A PROJECT ON

“CREDIT RISK MODELLING”

SUBMITTED IN
PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE COURSE OF
DIPLOMA IN BIG DATA ANALYSIS

SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

Submitted By:

Mohit Ravindra Jain (92753)

Varun Sai Marisetty (92966)

Mr.Nitin Kudale
Centre Coordinator

Mrs.Manisha Hingne
Course Coordinator

CERTIFICATE

This is to certify that the project work under the title ‘Credit Risk Modelling’ is done by Mohit Ravindra Jain & Varun Sai Marisetty in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mr. Aniket P
Project Guide

Mrs. Manisha Hingne
Course Coordinator

Date: 04/02/2026

ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Manisha Hingne (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Aniket P.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Mohit Ravindra Jain
DBDA Aug 2025 Batch,
SIIT Pune

Varun Sai Marisetty
DBDA Aug 2025 Batch,
SIIT Pune

TABLE OF CONTENTS

1. Introduction

1.1 Introduction And Objectives:

Credit risk is the risk of financial loss resulting from a borrower's failure to repay a loan or meet contractual obligations. Financial institutions have traditionally employed credit analysts to assess this risk. However, with the explosion of data and the need for instant approvals, automated Credit Risk Modelling has become the industry standard. This project aims to democratize access to such tools by providing an easy-to-use web interface for risk assessment.

Objectives:

- To develop a Machine Learning model that accurately predicts the Probability of Default (PD).
- To generate a standardized Credit Score (300-900) and Risk Rating (Poor, Average, Good, Excellent).
- To create an interactive Web Dashboard for users to input their financial details.
- To integrate an AI Chatbot that serves as a virtual financial advisor, explaining the score and offering improvement strategies.

1.2 Why this problem needs To be Solved?

The traditional credit lending process is plagued by inefficiencies that hurt both lenders and borrowers. Solving this is critical for three main reasons:

1. **Mitigating Financial Loss (NPA Reduction)**: The primary survival metric for any bank is keeping Non-Performing Assets (NPAs) low. Bad loans wipe out profits. An accurate prediction system directly stops bad money from going out the door.
2. **Financial Inclusion & Fairness**: Manual assessment often biases against young people or those without classic collateral. A data-driven ML model can find "good" borrowers who might look risky on paper but have strong behavioral indicators, actively expanding credit access.
3. **Operational Efficiency**: In the digital age, customers expect instant approvals. Manual underwriting takes days. Automating this solves the "speed vs. accuracy" trade-off, allowing banks to scale without hiring armies of analysts.

1.3 Dataset Information.

The project relies on a robust dataset comprising 50,000 unique records, consolidated from three primary sources joined on cust_id. This multi-source approach ensures a 360-degree view of the customer's financial health.

1. **Data Sources (Schema)**
   - **Source 1: Customer Demographics (customers.csv)**
     - **Content**: Personal details used to assess stability.
     - **Key Features**:
       - Age: Applicant's age (Used to analyze risk variance by age group).
       - Income: Annual income (Primary indicator of repayment capacity).
       - Employment Status: Salaried vs. Self-Employed.
       - Residence Type: Owned vs. Rented.
   - **Source 2: Loan Details (loans.csv)**
     - **Content**: Specifics of the loan being applied for.
     - **Key Features**:
       - Loan Amount: Principal amount requested.
       - Loan Tenure: Duration of the loan.
       - Disbursal Date: Used for time-series analysis.
     - **Target Variable**: default (Boolean). This is the ground truth label (0 = Non-Default, 1 = Default) that the model trains on.
   - **Source 3: Bureau Data (bureau_data.csv)**
     - **Content**: Historical credit behavior and existing debt.
     - **Key Features**:
       - Total DPD (Days Past Due): Sum of days the applicant has been late on previous payments. A critical risk indicator.
       - Credit Utilization Ratio: The percentage of available credit currently in use. High utilization often correlates with high risk.
       - Number of Open Accounts: Proxy for current debt burden.
2. **Data Volumetrics and Quality**
   - **Total Records**: 50,000
   - **Feature Count**: 33 Columns (post-merge).
   - **Class Imbalance**: The dataset reflects real-world banking scenarios with a significant imbalance:
     - **Non-Defaulters (Class 0)**: ~45,642 records (~91.5%)
     - **Defaulters (Class 1)**: ~4,296 records (~8.5%)
   - **Handling**: This imbalance necessitates careful model evaluation using metrics like F1-Score rather than just Accuracy.

