CREDIT RISK ASSESSMENT

AIMLCZG628T: Dissertation

by

Sudhir Suresh Zanje

2023AC05339

Dissertation work carried out at

HSBC Technology India, Pune

Submitted in partial fulfilment of **M.Tech (AIML)**

degree programme

Under the Supervision of

Tom Babu

HSBC Technology India, Pune

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE PILANI (RAJASTHAN)

Dec - 2025

CONTENTS

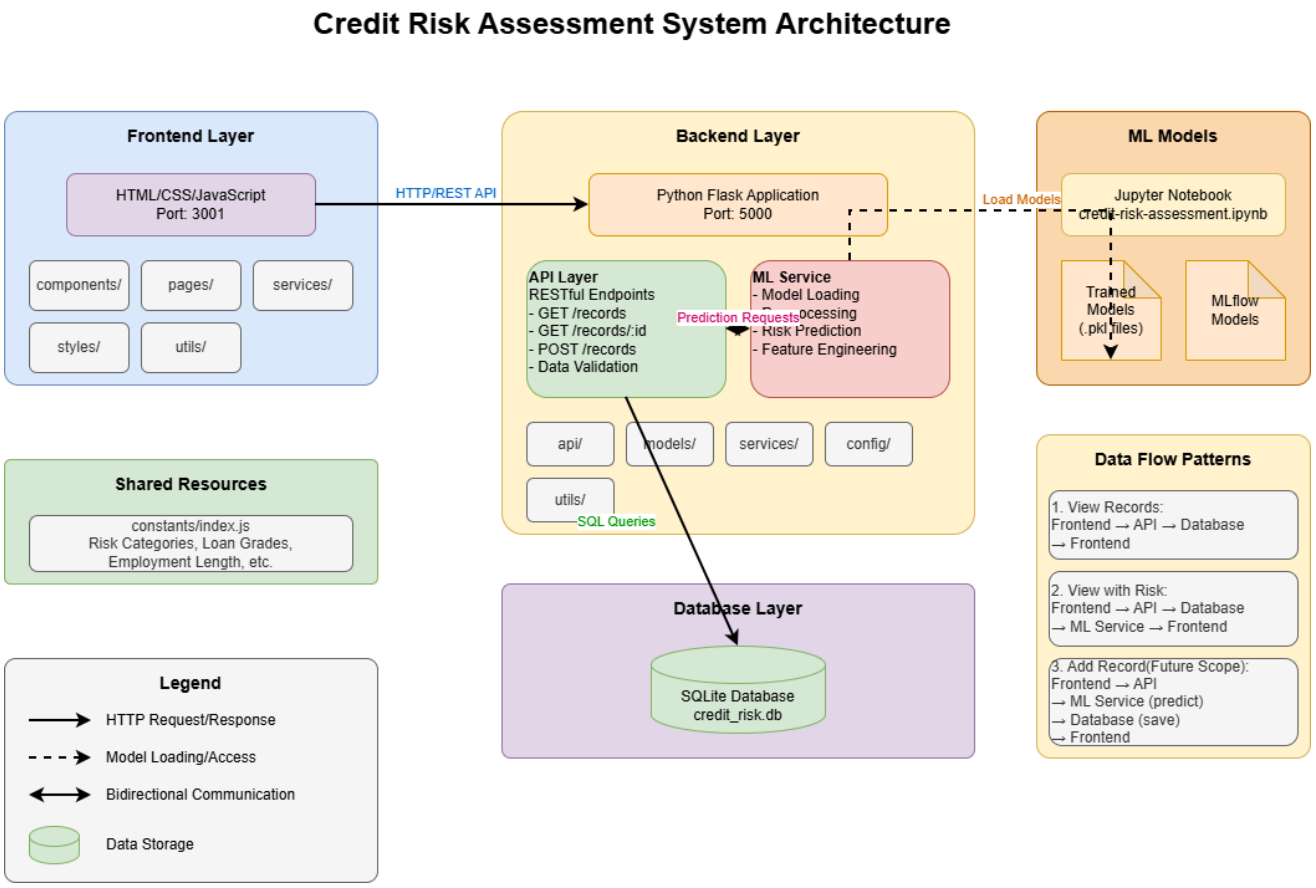
OBJECTIVE

1. Build a robust baseline for credit default prediction using multiple machine learning models on public credit risk datasets.
    a. Collect and preprocess at least one standardized, publicly available credit risk dataset (such as bank loan or credit card default data) and construct a clean modelling dataset with well-defined target (default/non-default) and feature set.
    b. Implement baseline classifiers including Logistic Regression, Random Forest, and Gradient Boosting (e.g., LightGBM or XGBoost) to quantify how different model families handle nonlinear patterns and high-dimensional financial features.
    c. Establish a comparative benchmark of these models using core metrics such as ROC-AUC, precision, recall, and F1-score to identify a strong candidate model for further refinement.
2. Systematically address class imbalance and improve discriminatory power of the models
    a. Analyze the degree of imbalance between good and bad borrowers in the chosen dataset and quantify its impact on default detection (recall for the minority "default" class).
    b. Apply and evaluate at least one resampling strategy (e.g., SMOTE or related techniques) and/or class-weighting to improve the model's ability to detect defaults without severely degrading precision.
    c. Tune key hyperparameters of the shortlisted models to increase AUC, recall, and F1-score for the default class, with a particular focus on minimizing false negatives from a risk-management perspective.
3. Design and evaluate an interpretable credit risk assessment pipeline suitable for banking deployment.
    a. Integrate explainable AI techniques (such as SHAP and, where feasible, LIME) with the best-performing model to generate both global feature importance and local, applicant-level explanations.
    b. Demonstrate how these explanations can support credit officers in understanding why a loan is classified as high or low risk, thereby improving transparency, auditability, and regulatory compliance.
    c. Define and document a deployable workflow—from raw data ingestion and preprocessing to model prediction and explanation generation—that could be integrated into a real banking environment with clear roles for each processing stage.
4. Quantify business impact through threshold calibration and metric selection
    a. Explore different decision thresholds on the predicted default probability to study the trade-off between approval rate, default detection rate (recall), and false positive rate.
    b. Propose one or two operating points (threshold + model) that are suitable for different business objectives, such as aggressive growth (higher approvals) versus conservative risk control (higher default capture), and justify them using the evaluation metrics.
    c. Relate technical metrics (ROC-AUC, precision, recall) to potential business outcomes such as reduction in expected losses, better portfolio quality, and improved consistency in credit decisions.

