Loan Prediction: Hypothesis Testing & Supervised Modeling Report

Abstract

This project focuses on predicting loan approval status using a combination of inferential statistical analysis and supervised machine learning models. Financial institutions must evaluate loan applications efficiently while minimizing risk. To support this decision-making process, statistical hypothesis testing techniques such as the T-Test and Chi-Square Test were applied to validate relationships between applicant attributes and loan approval outcomes.

Further, data preprocessing, correlation analysis, feature selection, and classification models including Logistic Regression and Decision Tree were implemented. Model performance was evaluated using Accuracy, Confusion Matrix, and ROC-AUC Score. The study demonstrates how statistical reasoning combined with machine learning can enhance financial risk assessment and automate loan approval prediction.

1. Introduction

Loan approval is a critical process in banking and financial services. Institutions must analyze applicant profiles carefully to ensure that loans are granted to reliable customers while minimizing default risk. Traditional loan approval systems rely heavily on manual verification and subjective judgment, which can lead to inefficiencies and inconsistencies.

With the growth of data analytics and machine learning, financial organizations are shifting toward automated decision systems. These systems analyze historical applicant data to identify patterns influencing loan approval. This project implements such a data-driven approach by integrating hypothesis testing with supervised machine learning to predict loan approval outcomes.

2. Problem Statement

The objective of this project is to analyze loan applicant data using statistical hypothesis testing and build supervised machine learning models to accurately predict loan approval status.

3. Objectives

The key objectives of this project are:

- Perform inferential statistical analysis on loan data
- Conduct T-Test and Chi-Square tests
- Study feature relationships using correlation analysis
- Prepare and encode the dataset
- Select top predictive features
- Train supervised classification models
- Evaluate model performance using standard metrics

4. Dataset Description

The Loan Prediction dataset contains demographic and financial details of loan applicants. It includes attributes related to income, education, employment, property area, and credit history.

Key features used in the study include:

- Gender
- Married
- Dependents
- Education
- Self_Employed
- ApplicantIncome
- LoanAmount
- Credit_History
- Property_Area
- Loan_Status (Target Variable)

The dataset contained 614 records with both categorical and numerical features.

5. Data Preprocessing

Data preprocessing was performed to ensure data quality and model readiness. The following steps were applied:

- Loan_ID column was removed as it is a unique identifier and has no predictive significance.
- Missing categorical values were filled using mode imputation.
- Missing numerical values were handled using median imputation.
- Categorical variables were encoded using Label Encoding.

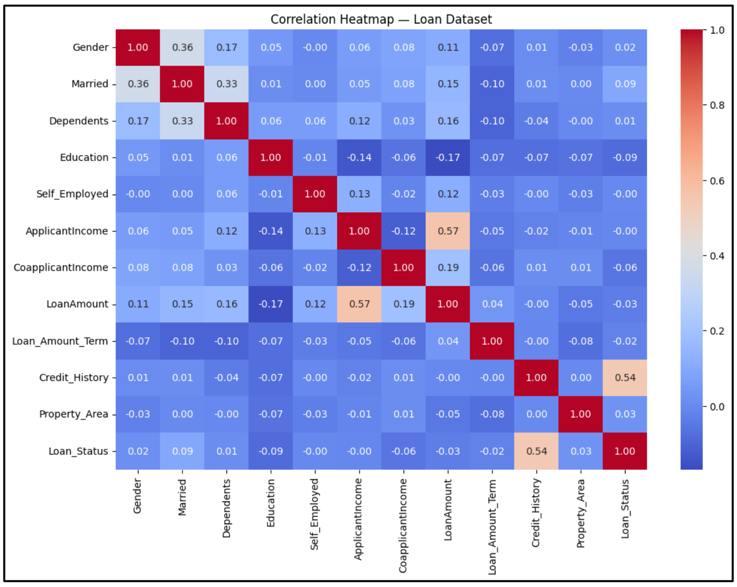
After preprocessing, the dataset was fully numerical and suitable for statistical and machine learning analysis.

6. Correlation Analysis

A correlation heatmap was generated to examine relationships between features and the target variable Loan_Status. Correlation analysis helps identify which variables strongly influence loan approval decisions.

Observation:

Credit_History showed the strongest positive correlation (0.54) with Loan_Status, indicating that past credit behavior plays a major role in loan approval. ApplicantIncome and LoanAmount showed moderate correlation among themselves. CoapplicantIncome and Loan_Amount_Term showed negligible correlation with Loan_Status.



Based on this analysis, weakly correlated features were removed to reduce noise and improve model performance.

Removed Features:

- CoapplicantIncome
- Loan_Amount_Term

7. Hypothesis Testing

7.1 T-Test

A two-sample T-Test was conducted to compare the mean ApplicantIncome between approved and rejected loan applicants.

