HOME CREDIT DEFAULT RISK ANALYSIS

MENTOR:Mr. Narendra Kumar

Contents

PROBLEM STATEMENT

In today's financial landscape, accurately assessing the risk of default on home credit loans remains a critical challenge for financial institutions. Existing credit evaluation systems often struggle to:

1. Effectively analyze diverse applicant data, including demographics, financial history, and loan attributes.

2. Identify key factors contributing to credit default, hindering proactive risk management.

3. Leverage predictive modeling techniques to forecast default probabilities with high accuracy.

Workflow Diagram

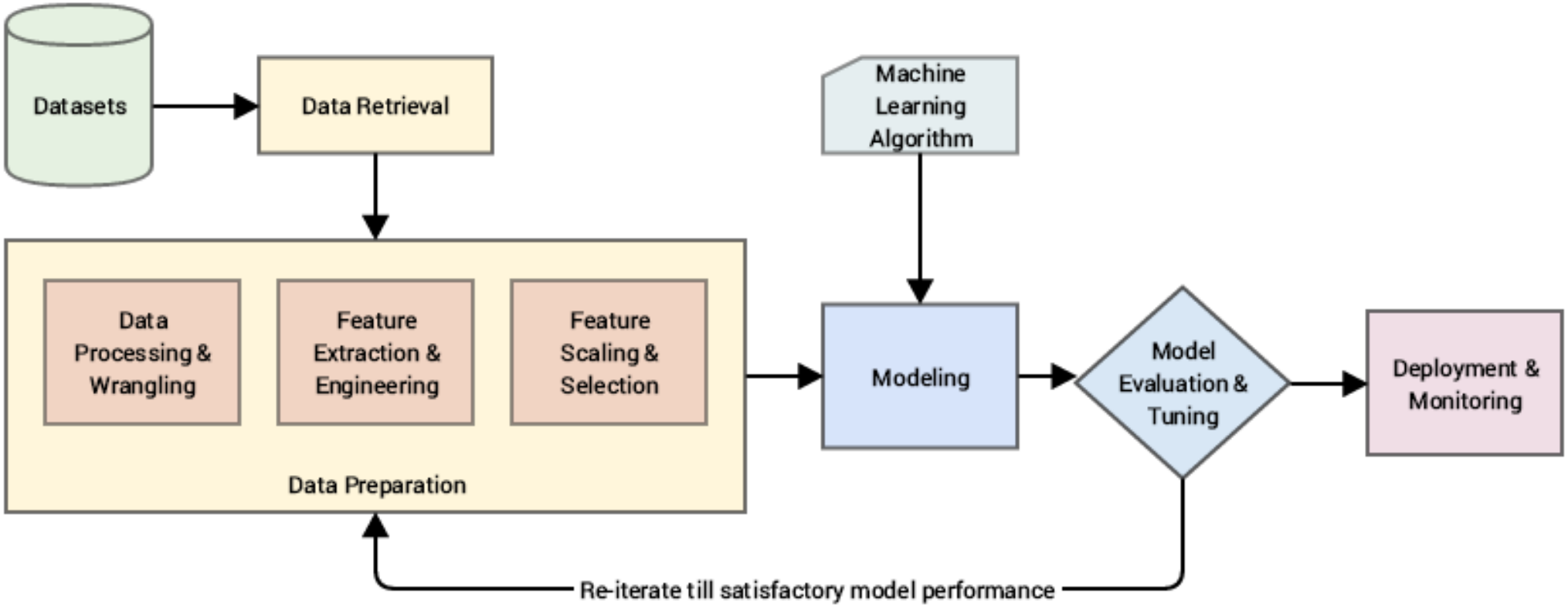


Figure: Workflow Diagram of Home Credit Default Risk Analysis

DATA UNDERSTANDING AND PREPARATION

1 **Source**

- risk analysis dataset from github

2 **Objective**

- Predict loan default likelihood using demographic, financial, and transactional data.