SYSTEM DESIGN

The Credit Risk Assessment System is designed to streamline the evaluation of credit risk for loan applicants. It consists of four primary layers: **Frontend, Backend, Database,** and **ML Models**, along with shared resources for constants and utilities. The system is modular, scalable, and integrates machine learning models for real-time risk prediction.

ARCHITECTURE DIAGRAM



ARCHITECTURE LAYERS

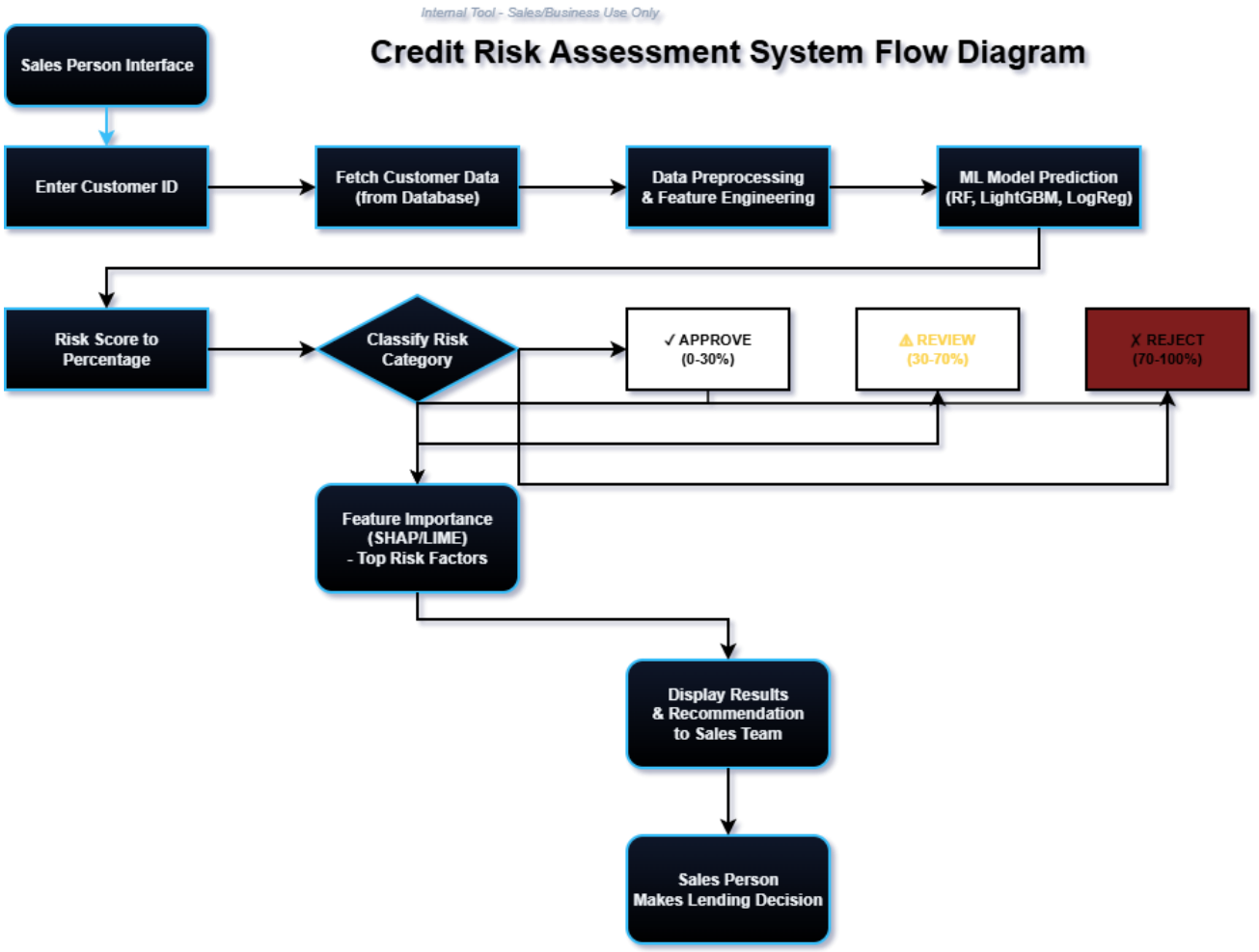
1. **Frontend Layer:**
   - Technology: HTML, CSS, JavaScript
   - Purpose: Provides a user-friendly interface for managing credit risk records.
   - Structure:
     - components/: Reusable UI components.
     - pages/: Page-level components for navigation.
     - services/: Handles API communication.
     - utils/: Utility functions for frontend logic.
     - styles/: CSS styles for consistent design.
   - Port: 3001
   - Communication: Sends HTTP requests to the backend for data retrieval and predictions.
2. **Backend Layer:**
   - Technology: Python Flask

- Purpose: Handles API requests, database operations, and ML model predictions.
- Structure:
  - api/: RESTful endpoints for CRUD operations.
  - models/: Database models for structured data storage.
  - services/: Business logic, including ML model inference.
  - config/: Configuration files for environment variables.
  - utils/: Helper functions for backend operations.
- Port: 5000
- Communication: Interacts with the database and ML models for predictions and data storage.

