Final Report

Loan Default Prediction Using Machine Learning

Abstract
Loan default prediction is critical for financial institutions to minimize risks and maintain operational stability. This report explores a machine learning approach to predict loan defaults using a real-world dataset from Kaggle. The methodology integrates exploratory data analysis (EDA), clustering, feature engineering, and regularization techniques to handle imbalanced data, extract meaningful features, and achieve robust predictive performance. Models evaluated include Logistic Regression, Gaussian Discriminant Analysis (GDA), and K-Nearest Neighbors (KNN), with enhancements through regularization. The project highlights key insights into model performance, showcasing their strengths, limitations, and practical implications.

Introduction
Loan default is a pressing challenge for financial institutions, directly affecting profitability, operational stability, and investor confidence. Accurate prediction of loan defaults allows institutions to:

- Minimize non-performing assets (NPAs).
- Optimize risk management strategies.
- Enhance decision-making in loan approvals.

Challenges

1. Class Imbalance: Non-defaulters significantly outnumber defaulters, introducing bias in predictions.
2. Feature Selection: High-dimensional data makes it difficult to identify critical predictors while avoiding overfitting.
3. Model Performance vs. Interpretability: Balancing advanced model accuracy with interpretability is crucial for real-world applicability.

Objective
This project aims to build an interpretable, efficient, and accurate machine learning pipeline that:

- Predicts loan defaults with high precision and recall.
- Mitigates class imbalance for equitable performance across classes.
- Provides actionable insights into borrower behavior for financial decision-making.

Problem Statement
The goal is to predict the likelihood of loan default (binary classification: Defaulted or Not Defaulted) using a set of borrower and loan attributes. The primary challenges include:

- Handling an imbalanced target variable.

- Extracting meaningful features from high-dimensional data.
- Selecting appropriate machine learning models for robust performance.

Dataset

1. Description: The dataset comprises 94,448 records and 16 features, including:
    - Numerical Features: Loan amount, EMI, term of loan, annual and monthly interest rates, household income, debt-to-income ratio, etc.
    - Categorical Features: Loan title, application type (individual/joint), grade, home ownership.
    - Target Variable: `Loan_Status` (1 = Defaulted, 0 = Not Defaulted).

2. Preprocessing
    - Data Integrity Checks: Ensured no missing values in the dataset.
    - Encoding Categorical Features: Applied One-Hot Encoding for categorical variables.
    - Scaling Numerical Features: Used StandardScaler to normalize numerical features, ensuring uniform scaling and faster model convergence.

Methodology

Exploratory Data Analysis (EDA):

1. **Initial Insights**

1. Target Variable Distribution:
    - Class imbalance was significant, with 80% non-defaulters and 20% defaulters.
    - Highlighted the need for resampling techniques to improve model sensitivity to defaulters.

2. Feature Distributions:
    - Features like `Loan_Amount`, `EMI`, and `Monthly_Household_Income` exhibited right-skewed distributions with notable outliers.
    - Histograms and boxplots provided insights into these distributions and outlier thresholds.

3. Correlation Analysis:
    - Moderate correlations were observed between `Debt_to_Income`, `Loan_Amount`, and the target variable.
    - Highly correlated features, such as `Annual_InterestRate` and `Monthly_InterestRate`, indicated redundancy.

4. Categorical Feature Analysis:
    - Chi-Square tests revealed `Loan_Title` as significantly associated with `Loan_Status`, making it a critical predictor.

2. Feature Engineering:

1. Clustering-Based Features:
    - K-Means Clustering: Segmented borrowers into three clusters based on financial attributes, adding a categorical feature (`KMeans_Cluster`).
    - Gaussian Mixture Models (GMM): Provided posterior probabilities for each borrower belonging to latent clusters, enriching the dataset with nuanced, probabilistic features (`GMM_Prob_0`, `GMM_Prob_1`, `GMM_Prob_2`).

