last 6 months.
18. Open Account: number of open credit line in representative's - credit line
19. Public Record: Number of derogatory public records
20. Revolving Balance: Total credit revolving balance
21. Revolving Utilities: Amount of credit a representative is using - relative to revolving balance.
22. Total Accounts: Total number of credit lines available in - representatives credit line
23. Initial List Status: Unique listing status of the loan - - W(Waiting), F(Forwarded)
24. Total Received Interest: Total interest received till date.
25. Total Received Late Fee: Total late fee received till date.
26. Recoveries: Post charge off gross recovery(Debt collectors recover as much as possible after a debt is charged of as loss).
27. Collection Recovery Fee: Post charge off collection fee(A fee charged by a debt collector agencies or a 3rd party collectors for their services in recovering delinquent debts).
28. Collection 12 months Medical: Total collections in last 12 months - excluding medical collections.
29. Application Type: Indicates when the representative is an individual or joint.
30. Last week Pay: Indicates how long (in weeks) a representative has paid EMI after batch enrolled.
31. Accounts Delinquent: Number of accounts on which the representative is delinquent.(Behind in payment).
32. Total Collection Amount: Total collection amount ever owed.
33. Total Current Balance: Total current balance from all accounts
34. Total Revolving Credit Limit: Total revolving credit limit (Max. amount one can borrow on a revolving credit account).
35. Loan Status: 1 = Defaulter, 0 = non-defaulters.
36. Visualizations

A pie chart with numbers

Description automatically generated

Figure 1 - Loan Term

A pie chart with a number and a triangle

Description automatically generated

Figure 2 - Delinquency Every Two Years

A drawing of a pie chart

Description automatically generated

Figure 3 - Inquiries Every Six Months

A circle with a point in the center

Description automatically generated

Figure 4 - Collections 12 Months Medical (dropped due to invariance)

A circle with a line in the middle

Description automatically generated

Figure 5 - Application Type (dropped due to huge invariance)

A pie chart with a blue triangle

Description automatically generated

Figure 6 - Loan Status

A colorful pie chart with different colored sections

Description automatically generated

Figure 7 - Loan Grades

A circular chart with different colored lines

Description automatically generated with medium confidence

Figure 8 - Loan Subgrades (dropped due to ineffectiveness)

A blue circle with a white line

Description automatically generated

Figure 9 - Initial List Status

A pie chart with a number of different options

Description automatically generated with medium confidence

Figure 10 - Verification Status

A pie chart with text on it

Description automatically generated

Figure 11 - Home Ownership

A pie chart with text on it

Description automatically generated

Figure 2 - Loan Title

A diagram of a number of items

Description automatically generated

A diagram of a diagram

Description automatically generated

A diagram of a graph

Description automatically generated

A graph of a bar graph

Description automatically generated with medium confidence

A graph with a line

Description automatically generated with medium confidence

A graph of a number of individuals

Description automatically generated with medium confidence

A diagram of a graph

Description automatically generated

A diagram of a recovery

Description automatically generated

A diagram of a graph

Description automatically generated

A diagram of a graph

Description automatically generated

A graph showing a number of accounts

Description automatically generated

A graph showing the salary and salary

Description automatically generated with medium confidence

A graph of a graph of a person

Description automatically generated with medium confidence

A graph of a graph of a credit card

Description automatically generated with medium confidence

A graph of a value by loan status

Description automatically generated

A graph of different colored rectangular shapes

Description automatically generated

A blue rectangular bars with white text

Description automatically generated

A graph of a loan histogram

Description automatically generated

A graph of a number of accounts

Description automatically generated

A diagram of a distribution of current balance

Description automatically generated

A graph of a graph with red dots

Description automatically generated

A blue square with red line

Description automatically generated

1. Preprocessing:

*Data cleaning and preparation:*

1. Checking for missing values:

* Printed the count for missing values for each column and found none in all.