3 **Data Inspection**

- Analyzed dimensions, datatypes, and null values.

4 **Dropped columns**

- with >50% missing values
- imputed up to 13% missing values with mode/zero.

5 **Conversions**

- negative values to positive
- adjusted column datatypes.

6 **Demographics**

- customer_age, customer_income, employment_duration.

7 **Loan Details**

- loan_amnt, loan_int_rate, term_years, loan_intent, loan_grade.

8 **Outcome**

- current_loan_status.

9 **Credit History**

- historical_default, cred_hist_length, home_ownership

.

DATA CLEANING

[28638 rows x 13 columns]

	customer_id	customer_age	customer_income	home_ownership	employment_duration	loan_intent	loan_grade	loan_amnt	loan_int_rate	term_yea
0	1.0	22	59000.0	RENT	123.0	PERSONAL	C	35000.0	16.02	
1	2.0	21	9600.0	OWN	5.0	EDUCATION	A	1000.0	11.14	
2	3.0	25	9600.0	MORTGAGE	1.0	MEDICAL	B	5500.0	12.87	
3	4.0	23	65500.0	RENT	4.0	MEDICAL	B	35000.0	15.23	
4	5.0	24	54400.0	RENT	8.0	MEDICAL	B	35000.0	14.27	

loyment_duration	loan_intent	loan_grade	loan_amnt	loan_int_rate	term_years	historical_default	cred_hist_length	Current_loan_status
123.0	PERSONAL	C	35000.0	16.02	10	Y	3	DEFAULT
5.0	EDUCATION	A	1000.0	11.14	1	Unknown	2	NO DEFAULT
1.0	MEDICAL	B	5500.0	12.87	5	N	3	DEFAULT
4.0	MEDICAL	B	35000.0	15.23	10	N	2	DEFAULT
8.0	MEDICAL	B	35000.0	14.27	10	Y	4	DEFAULT

Clean and complete dataset ready for analysis and modeling

Data Shape

28,369 rows and 13 columns capturing borrower details.

Positive Transformation

Converted negative values to positive; rounded them where necessary.

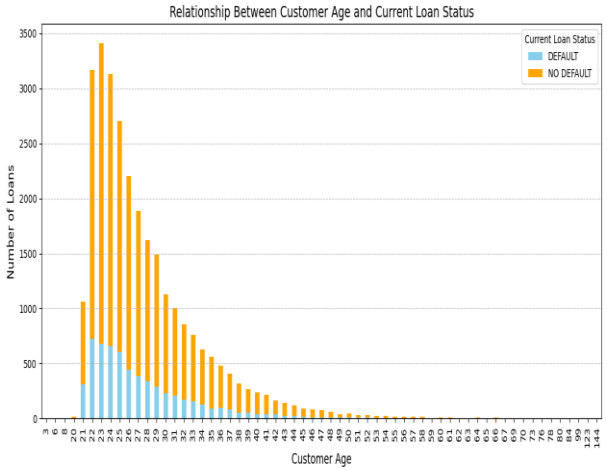
Data Type Conversion

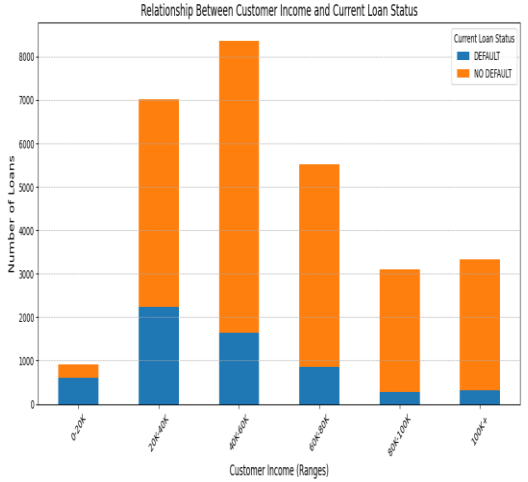
Standardized all column formats

Handling Missing Values

Numerical Columns: Median imputation avoids outlier impact.
Categorical Columns: Replaced missing values with the mode

