Loan Approval Prediction Using Machine Learning

A Comparative and Interpretable Modeling Approach

Presented by: Olatunji Ahmed

Introduction

Loan approval decisions represent one of the most critical processes in financial institutions, directly impacting both business profitability and customer satisfaction. Traditional manual review processes can be subjective and time-consuming, often leading to inconsistent decisions.

This project leverages **machine learning techniques** to predict loan approval outcomes using comprehensive applicant data including demographic information, financial profiles, and credit history. By applying advanced analytics and multiple modeling approaches, we aim to create a system that is both accurate and interpretable.

Primary Goal: To build accurate, interpretable, and robust predictive models that can support fair and efficient loan approval decisions while maintaining transparency for regulatory compliance.

Dataset Overview

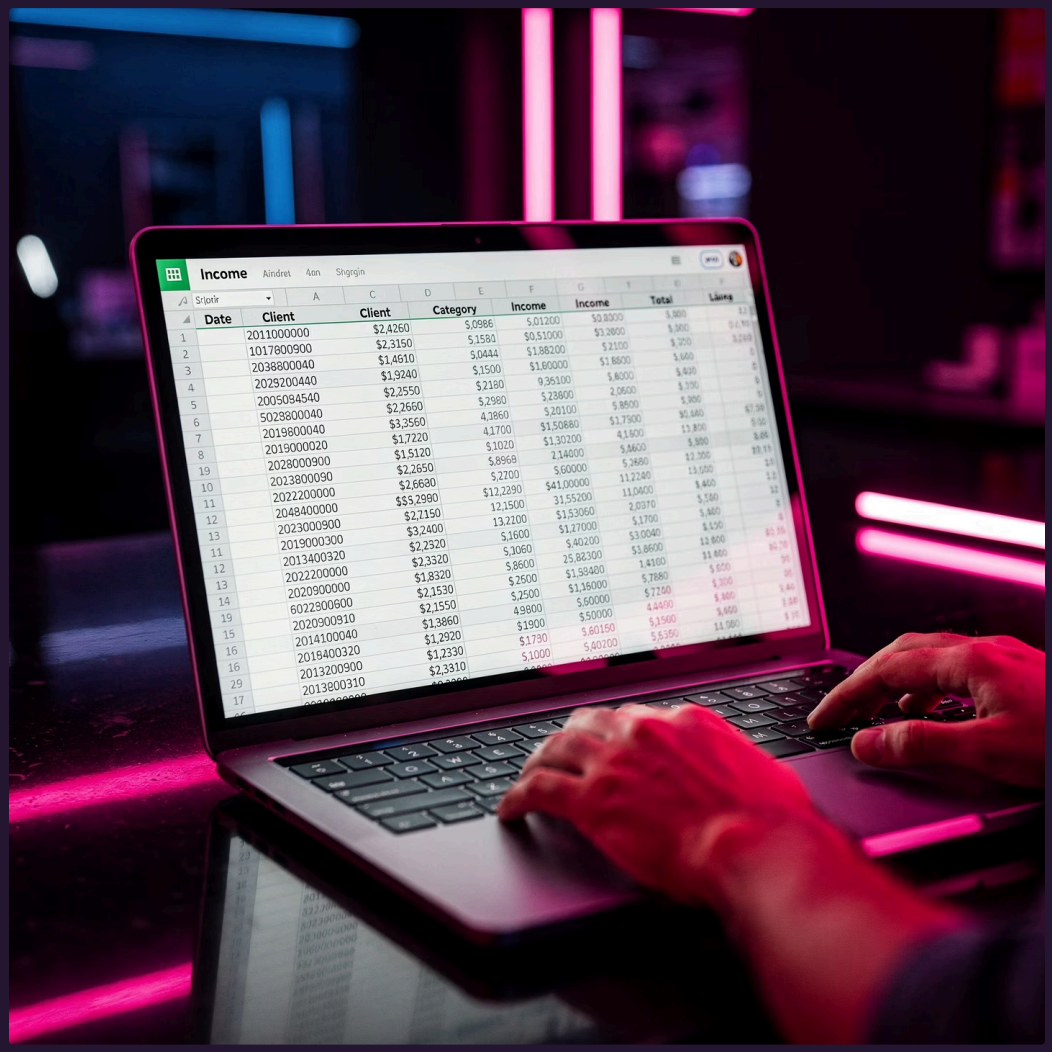
614	68.7%	31.3%
Total Samples	**Approved**	**Rejected**
Complete loan applications analyzed	Successful loan applications	Unsuccessful applications

The dataset contains 614 complete loan application records with the binary target variable **Loan Status** (Approved/Rejected). The class distribution shows moderate imbalance, with approximately two-thirds of applications approved.

