

Credit Scoring Project

DOCUMENTATION AND GUIDE

BY: DUKE KOJO KONGO



Introduction

- The goal of this project is to build an accurate and interpretable credit loan scoring system. The system will predict the likelihood of a borrower defaulting on a loan, aiding financial institutions in making informed lending decisions. This document outlines the end-to-end workflow of the project, from data acquisition to model deployment.

Data Requirement - Sources

01

Banking
transaction
history

02

Credit bureau
reports

03

Loan application
data

04

Customer
demographic
and employment
data

05

Macroeconomic
indicators
(inflation, GDP
growth)

06

Behavioral data
(e.g., mobile
money
transactions)

Data Requirements – Key Features



Demographic Data – Age, income, employment status, marital status, education.



Financial Data – Bank balance, loan history, savings, assets, liabilities.



Behavioral Data – Repayment history, frequency of defaults, transaction patterns.



Credit History – Previous loans, credit card usage, past defaults, credit score.



Macroeconomic Factors – Inflation, exchange rates, economic cycles.



Data Volume: Sufficient historical data (3-5 years) to ensure robust model training.

Technology Stack



Data Storage: PostgreSQL, BigQuery, AWS S3, or Azure Data Lake.



Data Processing: Python (Pandas, NumPy), SQL, Spark (for large datasets).



Modeling: LightGBM, XGBoost, CatBoost (for tabular data)



TensorFlow/PyTorch (for advanced neural networks if needed)



Scikit-learn for baseline models



Visualization: Tableau, Power BI, Matplotlib, Seaborn.



Deployment: Flask/FastAPI, Docker, Kubernetes, AWS Lambda, or GCP Functions.

Modeling and Methodology

Model Types:

- Logistic Regression (baseline)
- Gradient Boosting Models (LightGBM, XGBoost)
- Neural Networks (if necessary)
- Ensemble Models

Approach:

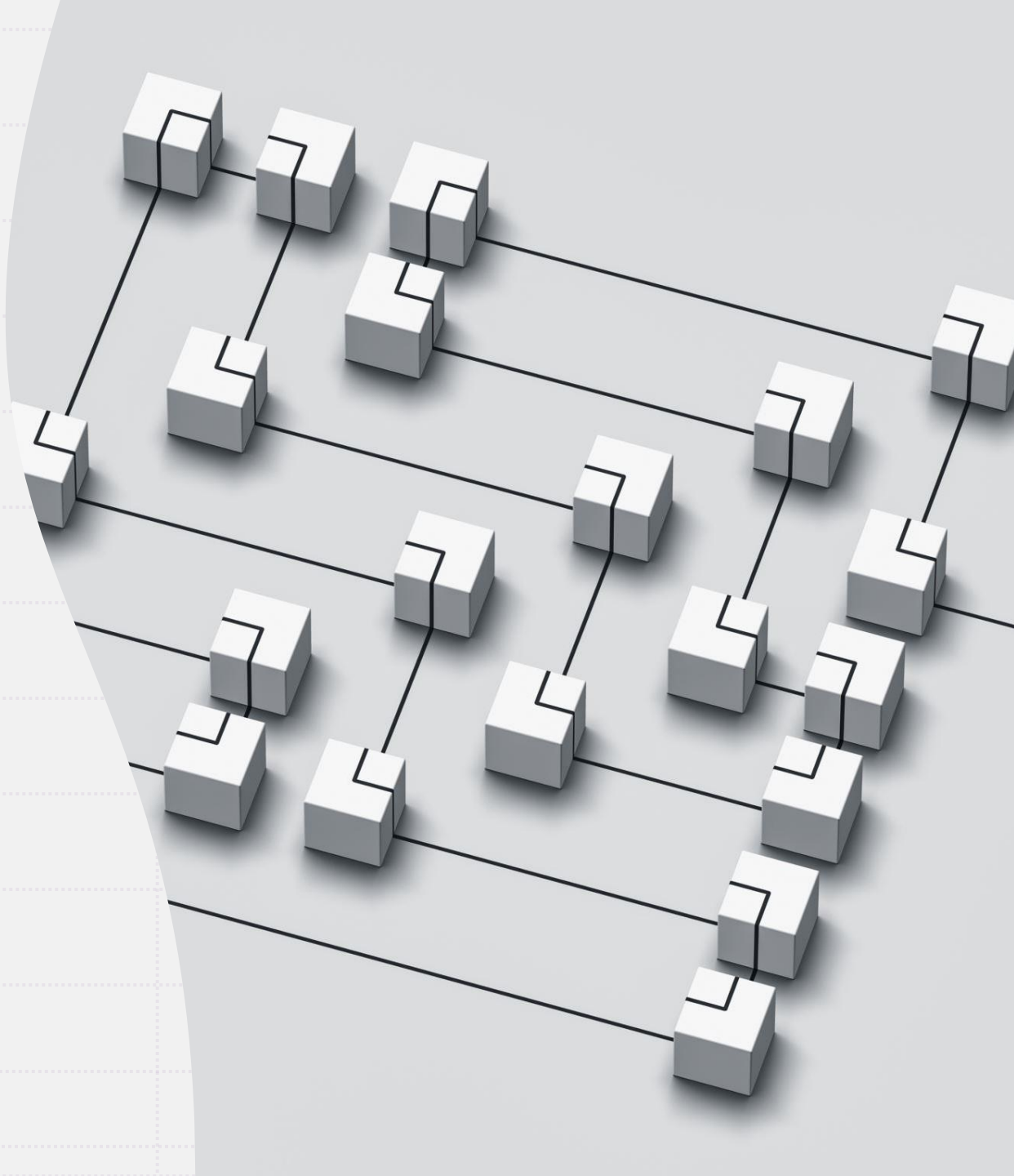
- Supervised learning with historical labeled data (approved/rejected loans).
- Time series analysis for financial indicators (if necessary).


