ML-Driven Loan Approval Prediction

A Data Science Approach

A supervised machine learning binary classification engine designed to predict borrower borrower solvency and provide data-driven repayment probabilities

98%
Precision & Recall

0.9997
AUC Score

XGBoost
Best Algorithm

Problem and Data Overview

The Business Problem

- Traditional lending workflows suffer from **operational friction** due to manual reviews
- Financial losses from **miscalculated risks** (Type I and Type II errors)

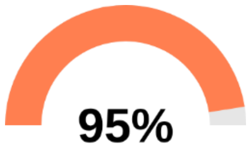
The ML Solution

- **Supervised learning** for borrower solvency prediction
- **Binary classification** to provide data-driven repayment probabilities

Dataset Overview

- **20,000 samples** from Kaggle
- Synthetic financial dataset covering **demographic, employment, and credit-specific domains**
- Simulates a master customer file for comprehensive analysis

Success Metrics

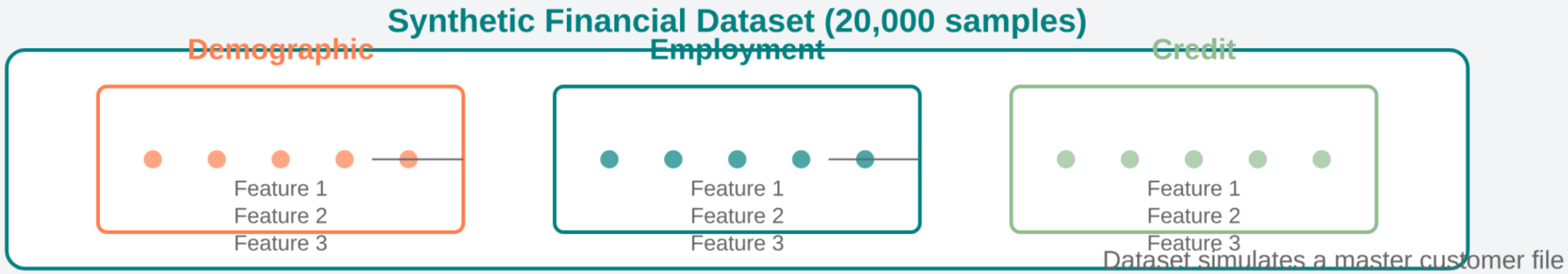


- **Precision:** >95%
- **Recall:** >95%
- **AUC:** >95%

Dataset Characteristics

Synthetic Financial Dataset

20,000 samples designed to simulate real-world lending data for analysis and model training



Demographic Domain

- Age groups and life stages
- Geographic distribution
- Family size and composition

Employment Domain

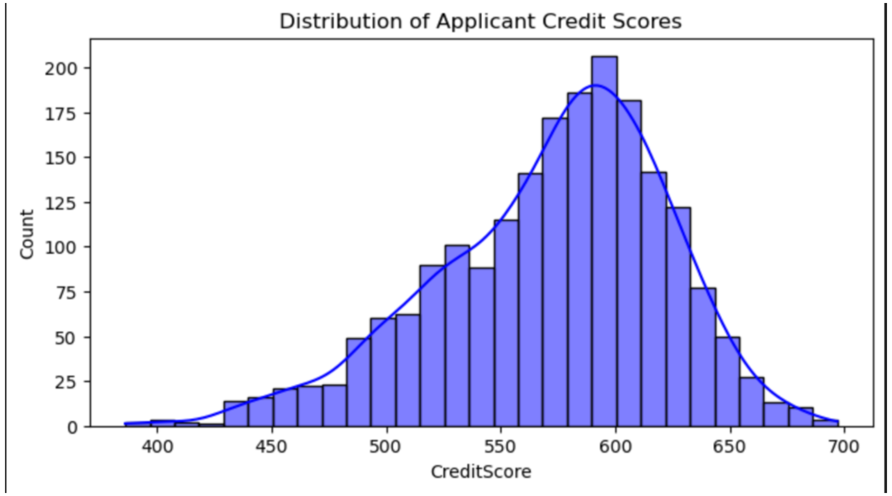
- Occupation categories
- Industry sector
- Years employed