Important Consideration: This moderate class imbalance was carefully addressed during model development through stratified sampling, appropriate evaluation metrics, and balanced training techniques to ensure fair predictions for both classes.

Features Used in Prediction

Numerical Variables



- **Applicant Income:** Primary borrower's income
- **Coapplicant Income:** Secondary borrower's income
- **Loan Amount:** Requested loan value
- **Loan Amount Term:** Repayment period in months

Categorical Variables

- **Gender:** Male/Female
- **Marital Status:** Married/Single
- **Dependents:** Number of dependents
- **Education:** Graduate/Not Graduate
- **Self-Employed:** Employment type
- **Property Area:** Urban/Semiurban/Rural

These features capture the comprehensive profile of each loan applicant, combining financial capacity indicators with demographic and situational factors that may influence creditworthiness and repayment ability.

Exploratory Data Analysis

Distribution Analysis

Comprehensive distribution analysis conducted for all numerical variables to understand data characteristics and identify potential issues.

Visualization Approach

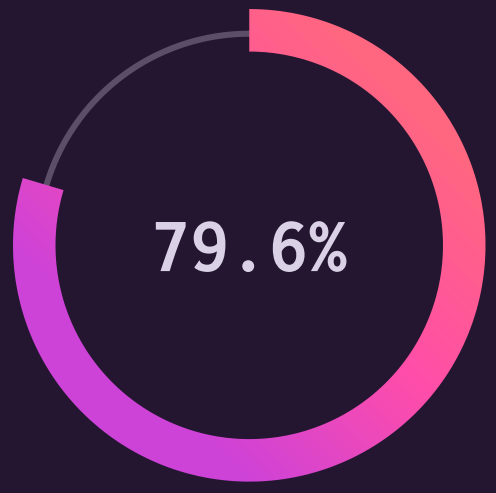
Kernel Density Estimates (KDE) employed to observe distribution shapes, detect multimodality, and assess normality assumptions.

Central Tendency

Median values highlighted throughout analysis to detect skewness, identify outliers, and understand typical applicant profiles.

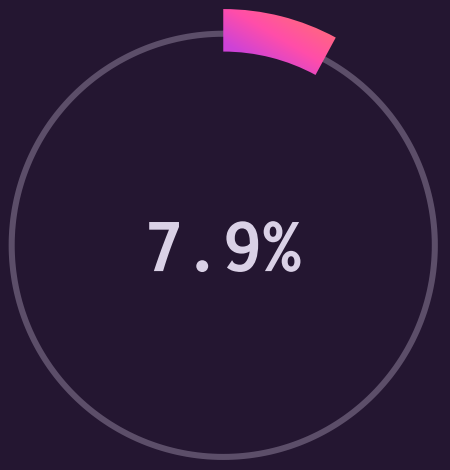
Key Observation: Most income and loan amount variables exhibited **right-skewed** distributions, indicating that a smaller number of high-value outliers pull the mean above the median. This finding justified the use of robust scaling techniques rather than standard normalization to prevent outliers from dominating the model training process.