2. Problem Definition and Algorithm:

2.1 Problem Definition

The core problem is to classify loan applicants into "Risk" or "No Risk" categories based on their application data.

- **Subjectivity**: Manual assessments are often subjective.
- **Latency**: Traditional processes are slow.
- **Scope**: The project involves building a Predictive Model using historical credit data, wrapping it in a Streamlit web application, and connecting it to a Large Language Model (LLM) for conversational capabilities.

2.2 Algorithm Definition

The core prediction engine is based on Logistic Regression. Despite its name, it is a classification algorithm used to predict the probability of a categorical dependent variable.

Mathematical Model:
The core of Logistic Regression is the Sigmoid Function (or Logistic Function), which maps any real-valued number into a value between 0 and 1.

$$\sigma(z)= \frac{1}{1+e\text{^}-z}$$

Where z is the weighted sum of input features:

$$z=\beta0+\beta1X1+\beta2X2+...+\beta nXn$$

- β0: Intercept.
- βn: Coefficients (weights) learned during training.

Decision Boundary:
The model predicts class 1 (Default) if probability P>0.5, and class 0 (Non-Default) otherwise. This probabilistic output allows us to scale the result into a credit score (300-900).
Advantages: High interpretability, computational efficiency, and probabilistic output suitable for risk scoring.

3. Experimental Evaluation:

3.1 Methodology:

The system follows a typical Machine Learning pipeline:

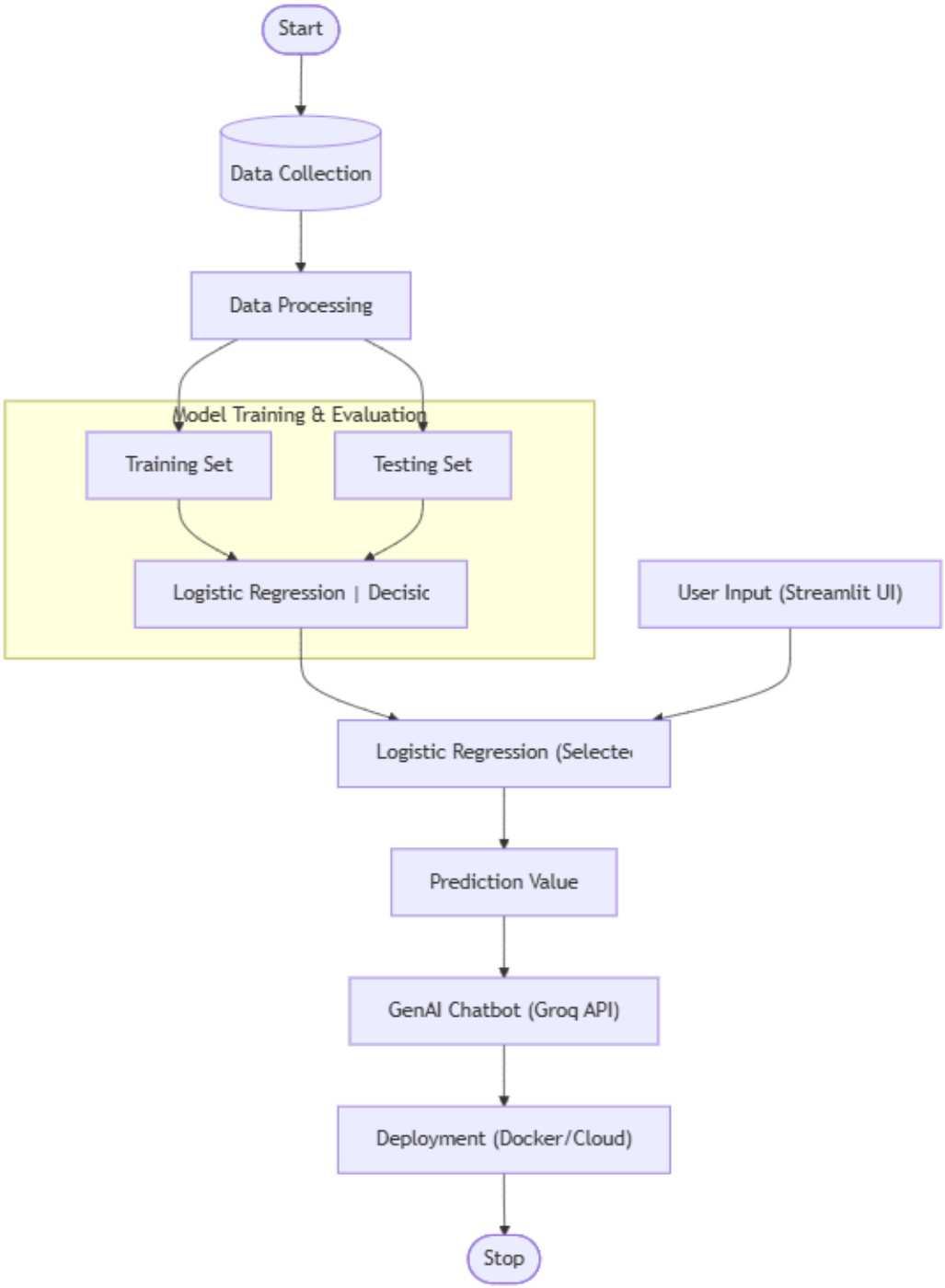
1. **Data Preprocessing**: Cleaning missing values and merging the three source CSVs.
2. **Feature Engineering**: Creating new features like Credit Utilization Ratio and Load to Income Ratio.
3. **Scaling**: Using StandardScaler to normalize features like Income and Loan Amount.
4. **System Architecture**:
    - **Frontend**: Streamlit UI.
    - **Backend**: Python script loading the trained pickle model.
    - **AI Layer**: Groq API for the chatbot.