Credit Domain

- Credit score ranges
- Debt-to-income ratio
- Credit history length

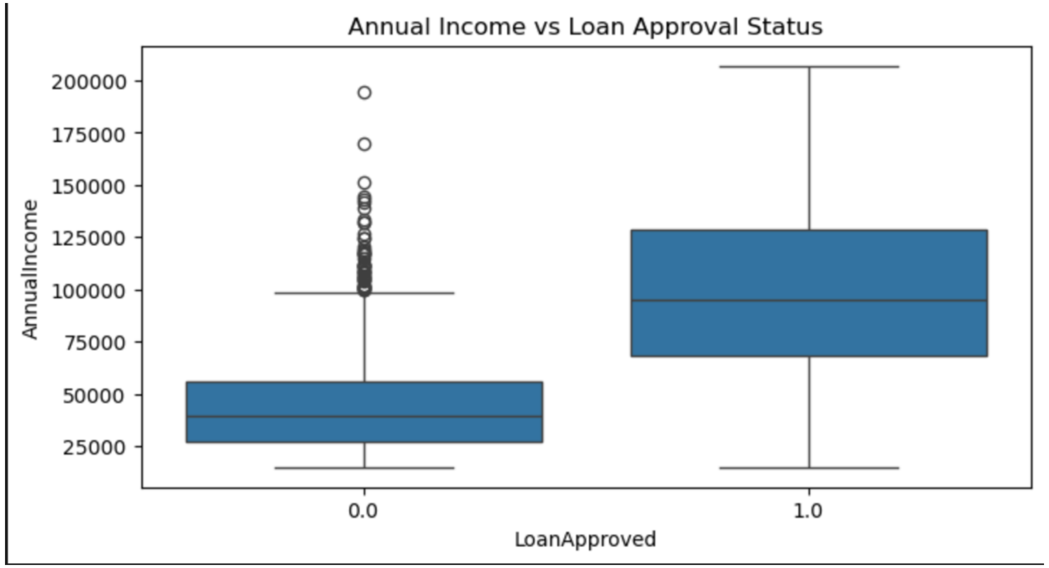
EDA Key Findings - Class Distribution & Correlations

Credit Scores Distribution Analysis



Key Insight: Applicant Credit Scores are approximately normally distributed, centered around the high-500s, with limited extreme values. The applicant population is predominantly moderate risk, making credit score thresholds and banding decisions particularly influential on loan approval outcomes.

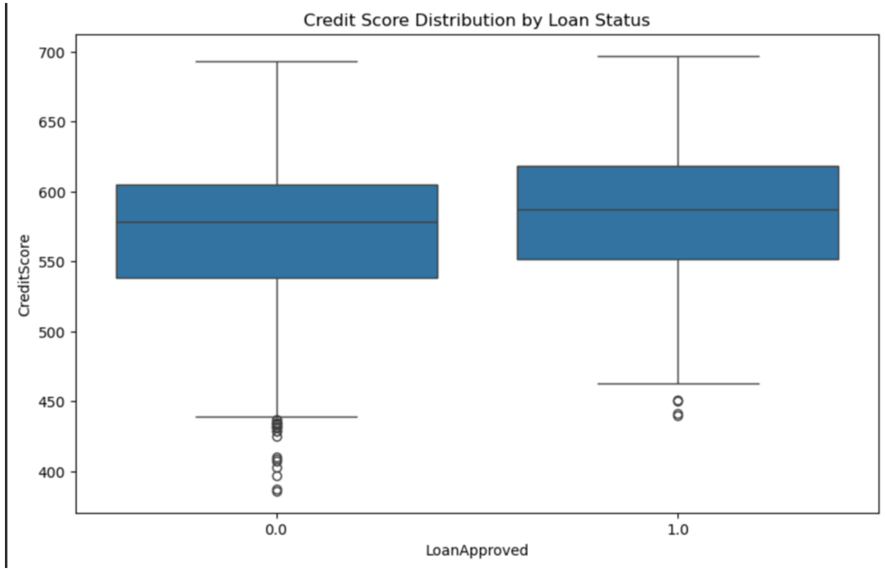
Credit Correlation Findings



Key Insight: Applicants with approved loans tend to have higher Annual Income than rejected applicants.

EDA Key Findings - Class Distribution & Correlations

Credit Correlation Findings



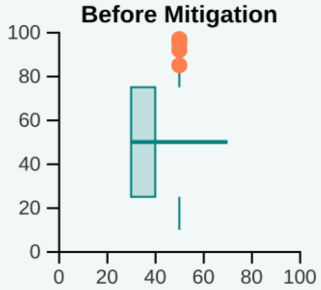
Key Insight: Applicants with approved loans tend to have higher and more stable credit scores than rejected applicants. The visible shift in medians and the concentration of low-score outliers among rejections confirm that credit score plays a decisive role in loan approval outcomes.