3. **Database Layer:**
   - Technology: SQLite
   - Purpose: Stores credit risk records and supports queries for data retrieval.
   - Structure:
     - scripts/: Database initialization scripts.
     - migrations/: Files for database schema updates.
   - Communication: Receives SQL queries from the backend.
4. **ML Models:**
   - Technology: scikit-learn, LightGBM, MLflow
   - Purpose: Provides trained models for risk prediction and feature engineering.
   - Structure:
     - credit-risk-assesment.ipynb: Jupyter notebook for model training.
     - models/: Directory for trained models (.joblib, MLflow models).
   - Communication: Models are loaded by the backend for inference.
5. **Shared Resources:**
   - Purpose: Centralized constants for risk categories, loan grades, and other shared values.
   - Structure:
     - index.js: Contains reusable constants for the system.

FLOW CHART DIAGRAM

The flowchart illustrates the Credit Risk Assessment System Flow for processing loan applications. It highlights the steps from data retrieval to risk classification and decision-making.



Steps in the Flowchart:

1. Enter Customer ID:
    - The sales team inputs the customer ID into the frontend interface.
2. Fetch Customer Data:
    - The backend retrieves customer data from the SQLite database.
3. Data Preprocessing & Feature Engineering:
    - The backend applies preprocessing steps (e.g., normalization, encoding) and extracts relevant features.
4. ML Model Prediction:
    - The trained ML models (e.g., Random Forest, LightGBM) predict the risk score based on customer data.
5. Risk Score to Percentage:
    - The risk score is converted into a percentage for easier interpretation.
6. Classify Risk Category:
    - The system categorizes the risk into:

        - Approve (0-30%): Low risk.
        - Review (30-70%): Medium risk.
        - Reject (70-100%): High risk.
7. Feature Importance Analysis:
    - Tools like SHAP or LIME identify the top risk factors influencing the prediction.
8. Display Results:
    - The frontend displays the risk score, category, and recommendations to the sales team.
9. Sales Team Decision:
    - Based on the results, the sales team makes lending decisions.

KEY DESIGN CONSIDERATIONS

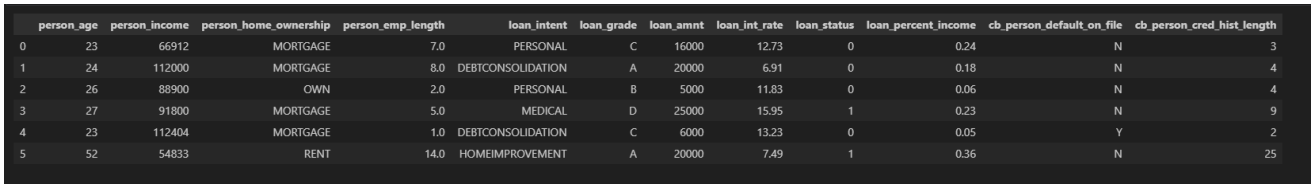
1. Scalability:
    - Modular architecture allows easy integration of new features (e.g., additional ML models, authentication).
2. Security:
    - Input validation, parameterized queries, and environment variables ensure secure operations.
3. Real-Time Predictions:
    - The system supports real-time risk assessment for quick decision-making.
4. Extensibility:
    - Shared resources and modular components simplify future enhancements.
5. Data Flow Patterns:
    - Clear communication paths between layers ensure efficient data processing and retrieval.

SAMPLE INPUT AND OUTPUT

The Credit Risk Assessment System processes loan applications by evaluating customer data and predicting credit risk. Below are examples of input data, API request formats, and corresponding outputs.

DATASET DETAILS :

DATA SAMPLE



	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
0	23	66912	MORTGAGE	7.0	PERSONAL	C	16000	12.73	0	0.24	N	3
1	24	112000	MORTGAGE	8.0	DEBTCONSOLIDATION	A	20000	6.91	0	0.18	N	4
2	26	88900	OWN	2.0	PERSONAL	B	5000	11.83	0	0.06	N	4
3	27	91800	MORTGAGE	5.0	MEDICAL	D	25000	15.95	1	0.23	N	9
4	23	112404	MORTGAGE	1.0	DEBTCONSOLIDATION	C	6000	13.23	0	0.05	Y	2
5	52	54833	RENT	14.0	HOMEIMPROVEMENT	A	20000	7.49	1	0.36	N	25

DATA COLUMNS DETAILS :

```
Data columns (total 12 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   person_age                  32581 non-null  int64
 1   person_income               32581 non-null  int64
 2   person_home_ownership       32581 non-null  object
 3   person_emp_length           31686 non-null  float64
 4   loan_intent                 32581 non-null  object
 5   loan_grade                  32581 non-null  object
 6   loan_amnt                   32581 non-null  int64
 7   loan_int_rate               29465 non-null  float64
 8   loan_status                 32581 non-null  int64
 9   loan_percent_income         32581 non-null  float64
 10  cb_person_default_on_file   32581 non-null  object
 11  cb_person_cred_hist_length  32581 non-null  int64
dtypes: float64(3), int64(5), object(4)
```

FEATURE DESCRIPTION:

person_age - Represents the applicant's age at the time of loan application.

person_income - The applicant's yearly earnings.