Credit History: The Dominant Factor



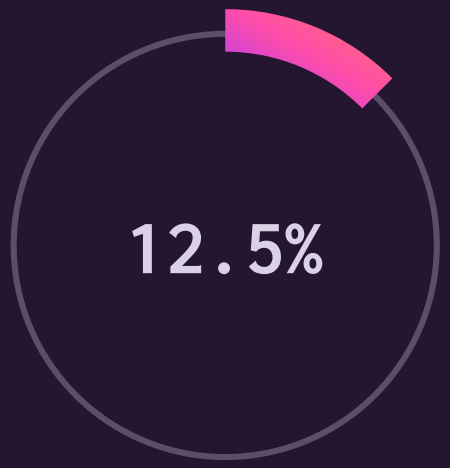
Good Credit History

Applicants with positive credit records



Poor Credit History

Applicants with negative credit records



Missing Data

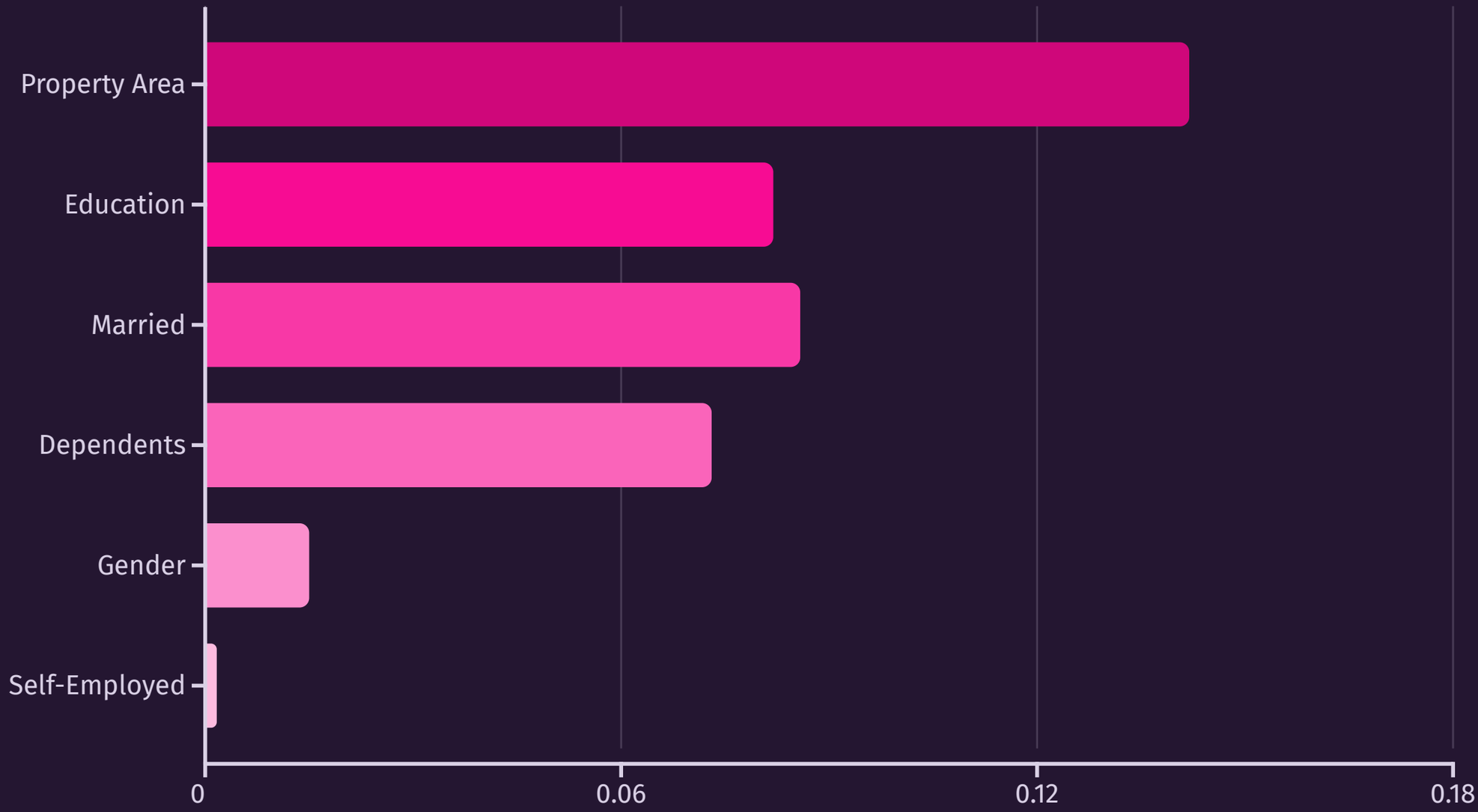
Records requiring imputation

🔑 Critical Insight

Credit History emerged as the single most important predictor of loan approval across all modeling approaches. Applicants with good credit history showed dramatically higher approval rates, underscoring the fundamental importance of past financial behavior in lending decisions.

Categorical Feature Relationships

Cramér's V Association Strength with Loan Approval



Cramér's V measures the strength of association between categorical variables, ranging from 0 (no association) to 1 (perfect association). The analysis reveals that **Property Area** shows the strongest categorical association with loan approval at 0.142 (moderate strength), suggesting that geographic location and property market characteristics play a meaningful role in approval decisions.

Most other categorical features show weak associations, with Gender and Self-Employment status having minimal predictive value on their own.

