Project Title	**EMIPredict AI - Intelligent Financial Risk Assessment Platform**
Skills take away From This Project	**Python, Streamlit Cloud Deployment, Machine Learning, Data Analysis, MLflow, Classification Models, Regression Models, Feature Engineering, Data Preprocessing**
Domain	**FinTech and Banking**

Problem Statement

Build a comprehensive financial risk assessment platform that integrates machine learning models with MLflow experiment tracking to create an interactive web application for EMI prediction.

Nowadays, people struggle to pay EMI due to poor financial planning and inadequate risk assessment. This project aims to solve this critical issue by providing data-driven insights for better loan decisions.

The platform should deliver:

- **Dual ML problem solving**: Classification (EMI eligibility) and Regression (maximum EMI amount)
- **Real-time financial risk assessment** using 400,000 records
- **Advanced feature engineering** from 22 financial and demographic variables

- **ML flow integration** for model tracking and comparison
- **Streamlit Cloud deployment** for production-ready access
- **Complete CRUD operations** for financial data management

Business Use Cases

Financial Institutions

- Automate loan approval processes and reduce manual underwriting time by 80%
- Implement risk-based pricing strategies for different EMI scenarios
- Real-time eligibility assessment for walk-in customers

FinTech Companies

- Instant EMI eligibility checks for digital lending platforms
- Integration with mobile apps for pre-qualification services
- Automated risk scoring for loan applications

Banks and Credit Agencies

- Data-driven loan amount recommendations based on financial capacity
- Portfolio risk management and default prediction
- Regulatory compliance through documented decision processes

Loan Officers and Underwriters

- AI-powered recommendations for loan approval decisions
- Comprehensive financial profile analysis in seconds
- Historical performance tracking and model accuracy monitoring

Approach

Step 1: Data Loading and Preprocessing

- Load the provided dataset of 400,000 realistic financial records across 5 EMI scenarios
- Implement comprehensive data cleaning for missing values, inconsistencies, and duplicates
- Apply data quality assessment and validation checks
- Create train-test-validation splits for model development

Step 2: Exploratory Data Analysis

- Analyze EMI eligibility distribution patterns across different lending scenarios
- Study correlation between financial variables and loan approval rates
- Investigate demographic patterns and risk factor relationships
- Generate comprehensive statistical summaries and business insights

Step 3: Feature Engineering

- Create derived financial ratios (debt-to-income, expense-to-income, affordability ratios)
- Generate risk scoring features based on credit history and employment stability
- Apply categorical encoding and numerical feature scaling
- Develop interaction features between key financial variables

Step 4: Machine Learning Model Development

Classification Models (EMI Eligibility Prediction) - Minimum 3 Models Required:

- **Logistic Regression** for baseline interpretable results
- **Random Forest Classifier** for feature importance analysis
- **XGBoost Classifier** for high-performance gradient boosting

- **Additional models**: Support Vector Classifier, Decision Tree, or Gradient Boosting Classifier
- Model evaluation using accuracy, precision, recall, F1-score, and ROC-AUC
- Best performing classification model will be selected for final deployment

Regression Models (Maximum EMI Amount Prediction) - Minimum 3 Models Required:

- **Linear Regression** for baseline performance
- **Random Forest Regressor** for ensemble-based predictions
- **XGBoost Regressor** for advanced gradient boosting
- **Additional models**: Support Vector Regressor, Decision Tree Regressor, or Gradient Boosting Regressor
- Model evaluation using RMSE, MAE, R-squared, and MAPE
- Best performing regression model will be selected for final deployment

Step 5: Model Selection and MLflow Integration

- Train minimum 3 models each for classification and regression problems
- Configure MLflow tracking server for organized experiment management
- Log comprehensive model parameters, hyperparameters, and performance metrics for all models
- Create systematic artifact storage for models, visualizations, and datasets
- Compare model performance using MLflow experiment tracking dashboard
- Select best performing models based on evaluation metrics for production deployment
- Implement model registry for version control and selected model storage