person_home_ownership - Housing status of the applicant, categorized as:

- *rent*: Currently leasing a residence
- *mortgage*: Property owner with an outstanding home loan
- *own*: Full property ownership without debt
- *other*: Alternative ownership arrangements

person_emp_length - Duration of current employment measured in years.

loan_intent - The purpose for which the loan is being requested.

loan_grade - Credit risk classification ranging from A (lowest risk) to G (highest risk):

- *A*: Excellent creditworthiness
- *B*: Good credit standing
- *C*: Average credit profile
- *D*: Below-average credit quality
- *E*: Poor credit standing
- *F*: Very poor creditworthiness
- *G*: Extremely high risk borrower

loan_amnt - Total loan amount sought by the applicant.

loan_int_rate - Annual percentage rate applied to the loan.

loan_status - Binary outcome variable (0 = repaid successfully, 1 = defaulted)

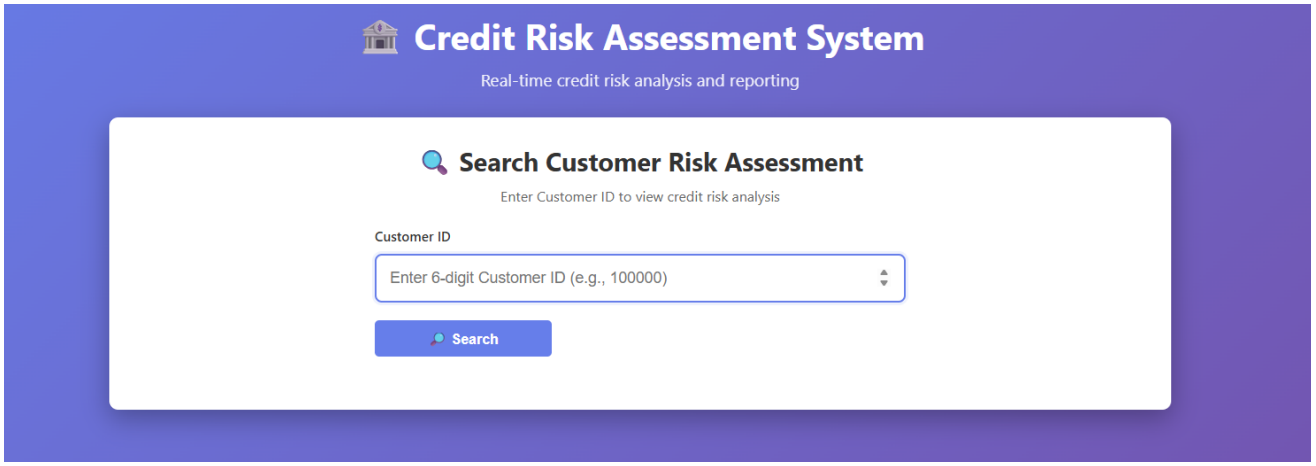
loan_percent_income - Debt-to-income ratio expressed as a percentage.

cb_person_default_on_file - Prior default history from credit bureau (Y = past defaults exist, N = clean record)

cb_preson_cred_hist_length - Number of years the individual has maintained credit accounts.

SAMPLE INPUT

UI INPUT:

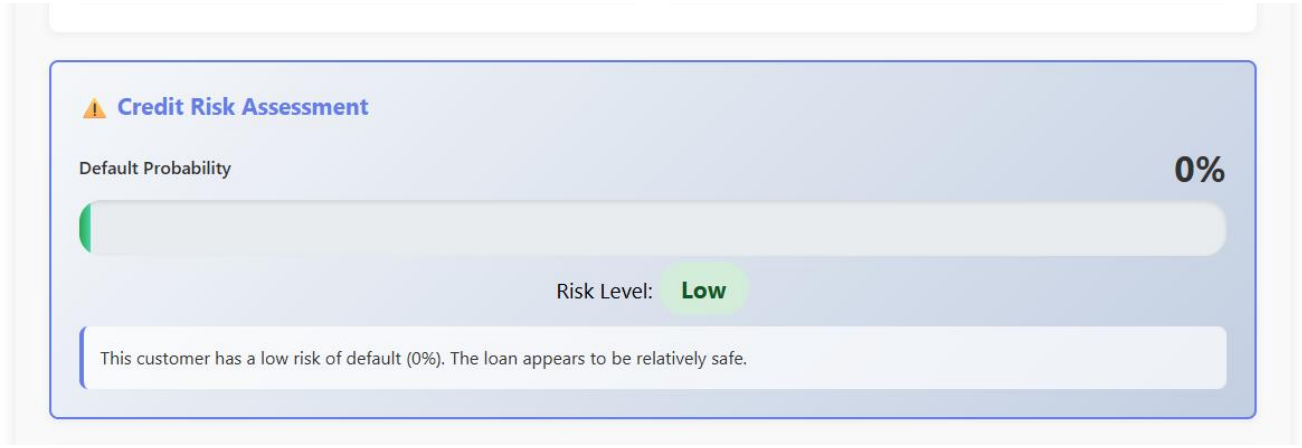


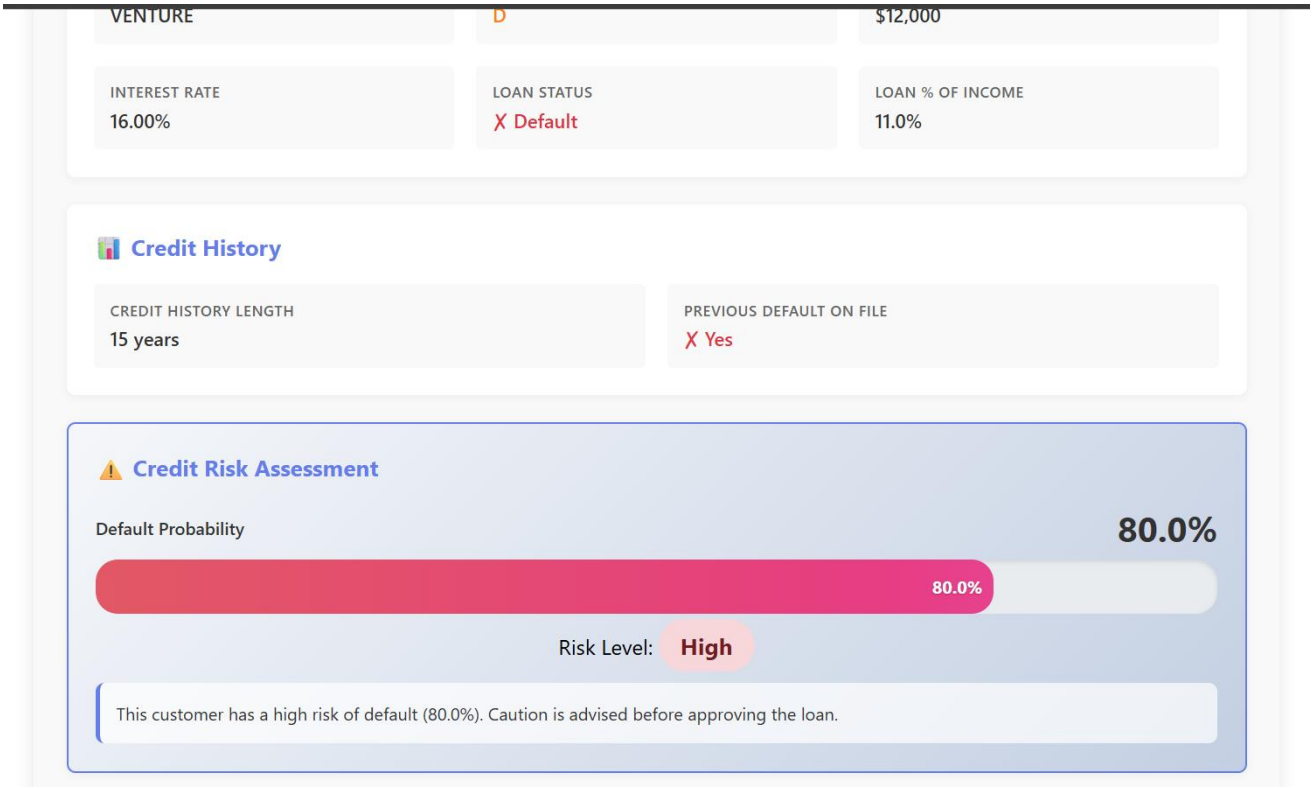
INPUT TO ML MODEL:

```
input data to prediction
   cb_person_default_on_file_N  cb_person_default_on_file_Y  loan_grade_A  loan_grade_B  loan_grade_C  loan_grade_D  loan_grade_E  loan_grade_F  loan_g
rade_G  person_home_ownership_MORTGAGE  person_home_ownership_OTHER  person_home_ownership_OWN  person_home_ownership_RENT  loan_intent_DEBTCONSOLIDATI
ON  loan_intent_EDUCATION  loan_intent_HOMEIMPROVEMENT  loan_intent_MEDICAL  loan_intent_PERSONAL  loan_intent_VENTURE  income_group_high  income_group
_high-middle  income_group_low  income_group_low-middle  income_group_middle  age_group_20-25  age_group_26-35  age_group_36-45  age_group_46-55  age_g
roup_56-65  age_group_nan  loan_amount_group_large  loan_amount_group_medium  loan_amount_group_small  loan_amount_group_very large  index  person_age
 person_income  person_emp_length  loan_amnt  loan_int_rate  loan_percent_income  cb_person_cred_hist_length  loan_to_income_ratio  loan_to_emp_length_
ratio  int_rate_to_loan_amt_ratio
0                      1.0                      0.0         0.0         1.0         0.0         0.0         0.0         0.0
   0.0                      1.0                      0.0                      0.0                      0.0                      0
.0                1.0                      0.0               0.0               0.0               0.0               0.0
       0.0               0.0                0.0               1.0               0.0               1.0               0.0               0.0
      0.0               0.0                1.0               0.0                    0.0                    0.0                    0         32
      60000.0             11.0      15000         11.86               0.25                    6                0.25                0.0
00733                0.000791
```

SAMPLE OUTPUT

CREDIT RISK PARAMETER UI :





EVALUATION MATRIX USED

EVALUATION MATRIX EXPLAINED

1. **Accuracy**
   - Definition: Proportion of correctly classified instances out of total instances
   - Formula: (TP + TN) / (TP + TN + FP + FN)
   - Range: 0 to 1 (higher is better)
   - Purpose: Overall model performance measure

2. **Precision**
   - Definition: Proportion of true positive predictions among all positive predictions
   - Formula: TP / (TP + FP)
   - Range: 0 to 1 (higher is better)
   - Purpose: Measures how many predicted defaults were actually defaults (reduces false alarms)

3. **Recall (Sensitivity)**
   - Definition: Proportion of actual positive cases correctly identified
   - Formula: TP / (TP + FN)
   - Range: 0 to 1 (higher is better)
   - Purpose: Measures how many actual defaults were correctly identified (reduces missed defaults)

4. **Specificity**
   - Definition: Proportion of actual negative cases correctly identified
   - Formula: TN / (TN + FP)
   - Range: 0 to 1 (higher is better)
   - Purpose: Measures how many non-defaults were correctly identified as non-defaults

Why These Metrics for Credit Risk?