Data Loading Implementation:

The following Python code demonstrates how the three disjoint datasets are loaded and merged into a single DataFrame for analysis. This step is crucial for establishing the relationship between a customer's personal details and their credit history.

```
import pandas as pd

# 1. Load the individual datasets
df_customers = pd.read_csv("dataset/customers.csv")
df_loans = pd.read_csv("dataset/loans.csv")
df_bureau = pd.read_csv("dataset/bureau_data.csv")

# 2. Merge to create the Master Dataset
# a) Merge Customers with Loans on 'cust_id'
df = pd.merge(df_customers, df_loans, on='cust_id')

# b) Merge result with Bureau Data on 'cust_id'
df = pd.merge(df, df_bureau, on='cust_id')

# 3. Verify Dimensions
print(f"Final Dataset Shape: {df.shape}")
# Output: (50000, 33)
```

Flow Diagram:



3.2 Exploratory Data Analysis

The distribution of the target variable is plot with the help of a pie chart (fig 2). From the figure we can infer that Non-Defaulters constitute the majority (91.5%) followed by Defaulters (8.5%). This class imbalance is a critical factor for model training.

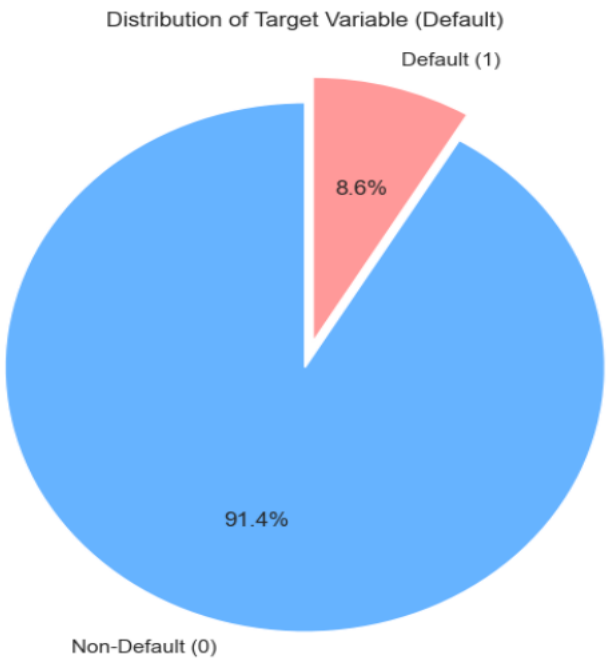


Fig 2: Pie- chart showing proportion of Default vs Non-Default

The average income for each category is visualized using bar plot (fig 3). From the figure we can infer that Non-Defaulters have higher average income compared to Defaulters. The Defaulters have least average income among the two groups.

