PROJECT REPORT

Academy of Skill Development

ASD
Skills For Employment

LOAN APPROVAL DATA ANALYSIS

Prepared by
Debolina Chakraborty

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my mentor Mr. Tridip Kundu, for his invaluable guidance and unwavering support throughout this data analysis project. His expertise in methodology and statistical analysis, insightful feedback on coding approaches, and continual encouragement have been instrumental in shaping both the direction and the quality of this work.

INTRODUCTION:

Loan risk assessment plays a pivotal role in the financial industry by enabling organizations to evaluate the creditworthiness of borrowers and effectively asses risk such as loan defaulting. This process involves analysing various borrower-related factors such as income, credit history, education, employment status and existing debt obligations to predict the likelihood of loan repayment. Effective risk assessment helps lenders make informed lending decisions, protect assets and maintain financial stability. Exploratory Data Analysis (EDA) was conducted to uncover trends and relationships among features. The analysis provides actionable insights for financial institutions to enhance their decision-making processes, reduce default risk and improve fairness in loan approvals. This study tries to determine features which highly influence loan approval.

DATA:

The dataset is collected from Kaggle at:
https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data
The dataset originally contained 45,000 records but for the simplicity of this project, the beginning 4000 records were selected.

METHODOLOGY:

The dataset contains **features** namely - 'person_age', 'person_gender', 'person_education', 'person_income', 'person_emp_exp', 'person_home_ownership', 'loan_amnt', 'loan_intent', 'loan_int_rate', 'loan_percent_income', 'cb_person_cred_hist_length', 'credit_score', 'previous_loan_defaults_on_file' with the **target** named - 'loan_status'

DATA DESCRIPTION:

Attribute Name	Attribute Description	Attribute Type
person_age	Age of the person	Float
person_gender	Gender of the person	Categorical
person_education	Highest education level	Categorical
person_income	Annual income	Float
person_emp_exp	Years of employment experience	Integer
person_home_ownership	Home ownership status	Categorical
loan_amnt	Loan amount requested	Float
loan_intent	Purpose of the loan	Categorical
loan_int_rate	Loan interest rate	Float
loan_percent_income	Loan amount as a percentage of annual income	Float
cb_person_cred_hist_length	Length of credit history in years	Float
credit_score	Credit score of the person	Integer
previous_loan_defaults_on_file	Indicator of previous loan defaults	Categorical
loan_status (target variable)	Loan approval status: 1 = approved; 0 = rejected	Integer

STEP 1: Checking and Handling null values

The data contains no values.

STEP 2: One-Hot Encoding the categorical features

The categorical variables namely – ‘person_gender’, ‘person_education’, ‘person_home_ownership’, ‘loan_intent’, ‘previous_loan_defaults_on_file’ are one-hot encoded using pd.get_dummies() function of pandas.

STEP 3: Handling Imbalance in Data

The original dataset has an imbalance in the target variable where the number of records with value ‘1’ (Loan Approved) is 1342 and the number of records with value ‘0’ (Loan Declined) is 2657.

Data imbalance can cause the model to be skewed towards the majority class (0 in this case).

The dataset should be balanced to reduce biasness, enhance accuracy and make fairer predictions. The data is balanced using Synthetic Minority

Oversampling Technique (SMOTE). The balanced data contains 2657 records for each class (1 and 0).

STEP 4: Handling Outliers

Boxplots are used to determine outliers.

- Checking outliers in ‘person_age’:

  Four outliers with age 144, 144, 123, 123 are observed which is very unusual and is likely caused by the entry of wrong data so these records are deleted.

- Checking outliers in ‘person_income:

  Outliers are present in income which fall into valid income range and is likely caused due to the diverse background of population so these are retained.

- Checking outliers in ‘person_emp_exp':

  Outliers are present in person’s employment experience but they are valid in real-life scenarios. Some people can have a relatively higher number of years as employment experience from the majority of population. Therefore, these outliers are retained.

- Checking outliers in ‘loan_amnt’:

  There are no outliers.

- Checking outliers in ‘loan_int_rate’:

  There is an outlier which is very close to the upper fence. As a result, it can be considered as a mild outlier. Also, a loan interest of 20 percent is realistic. Therefore, this outlier is retained

- Checking outliers in ‘credit_score’:

  There are outliers in credit score but they are valid credit scores which can be classified as poor. Therefore, the outliers are retained.