Missing Data Handling Strategy

Missing Value Distribution

8.1%

Credit History

5.2%

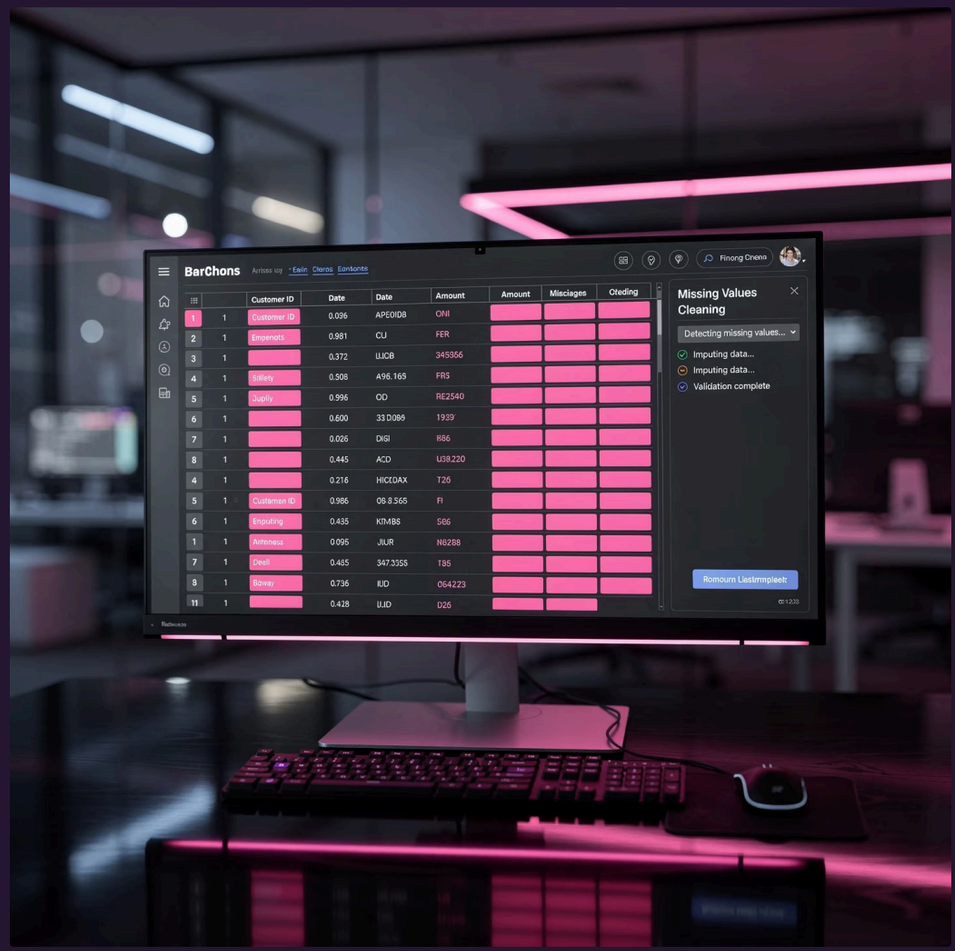
Self-Employed

3.6%

Loan Amount

<3%

Other Variables



1

Numerical Features

Median imputation used to preserve robust central tendency without influence from outliers