2. Regularization for Feature Selection:
    - L1 Regularization (Lasso): Reduced dimensionality by selecting the most predictive features, improving interpretability without significant accuracy loss.
    - Elastic Net: Balanced L1 and L2 penalties, particularly effective for correlated features.

Handling Class Imbalance

To address the imbalance in `Loan_Status`, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generated synthetic samples for the minority class, creating a balanced dataset and improving recall for defaulters.

Models Evaluated

1. Logistic Regression: Used as a baseline and enhanced with L1 and L2 regularization.
2. Gaussian Discriminant Analysis (GDA): Leveraged class-specific Gaussian distributions.
3. K-Nearest Neighbors (KNN): Explored non-linear relationships between features.
4. Naive Bayes: Assumed feature independence for simplicity.

Results

1. **Model Performance**

Model	Accuracy (%)	Precision (Defaulters) (%)	Recall (Defaulters) (%)	F1-Score (Defaulters)
Logistic Regression	82.2	96	68	0.79
GDA	82.4	98	66	0.79
KNN	82.8	97	65	0.79

Naive Bayes	62.3	70	55	0.68

2. Analysis of Results

Logistic Regression

1. Achieved **82.2% accuracy**, providing a balanced trade-off between precision (96%) and recall (68%) for defaulters.
2. Balanced class weights (`class_weight='balanced'`) effectively mitigated the issue of class imbalance, ensuring the minority class (defaulters) was not overlooked.
3. Logistic Regression's linear decision boundary generalized well to unseen data, reducing overfitting. This was because of interpretable coefficients allowing understanding of the relative importance of features.
4. The linear nature of Logistic Regression limits its ability to capture complex non-linear relationships between features.
5. While Logistic Regression slightly underperformed in precision compared to GDA, its recall was higher (68% vs. 66%), making it better at identifying actual defaulters.

Gaussian Discriminant Analysis (GDA)

1. Achieved **82.4% accuracy**, with **high precision (98%)** for defaulters. This means nearly all flagged defaulters were actual defaulters.
2. Recall for defaulters was lower (66%), indicating the model missed a significant number of true defaulters.
3. High precision ensured that flagged loans were likely actual defaulters, minimizing unnecessary interventions and reducing false positives.
4. GDA assumes shared covariance across classes, which may not align well with complex datasets where class distributions differ significantly.
5. GDA had slightly higher precision but lower recall, suggesting it is better suited for minimizing false positives than maximizing defaulter detection.

K-Nearest Neighbors (KNN)

1. Achieved the **highest accuracy (82.8%)**, leveraging its ability to model non-linear decision boundaries.
2. Precision for defaulters was **97%**, while recall was **65%,** slightly lower than Logistic Regression and GDA.
3. KNN's proximity-based algorithm excels in datasets with non-linear separations between classes. However, its high computational cost and sensitivity to scaling Limitations significantly degrades performance.

4. The inclusion of clustering-based features (e.g., KMeans_Cluster and GMM_Probabilities) complemented KNN's ability to group similar instances, boosting its performance.
5. KNN significantly outperformed Naive Bayes in all metrics, demonstrating its superior capability to capture complex data patterns.

Naive Bayes

1. Achieved **62.3% accuracy**, significantly lower than other models. Precision and recall for defaulters were **70%** and **55%,** respectively, indicating its inability to effectively separate the classes.
2. The model's fast training time, simplicity and assumption of feature independence make it suitable for quick preliminary analysis but not for real-world datasets.
3. The model struggled to differentiate defaulters from non-defaulters due to its simplistic assumptions.
4. Naive Bayes was outperformed by all other models due to its simplistic assumptions. Logistic Regression and KNN provided far superior results in both precision and recall.

Comparative Insights

Aspect	Logistic Regression	GDA	KNN	Naive Bayes
Accuracy (%)	82.2	82.4	82.8	62.3
Precision (Defaulters)	96	98	97	70
Recall (Defaulters)	68	66	65	55
Computational Efficiency	High	High	Low	Very High
Interpretability	High	Moderate	Low	High
Complex Relationships	Limited	Moderate	Excellent	Limited

Key Observations:

1. **Overall Best Performer:**
   - Logistic Regression provided the best balance of precision, recall, interpretability, and computational efficiency. Its use of class_weight='balanced' ensured fair treatment of minority and majority classes.
2. **Best for Precision:**

- GDA excelled in precision (98%), making it suitable for scenarios where minimizing false positives is critical, such as high-value loans or stringent risk-averse policies.