- Checking outliers in 'cb_person_cred_hist_length':

  There are no outliers in person’s credit history length in years.

STEP 5: Visualizing the distribution of Person Education, Person Home Ownership, Loan Intent and Gender through Pie Charts

STEP 6: Visualizing various combination of attributes to analyse which of them are the most important to be considered for risk assessment

ANALYSIS:

- Males have a higher loan rejection rate than males according to this dataset but gender is not a strong determinant for risk assessment. [Fig 5]
- The Associate educational degree has the highest loan rejection rate followed by Doctorate, Bachelor, High School and Master degrees. [Fig 6]
- Borrowers having their own house have the highest loan rejection rate followed by the categories of houses on mortgage, others and rent. [Fig 8]
- Educational loans have the highest loan rejection rate followed by venture loans, personal loans, medical loans, home improvement loans and debt consolidation loans. [Fig 7]
- Loan Default rate is highest for the Associate educational degree followed by the High School, Bachelor, Master and Doctorate degrees. [Fig 9]
- Loan Default rate is highest for borrowers having house on mortgage, followed by borrowers having own house, house on rent and others. [Fig 10]
- Loan Default rate is highest for educational loans followed by personal loans, venture loans, medical loans, debt consolidation loans and house improvement loans. [Fig 11]

- Loans are approved only for borrowers with no previous loan defaults on file. [Fig 16]

- Borrowers with higher credit score have lower loan defaulting and hence are more likely to get loan approvals. [Fig 12 and Fig 13]

- Borrowers with higher income are less likely to default and hence loan approval is higher. [Fig 14 and Fig 15]

- The credit scores range from 450 to 730 with majority of them in the range 550 – 700. [Fig 18]

- Majority of loans approved have person's income in the range 13000 – 120000 and credit score in the range 550 – 650. [Fig 18]

- Count of loan approval decreases with increase in credit history length from 2 years to 4 years. [Fig 21]

- Features with higher positive correlation: [Fig 22]
    1. 'loan_percent_income' and 'loan_amount'
    2. 'person_emp_exp' and 'person_age'
    3. 'loan_amount' and 'person_income'

  Features with higher negative correlation: [Fig 22]
    1. 'loan_status' and 'previous_loan_defaults_on_file_Yes'
    2. 'person_home_ownership_RENT' and 'person_home_ownership_MORTGAGE'

RESULTS:

- Borrowers with Associate degree as their highest educational qualification are frequently rejected as they have a higher risk of loan defaulting.
- Borrowers having house on mortgage are risky as they have a higher rate of loan defaulting.
- Educational loans are rejected most frequently rejected as they have a higher risk of loan defaulting.
- Borrowers with no history of loan defaulting are preferred.
- Borrowers with higher credit score are less likely to default.
- Higher salaries may guarantee safe borrowers. Several other factors like stable source of income, good DTI, good credit score and number of loan defaulting records must also be considered.

CONCLUSION:

The data analysis of loan approvals reveals key factors that help identify risk applicants. Our findings show that education, type of home ownership, loan intent, credit history and previous loan default records are significant predictors of risky loan applicants. Applicants having Associate degree, having house on mortgage with lower credit scores and previous loan default records showcase higher risk.

APPENDICES:

➢ USEFUL CHARTS:

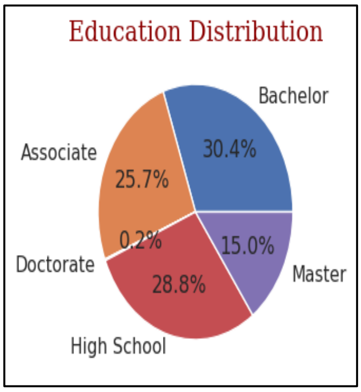


Fig 1

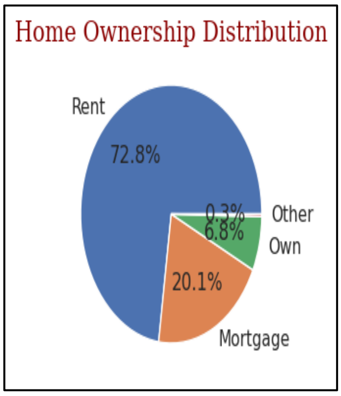


Fig 2

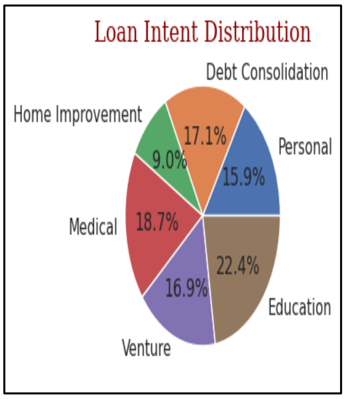


Fig 3

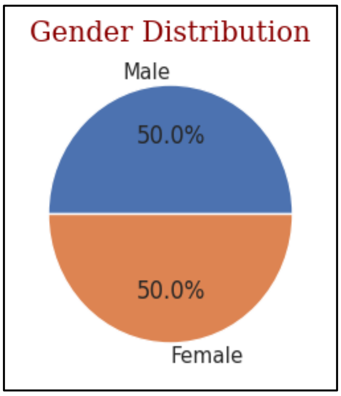


Fig 4

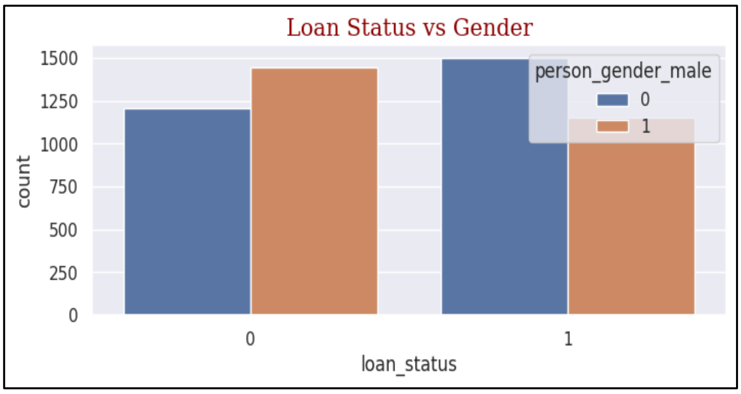


Fig 5

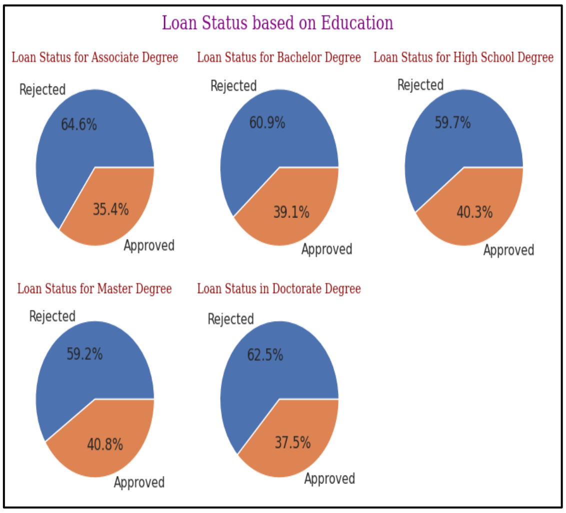


