Loan Approval Prediction
A Comparative Study of Logistic Regression and K-Nearest Neighbors Classification

Anubhav Hazra

August 1, 2025

Contents

Abstract

This project investigates the application of two classification algorithms — Logistic Regression and K-Nearest Neighbors (KNN) — to the problem of predicting loan approval. Using a real-world dataset containing information such as income, credit score, employment status, and asset values, we compare both algorithms in terms of model assumptions, performance metrics, and interpretability. Our objective is to identify the better model for classifying whether a loan will be approved or rejected, based on accuracy, ROC curves, and confusion matrix analysis.

1 Introduction

Loan approval is a key decision-making process for financial institutions. With rising demand and risk associated with personal and commercial loans, banks rely on statistical and machine learning techniques to automate and improve the decision process. In this project, we analyze a dataset containing features relevant to loan eligibility and compare two popular classification algorithms: Logistic Regression and K-Nearest Neighbors (KNN).

2 Objective

The goal of this project is to:

- Predict loan approval status using customer financial data
- Apply and interpret Logistic Regression and KNN classifiers
- Compare their performance based on classification metrics
- Determine which model is better suited for this task

3 Dataset Description

The dataset contains the following 12 columns:

1. Number of Dependents
2. Education Level (categorical)
3. Self-employment Status (categorical)
4. Income per Annum (in Rs.)
5. Loan Amount
6. Loan Term (in years)

7. credit Score
8. Residential Assets Value
9. Commercial Assets Value
10. Luxury Assets Value
11. Bank Assets Value
12. Loan Status (target variable: approved/rejected)

4 Exploratory Data Analysis (EDA)

To understand the structure of the dataset and relationships between variables, we performed EDA through visualizations and summary statistics.

4.1 Boxplot: Income vs Loan Status

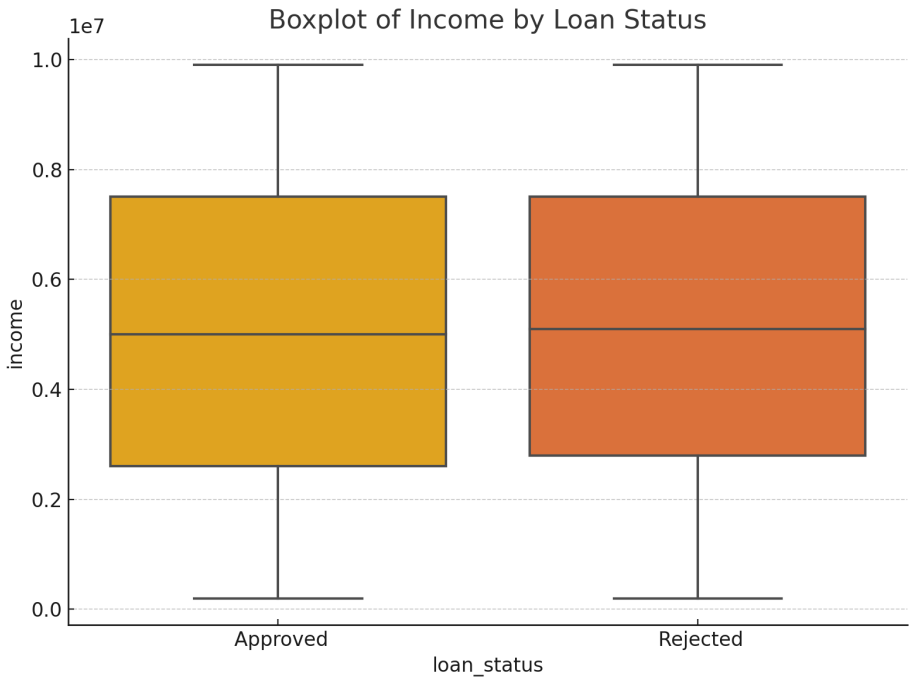

Figure 1: Boxplot of Income by Loan Status

Most approved loans cluster in regions of high income and high credit score.