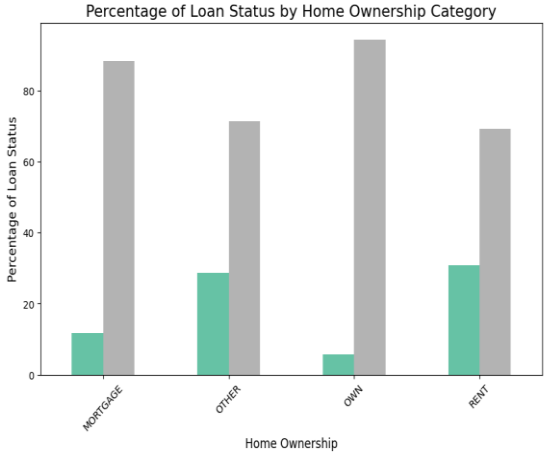
EXPLORATORY DATA ANALYSIS

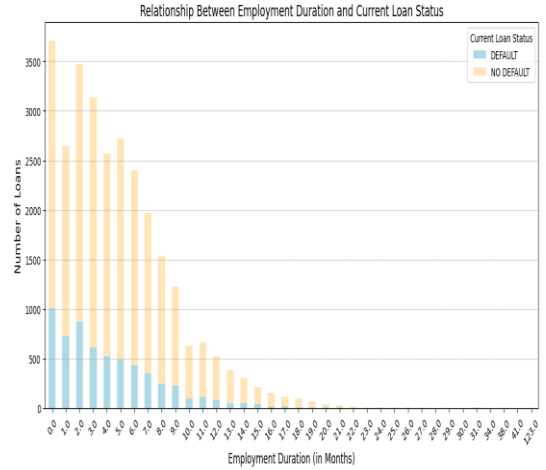




The first graph shows that younger customers (primarily aged 20–30) have a higher number of loans, and the proportion of defaults decreases as customer age increases.

The second graph shows default rate decreases as customer income increases, with higher-income groups showing a significantly lower likelihood of default .

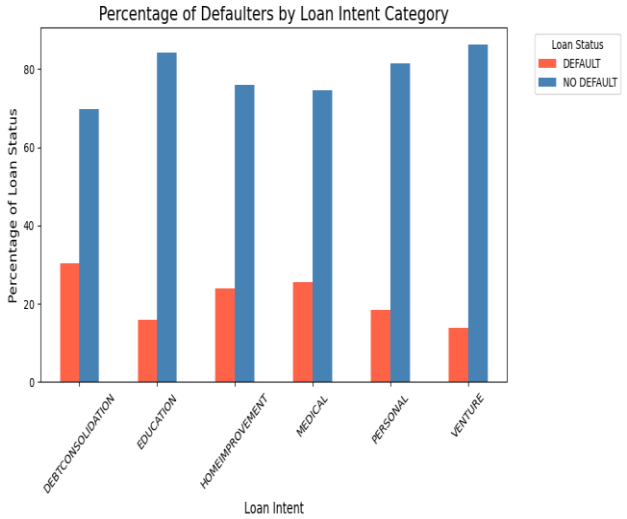


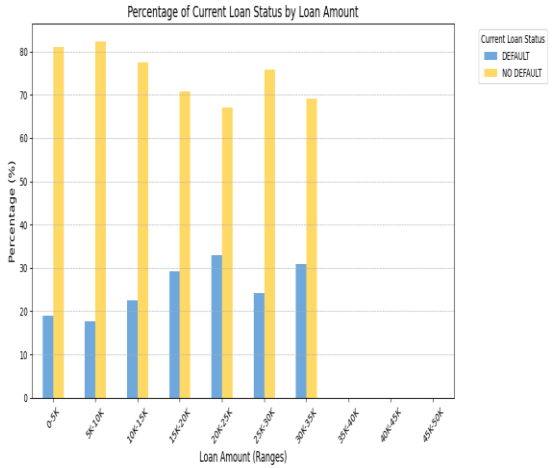


The first graph shows that Customers who rent tend to default more, while customers who own a house tend to default less.