Fig 6

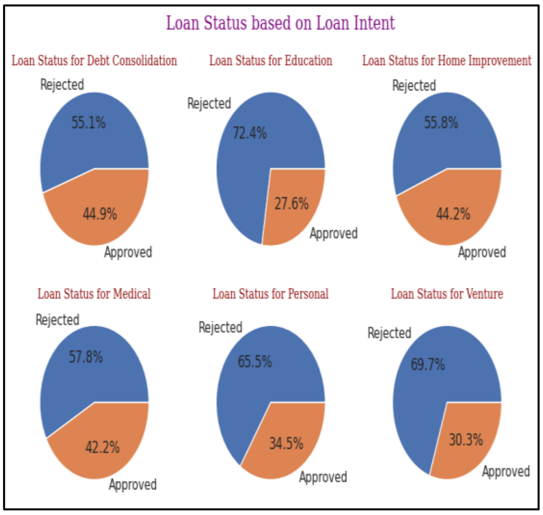


Fig 7

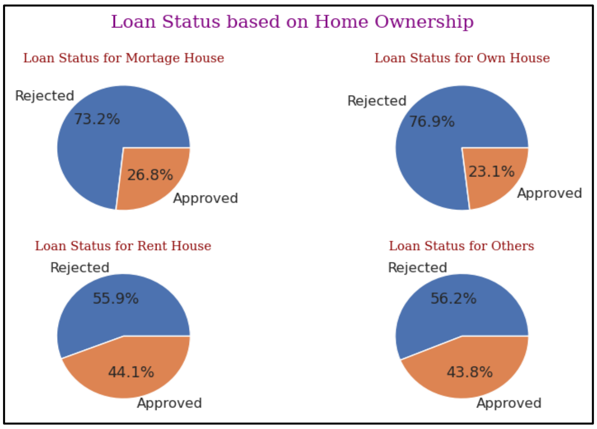


Fig 8

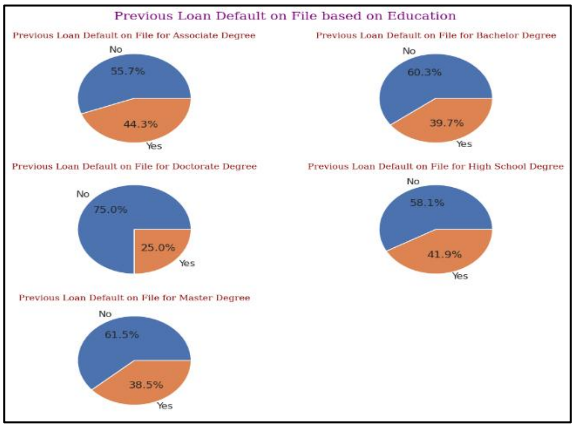


Fig 9

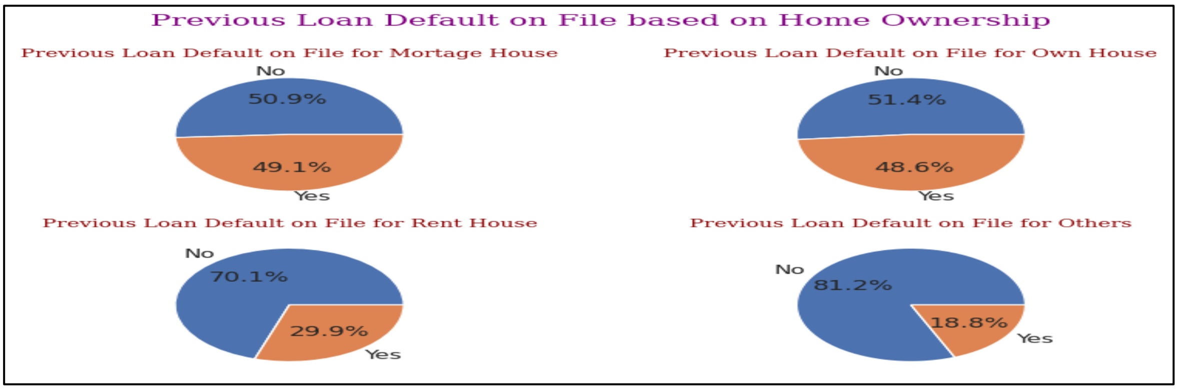


Fig 10

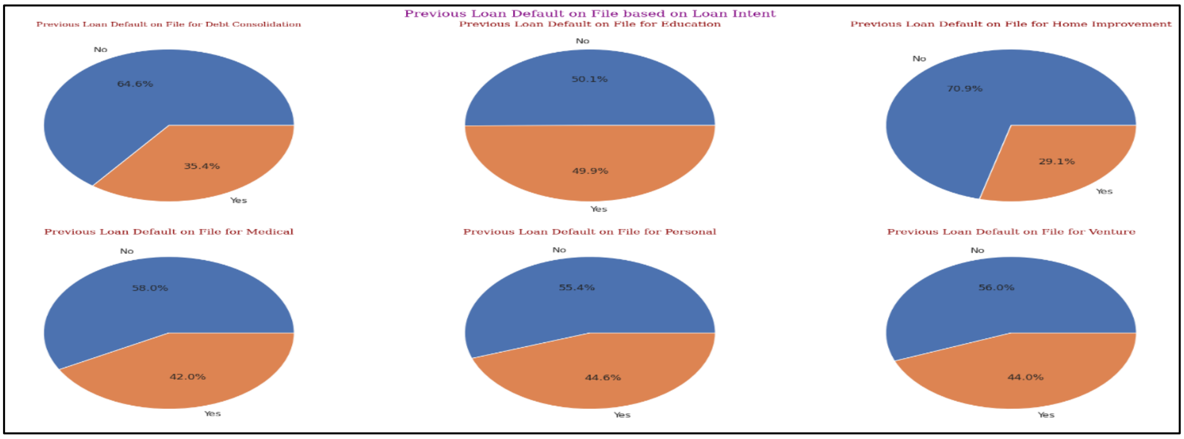
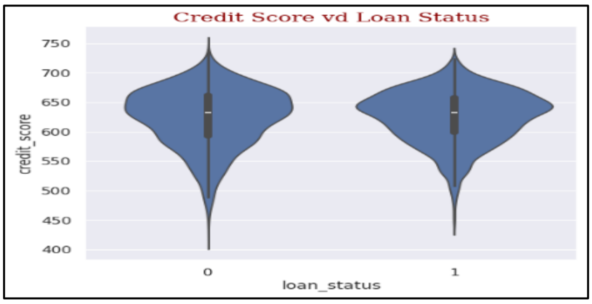