Step 6: Streamlit Application Development

- Multi-page web application with intuitive user interface
- Real-time prediction capabilities for both classification and regression

- Interactive data exploration and visualization components
- Model performance monitoring and MLflow integration dashboard
- Administrative interface for data management operations

Step 7: Cloud Deployment and Production

- Deploy complete application on Streamlit Cloud platform
- Implement responsive design for cross-platform accessibility
- Configure automated deployment pipeline from GitHub repository
- Ensure proper error handling and user feedback mechanisms

Data Flow and Architecture

Dataset (400K Records)

↓

Data Quality Assessment & Preprocessing

↓

Feature Engineering & Exploratory Analysis

↓

ML Model Training & MLflow Tracking

↓

Model Evaluation & Selection

↓

Streamlit Application Development

↓

Cloud Deployment & Performance Testing

↓

Production-Ready Financial Platform

Architecture Components:

- **Data Layer: Structured financial data following domain rules**
- **Processing Layer**: Data cleaning, feature engineering, and ML pipelines
- **Model Layer**: Classification and regression models with MLflow experiment tracking
- **Application Layer**: Multi-page Streamlit web application with real-time predictions
- **Deployment Layer**: Streamlit Cloud hosting with GitHub integration and CI/CD

Dataset: EMI_dataset

Dataset Scale:

- **Total Records**: 400,000 financial profiles
- **Input Features**: 22 comprehensive variables
- **Target Variables**: 2 (Classification + Regression)
- **EMI Scenarios**: 5 lending categories with realistic distributions

EMI Scenario Distribution:

- **E-commerce Shopping EMI** (80,000 records) - Amount: 10K-200K, Tenure: 3-24 months
- **Home Appliances EMI** (80,000 records) - Amount: 20K-300K, Tenure: 6-36 months
- **Vehicle EMI** (80,000 records) - Amount: 80K-1500K, Tenure: 12-84 months
- **Personal Loan EMI** (80,000 records) - Amount: 50K-1000K, Tenure: 12-60 months
- **Education EMI** (80,000 records) - Amount: 50K-500K, Tenure: 6-48 months

Dataset Explanation

Input Features (22 Variables):

Personal Demographics:

- **age**: Customer age (25-60 years)
- **gender**: Gender classification (Male/Female)
- **marital_status**: Marital status (Single/Married)
- **education**: Educational qualification (High School/Graduate/Post Graduate/Professional)

Employment and Income:

- **monthly_salary**: Monthly gross salary (15K-200K INR)
- **employment_type**: Employment category (Private/Government/Self-employed)
- **years_of_employment**: Work experience duration
- **company_type**: Organization size and type

Housing and Family:

- **house_type**: Residential ownership status (Rented/Own/Family)
- **monthly_rent**: Monthly rental expenses
- **family_size**: Total household members
- **dependents**: Number of financial dependents

Monthly Financial Obligations:

- **school_fees**: Educational expenses for dependents
- **college_fees**: Higher education costs
- **travel_expenses**: Monthly transportation costs
- **groceries_utilities**: Essential living expenses
- **other_monthly_expenses**: Miscellaneous financial obligations

Financial Status and Credit History:

- **existing_loans**: Current loan obligations status
- **current_emi_amount**: Existing monthly EMI burden
- **credit_score**: Credit worthiness score (300-850)
- **bank_balance**: Current account balance
- **emergency_fund**: Available emergency savings

Loan Application Details:

- **emi_scenario**: Type of EMI application (5 categories)
- **requested_amount**: Desired loan amount
- **requested_tenure**: Preferred repayment period in months

Target Variables:

Classification Target:

- **emi_eligibility**: Primary classification target with 3 classes
    - **Eligible**: Low risk, comfortable EMI affordability
    - **High_Risk**: Marginal case, requires higher interest rates
    - **Not_Eligible**: High risk, loan not recommended

Regression Target:

- **max_monthly_emi**: Primary regression target
    - Continuous variable representing maximum safe monthly EMI amount (500-50000 INR)
    - Calculated using comprehensive financial capacity analysis

Expected Results

Technical Deliverables:

- Successfully process and analyze 400,000 financial records with comprehensive quality assessment

- Achieve classification accuracy above 90% and regression RMSE below 2000 INR
- Complete MLflow integration with organized experiment tracking and model registry
- Deploy fully functional Streamlit Cloud application with real-time prediction capabilities

Business Impact:

- Automated financial risk assessment reducing manual processing time by 80%
- Standardized loan eligibility criteria across different EMI scenarios
- Data-driven decision making framework for financial institutions
- Scalable platform architecture supporting high-volume loan applications

Project Evaluation Metrics

Technical Performance Evaluation (70%):

- Data preprocessing completeness and quality assessment accuracy (15%)
- Machine learning model development: minimum 3 models each for classification and regression (25%)
- Best model selection process and performance justification (15%)
- ML flow integration effectiveness: experiment tracking for all models and model registry usage (15%)

Application Development and Deployment (30%):

- Streamlit application functionality using best selected models for real-time predictions (20%)
- Cloud deployment stability, performance optimization, and accessibility (10%)

Technical Tags

Python, **Data Preprocessing**, **Feature Engineering**, **Machine Learning**, **Classification**, **Regression**, **MLflow**, **Streamlit Cloud**, **FinTech**, **Risk Assessment**, **XGBoost**, **Random Forest**, **Financial Analytics**, **Big Data Processing**, **Model Registry**, **Experiment Tracking**

Deliverables

Data Processing and Analysis Scripts:

- Data preprocessing and cleaning pipeline for the provided 400K dataset
- Feature engineering and transformation modules
- Exploratory data analysis scripts and visualizations

Machine Learning Models and Analysis:

- Minimum 3 trained classification models for EMI eligibility prediction with performance comparison
- Minimum 3 trained regression models for maximum EMI amount calculation with evaluation metrics
- Best model selection process and justification based on performance metrics
- Model performance evaluation and comparison reports across all trained models
- MLflow experiment tracking and model registry implementation with all model variants

Web Application and Deployment:

- Multi-page Streamlit application with interactive user interface
- Real-time prediction capabilities for both classification and regression
- Cloud deployment on Streamlit Cloud platform with public URL access
- GitHub repository with complete project codebase and documentation

Documentation and Reports:

- Comprehensive technical documentation covering methodology and architecture
- Exploratory data analysis report with business insights and visualizations
- Model performance analysis and MLflow experiment comparison
- Business impact assessment and recommendations for financial institutions

Timeline

14 Days

References:

TOPIC	LINK
Project Live Evaluation	Project Live Evaluation
EDA Guide	Exploratory Data Analysis (EDA) G…
Capstone Explanation Guideline	Capstone Explanation Guideline
GitHub Reference	How to Use GitHub.pptx
Project Orientation (English)	Project Orientation Session_EMIPr…

Project Orientation (Tamil)	Project Orientation Session_EMI_P…
STREAMLIT RECORDING (English)	Special session for STREAMLIT(11/…
STREAMLIT DOCUMENTATION	Install Streamlit
ML FLOW Tutorial 1	ML FLOW 1
ML FLOW Tutorial 2	ML FLOW 2
MLflow DOCUMENTATION:	Getting Started with MLflow
Project Excellence Series [[Machine learning] (English)	**Project Excellence Series: Guided L..**
Project Excellence Series [[Machine learning] (Tamil)	**Project Excellence Series: Guided L..**
Project Excellence Series [Machine learning-Unsupervised learning] (English)	**Project Excellence Series: Guided L..**

Project Excellence Series [Machine learning-Unsupervised learning] (Tamil)	**Project Excellence Series: Guided L..**
Project Excellence Series [EDA] (English)	**Project Excellence Series: Guided L..**
Project Excellence Series [EDA] (Tamil)	**Project Excellence Series: Guided L..**