2

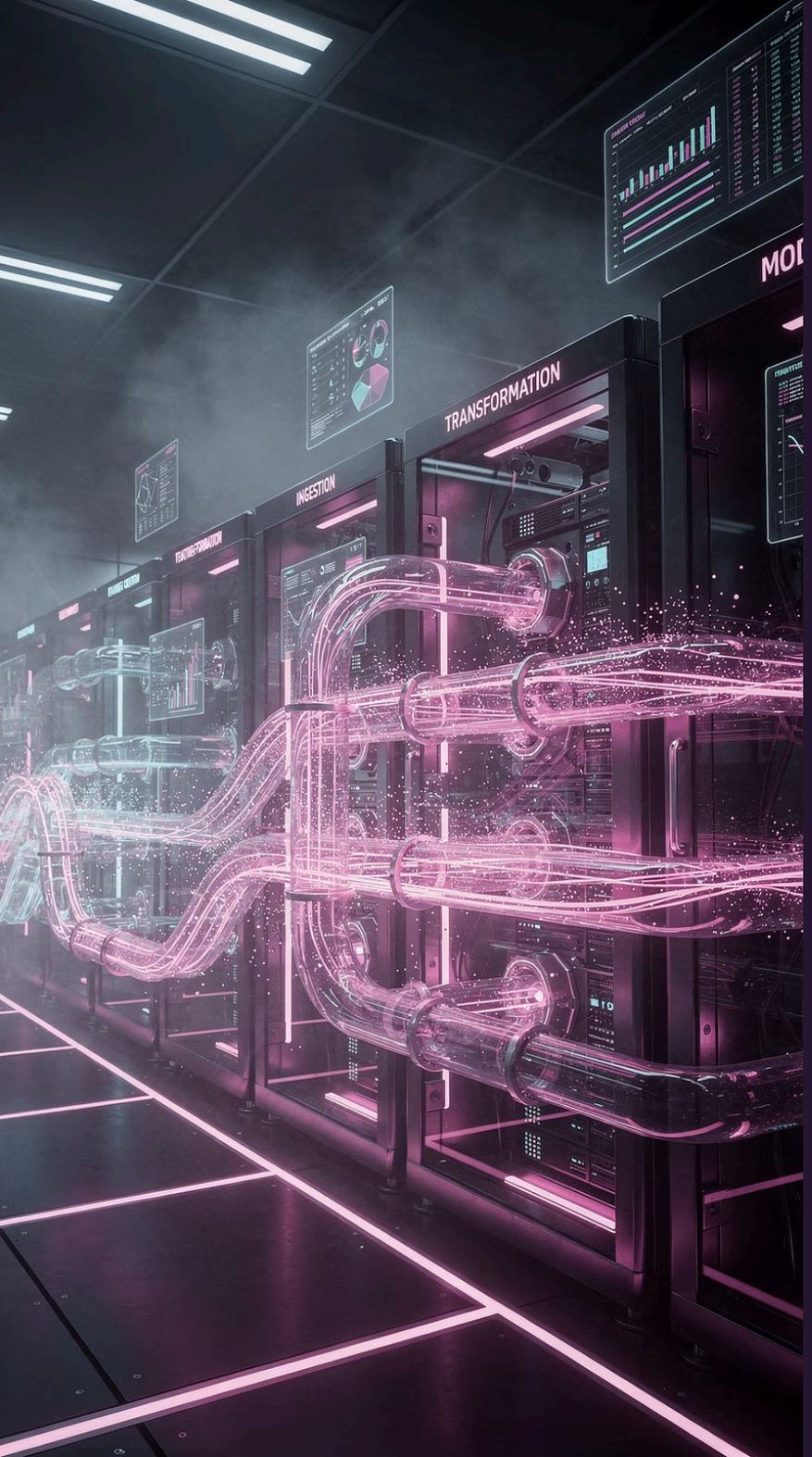
Categorical Features

Mode imputation applied to fill missing values with most frequent category

3

Validation

Zero missing values confirmed after imputation process completion



Feature Engineering

To enhance predictive power beyond raw input variables, we created sophisticated derived features that capture complex relationships and domain-specific insights about loan applicants.

Financial Ratios

- Total Income (applicant + coapplicant)
- Income per Dependent
- Loan-to-Income Ratio
- Debt-to-Income Ratio

Interaction Terms

- Credit History × Income
- Credit History × Loan Amount
- Property Area × Income Level
- Education × Employment Type

Derived Attributes

- Family Size (dependents + married)
- Binary Encodings
- One-Hot Encodings
- Bucketed Variables

36

Total Features

After engineering process

21

Final Selected

Used in modeling after selection

Statistical Significance Testing

Chi-Square Tests (Categorical)

Marital Status

Significant (p = 0.034)

Education Level

Significant (p = 0.043)

Other Categories

Not statistically significant

T-Tests (Numerical)

Independent sample t-tests were conducted to compare mean values of numerical features between approved and rejected loan applications.

Result: All numerical variables showed **no statistical significance** at conventional levels (p > 0.05) when tested individually, suggesting that their predictive power emerges through complex interactions rather than simple mean differences.

Modeling Setup and Approach

01

Feature Selection

Final set of 21 features selected based on importance, correlation analysis, and domain knowledge

02

Data Split

Train-Test split with 491 training samples and 123 test samples (80/20 ratio with stratification)

03

Scaling

RobustScaler applied to numerical features to handle outliers and maintain scale consistency