Fig 11



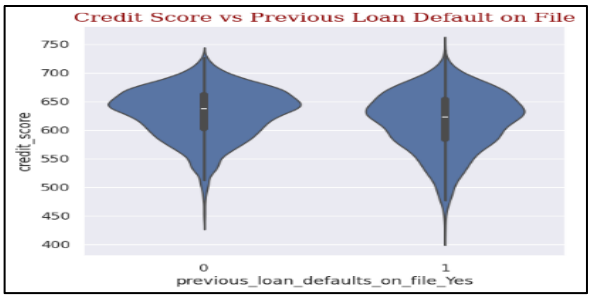


Fig 12

Fig 13

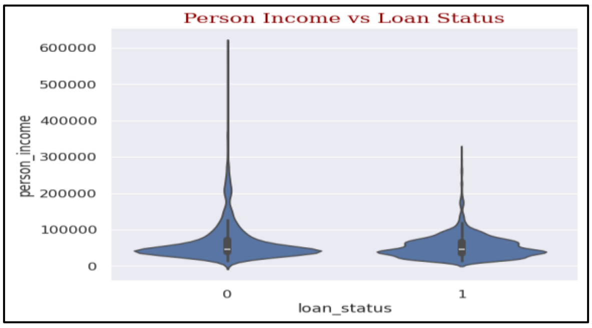


Fig 14

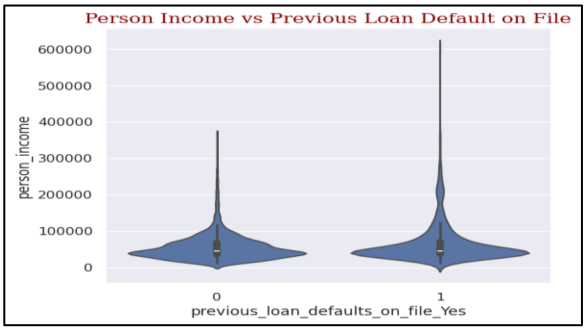


Fig 15

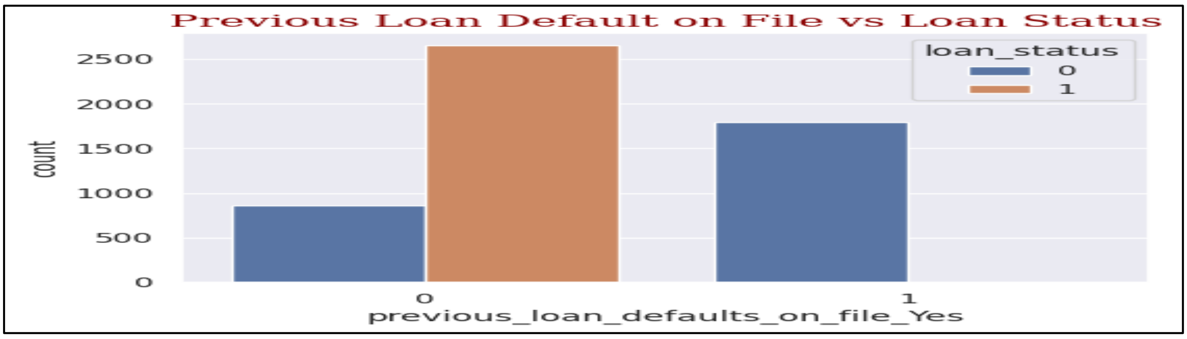


Fig 16

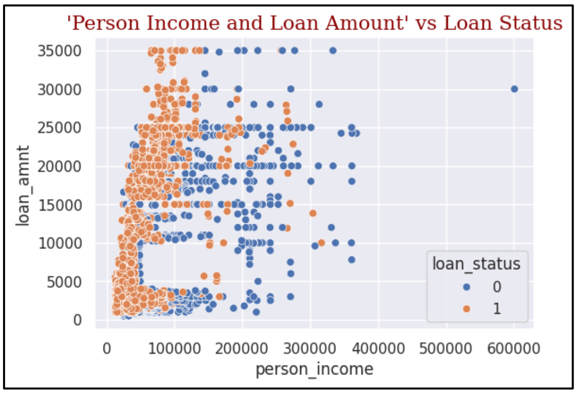


Fig 17

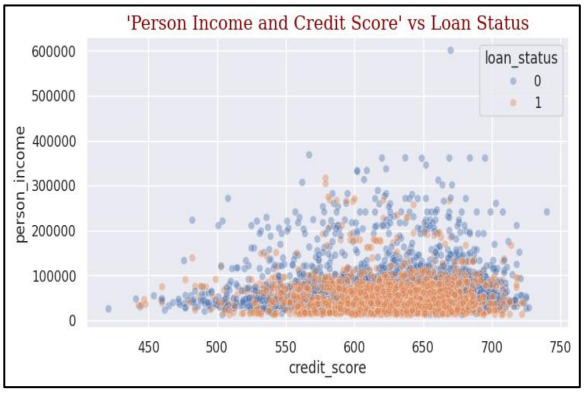