3. **Best for Non-Linear Data:**
    - KNN leveraged non-linear relationships to achieve the highest accuracy (82.8%) but at a computational cost.
4. **Least Effective:**
    - Naive Bayes was the least effective, largely due to its independence assumption, which failed to capture complex relationships in the data.

Discussion

1. **Feature Selection**

a. L1 Regularization:

- By shrinking coefficients of less important features to zero, L1 Regularization reduces model complexity, making predictions more interpretable for financial analysts.
- Features like `Debt_to_Income` and `Loan_Amount` were consistently selected, highlighting their predictive power in determining loan default risks.
- L1 Regularization is particularly beneficial in datasets with many irrelevant or redundant features, streamlining the modeling process.
- Its sparsity-promoting nature helps avoid overfitting by focusing only on the most informative features.

b. Elastic Net:

- Elastic Net combines the strengths of L1 and L2 penalties, balancing feature selection with stability in the presence of multicollinearity.
- It handled correlated features like `Annual_InterestRate` and `Monthly_InterestRate` effectively by penalizing their combined impact rather than excluding one arbitrarily.
- Elastic Net provided a middle ground, allowing some features to contribute partially instead of enforcing complete exclusion as in L1.
- The approach proved useful for high-dimensional datasets, where multicollinearity among financial indicators is common.

2. **Addressing Class Imbalance**
    - The Synthetic Minority Over-sampling Technique (SMOTE) created synthetic data points for the minority class (defaulters) by interpolating between nearest neighbors.

- This approach helped balance the dataset, improving the recall for defaulters by ensuring the model learned from a more representative sample of minority class instances.
- However, SMOTE's synthetic nature introduced the risk of overfitting, as the new data points closely resembled existing ones, emphasizing the need for rigorous validation.
- The technique works well with linear models like Logistic Regression but may lead to degraded performance in non-parametric models like KNN if not carefully applied.

3. Business Implications

1. Actionable Insights:
    - Clustering features such as KMeans_Cluster and GMM_Probabilities segmented borrowers into meaningful categories, enabling lenders to design targeted strategies for high-risk groups.
    - Predictors like Debt_to_Income and Loan_Amount offered clear, interpretable metrics for assessing borrower risk, aligning closely with industry practices.
    - These insights help financial institutions prioritize resources, offering more stringent terms to riskier profiles and favorable terms to low-risk groups.
    - The ability to pinpoint high-risk features allows loan officers to justify decisions with data-backed explanations, improving transparency and trust.
2. Operational Utility:
    - Models with high precision, such as GDA, ensure that flagged loans are likely actual defaulters, minimizing false alarms and unnecessary interventions.
    - Moderate recall, while acceptable for initial screening, highlights the need for supplementary risk assessment tools to catch additional defaulters missed by the models.
    - The use of machine learning for predictive insights complements traditional financial metrics, offering a scalable solution for large loan portfolios.
    - Combining these models with real-time systems could streamline decision-making, reducing processing times and improving customer satisfaction.

Conclusion

This project successfully implemented a machine learning pipeline to predict loan defaults, addressing challenges like class imbalance and feature selection. Logistic Regression emerged as the most practical model, balancing accuracy, interpretability, and computational efficiency. The clustering features and regularization techniques significantly enriched the dataset and improved model robustness.

Key Takeaways

1. Model Selection: Logistic Regression is recommended for balanced performance, while GDA excels in high-precision applications.
2. Feature Engineering: Clustering features provided actionable borrower profiles, enhancing model insights.
3. Future Work: Explore ensemble methods like Random Forests and Gradient Boosting, along with explainability techniques such as SHAP and LIME.