- Precision: Critical to avoid falsely labeling good customers as high-risk
- Recall: Important to catch actual defaulters to minimize financial losses
- Specificity: Ensures good customers aren't denied loans unnecessarily
- Accuracy: Overall performance indicator across all predictions

Confusion Matrix Elements

- TP (True Positive): Correctly predicted defaults
- TN (True Negative): Correctly predicted non-defaults
- FP (False Positive): Incorrectly predicted defaults (Type I error)
- FN (False Negative): Missed actual defaults (Type II error)

ACTUAL RESULT

```
Parent Run ID: 7ec775ad11704619a7028289ed874cb3
    Child Run ID for SVC: 639375fed7384eb2b6c32b6fe3ccc876
Registered model 'CreditRiskModel_SVC' exists. Creating a new version...
Created version '6' of model 'CreditRiskModel_SVC'.
     Accuracy -  0.7827833071416099
     Precision -  0.0
     Recall -  0.0
```

```
     Specificity -  1.0
    SVC Training and Evaluation Run completed.
```

```
    Child Run ID for KN: fbb0530a780f4c508b0480d6064ac098
Registered model 'CreditRiskModel_KN' exists. Creating a new version...
Created version '6' of model 'CreditRiskModel_KN'.
     Accuracy -  0.7943076654443862
     Precision -  0.5565068493150684
     Recall -  0.2612540192926045
     Specificity -  0.9422261878206558
    KN Training and Evaluation Run completed.
```

```
    Child Run ID for DT: bce55d137de745c69cfd1d0be8cf2e8d
Registered model 'CreditRiskModel_DT' exists. Creating a new version...
Created version '6' of model 'CreditRiskModel_DT'.
     Accuracy -  0.8885978697398289
     Precision -  0.737460815047022
     Recall -  0.7564308681672026
     Specificity -  0.9252732545170644
    DT Training and Evaluation Run completed.
```

```
    Child Run ID for LR: 041bac964d2b4783a462d881aee04212
Registered model 'CreditRiskModel_LR' exists. Creating a new version...
Created version '6' of model 'CreditRiskModel_LR'.
     Accuracy -  0.8636284267504801
     Precision -  0.7508125677139762
     Recall -  0.5570739549839229
     Specificity -  0.9486950702654472
    LR Training and Evaluation Run completed.
```

```
    Child Run ID for RF: 782aa72b03f94bac9ae1e45bff62265e
Registered model 'CreditRiskModel_RF' exists. Creating a new version...
Created version '6' of model 'CreditRiskModel_RF'.
     Accuracy -  0.9312030731622141
     Precision -  0.9701327433628318
     Recall -  0.704983922829582
     Specificity -  0.9939772473789873
    RF Training and Evaluation Run completed.
```

```
    Child Run ID for AdaBoost: 0a69c17ebc7a4b1faf46643db7cd98cb
Registered model 'CreditRiskModel_AdaBoost' exists. Creating a new version...
Created version '6' of model 'CreditRiskModel_AdaBoost'.
     Accuracy -  0.8784704033525406
     Precision -  0.7644787644787645
     Recall -  0.6366559485530546
     Specificity -  0.9455721614989963
    AdaBoost Training and Evaluation Run completed.
```

```
    Child Run ID for BgC: 215e511bd1a24d4393ea6eee8e87dd4a
Registered model 'CreditRiskModel_BgC' exists. Creating a new version...
Created version '6' of model 'CreditRiskModel_BgC'.
     Accuracy -  0.9329491880565741
     Precision -  0.9663774403470716
     Recall -  0.7162379421221865
     Specificity -  0.9930849877314298
    BgC Training and Evaluation Run completed.
```

```
    Child Run ID for ETC: fd73d323ac624f9e8df5faa4d687be5a
Registered model 'CreditRiskModel_ETC' exists. Creating a new version...
Created version '6' of model 'CreditRiskModel_ETC'.
     Accuracy -  0.9257901169896979
     Precision -  0.9379679144385027
     Recall -  0.704983922829582
     Specificity -  0.9870622351104171
    ETC Training and Evaluation Run completed.
```

```
    Child Run ID for GBDT: c6db22d49d31438394a48e31c603d480
Registered model 'CreditRiskModel_GBDT' exists. Creating a new version...
Created version '6' of model 'CreditRiskModel_GBDT'.
     Accuracy -  0.92788545486293
     Precision -  0.9560922063666301
     Recall -  0.7001607717041801
     Specificity -  0.9910774035244256
    GBDT Training and Evaluation Run completed.
```

RESEARCH REFERENCES

PAPER 1: RESEARCH ON MACHINE LEARNING-BASED CREDIT RISK PREDICTION MODELS AND ALGORITHMS

Focus: Sichuan Provincial Data Science and Statistical Modeling Competition (2024)

Key Contribution:
This paper develops a credit risk prediction model using **LightGBM** on customer credit data with features including demographic statistics and borrowing behavior. The study achieved an **AUC score of 0.704523** (ranking 3rd in competition).