Fig 18

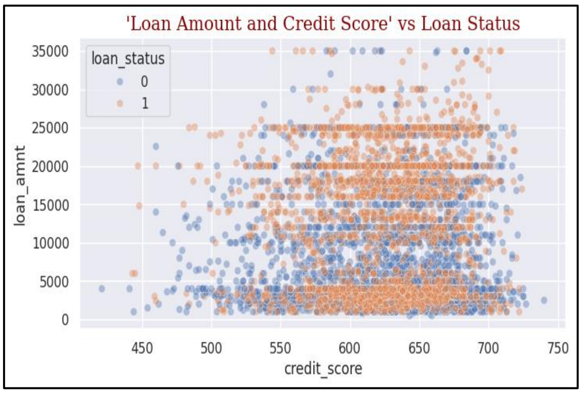


Fig 19

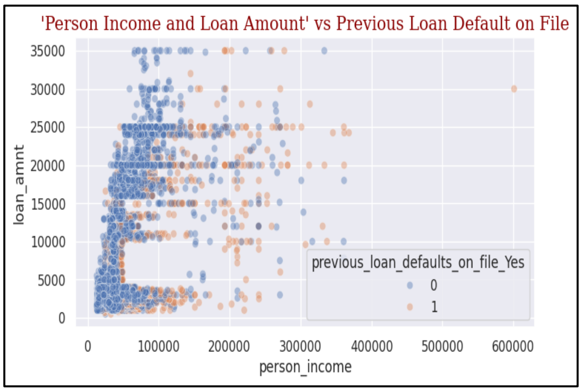


Fig 20

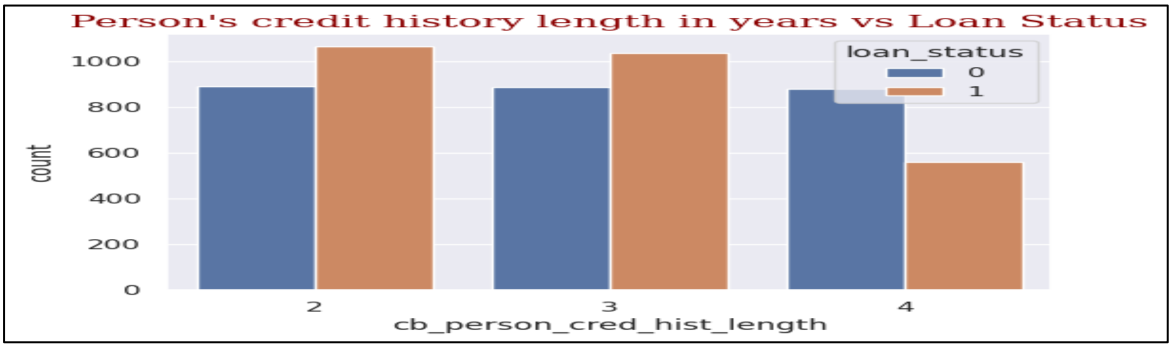


Fig 21

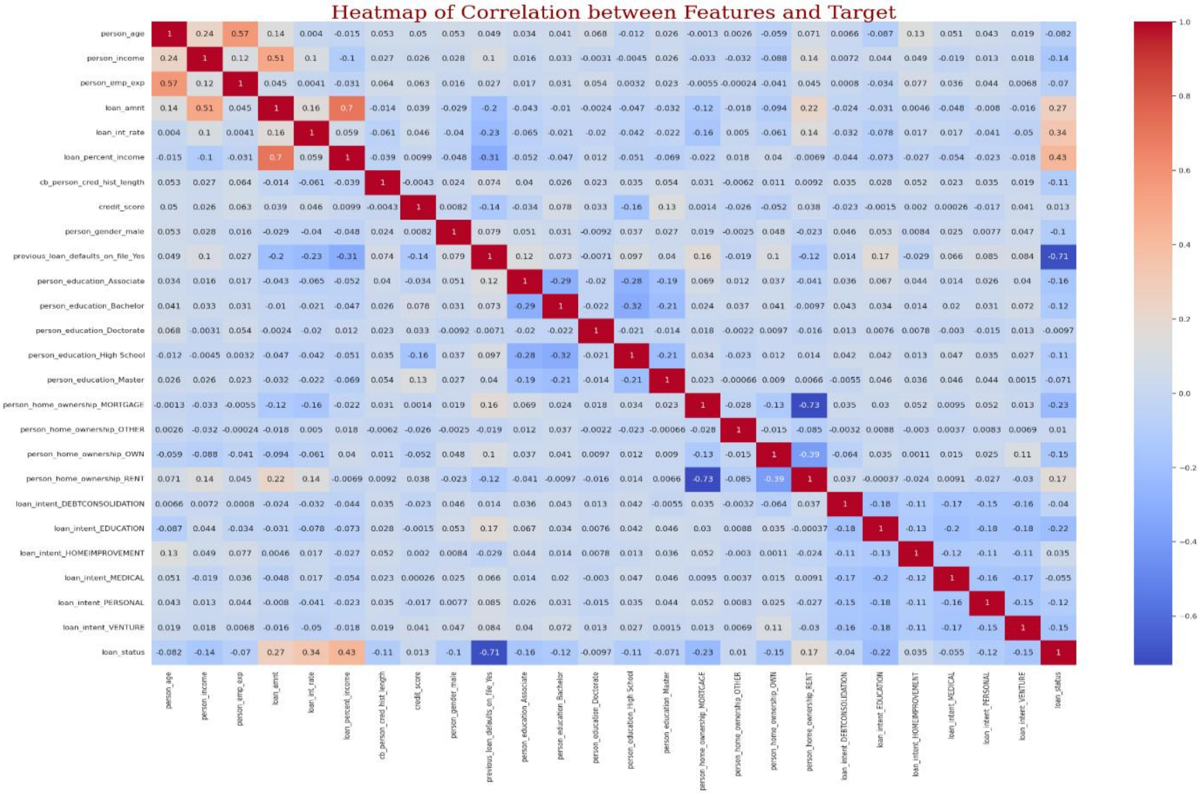


Fig 22

- **CODE:**

  Github: https://github.com/AI-fanatic24/ASD-DS-AI-ML-Internship/blob/main/Loan.ipynb

REFERENCES:

- Lo, Ta-wei. "Loan Approval Classification Dataset." *Kaggle.com*, 2024, www.kaggle.com/datasets/taweilo/loan-approval-classification-data.

- "GeeksforGeeks." *GeeksforGeeks*, 11 June 2018, www.geeksforgeeks.org/dbms/kdd-process-in-data-mining/.