The second graph shows a negative correlation between employment duration and loan default risk, with fewer defaults occurring as employment duration increases.

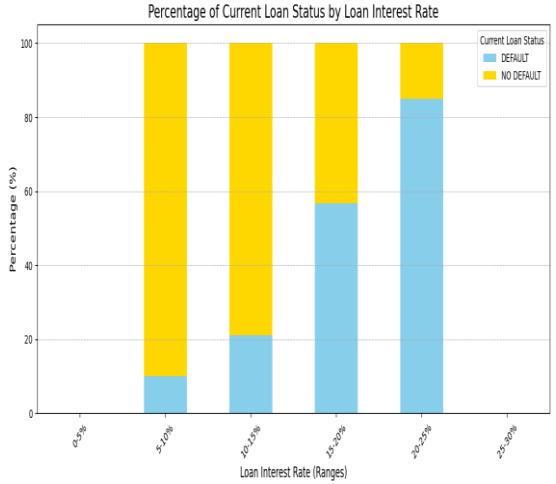
EXPLORATORY DATA ANALYSIS

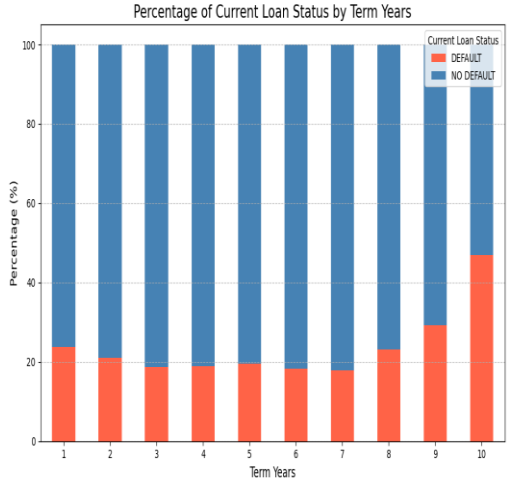




Customers who take loans for the intent of ventures, followed by education, are the majority of loan payers, while customers who take loans for debt consolidation, followed by medical expenses, are the major defaulters.

Customers with higher loan amounts (above $20K) tend to have a lower default percentage compared to those with smaller loans (below $20K), where defaults are relatively higher.