04

Model Training

Multiple algorithms trained and evaluated for comparison

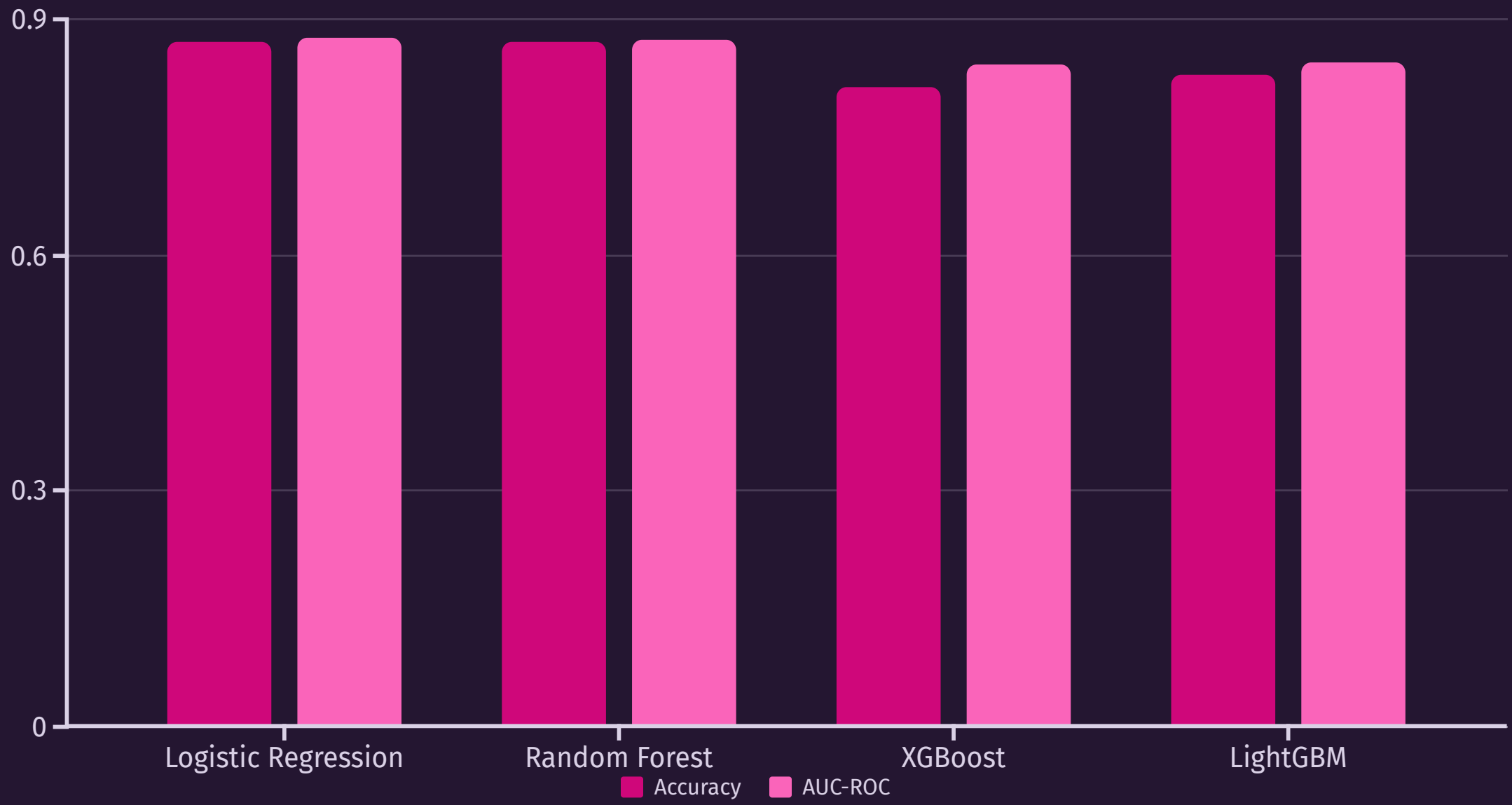
Models Trained

- **Logistic Regression**
- **K-Nearest Neighbors (KNN)**
- **Decision Tree**

- **Random Forest**
- **XGBoost**
- **LightGBM**
- CatBoost

Baseline Model Performance

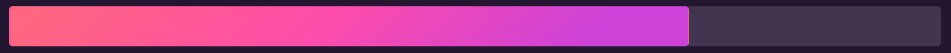
Initial model training without hyperparameter tuning established performance benchmarks and identified promising algorithms for further optimization.



📌 **Notable Finding:** Logistic Regression demonstrated remarkably strong performance even before hyperparameter tuning, achieving 87% accuracy and an AUC of 0.876. This suggests that the feature engineering and data preprocessing steps created a relatively linear decision boundary, allowing this simpler model to compete effectively with more complex ensemble methods.