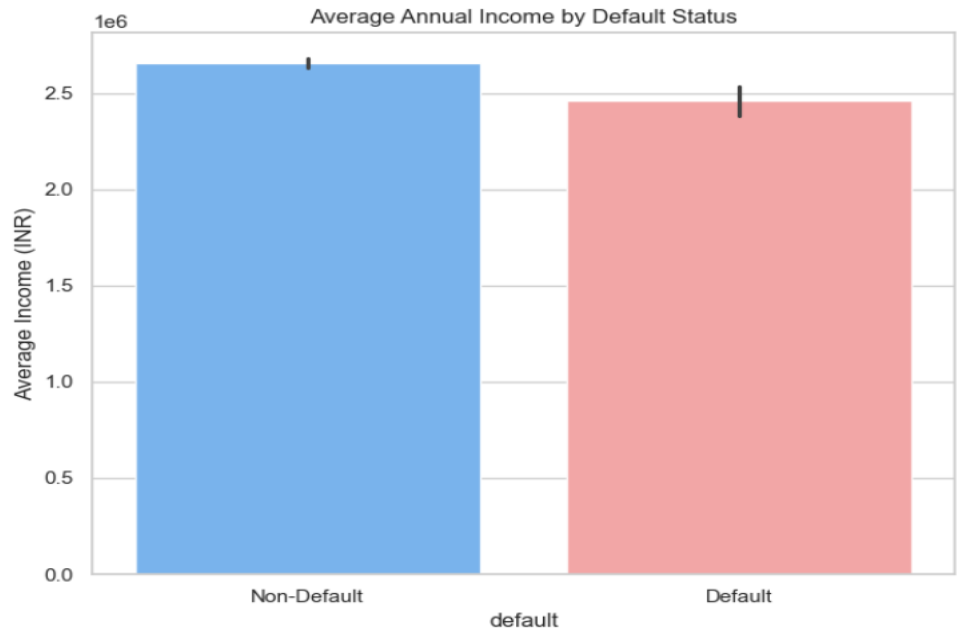


Fig 3: Default Status Vs Average Income

The Loan Purpose wise distribution is plotted for each category (fig 4). The plot shows that Education and Personal loans witnessed the highest default rates while for Secured loans like Home and Auto the default probability is lower. From the figure we can infer that unsecured loans generally carry higher risk.

