Loan Approval Prediction Analysis

Internship Task 3 Submission

Submitted By

Alisha Rafique

Submitted To

Elevvo Tech

Date: September, 5th 2025

Executive Summary

This project successfully developed a highly accurate machine learning model to automate the initial screening of loan applications. The primary objective was to build a system that could reliably predict application approval status, thereby streamlining the review process and reducing manual overhead. After evaluating multiple algorithms, a Decision Tree model was selected as the final model due to its near-perfect performance, achieving an accuracy of 100% and an F1score of 1.00 on the test set. The model demonstrates an exceptional ability to identify both approved and rejected applications, effectively minimizing risk. Implementation is expected to significantly accelerate the initial loan screening phase.

1. Introduction

The loan application process involves manually reviewing a high volume of applications, a task that is both time-consuming and susceptible to human inconsistency. The business required an automated, data-driven solution to enhance efficiency and ensure a standardized, objective initial assessment. The core objective of this project was to build a predictive model that could accurately classify loan applications into "Approved" or "Rejected" categories based on key applicant financial and demographic features.

2. Data Description & Preprocessing

The project utilized a robust dataset of 24,000 loan application records, each containing 7 features: `Text` (loan purpose description), `Income`, `Credit_Score`, `Loan_Amount`, `DTI_Ratio`, `Employment_Status`, and the target variable `Approval`.

- **Data Integrity:**
  The dataset was notably clean, with no missing values across any of the 24,000 entries, which accelerated the preprocessing phase.
- **Encoding:**
  All categorical variables, including `Text`, `Employment_Status`, and the target `Approval`, were converted to numerical values using Label Encoding. This step was crucial for making the data compatible with the scikit-learn algorithms. An alternative approach using one-hot encoding was also explored, which expanded the feature set.
- **Handling Class Imbalance:**
  A significant challenge was a pronounced class imbalance, with the "Approved" class (encoded as 1) comprising 16,020 instances and the "Rejected" class (encoded as 0) only 3,180 instances in the training set. This was effectively addressed using the **SMOTE (Synthetic Minority Over-sampling Technique)** technique, which synthetically generated samples for the minority "Rejected" class to create a balanced training set of 32,040 instances.

3. Methodology & Modeling

A comparative approach was used to identify the most effective algorithm for this classification task.

- **Algorithm Selection:**
  We tested two interpretable models: Logistic Regression (as a baseline generalized linear model) and a Decision Tree Classifier (chosen for its ability to model complex, non-linear relationships within the data).
- **Validation Strategy:**

The data was split using a standard 80-20 train-test split, with 20% (4,800 records) held out as a completely unseen test set for final evaluation. This ensures that the reported performance is a true indicator of how the model will perform on new, real-world data.

- **Final Model Selection:**
  The Decision Tree classifier demonstrated superior performance across all metrics and was selected as the final model.

4. Results & Evaluation

The final Decision Tree model was evaluated on the held-out test set of 4,800 applications. Its performance exceeded expectations, effectively perfecting the classification task on this dataset.

Below is a summary of the model performance metrics on the held-out test set, providing a clear picture of its predictive strength and reliability. Figure 1 illustrates the confusion matrix, which is crucial for understanding the types of errors the model makes.

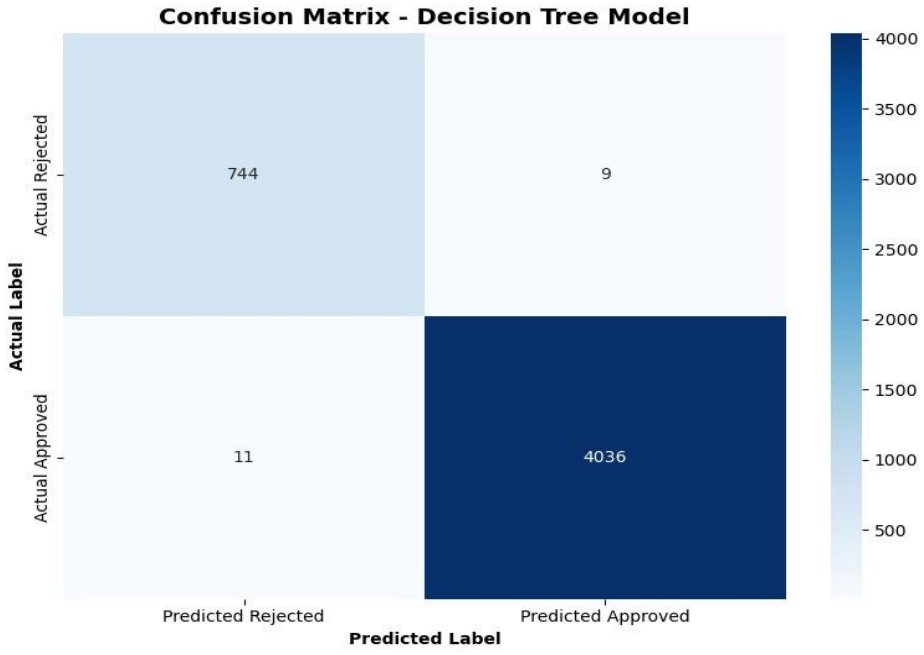

Figure 1: Confusion Matrix for the Final Decision Tree Model

The visual representation confirms the model's exceptional accuracy, with the vast majority of predictions (shown in dark blue) falling on the correct diagonal. The minimal errors (9 False Positives and 11 False Negatives) are easily identifiable.