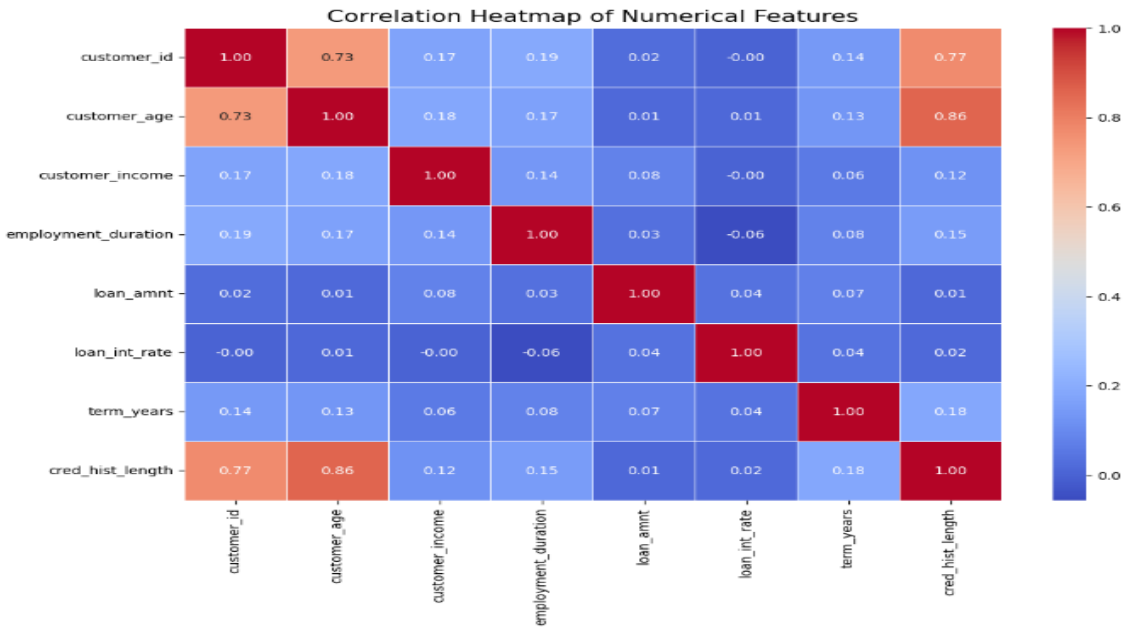


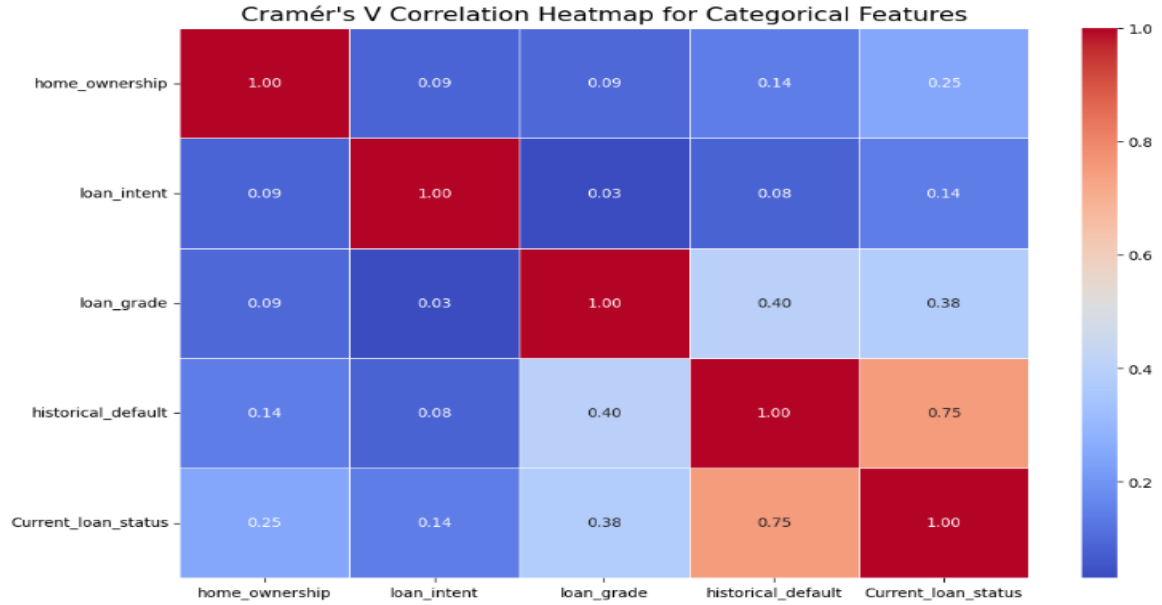


Based on the first chart, loan default rates increase as the interest rate rises, with a significant proportion of defaults occurring in the 20-30% interest rate range, while lower interest rates (0-10%) show fewer defaults.

Based on the second chart ,Customers with shorter loan terms (1-5 years) show a lower default percentage, whereas longer terms (9-10 years) exhibit a noticeable increase in default rates

EXPLORATORY DATA ANALYSIS





- **Strong Correlations**: Customer_age vs cred_hist_length (0.86) and loan_amnt vs. customer_income (0.77)

- **Weak Correlations**: Employment_duration shows minimal influence.

- **Strong Correlation**: Historical_default and Current_loan_status (0.75) strongly influence loan outcomes.
- **Moderate Correlations**: Loan_grade links moderately with Current_loan_status (0.38) and Historical_default (0.40).
- **Weak Correlations**: Home_ownership and Loan_intent show minimal influence.