Evaluation Metrics:

- AUC-ROC, F1-score, Precision-Recall, Gini coefficient.
- Business KPIs – Default rate, approval rate, profitability.

Project Workflow Overview

- The workflow is divided into six key phases:
 1. Data Acquisition and Quality Assessment
 2. Data Cleaning and Preprocessing
 3. Feature Engineering
 4. Model Development and Evaluation
 5. Model Deployment and Monitoring
 6. Continuous Improvement





Data Acquisition and Quality Assessment

- Data is collected from multiple sources such as customer demographics, loan details, credit history, and financial indicators. The goal is to ensure the completeness, consistency, and accuracy of the data.

Workflow Diagram



Data Acquisition



Data Cleaning



Feature Engineering



Model Development



Deployment

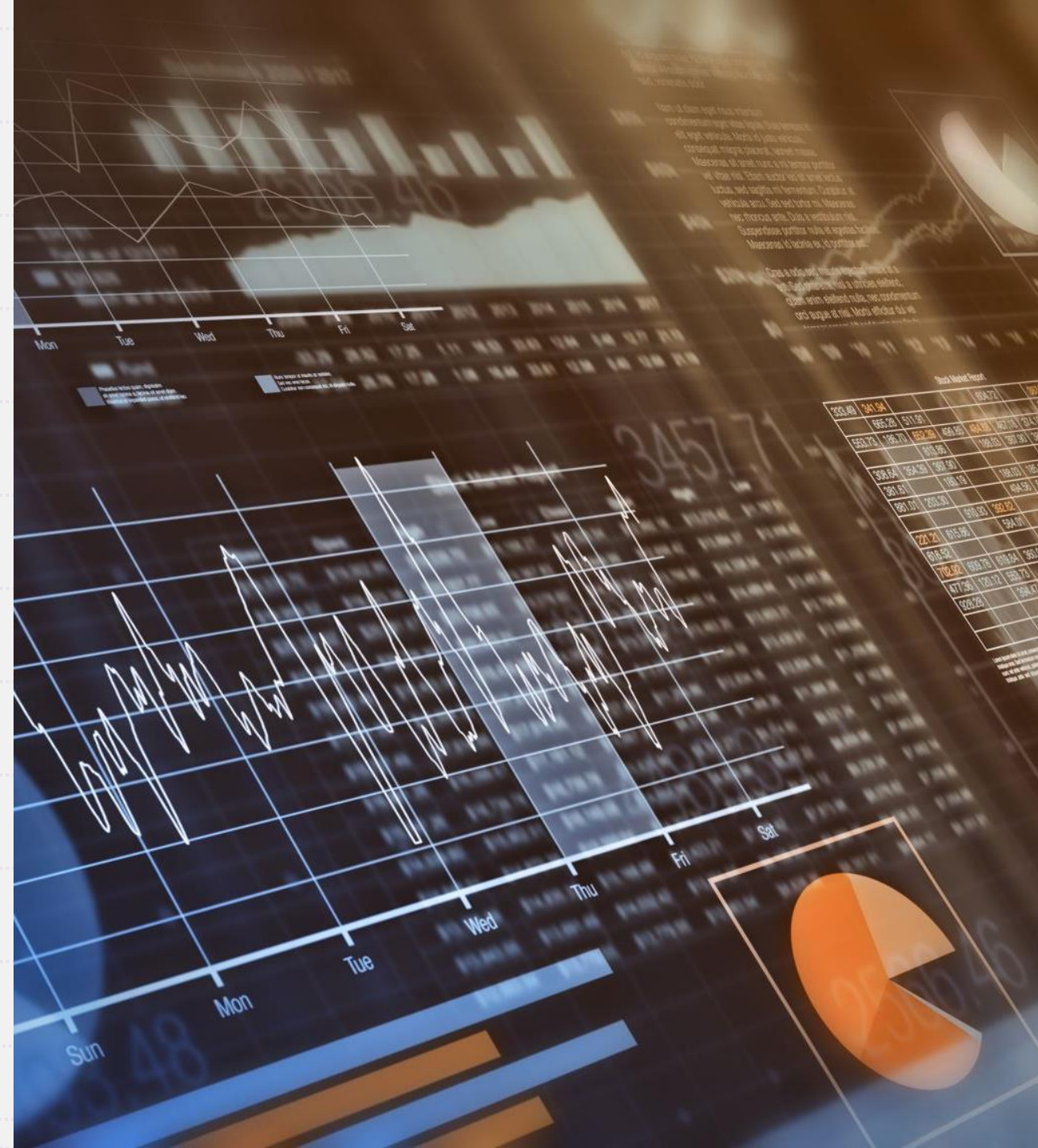
Data Cleaning and Preprocessing

- Raw data often contains errors and missing values. The data is processed to remove outliers, impute missing values, and scale numerical features for better model performance.