EDA Key Findings - Outlier Detection

Initial inspection identified significant outliers in key financial features that required mitigation to prevent model distortion.

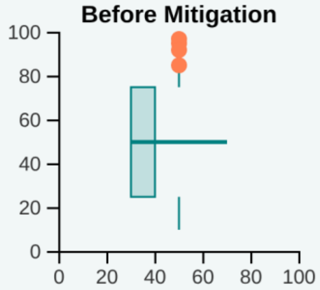
Outlier Detection Results

Annual Income



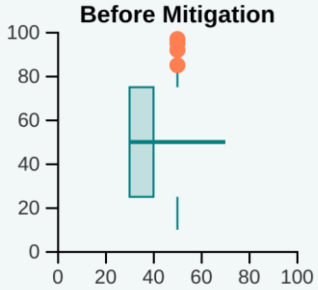
⚠**Outliers:** High skew

Loan Amount



⚠**Outliers:** Extreme values

Monthly Debt



⚠**Outliers:** High variation

Mitigation Strategies

Winsorization
Capping at 1st and 99th percentiles

Trimming
Removing extreme values

Transformation
Log scaling to reduce skew

Result

Improved model reliability and performance

Feature Engineering - Data Cleaning

Key data cleaning techniques applied to prepare the dataset for modeling:

Median Imputation

- Applied to **numeric** features with null values
- Replaces missing values with **median** to reduce skewness
- Maintains distribution shape better than mean imputation

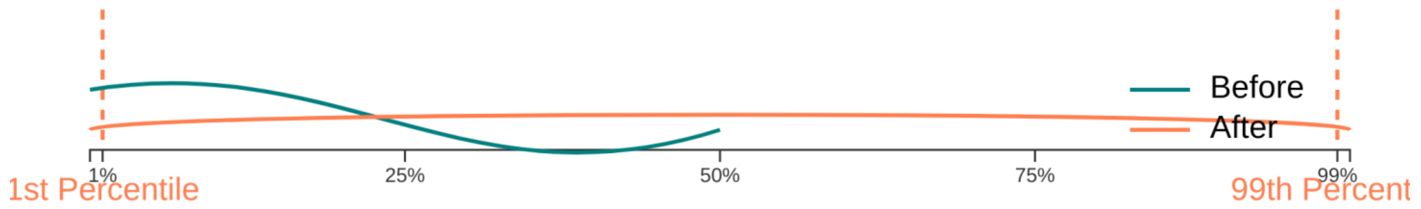
Mode Imputation

- Applied to **categorical** features with null values
- Replaces missing values with **mode** (most frequent value)
- Preserves category distribution in the dataset

Winsorization

- Capping outliers at **1st** and **99th** percentiles
- Prevents **model distortion** caused by extreme values
- Retains data integrity while reducing impact of outliers

Winsorization Visualization



Winsorization caps extreme values at specified percentiles, reducing their influence on the model while maintaining data distribution characteristics.

Feature Engineering - Transformations & Risk Score

Data Transformations

ApplicationDate Standardization
Cutoff date of 2025 for temporal analysis

Categorical Encoding
Converting categorical variables into numerical representations

Feature Scaling
Normalization using StandardScaler for consistent impact

Risk Score Feature

Composite feature derived from multiple variables to assess borrower solvency

Model Selection & Methodology

XGBoost

- Tree-based boosting
- Handles structured data
- Regularization to prevent overfit

Random Forest

- Ensemble of trees
- Reduces overfitting risk
- Handles mixed data types

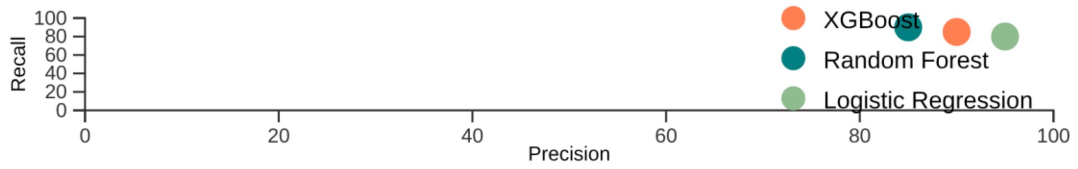
Logistic Regression

- Linear model for classification
- Computational efficiency
- Interpretable coefficients