Dataset:

- Multiple dimensions of customer credit data
- Features: Demographics + borrowing/repayment behavior + risk performance post-loan

Methodology:

- Data preprocessing
- Feature selection with memory optimization
- LightGBM model with early stopping mechanism
- Hyperparameters: learning_rate=0.02, num_leaves=128, feature_fraction=0.8, bagging_fraction=0.8

Performance Results:

Model	Precision	Recall	F1	AUC
LightGBM	0.74	0.51	0.60	0.6885
XGBoost	0.68	0.46	0.54	0.6464
CNN	0.69	0.49	0.57	0.6794
DRL	0.71	0.50	0.58	0.6843

Novelty:

- AUC stability index for model robustness across economic cycles
- Optimization considering temporal stability rather than just static accuracy

PAPER 2: CREDIT RISK PREDICTION BASED ON MACHINE LEARNING METHODS

Focus: XGBoost vs. Logistic Regression comparison for credit scoring

Key Contribution:
Comprehensive comparison demonstrating **XGBoost significantly outperforms traditional logistic regression** in credit risk prediction with superior discrimination and stability.

Dataset:

- 50,000 samples: 45,000 good (9:1 ratio), 5,000 bad
- 15-month performance period (90+ days overdue = bad)
- 700+ derived features from original variables

Methodology:

- Feature engineering: continuous, categorical, and date variables
- Multiple time window aggregations
- Logistic Regression vs. XGBoost comparison

Performance Results:

Metric	Logistic Regression (Test)	XGBoost (Test)	Improvement
AUC	0.8585	0.9079	+5.7%
KS	0.5710	0.6617	+16%
Validation AUC	0.8350	0.8806	+5.5%

Stability Metrics (PSI):

- Logistic Regression: 0.0458 (very stable)
- XGBoost: 0.0565 (very stable)

Key Insight:
XGBoost not only improves discrimination ability (KS +16%) but maintains exceptional model stability, with PSI <0.1 on validation set.

PAPER 3: RESEARCH ON CREDIT RISK PREDICTION MODELS BASED ON MACHINE LEARNING

Focus: Comparative analysis of logistic regression, decision tree, random forest, and SVM

Key Contribution:
Random Forest emerges as the best performer with comprehensive performance analysis across 4 classical algorithms.

Performance Comparison:

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	**88.3%**	**87.5%**	**86.8%**	**87.1%**
SVM	84.5%	82.1%	83.4%	82.7%
Decision Tree	81.6%	80.2%	81.1%	80.6%
Logistic Regression	78.4%	76.8%	77.5%	77.1%

Key Findings:

- **Random Forest:** Handles complex data patterns and high-dimensional features effectively
- **SVM:** Good for high-dimensional data but computationally expensive at scale
- **Decision Tree:** Highly interpretable but prone to overfitting on high-dimensional data
- **Logistic Regression:** Limited for nonlinear relationships despite computational efficiency

Novelty:
Emphasizes critical importance of **data preprocessing and feature selection** in improving model performance across all algorithms.

PAPER 4: DESIGN OF A PERSONAL CREDIT RISK PREDICTION MODEL AND LEGAL PREVENTION OF FINANCIAL RISKS

Focus: GAN (WGAN-GP) + LightGBM combination for handling imbalanced credit data

Key Innovation:
First to combine **Wasserstein GAN with Gradient Penalty (WGAN-GP)** for synthetic minority class generation with **LightGBM** for prediction.

Dataset:

- Consumer finance company data: 559 features
- Severe class imbalance (default vs. normal samples)

Methodology:

- **WGAN-GP** for data oversampling: Generates synthetic default samples
- **LightGBM** for classification
- Addresses class imbalance without traditional SMOTE limitations

Performance Results:

Metric	WGAN-GP + LightGBM
Accuracy	86.7%
AUC (avg)	0.86
Kolmogorov-Smirnov (avg)	0.87

WGAN-GP Advantages:

- Uses Wasserstein distance for stable training
- Gradient penalty prevents mode collapse
- Better sample quality than traditional GAN
- Mitigates vanishing gradient problem

Key Findings:

- GAN significantly improves risk classification on imbalanced datasets

Consumer behavior data identified as important predictor despite weak linear correlation

PAPER 5: EXPLAINABLE ARTIFICIAL INTELLIGENCE CREDIT RISK ASSESSMENT USING MACHINE LEARNING

Focus: XAI implementation (SHAP + LIME) with ensemble methods for transparent credit risk assessment

Key Innovation:
Comprehensive framework combining **XGBoost, LightGBM, Random Forest with SHAP and LIME** for interpretable predictions.

Methodology:

- **Preprocessing:** Custom imputation, one-hot encoding, standardization
- **Class Imbalance:** SMOTE
- **Hyperparameter Tuning:** GridSearchCV

- **XAI Methods:** SHAP (global feature importance) + LIME (local instance explanations)

Model Comparison Results:

Metric	XGBoost	LightGBM	Random Forest
Accuracy	94.2%	**96.1%**	93.8%
Precision	89.5%	**92.3%**	88.7%
Recall	87.6%	**90.1%**	86.9%
ROC-AUC	0.92	**0.94**	0.91

Key Insight:
LightGBM = Most business-optimal with highest accuracy and best approval/default rate trade-off. SHAP and LIME provide applicant-specific explanations for regulatory compliance.

XAI Features:

- **SHAP Summary Plot:** Feature contribution visualization
- **LIME Feature Importance:** Per-applicant prediction explanation

Risk categorization (Low/Moderate/High) with transparent decision rules

PAPER 6: FINANCIAL TECHNOLOGY CREDIT RISK MODELING AND PREDICTION BASED ON RANDOM FOREST ALGORITHM

Focus: Random Forest for FinTech credit risk prediction with feature importance analysis

Key Contribution:
Demonstrates **Random Forest achieves 90-98% accuracy** in credit risk prediction with superior robustness to outliers and nonlinear relationships.