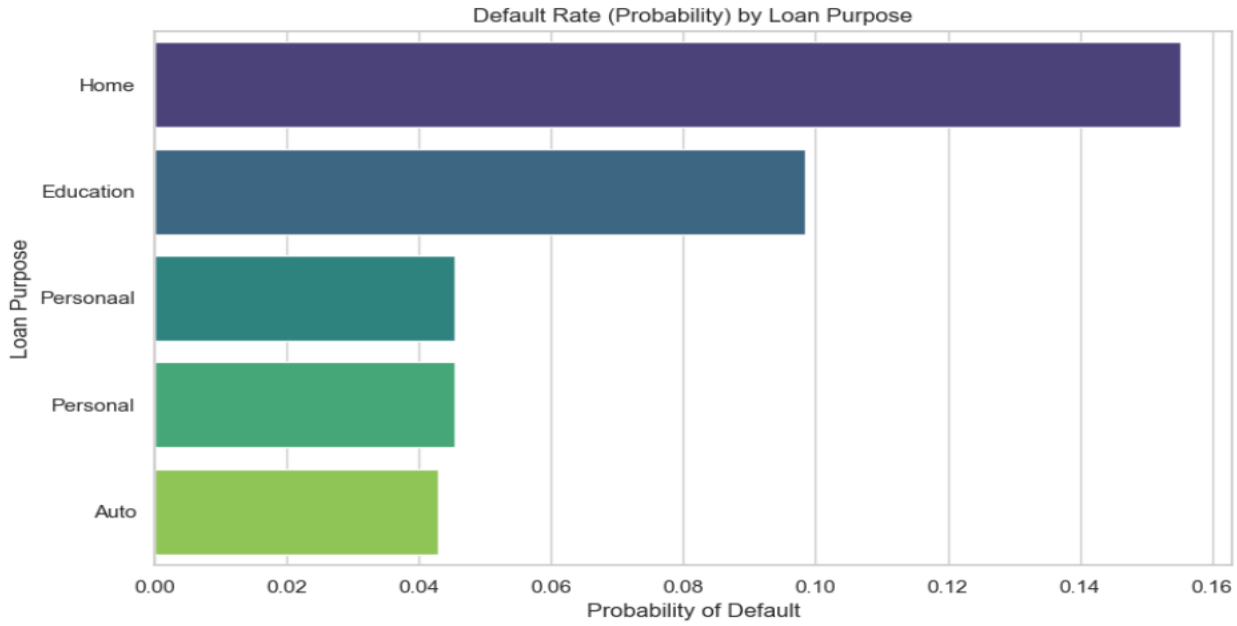


Fig 4: Loan Purpose wise Risk Analysis

The impact of credit history is analyzed for Days Past Due (DPD) on risk. This shows that the DPD is comparatively higher for Defaulters. This information is useful to further improve the risk model. Despite being a small percentage of total applicants, those with high DPD are on average significantly more likely to default.