Cross-Validation Results

5-Fold Cross-Validation AUC-ROC Scores



73.0%

Logistic Regression

± 9.3% standard deviation

76.0%

Random Forest

± 6.3% standard deviation

74.8%

XGBoost

± 7.6% standard deviation

75.5%

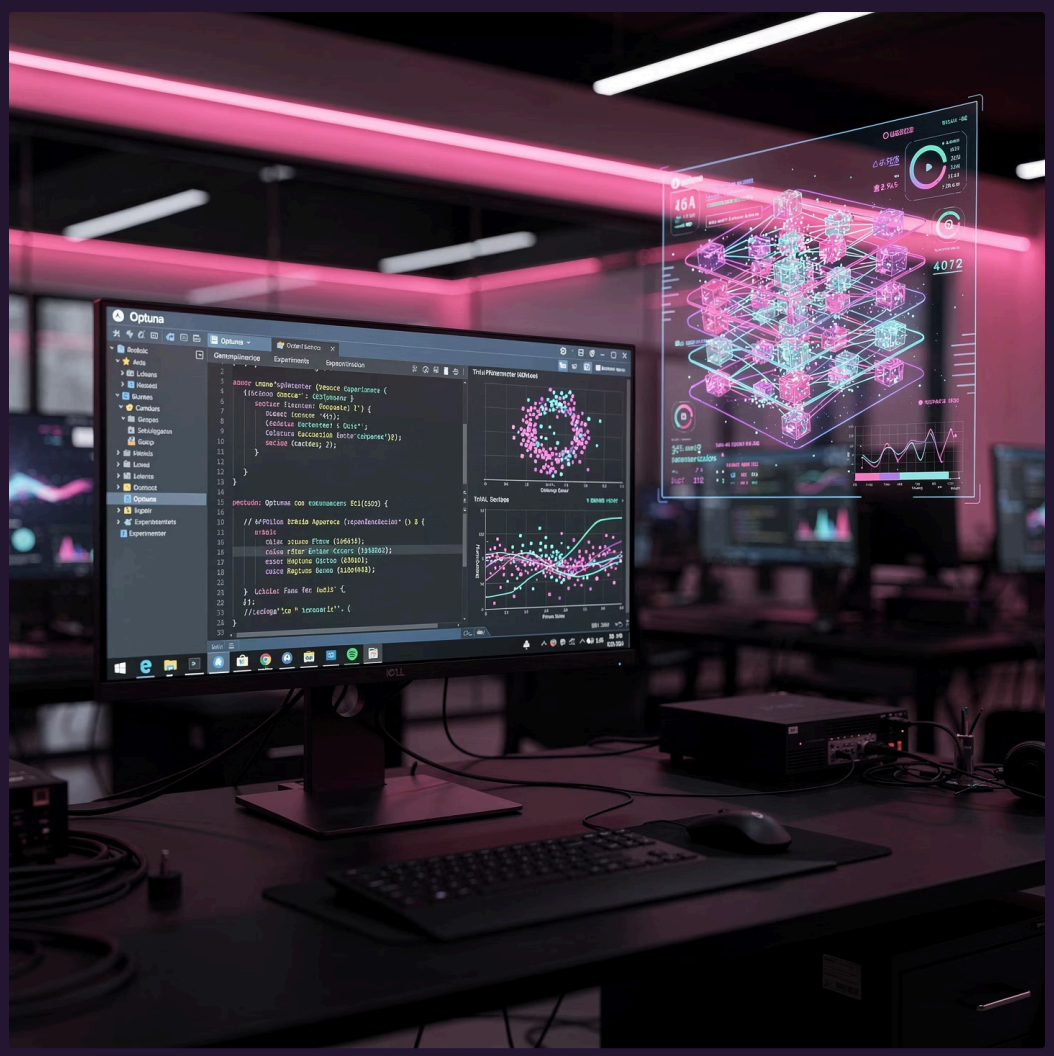
LightGBM

± 6.5% standard deviation

Cross-validation provides a more robust assessment of model generalization by evaluating performance across multiple train-test splits. The results reveal that **ensemble models** (Random Forest, XGBoost, LightGBM) demonstrate improved generalization with higher mean AUC scores and relatively lower variance compared to the baseline logistic regression.

The lower standard deviations for ensemble methods indicate more consistent performance across different data subsets, suggesting better stability and reliability for production deployment.

Hyperparameter Optimization



Systematic hyperparameter tuning was conducted using **Optuna**, a state-of-the-art optimization framework that employs sophisticated algorithms to efficiently search the hyperparameter space.

50

Optimization Trials

Per model

AUC

Target Metric

ROC-AUC optimized

Models Optimized

- Logistic Regression
- Random Forest
- XGBoost
- LightGBM
- CatBoost

Optuna's Tree-structured Parzen Estimator (TPE) algorithm intelligently explored the hyperparameter space, focusing computational resources on promising regions while maintaining diversity to avoid local optima.

Tuned Model Performance Comparison

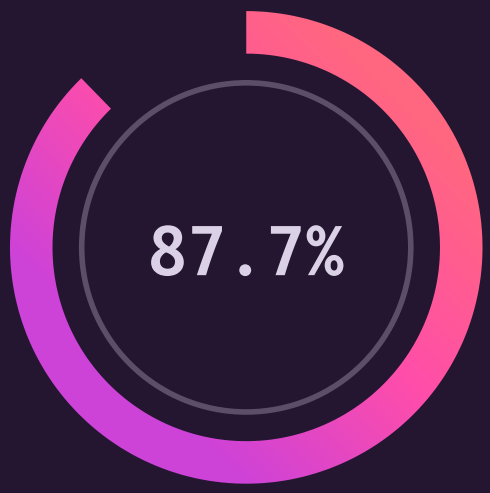
Rank	Model	Accuracy	F1-Score	AUC-ROC
🥇	**XGBoost (Tuned)**	**0.878**	0.916	**0.877**
🥈	Logistic Regression (Tuned)	0.870	0.913	0.876
🥉	Random Forest (Tuned)	**0.894**	**0.926**	0.874
4	LightGBM (Tuned)	0.862	0.908	0.869
5	CatBoost (Tuned)	0.854	0.902	0.865

After hyperparameter optimization, all models showed improved performance. The rankings reveal interesting trade-offs: while Random Forest achieved the highest accuracy (89.4%) and F1-score (0.926), **XGBoost secured the highest AUC-ROC** (0.877), indicating superior discrimination ability across all classification thresholds.