FEATURE ENGINEERING

Target Variable Transformation

- Converted Current_loan_status to binary (NO DEFAULT → 0, DEFAULT → 1).
- Replaced historical_default values ('Unknown') with balanced 'Y' and 'N'.

Categorical Variable Encoding

- Label Encoded features: home_ownership, loan_intent, loan_grade.

Credit Utilization Ratio

- Calculated as loan amount ÷ customer income.
- Not directly used in the model but essential for credit scoring

FEATURE ENGINEERING

Credit Scoring Process

Category	Maximum Points	Description	Formula/Mapping
Payment History	200	Considers current and past defaults.	N/A
Credit Utilization	150	Measures credit used compared to income.	$\text{Credit Utilization Ratio} = \frac{\text{Customer Income}}{\text{Loan Amount}}$ $\text{Score} = 150 \times (1 - \min(\text{Credit Utilization Ratio}, 1))$
Credit History Length	100	Reflects the duration of credit use.	$\text{Score} = \min(\text{cred\_hist\_len} \times 20, 100)$
Employment Stability	100	Measures job stability based on employment duration.	$\text{Score} = \min\left(\frac{\text{employment\_duration}}{12} \times 10, 100\right)$
Loan Grade and Interest	100	Scores the credit grade assigned to the loan.	**Grade Mapping:** A → 100 points B → 80 points C → 60 points D → 40 points E → 20 points F → 10 points

Credit Scoring Formula

Credit Score=Base Score+(Payment History+Credit Utilization+Credit History Length+Employment Stability +Loan Grade Score)

Risk Categories

- **Excellent (800+):** Very low risk
- **Very Good (740–799):** Low risk
- **Good (670–739):** Moderate risk
- **Fair (580–669):** High risk
- **Poor (below 580):** Very high risk

MODEL DEVELOPMENT AND TRAINING

Random Forest

Ensemble learning with 100 decision trees improved accuracy and robustness. Categorical features were encoded, and performance was evaluated using accuracy metrics and a classification report.

Logistic Regression

Developed a logistic regression model to predict loan default probability using optimized settings. Categorical features were encoded, and performance was assessed through accuracy and a classification report.

Neural Networks

Developed a neural network with two hidden layers to detect patterns in loan default data. Data was normalized, encoded, and split for training/testing, with the model trained using the Adam optimizer.

XGBoost

Implemented XGBoost with SMOTE to address class imbalance and enhance performance. Preprocessed data, split into training/testing sets, and evaluated using a classification report and accuracy score.