Tuning Strategy

GridSearchCV

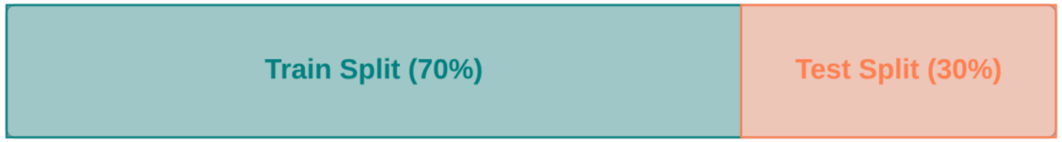
- Systematic hyperparameter optimization
- Balance between precision and recall
- Exhaustive search across parameter grid



Cross-Validation

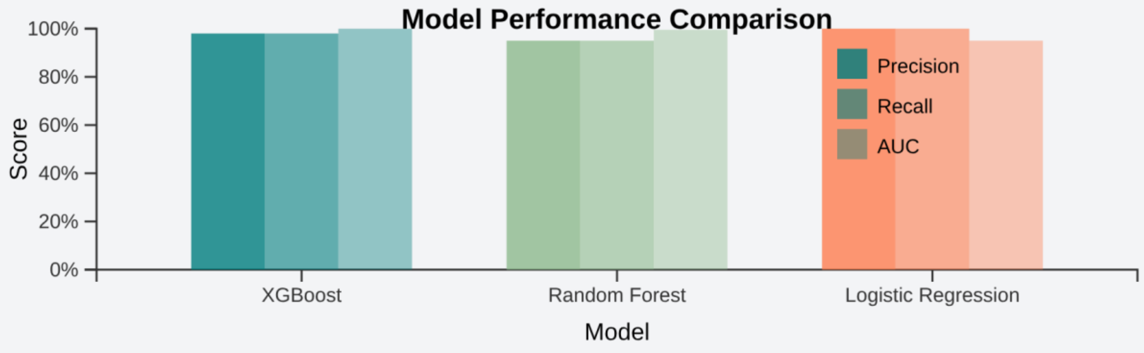
Train-Test Validation

- Standard train-test splits
- Ensures generalizability
- Reduces overfitting risk



Validation across multiple splits

Results & Model Performance Comparison



Key Findings

- **XGBoost** achieved the best balance with Precision & Recall of 0.98 and AUC of 0.9997
- **Random Forest** was highly accurate but more conservative, missing 5% of good applicants vs XGBoost's 2%
- **Logistic Regression** flagged as "Red Flag" despite perfect AUC of 1.0000 (data leakage suspected)

XGBoost

0.98

Precision & Recall

0.9997

AUC Score

- Best balance of metrics
- Selected as production model

Random Forest

0.95

Precision & Recall

0.9950

AUC Score

- High accuracy
- More conservative than XGBoost

Logistic Regression

1.00

AUC Score

Red Flag

Data Leakage?

- Perfect score suspicious
- Features like InterestRate likely leaked

Data Leakage Detection

The Red Flag

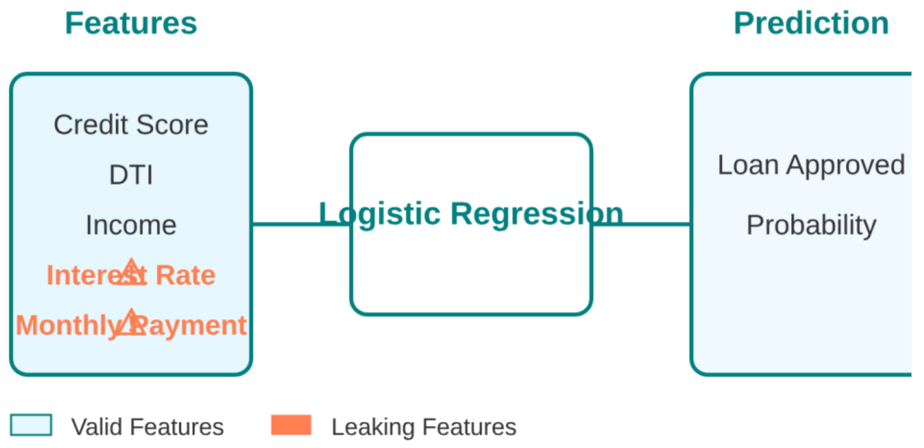
Logistic Regression model achieved a perfect AUC score of **1.0000**, which should raise concerns about data integrity.