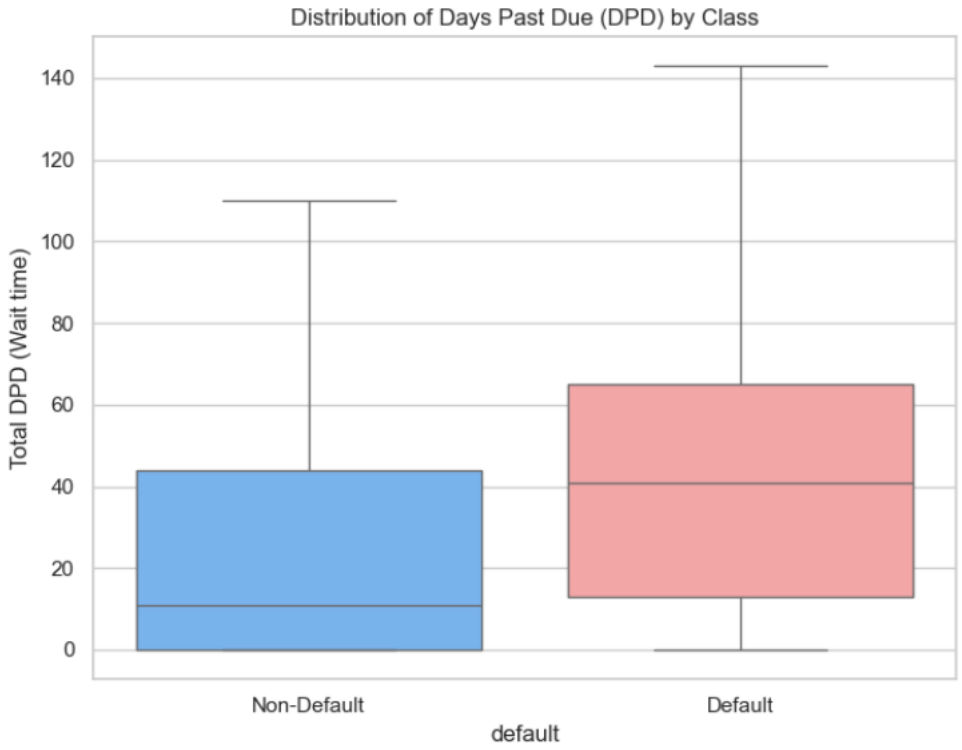


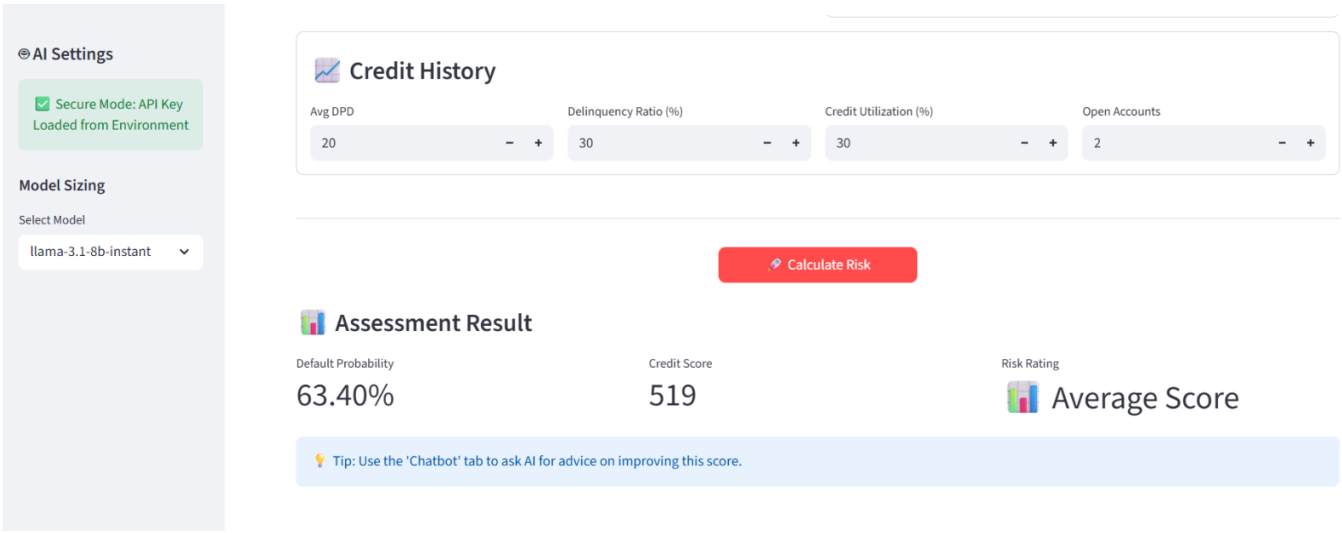
Fig 5: Analysis of Days Past Due (DPD) on Risk

4. Results and discussion:

Logistic Regression, Decision Tree and Random Forest were used to predict the default probability of the loan applicants. Among the given algorithms Logistic Regression was the best performing one as it provided the highest Accuracy of 0.91 and F1-score of 0.88.

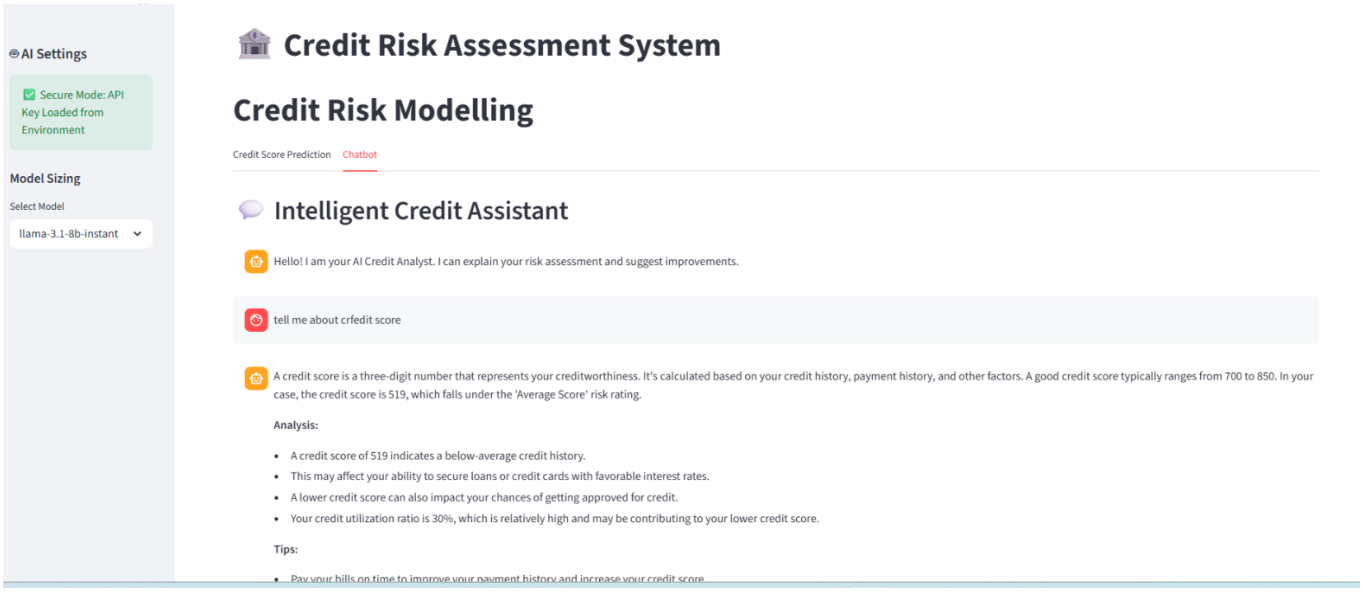
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, mean_absolute_error