2.Dropping unnecessary columns:

* **“Accounts. Delinquent”:** Redundant, as other variables in the dataset capture similar information more effectively.
* **“Application. Type” :** 99.9% have individual accounts and the rest have joint accounts.
* **“Payment. Plan”:** All values ‘n’.
* **“ID”:** Doesn’t provide any meaningful information related to the likelihood of a loan default.
* **“Batch.Enrollement”:** Batch numbers assigned to representatives are unlikely to have a direct impact on whether a person will default a loan.
* **“Collection.12.months.Medical”:** The analysis is not specifically targeting medical collections.

3.Data Transformation:

* Replaced specific values in the **"Verification. Status"** column.
* Modified values in the **"Loan. Title"** column based on patterns (e.g., merging similar categories into one).
* Converted character labels to numeric values using label encoding for machine learning compatibility.

4.Feature Engineering:

* Created new features or modified existing ones based on domain knowledge or data exploration, like modifying **"Loan. Title"** values or creating quartile-based ranges for **"Total.Current.Balance".**

5.Statistical Analysis:

* Conducted **ANOVA** and **Kruskal-Wallis** tests to analyze the relationship between **"Interest. Rate"** and **"Grade"** or **"Loan. Status".**
* Performed **t-tests** and **Wilcoxon rank-sum tests** to compare interest rates between **defaulted** and **non-defaulted loans**.

1. Applied Hypothesis and Testing:

A good number of hypothesis testing was done between multiple columns using ANOVA to compare means across multiple groups as a way of feature selection.

* Interest Rate vs Grade with a p value of: 0.000131. Helps us test the Interest Rate across different "Grade" categories to see if the loan grade affects the interest rate and consequently the likelihood of default.
* Interest Rate vs Ownership yielded a p value of **0.000388**. The significant difference in mean interest rates across different types of home ownership suggests that home ownership status is associated with variations in interest rates.
* Revolving Utilities vs Grade with a p value of **2e-16**
* Total Accounts vs Grade and its p value is **8.44e-10**. Its Tukey test: diff lwr upr p adj
* B-A -0.580025340 -0.8662900 -0.29376072 0.0000000
* C-A -0.615072391 -0.9003283 -0.32981648 0.0000000
* D-A -0.585675072 -0.9359067 -0.23544339 0.0000170
* Total Accounts vs Ownership p value: **0.00436**
* Funded Amount Investor vs Ownership p value: **2.03e-11**

More tests conducted also yielded similar p values.

1. Machine Learning Model:

After testing with multiple models, we ended up with Naïve Bayes as it held the best metrics. We tested with logistic regression, SVM, decision tree, and k-means. Naïve Bayes’s metrics are:

* **Accuracy:** 0.8671714
* **Precision:** 0.1013917
* **Recall:** 0.05368421
* **F1 Score:** 0.07019959

1. Discussion/Quantification for Relevant Project Findings:

The data was easy to work with as it contained no nulls and no duplicate values. It consists of over 67000 rows and 35 columns. There was a lot of issues within the data itself though. For example, a column indicating the type of ownership of one’s house was named Employment Duration and another column for one’s salary was called Home Ownership. Another categorical column, Loan Titles, had 10 different variations for a single value. For example, credit card refinancing had names like credit card, credit card refinancing, Credit Card Refinancing (capital letters), credit card card loan, etc. And many more for other values. The values of this column and another column, verification type, were cleaned internally and grouped together under one single value.

Visualizations were done to show the distribution of values, the minimum, the maximum, the outliers, and the relationship between columns. Hypothesis testing using ANOVA and Tukey’s test were also applied as a way of feature selection. Multiple hypotheses were conducted like seeing if the grade of a loan affects the interest rate and if the type of ownership also affects the interest rate. Then, after testing with multiple models, we concluded that Naïve Bayes yielded better results than the rest of the models with an accuracy of 0.867, precision of 0.10, recall of 0.05, and an F1 score of 0.07.