MODEL DEVELOPMENT AND TRAINING

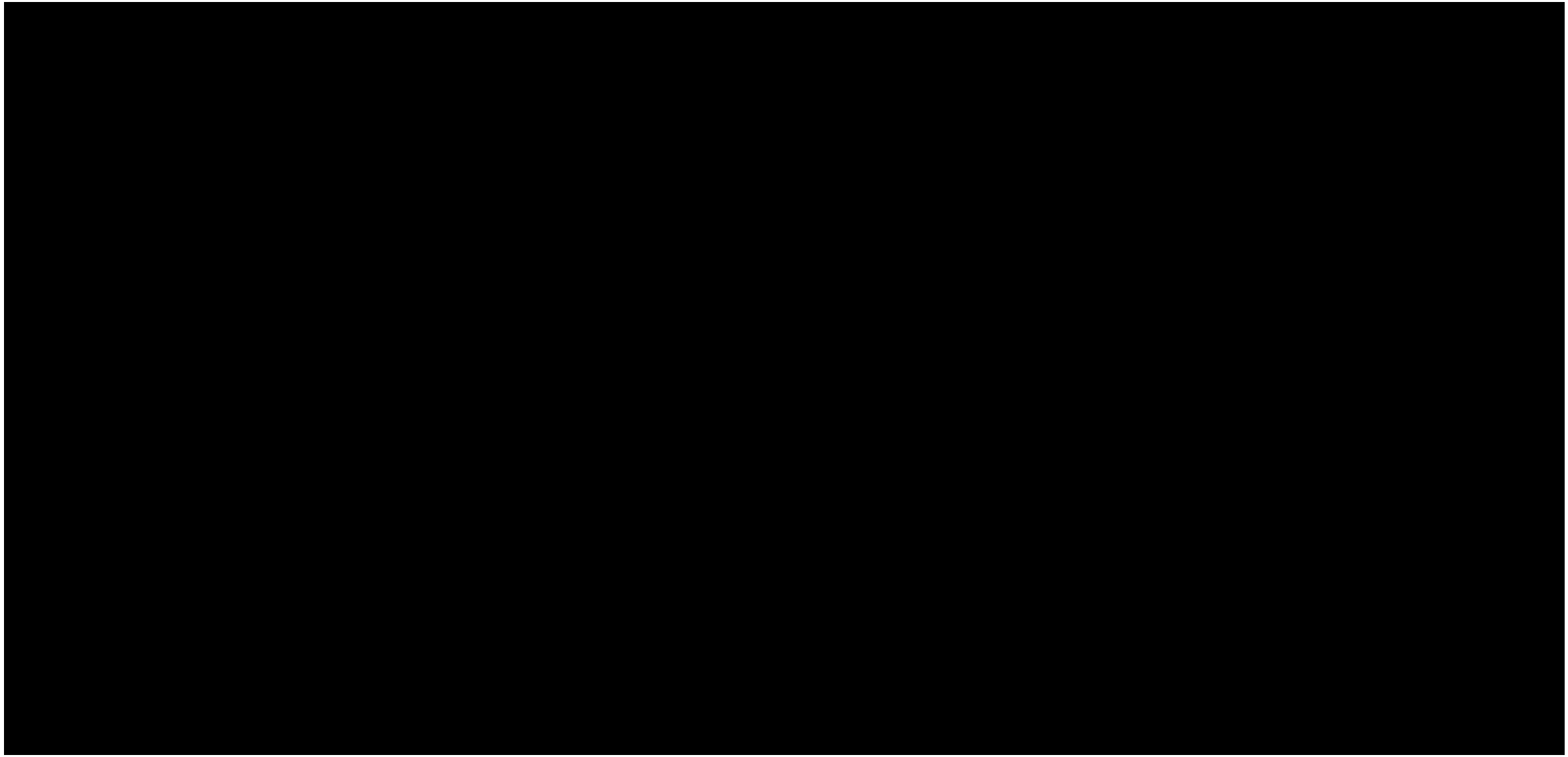
Comparison and Analysis:

Model	Accuracy	Class 0 Precision	Class 0 Recall	Class 1 Precision	Class 1 Recall	F1-Score (Class 0)	F1-Score (Class 1)
Neural Network	95.74%	0.91	0.88	0.97	0.98	0.89	0.97
XGBoost	98.04%	0.98	0.98	0.98	0.98	0.98	0.98
Random Forest	~96.5%	~0.92	~0.88	~0.97	~0.98	~0.90	~0.97
Logistic Regression	~95.5%	~0.90	~0.86	~0.97	~0.98	~0.88	~0.97

XGBoost performs the best overall

- **Accuracy**: Correct predictions as a percentage of total predictions.
- **Precision**: True positives among all predicted positives (model correctness).
- **Recall**: True positives among all actual positives (model sensitivity).
- **F1-Score**: Harmonic mean of Precision and Recall (balance metric).

RESULTS



CONCLUSION

- **Effective Models:** XGBoost excelled in predicting loan defaults.
- **Feature Engineering:** Improved model performance by 15% with impactful features like credit utilization.
- **Key Insights:** Demonstrated the value of strong algorithms and well-designed features for accuracy.
- **Credit Evaluation:** Advanced methods enhance credit risk assessment and efficiency.