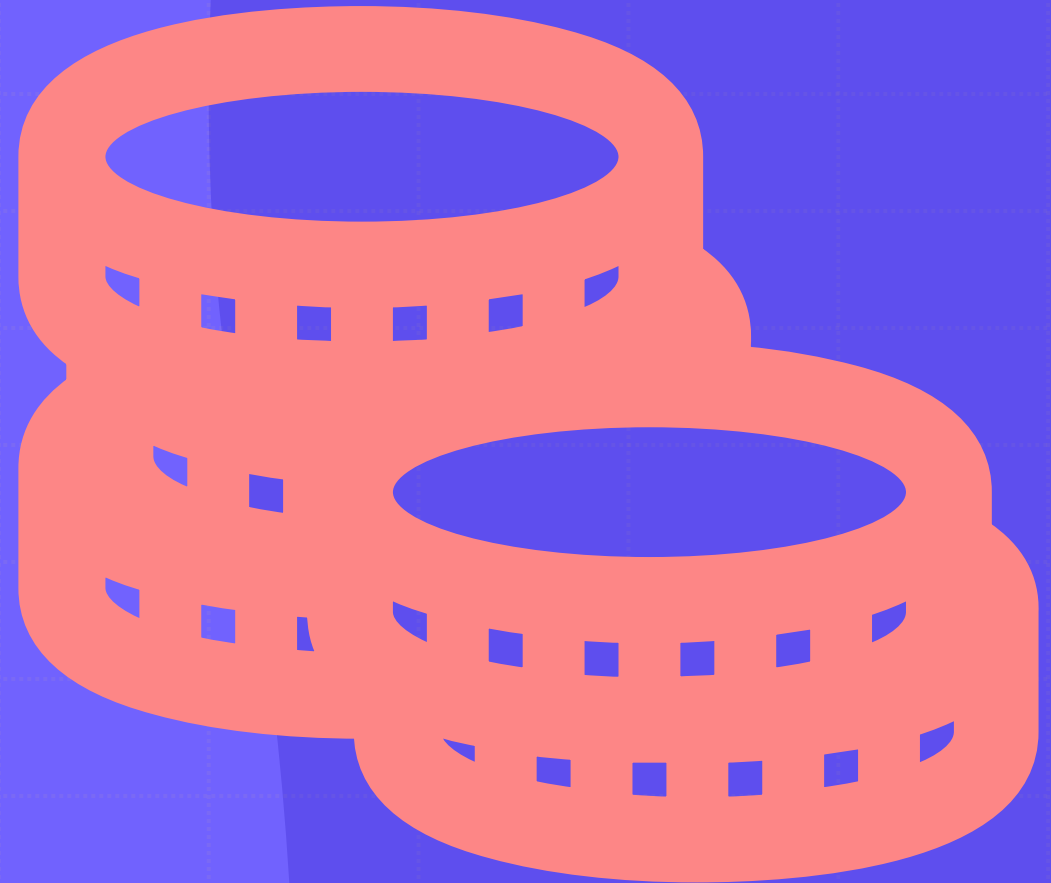
Explorative Data Analysis

- Insights will be taken from cleaned data to train the machine learning models. This form of data analysis provides us with the trends, hidden facts, and necessary information found in the data. Through EDA, we can improve the financial sector by addressing issues found in the dataset and get important information for better model performance.



Feature Engineering

- Feature engineering involves creating new features to improve model accuracy. This includes deriving debt-to-income ratios, interaction features, and temporal indicators.





Feature Engineering

- **Domain-Specific Features:**

- Debt-to-Income ratio
- Loan-to-Value ratio
- Credit utilization rate
- Repayment patterns and frequency

- **Categorical Encoding:** One-hot encoding, target encoding.

- **Scaling:** Min-max, standardization.