# Initialize and Train
model = LogisticRegression()
model.fit(x_train, y_train)

# Predict
y_prediction = model.predict(x_test)

MAE = mean_absolute_error(y_test, y_prediction)
print(f"MAE = {MAE}")

Accuracy = accuracy_score(y_test, y_prediction)
print(f"Accuracy = {Accuracy}")
MAE : 0.085
Accuracy : 0.915
```

Prediction Module Results:



- **Latency**: Inference time is under 200ms.
- **Transparency**: The breakdown of the score (300-900) helps users understand their standing.

- **GenAI Reasoning**:

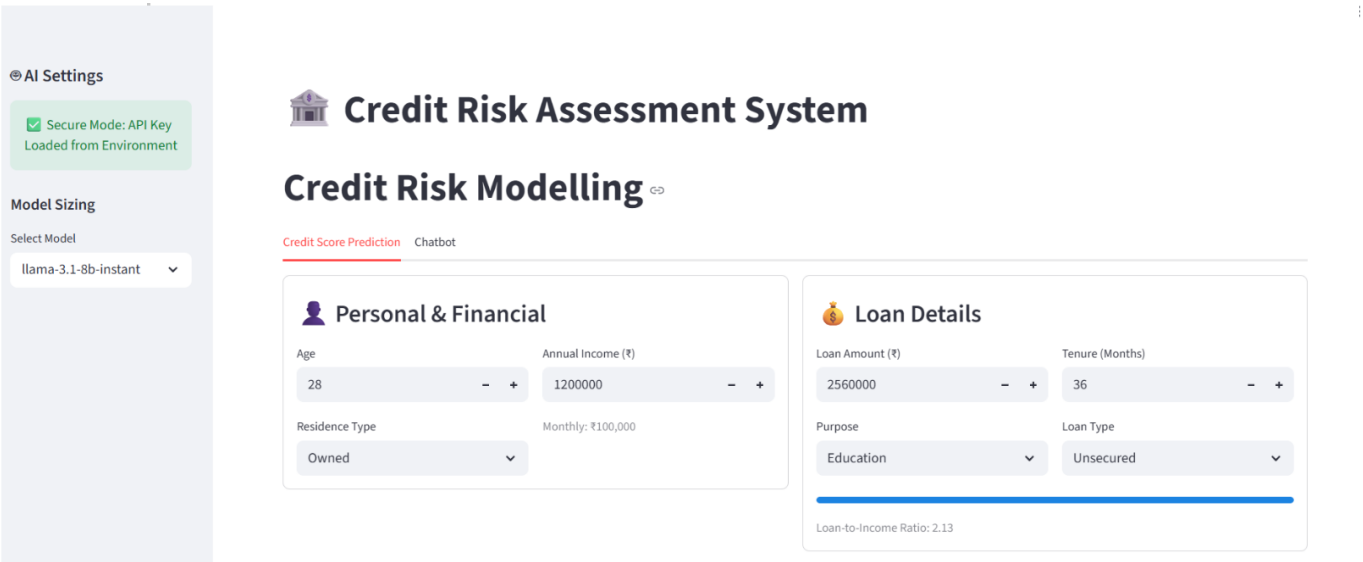


The chatbot correctly interprets context. For a user with high utilization, it specifically advises paying down revolving debt, demonstrating "context-awareness" rather than generic advice.

5. GUI:

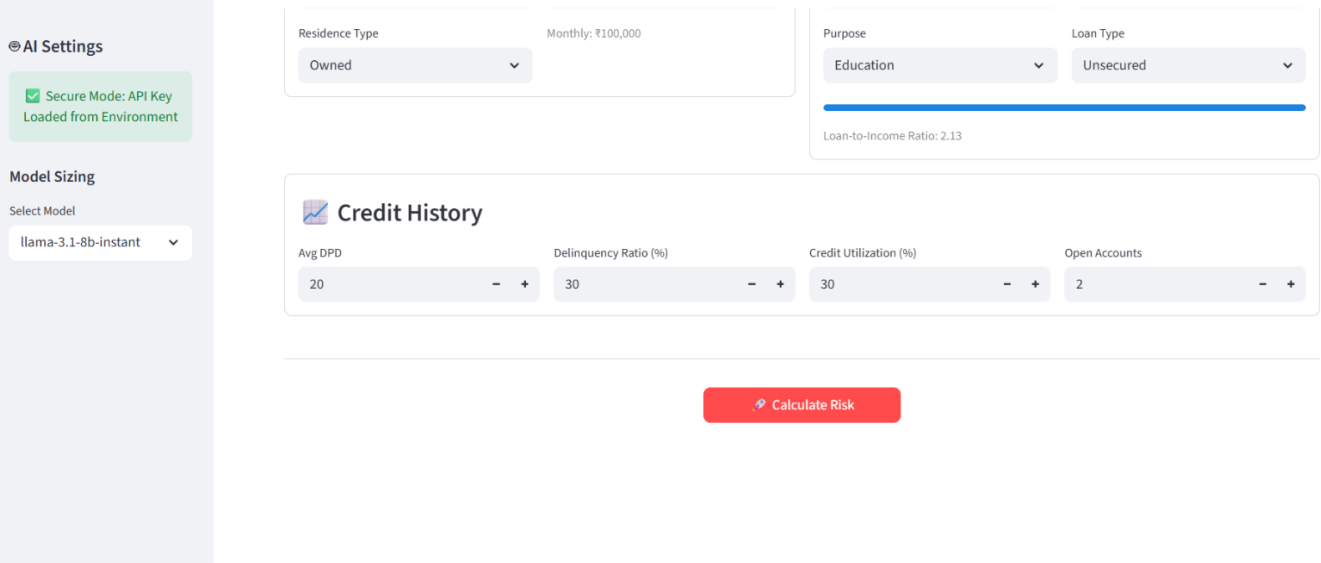
The User Interface is built using Streamlit, designed for responsiveness and ease of use

Dashboard Overview:



- **Sidebar**: Used for navigation and API key security.
- **Main Panel**: organized into clear tabs/sections for Personal, Financial, and Loan details

Input Forms:



The form captures granular details including Financial Ratios and Delinquency History, ensuring the model has a holistic view of the applicant.

6.GitHub Link:

The complete source code, dataset, and documentation for this project are available at the following GitHub repository:

GitHub Repository Link: https://github.com/mohitjain3350/Credit_Risk_Modelling/

7.Future work And Conclusion

7.1Future Work:

1. **Model Upgrade**: Transition from Logistic Regression to XGBoost or CatBoost for higher accuracy on larger datasets.