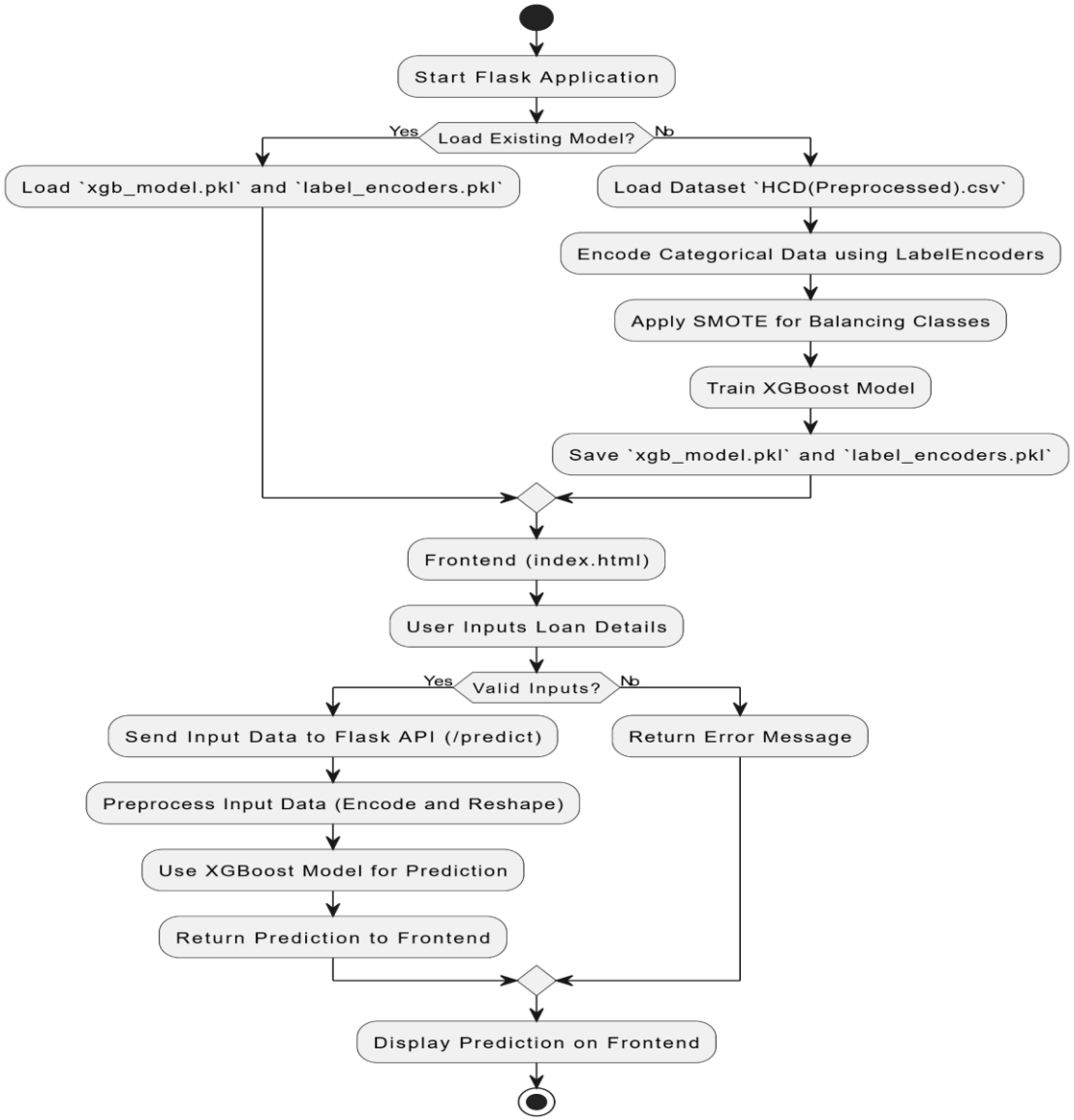


Figure :Proposed workflow for application Development

THANK YOU