> **Perfect AUC is often a sign of data leakage**

Investigation Findings

- Features like **InterestRate** and **MonthlyLoanPayment** showed suspicious correlation with the target
- These features are **generated after the loan decision**, making them invalid for prediction

Data Leakage Visualization

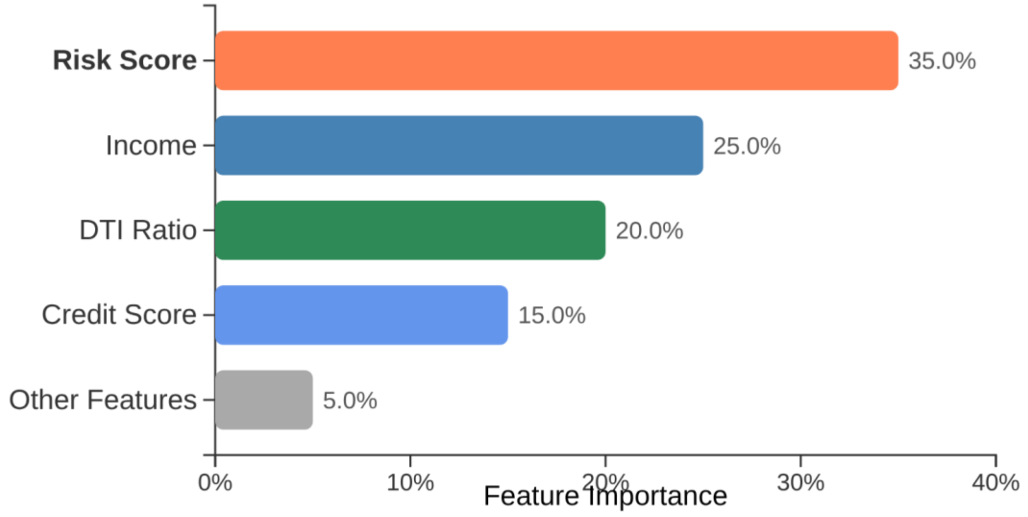


How to Address Data Leakage

- **Remove** post-decision features like InterestRate and MonthlyLoanPayment
- Implement **strict feature selection** to ensure only pre-decision variables are used
- Re-train models with **cleaned feature set** to obtain realistic performance metrics

Feature Importance & Explainability

Feature Importance Analysis



Tree-based models used for importance calculation

Key Insights

Primary Driver: Risk Score

- Risk Score emerged as the **most important feature**
- Composite feature from Credit Score, DTI, payment history

Non-Linear Relationships

- Risk Score showed **lower linear correlation**
- Model captured **complex, non-linear thresholds**

Supporting Features

Income
Strong linear correlation

DTI Ratio
Critical debt indicator

Credit Score
Established risk factor

Model Limitations & Dependencies

Current Model Limitations

High Dependency on RiskScore

The model is currently highly dependent on the synthesized RiskScore feature

Feature Generation Timing

Need to remove features generated after loan decision to ensure real-world deployability

Data Leakage Concerns

Future iterations must strictly remove post-decision features to prevent data leakage

Implementation Considerations

Feature Dependency Relationship

RiskScore — Model Prediction

Current implementation shows high dependency

Institutional Risk Appetites

In production, model must account for varying "risk appetites" of different financial institutions

Fairness Analysis

Implement FNR/FPR checks across demographic groups to ensure algorithmic fairness

Future Work & Improvements

Leakage Mitigation

Feature Removal Protocol

Strictly remove features generated after loan decision (e.g., specific interest rates) to ensure real-world deployability

Risk Score Dependency

Reduce dependency on synthesized RiskScore feature by incorporating more direct financial indicators

External Factors

Implement adaptability for varying "risk appetites" of different financial institutions

Implementation Progress

Current

Target

Fairness Analysis & Compliance

Demographic Group Analysis

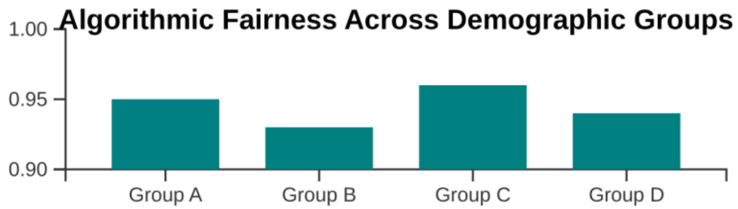
Implement FNR/FPR (False Negative Rate / False Positive Rate) checks across demographic groups to ensure algorithmic fairness

Regulatory Compliance

Design compliance framework to meet regulatory requirements for algorithmic decision-making in lending

Continuous Monitoring

Establish ongoing evaluation of model performance across different demographic segments to identify and address potential biases



Thank you for your attention