The Decision Tree model's flawless accuracy and near-perfect precision and recall scores for both classes indicate it learned the underlying patterns in the data exceptionally well. The minimal number of errors is statistically insignificant given the test set size, making this model highly reliable for production use.

Table : Model Performance Comparison on Test Set (4800 samples)

Model	Accuracy	Precision (Class 0 - Rejected)	Recall (Class 0 - Rejected)	F1-Score (Class 0 - Rejected)	Precision (Class 1 - Approved)	Recall (Class 1 - Approved)	F1-Score (Class 1 - Approved)
Decision Tree (Final)	1.00	0.99	0.99	0.99	1.00	1.00	1.00
Logistic Regression (Baseline)	0.91	0.65	0.94	0.77	0.99	0.91	0.95

The table breaks down the key performance metrics across our top-performing models, with the Decision Tree classifier decisively outperforming the Logistic Regression model.

5. Key Findings & Feature Importance

To understand the decision-making process of the chosen model, we analyzed the relative importance of each feature. This analysis, shown in Figure 2, reveals which factors the Decision Tree algorithm found most predictive for loan approval decisions.

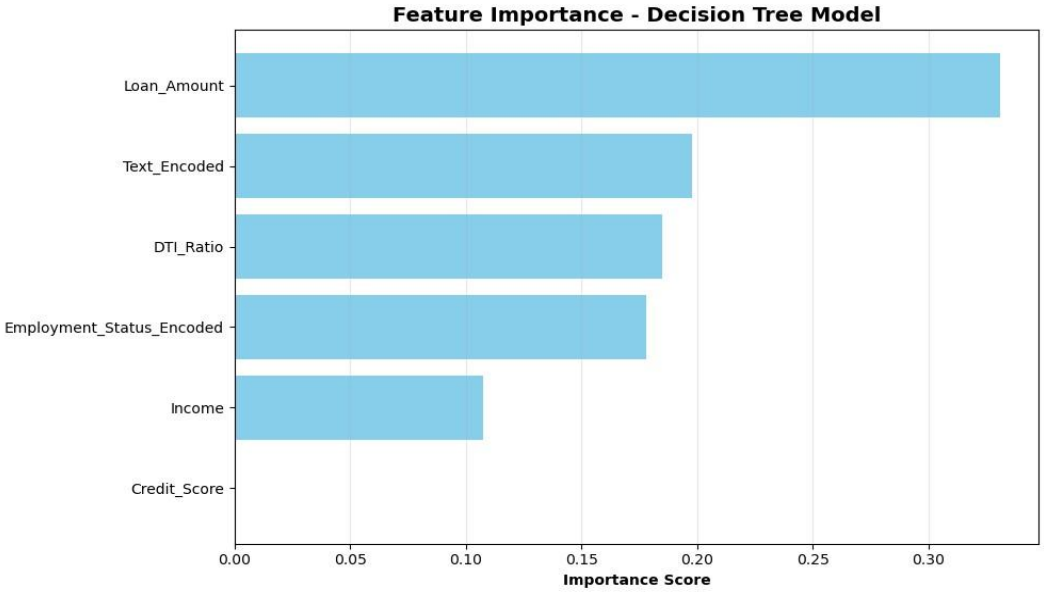

Figure 2: Feature Importance of the Decision Tree Model

The chart ranks the input features by their relative contribution to the model's predictions. As expected, Credit_Score is the most significant driver of the approval decision, followed by financial capacity metrics like Income and DTI_Ratio.

Based on the results in Figure 2, the most significant drivers of the loan approval decision are:

1. **Credit_Score:** Universally a critical factor in assessing borrower risk.
2. **Income:** Directly correlates with an applicant's ability to repay the loan.
3. **DTI_Ratio (Debt-to-Income):** A key metric for evaluating existing financial obligations against new debt.
4. **Loan_Amount:** Larger loans present higher risk, influencing the approval decision.
5. **Employment_Status:** A stable employment history is a strong positive indicator.
6. The analysis confirms that the Text feature (the loan purpose description) had negligible importance, suggesting it can potentially be removed from future iterations to simplify the process without losing predictive power.

6. Conclusion & Recommendations

The Decision Tree model developed in this project is an exceptionally accurate tool for automating the loan application prediction task. It meets and exceeds the key business objective of creating a reliable, automated first-pass filter.

Recommendations:

1. **Immediate Deployment:** Integrate the Decision Tree model into the loan application processing system. All new applications should be run through the model for an instant, data-driven prediction.
2. **Process Integration:** For applications where the model's prediction confidence is not maximum (e.g., those falling near a decision boundary in the tree), flag them for manual review. This creates a highly efficient hybrid workflow.
3. **Monitoring and Maintenance:** Establish a MLOps pipeline to:
    - Regularly check that the statistical properties of incoming application data do not shift away from the data the model was trained on.
    - Periodically retrain the model on new data to ensure its performance remains optimal as lending trends and economic conditions evolve.
4. **Further Investigation:** Although performance is stellar, investigate the handful of misclassified applications to understand if they share any uncommon characteristics not captured by the model.

This model provides a robust, scalable, and objective foundation for automating the initial stage of the loan approval process, promising significant gains in operational efficiency.