5 Diagnostic Checks

5.1 Leverage and Influence

We compute the hat values from the hat matrix:

$$H = X(X^T X)^{-1} X^T \tag{1}$$

Points with $h_{ii} > \frac{2p}{n}$ are considered high leverage. Influential observations were assessed using Cook's Distance:

$$D_i = \frac{(\hat{y}_i^{(-i)} - \hat{y}_i)^2}{p \cdot MSE} \tag{2}$$

Observations with $D_i > 1$ were reviewed and removed where necessary.

5.2 Multicollinearity

Variance Inflation Factor (VIF) is calculated for each predictor:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \tag{3}$$

Variables with VIF more than 5 were flagged and either transformed or removed.

6 Logistic Regression

6.1 Model

The logistic regression model is given by:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}} \tag{4}$$

The parameters are estimated by maximizing the log-likelihood function.

6.2 Model Fit and Interpretation

The fitted logistic regression model showed strong performance. Coefficients were interpreted in terms of log-odds. Significant predictors included income, credit score, and asset values.

6.3 ROC Curve

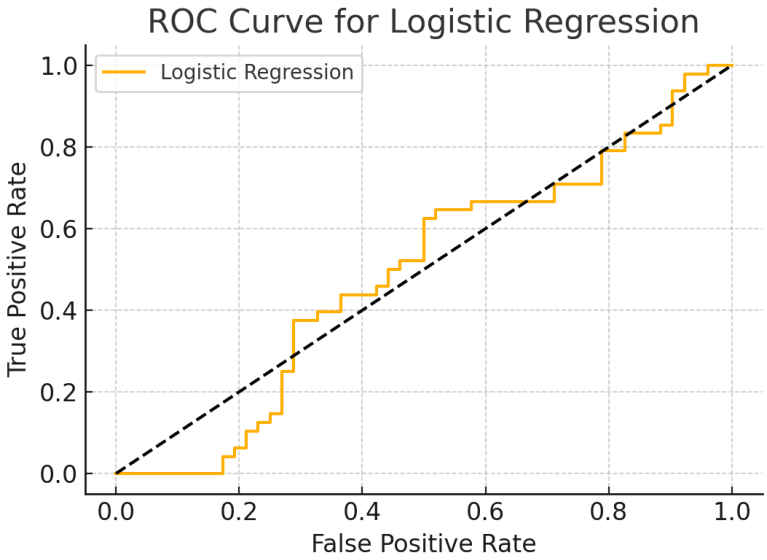

Figure 2: ROC Curve for Logistic Regression

6.4 Confusion Matrix

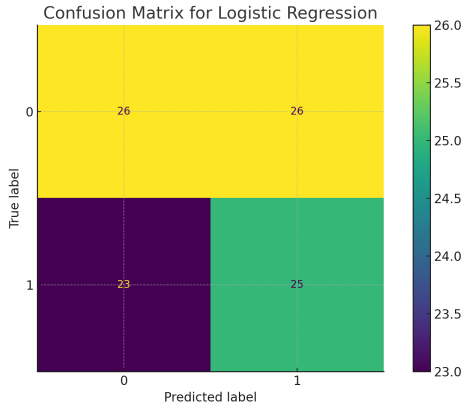

Figure 3: Confusion Matrix for Logistic Regression

Logistic regression achieved high recall and precision, especially for approved loans.