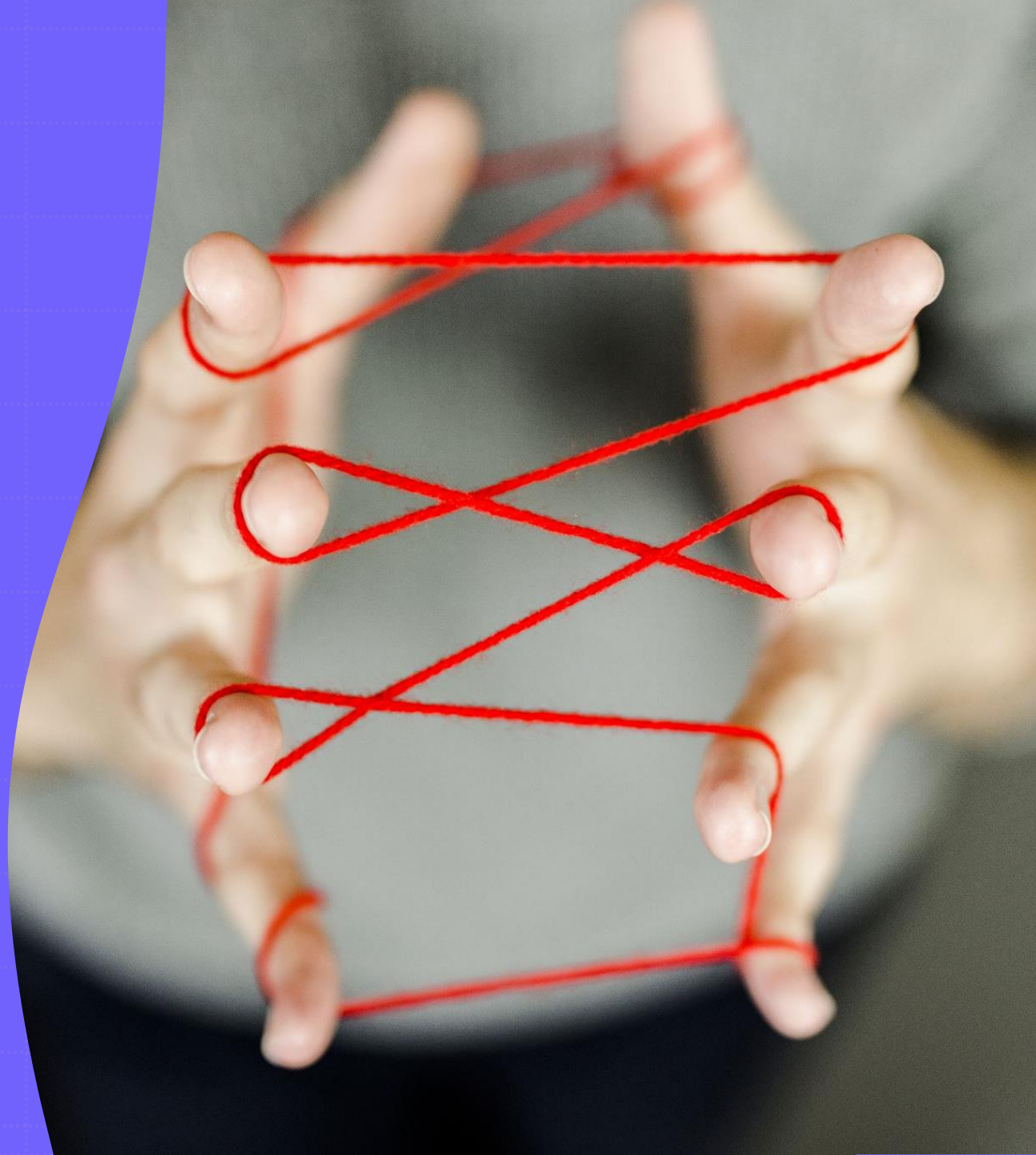


Model Development and Evaluation

- Various models are tested to determine the best-performing algorithm. Models like Logistic Regression, Decision Trees, and Gradient Boosting are evaluated based on accuracy, AUC-ROC, and precision-recall curves.

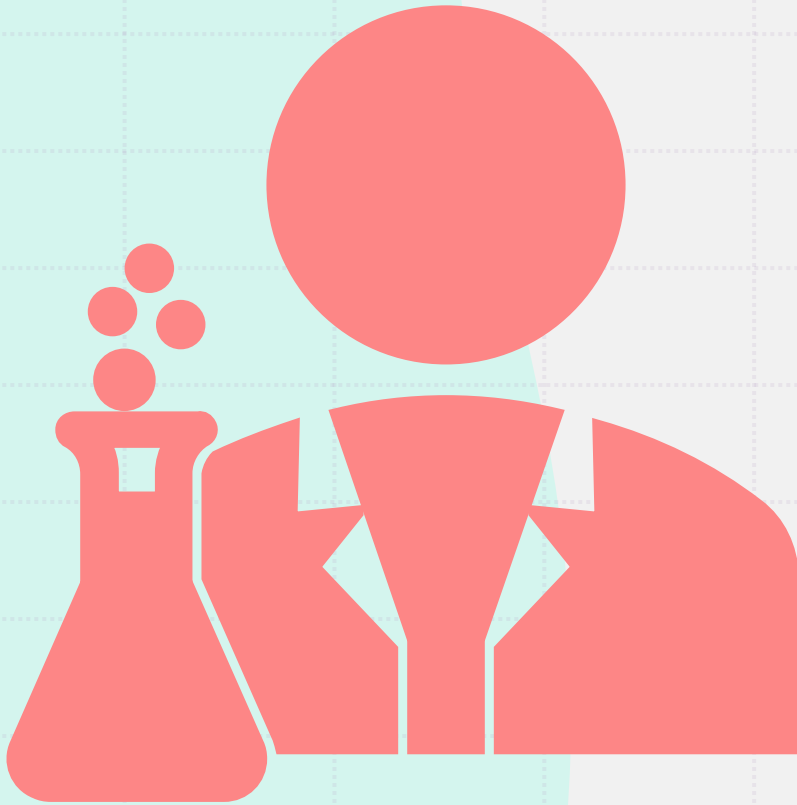
Validation and Testing

- **Train-Validation Split:** 70%-30% (stratified sampling).
- **Cross-Validation:** K-fold cross-validation (5 or 10 folds).
- **Out-of-Time Validation:** Evaluate performance on the most recent data to assess generalization.
- **Testing:** Perform A/B testing during deployment with live customer data.