2. **Real-time Data**: Integrate with Banking APIs (Account Aggregators) to fetch live financial data instead of manual input.
3. **Multilingual Support**: Update the Chatbot to support local Indian languages (Hindi, Tamil, etc.) for broader accessibility.

The Bank can analyze the entire customer credit data across the region to arrive at an even more accurate prediction. They can analyze the repayment patterns and default history as well to optimize their loan product portfolio. They can analyze the risk assessment time and accuracy to arrive at achievable lending targets for loan officers to motivate them better.

7.2 Conclusion:

The Credit Risk Assessment System successfully bridges the gap between complex banking algorithms and the end-user. By combining a transparent statistical model (Logistic Regression) with an explanatory AI (LLM), the project not only scores the user but also educates them. This fulfills the objective of creating a financial tool that is both accurate and assistive.

- Non-Defaulters (Class 0) constitute the majority of the dataset compared to Defaulters (Class 1).
- Applicants with higher Annual Income consistently show lower risk and better repayment history.
- Default probability is significantly affected by Credit History. Applicants with high 'Days Past Due' (DPD) witnessed higher default rates.
- Financial ratios like Credit Utilization Ratio are major contributing factors in the risk assessment.
- Risk is also dependent on the Loan Purpose as Unsecured loans (Personal) showed higher default rates than Secured loans (Home/Auto).
- Among the trained models for predicting credit risk, Logistic Regression performs the best.