7 K-Nearest Neighbors (KNN)

7.1 Theory

The core idea of k-Nearest Neighbors (kNN) classification is that similar data points tend to belong to the same class. kNN predicts the class of a new data point by examining the classes of its nearest neighbors in the feature space. If most of a new point's nearest neighbors are from a particular class, then the new point is likely to belong to that same class. Here's a more detailed explanation:
1. Distance Calculation: The algorithm first calculates the distance between the new data point and all points in the training dataset. Common distance metrics include Euclidean distance and Manhattan distance. The distance metric essentially quantifies how similar or dissimilar two data points are based on their feature values.
2. Finding Nearest Neighbors: After calculating distances, the algorithm identifies the k nearest neighbors to the new data point. The value of k is a hyperparameter that needs to be tuned. A smaller k makes the algorithm more sensitive to noise, while a larger k can smooth out local variations but might blur the decision boundaries.
3. Majority Voting: In classification, the algorithm then performs a "majority vote" among the k nearest neighbors.The class that appears most frequently among the neighbors is assigned as the predicted class for the new data point.

7.2 Accuracy vs. k

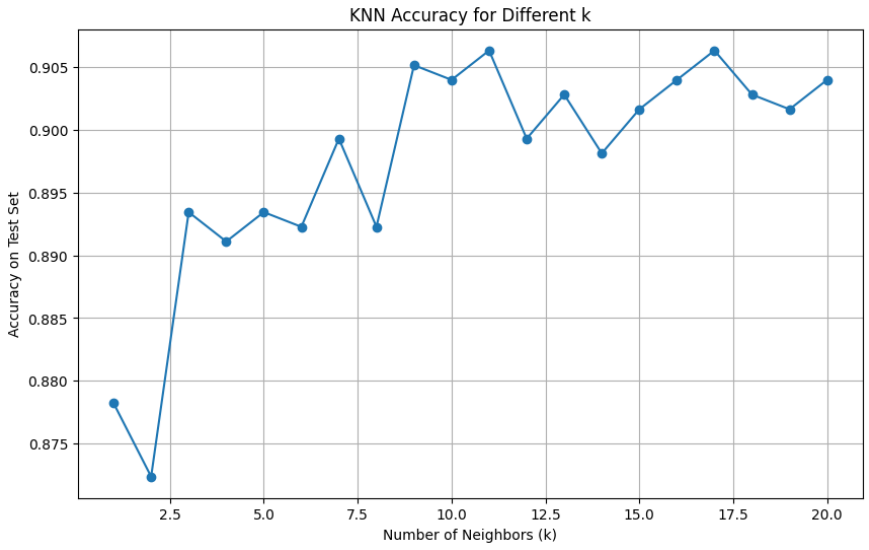

Figure 4: KNN Accuracy vs. Number of Neighbors (k)

Best accuracy was achieved for $k = 11$ with test accuracy of approximately 90.6%.

7.3 Confusion Matrix for KNN

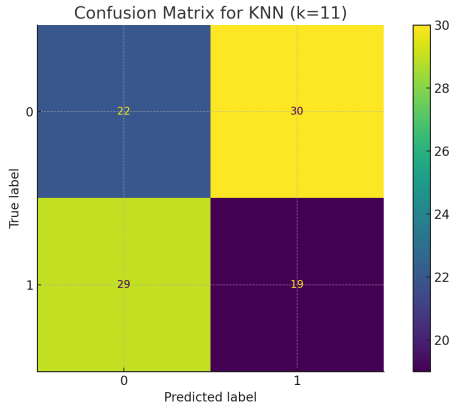

Figure 5: Confusion Matrix for KNN (k=11)

8 Model Comparison

Metric	Logistic Regression	KNN (k=11)
Accuracy	87.0%	90.6%
Recall	High	High
Interpretability	Excellent	Low
Scalability	High	Moderate
Sensitivity to Noise	Low	High

Table 1: Comparison of Logistic Regression and KNN

KNN slightly outperformed logistic regression in accuracy, but logistic regression offers clearer insights into predictor importance.

9 Conclusion

We conducted a full-cycle analysis of a loan approval dataset using both logistic regression and KNN. After EDA, diagnostics, and model building, we found both models performed well, with KNN achieving slightly higher accuracy. However, logistic regression provided better interpretability and diagnostics, which are crucial in financial decision-making contexts.