**Credit Defaulter Analysis and Prediction**

Ajith Kumar Ethirajulu, Deepak Raj Mohan Raj, Madhu Babu Sikha, Zeming Zhang

**Abstract**

This project uses exploratory data analysis to identify patterns that can be used to detect potential credit defaulters. The data is used to identify the driving factors behind loan default and to assess the risk associated with lending to customers. The goal is to ensure that the applicants capable of repaying the loan are not rejected, while also minimizing the risk of loss for the company.

1. **Introduction**

Credit defaulters fail to make timely payments on a loan, negatively impacting their credit score. Banks mitigate this risk by requiring larger down payments, higher interest rates, stricter repayment terms and additional collateral. They may also monitor defaulters more closely and report them to credit bureaus. Lastly, banks may limit the credit they offer defaulters or only offer secured loans.

This project study aims to identify patterns which indicate a client's difficulty in paying their installments, using data from loan applications, previous loan applications, and their respective descriptions. The objective is to help the company take actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. to ensure that customers who can repay the loan are not rejected.

1. **Model**

This flow chart outlines the process of building a machine learning model. It starts with loading the data into an SQLite database, followed by removing null values and feature selection. Then the database is normalized, and feature engineering is used to create new features. Exploratory Data Analysis is then carried out to gain insights about the data. Finally, a machine learning model is built and evaluated to determine its accuracy.

Diagram

Description automatically generated

1. **Dataset**

This repository contains 3 records that explain the data of a specific client. The 'application\_data.csv' file consists of the data of the customer at the time of application, which includes their payment status. The 'previous\_application.csv' file contains the records of the customer's preceding loan and whether the loan had been accepted, cancelled, denied or an unused offer. Lastly, the 'columns\_description.csv' file is a lexicon that explains the interpretation of the variables.

|  |  |  |
| --- | --- | --- |
| File Name | Number of features | Number of columns |
| application\_data | 122 | 3,07,511 |
| Previous\_application | 37 | 16,70,214 |

1. **Exploratory Data Analysis**

With consideration to domain knowledge and the proportion of missing values in the dataset, we selected several columns, containing both numerical and categorical features. Analyzing the numerical column data using a bar graph [1], we detected outliers in the numerical columns and removed them. We also segregated the categorical columns into separate tables by listing the unique values in the respective columns and referencing their ID from the newly created tables to the normalized table as a numerical column. This preserved the integrity of the database, forming a normalized database.

Utilizing the target value which holds either 1 for defaulters or 0 for non-defaulters, we observed an imbalance in the data skewed towards non-defaulters when plotting a bar chart [2]. We employed bar [3] charts and pie charts to understand the correlation of all the categorical columns in our list such as NAME\_CONTRACT\_TYPE, CODE\_GENDER, FLAG\_OWN\_CAR, etc. Results indicate that a greater proportion of defaulters (Target = 1) have realty ownership, are female, and are married. The same methodology was applied to numerical features [4], and it was observed that a greater number of defaulters have more children. A word cloud [5] was employed to evaluate the category distribution within the categorical column, yielding results like those of the above plots.

A Sankey graph [6] was generated which compared numerical columns such as age and employment days, alongside select categorical columns. The word cloud [7] analysis revealed that the working category was the largest group of applicants. Additionally, it showed that the working category and commercial associates were the highest among the defaulters. A Pareto chart [8] was developed which showed that applicants younger than 50 years of age are disproportionately represented among defaulters, and that applicants aged between 30 and 39 are particularly vulnerable. A chart [9] was constructed which showed that the proportion of commercial associates and working category applicants among defaulters is comparable across age groups. A Violin [10] chart was produced which revealed that defaulters are concentrated at the beginning of the business hours, and that those who recently changed their phone number or ID document were more likely to default.

1. **Classification Model and Results**

The classification machine learning models that we built, such as Support Vector Model, K Nearest Neighbor, and Random Forest, did not yield good results in terms of accuracy, recall, and F1 score. This is due to the overfitting of the model caused by the high number of features.

To improve these models, we can investigate reducing the number of features, introducing regularization, or using other techniques such as feature selection or dimensionality reduction. We can also use more advanced models such as neural networks or ensembles, which can help to reduce the overfitting. Additionally, we can tune the hyperparameters of the models to better optimize their performance.

1. **Conclusions** In conclusion, using EDA techniques, we were able to identify some of the features that contributed more towards identifying the defaulters. To name a few, DAYS\_LAST\_PHONE\_CHANGE, WEEKDAY\_APPR\_PROCESS\_START, DAYS\_ID\_PUBLISH, HOUR\_APPR\_PROCESS\_START, some categories in NAME\_INCOME\_TYPE such as “Working” and “Commercial Associate”. The bank needs to focus more on these features before lending out a credit or loan to the customer.
2. **Future Work** The feature selection needs to be done more rigorously with advanced techniques since there are a lot of features, and the conventional models are clearly overfitting on the training data. Due to the huge imbalance in the target class, we need to employ techniques that overcome this challenge.
3. **Reference**1. https://www.kaggle.com/datasets/venkatasubramanian/credit-eda-case-study
4. **Diagrams**

|  |
| --- |
| **Database Entity Diagram after creating normalized tables** |



**Exploratory Diagrams:**

|  |
| --- |
| **[1] Box plots to detect outliers in each numerical feature** |

Graphical user interface, diagram, application

Description automatically generated

A picture containing text, clock, screenshot

Description automatically generated

Graphical user interface, application

Description automatically generated

Chart, box and whisker chart

Description automatically generated

A picture containing text, clock, screenshot

Description automatically generated

**[2] Target class distribution to show imbalance in data**

Chart, bar chart

Description automatically generated

**[3] Distribution of categories and target class in categorical features**

Chart, waterfall chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, waterfall chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, waterfall chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, waterfall chart

Description automatically generatedChart, bar chart

Description automatically generated

**[4] Distribution of data in numerical features**

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart, histogram

Description automatically generated

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generated

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generated

**[5] Word cloud to depict dominant categories in each categorical feature**

A picture containing text

Description automatically generated

|  |
| --- |
| **[6] Sankey Diagram**  A picture containing text  Description automatically generated |

**[7] Charts showing valuable insights**

**Working category are the maximum number of applicants**

A picture containing logo

Description automatically generated

**Among the defaulters working category and commercial associates turn out to be high in number**

Graphical user interface

Description automatically generated with medium confidence

|  |
| --- |
| **[8] Pareto chart showing applicants less than 50 years old are high in number** |

Chart, line chart

Description automatically generated

**Pareto chart showing applicants between 30 and 39 years old are high in defaulters**

Chart, line chart

Description automatically generated

**[9] Chart showing commercial associates and working category applicants are equally distributed across ages among defaulters**

Chart, bar chart

Description automatically generated

|  |
| --- |
| **[10] Violin chart displaying the defaulters are accumulated more at the start of the business hours** |

Chart, funnel chart

Description automatically generated

**Violin plot showing defaulters who changed their phone number recently are prone to default.**

**(It also shows that people who changed their ID document recently are also prone to default)**

Chart

Description automatically generated