Best Model Selection

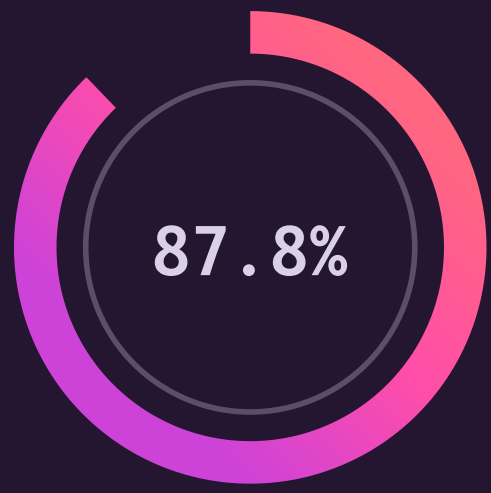
XGBoost (Tuned)

Selected as Best Overall Model



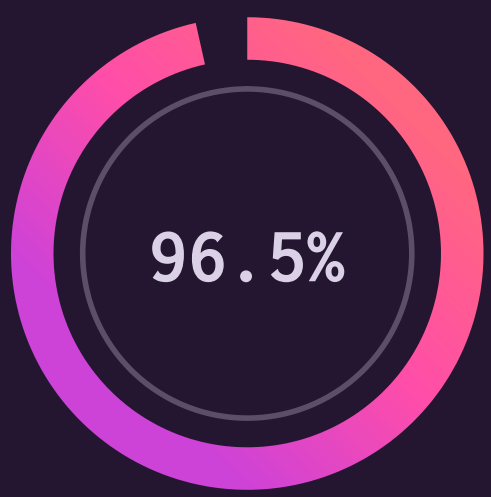
AUC-ROC Score

Best discrimination ability



Accuracy

Correct predictions



Recall

Approved loans captured

Why XGBoost? This model provides the optimal balance between **high predictive power** and **robust generalization**. The AUC-ROC metric was prioritized because it measures performance across all classification thresholds, making it ideal for financial applications where the decision threshold may need adjustment based on risk appetite and business objectives. The exceptional recall rate minimizes false negatives, reducing the risk of rejecting creditworthy applicants.

Model Interpretability with SHAP

While XGBoost delivers excellent predictive performance, understanding *why* the model makes specific predictions is crucial for financial applications, regulatory compliance, and building stakeholder trust.

We employed **SHAP (SHapley Additive exPlanations)**, a cutting-edge interpretability framework based on game theory, to explain both individual predictions and global feature importance.

Individual Explanations

Explain why specific applicants were approved or rejected

Global Importance

Identify which features matter most across all predictions

Trust & Transparency

Build confidence through explainable AI

SHAP values quantify each feature's contribution to pushing a prediction away from the baseline average, enabling clear, actionable insights for loan officers and compliance teams.

Key Findings and Insights

1

Credit History Dominance

Credit history emerged as the overwhelmingly dominant predictor of loan approval, confirming that past financial behavior remains the strongest indicator of future repayment probability.

2

Feature Engineering Impact

Sophisticated feature engineering, including financial ratios and interaction terms, significantly improved model performance beyond raw input variables.

3

Ensemble Model Advantage

Ensemble methods (particularly XGBoost and Random Forest) benefited most from hyperparameter tuning, achieving substantial performance gains.

4

Logistic Regression Resilience

Despite its simplicity, Logistic Regression remained surprisingly robust and competitive, demonstrating that well-engineered features can enable simpler models to perform effectively.

5

Balanced Performance

XGBoost provided the optimal balance of accuracy, AUC-ROC, and generalization, making it the ideal choice for production deployment in loan approval systems.

Conclusion and Business Impact

Project Demonstrates:

Reliable Prediction

Machine learning can accurately and consistently predict loan approval outcomes with 87.7% AUC-ROC

Superior Performance

Advanced ML models substantially outperform simple heuristics and rule-based systems

Essential Interpretability

Explainability tools like SHAP are critical for financial applications requiring transparency and regulatory compliance

Strategic Value

📌 This framework can support **fairer, more efficient, and data-driven lending decisions**, reducing bias, accelerating approval processes, and optimizing risk management while maintaining transparency for stakeholders and regulators.

Thank You

Questions & Discussion

Presented by: Olatunji Ahmed

This comprehensive analysis demonstrates the power of combining rigorous statistical methods, advanced machine learning techniques, and interpretability frameworks to solve real-world financial challenges. I welcome your questions and look forward to discussing potential applications and extensions of this work.