Model Deployment and Monitoring

- The best model is deployed as an API using Flask or FastAPI, hosted on cloud platforms like AWS or GCP. Continuous monitoring ensures that the model performance remains stable over time.





Tools For Data Collection & Storage

Tools: SQL, PostgreSQL, MongoDB, AWS S3, Google BigQuery

Purpose: Store and manage large datasets from multiple sources.

Tools For Data Exploration & Preprocessing



Tools: Python (Pandas, NumPy, Scikit-learn), R, Jupyter Notebooks



Purpose: Perform EDA, clean data, handle missing values, and engineer features.



Tools For Development

- **c. Model Development:**

- **Tools:**

- **Machine Learning:** LightGBM, XGBoost, CatBoost, Scikit-learn
- **Deep Learning (if applicable):** TensorFlow, PyTorch

- **Purpose:** Build, train, and evaluate models.



Tools For Model Tuning & Validation

Tools: Optuna, Hyperopt, GridSearchCV, Bayesian Optimization

Purpose: Optimize model performance with hyperparameter tuning.

Tools for Performance Metrics & Visualization

Tools: Matplotlib, Seaborn, Plotly, SHAP (Explainability), LIME

Purpose: Visualize model performance and explain model decisions.

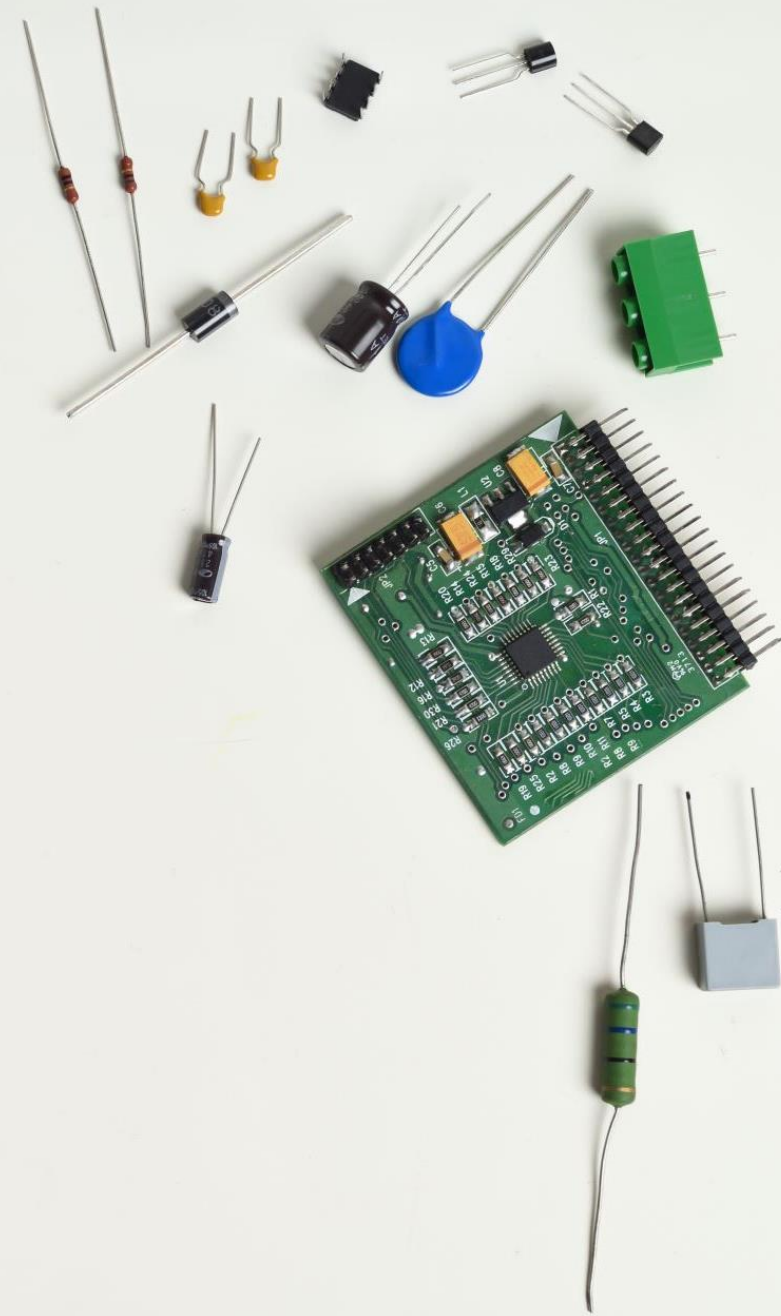
Tools for Model Packaging

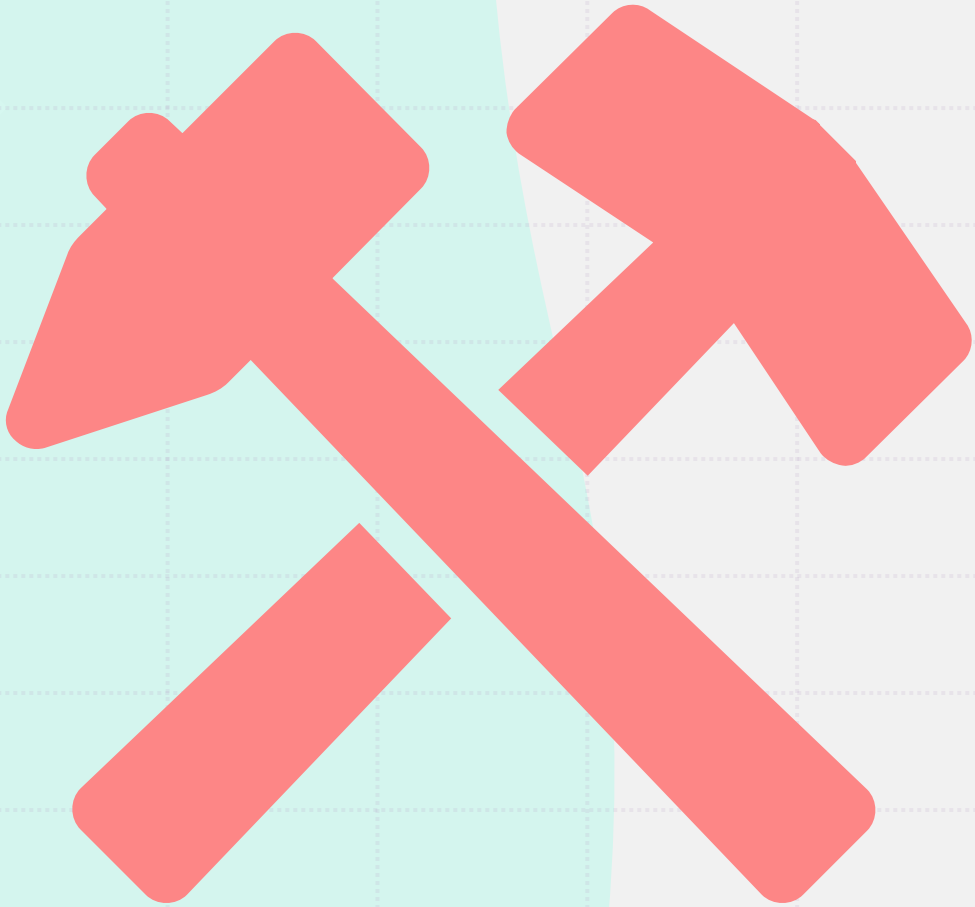
- Tools:** Docker, Conda, MLflow
- Purpose:** Containerize the model to ensure consistency across environments.



Tools for Model Serving (API):

- Tools:** FastAPI, Flask, TensorFlow Serving, TorchServe
- Purpose:** Serve the model as an API for real-time scoring.





Tools for Collaboration & Documentation

- **Tools:** Git (GitHub, GitLab), Confluence, Notion
- **Purpose:** Version control and document model development processes.

Conclusion

- This credit loan scoring project provides a structured approach to developing predictive models for assessing borrower risk. Financial institutions can use this approach to improve lending decisions and reduce defaults.