Dataset:

- Large-scale credit data with high-dimensional features
- Multiple decision trees ensemble approach

Random Forest Parameters:

- n_estimators: 200
- max_depth: 10
- max_features: "sqrt"

Performance Results:

Algorithm	Accuracy (Min)	Accuracy (Max)	Recall Rate (Min)	Recall Rate (Max)
Random Forest	**90%**	**98%**	**88%**	**96%**
Logistic Regression	85%	89%	79%	88%

Key Advantages:

- **Handles nonlinearity:** Captures complex feature interactions
- **Robust to outliers:** Multiple tree aggregation reduces noise impact

- **Feature importance ranking:** Interpretable variable importance scores
- **Imbalanced data handling:** Out-of-bag samples for balanced predictions

Limitation Noted:
Traditional methods don't capture dynamic customer behavior changes over time.

PAPER 7: ADVANCED USER CREDIT RISK PREDICTION MODEL USING LIGHTGBM, XGBOOST AND TABNET WITH SMOTEENN

Focus: SMOTEENN + PCA + ensemble models (LightGBM, XGBoost, Tabnet) for high-dimensional imbalanced data

Key Innovation:
Combines **PCA dimensionality reduction** with **SMOTEENN** (hybrid oversampling/undersampling) for optimal ensemble model performance.

Dataset:

- 40,000+ commercial bank records
- Extreme class imbalance (667 approve vs. 45,318 not approve = ~1.5% positive class)

Data Processing Pipeline:

1. **Information Value (IV)** feature selection
2. **Random UnderSampling** for class balance
3. **PCA** for dimensionality reduction
4. **SMOTEENN** for synthetic sampling

Performance Results:

Model	Accuracy	Precision	Recall	ROC-AUC
LightGBM + PCA + SMOTEENN	**92.4%**	**88.7%**	**85.3%**	**0.91**
XGBoost + PCA + SMOTEENN	91.8%	87.2%	84.1%	0.89
Tabnet + PCA + SMOTEENN	90.6%	85.9%	82.8%	0.88

Key Finding:
LightGBM shows most outstanding performance, with PCA + SMOTEENN combination addressing both curse of dimensionality and severe class imbalance simultaneously.

PAPER 8: DYNAMIC ENSEMBLE MACHINE LEARNING CLASSIFIER BASED CREDIT CARD FINANCIAL RISK MANAGEMENT AND PREDICTION

Focus: Dynamic Ensemble (RF + XGBoost + SVM + ANN) for credit card fraud risk prediction

Key Contribution:
Proposes **Dynamic Ensemble Machine Learning (DEML)** combining 4 complementary algorithms with min-max normalization for imbalanced financial data.

Dataset:

- CCF Kaggle dataset: 284,807 European transactions (Sept 2013)
- 492 fraud transactions, 30 features

- Severe imbalance (0.17% positive class)

Preprocessing:

- **Min-max normalization:** Scales to range[1]
- Prevents single feature dominance due to scale differences

Performance Results:

Classifier	F-Measure	Precision	Accuracy	Recall
DEML	**98.82%**	**98.37%**	**99.16%**	**98.73%**
ANN (single)	96.14%	96.34%	97.65%	95.97%
SVM (single)	95.42%	95.28%	96.47%	95.31%
RF (single)	93.71%	93.62%	94.28%	93.46%

Comparative Performance (K=5):

Method	F-Measure	Accuracy
DEML	98.82%	99.16%
DBN [Prior]	47.64%	78.28%
MLP-WOA	98.56%	98.56%

Key Insight:
Ensemble voting outperforms individual models by 2-4%, with DEML achieving near-perfect performance through aggregated predictions reducing false positives/negatives.

INSIGHT FROM RESEARCH PAPERS

Comparative Summary Table

Paper	Algorithm(s)	Best Accuracy/AUC	Unique Contribution	Data Size
1	LightGBM	0.705 AUC	AUC stability index	Large (Sichuan competition)
2	XGBoost	90.79% AUC	Feature engineering (700 vars)	50,000 samples
3	Random Forest	88.3%	Algorithm comparison (4 methods)	Large dataset
4	WGAN-GP + LightGBM	86.7% accuracy	GAN for imbalance + LightGBM	559 features
5	LightGBM + SHAP + LIME	96.1%	**Explainability focus**	Large bank dataset
6	Random Forest	90-98%	Feature importance analysis	High-dimensional
7	LightGBM + PCA + SMOTEENN	92.4%	**Dimensionality reduction**	40,000 records
8	DEML (RF+XGB+SVM+ANN)	**99.16%**	**Ensemble voting**	284,807 transactions

Algorithm Hierarchy for Credit Risk:

1. **Best Overall:** LightGBM (Papers 1, 4, 5, 7)
2. **Close Second:** XGBoost (Papers 2, 7, 8)
3. **Solid Alternative:** Random Forest (Papers 3, 6, 8)
4. **Emerging:** Tabnet (Paper 7), WGAN-GP (Paper 4)

Data Handling Trends:

- **Imbalance Solutions:** SMOTE, SMOTEENN, GAN, Class weighting, Threshold tuning
- **Feature Engineering:** 700+ derived variables (Paper 2), IV-based selection (Paper 7)
- **Dimensionality:** PCA effective (Paper 7), 559-3805 features across studies

Performance Metrics Focus:

- **Financial institutions prioritize:** AUC/ROC (0.72-0.99 range), Recall (catching defaults), Precision (reducing false approvals)
- **Business metrics:** Default rate prediction, approval rates, FPR/FNR trade-offs

This comprehensive review demonstrates that **modern credit risk prediction requires ensemble approaches combining LightGBM/XGBoost with proper class imbalance handling (SMOTE/GAN/Weighting) and explainability (SHAP/LIME) for production deployment.**