Results:

T-Statistic: −0.116
P-Value: 0.907

Since the p-value is greater than 0.05, we fail to reject the Null Hypothesis. This indicates that there is no statistically significant difference in applicant income between approved and rejected loans. Therefore, ApplicantIncome alone is not a decisive factor in loan approval.

7.2 Chi-Square Test

The Chi-Square test was performed to examine the association between Education and Loan_Status.

Results:

Chi-Square Value: 4.091
P-Value: 0.043

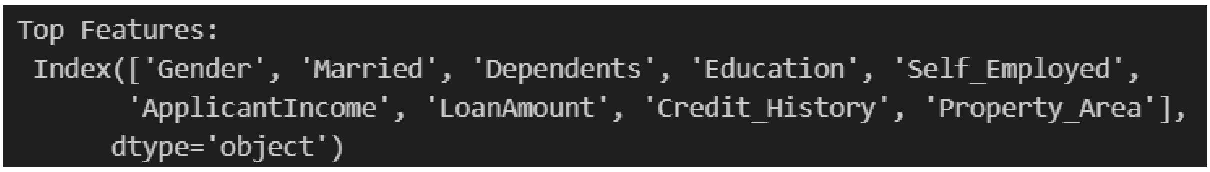
Since the p-value is less than 0.05, the Null Hypothesis was rejected. This confirms a statistically significant relationship between education level and loan approval status.

7.3 ANOVA (Conceptual Discussion)

ANOVA is used when comparing mean values across three or more groups. For example, if applicants were divided into low, medium, and high-income groups, ANOVA would determine whether mean loan amounts differ significantly across these groups.

8. Feature Selection

Feature selection was performed using the SelectKBest method with Chi-Square scoring to identify the most influential predictors of loan approval.



```
Top Features:
 Index(['Gender', 'Married', 'Dependents', 'Education', 'Self_Employed',
       'ApplicantIncome', 'LoanAmount', 'Credit_History', 'Property_Area'],
      dtype='object')
```

Top features included:

- Credit_History
- ApplicantIncome
- LoanAmount
- Education
- Property_Area

Feature selection improved model efficiency by removing redundant predictors.

9. Model Development

Two supervised classification models were implemented:

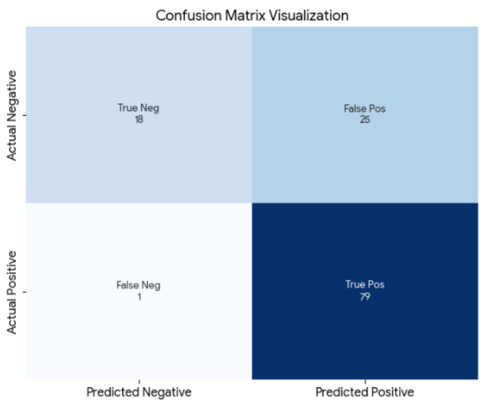
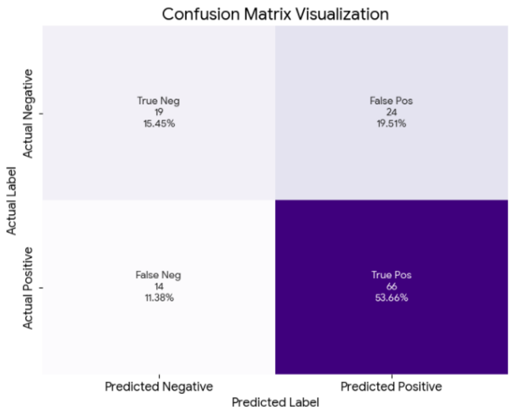
Logistic Regression

A linear model used as the baseline classifier for loan approval prediction.

Decision Tree Classifier

A tree-based model capable of capturing nonlinear decision boundaries.

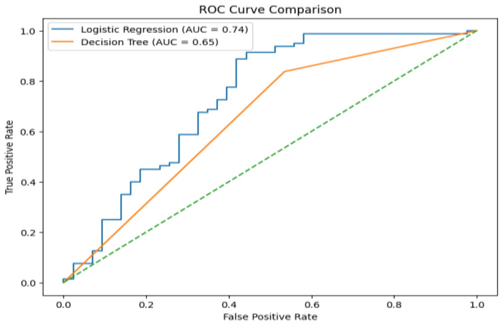
10. Model Evaluation

	Logistic Regression	**Decision Tree**
Accuracy	78.86%	69.10%
ROC-AUC Score	0.703	0.633
Confusion Matrix		

11. Model Comparison

Logistic Regression outperformed the Decision Tree model across evaluation metrics. It achieved higher accuracy and better ROC-AUC score, indicating stronger classification capability and generalization performance.

Hence, Logistic Regression was selected as the best model for this loan prediction task.



12. Key Findings

- Credit_History is the most influential predictor of loan approval.
- Education has a statistically significant association with approval status.
- ApplicantIncome alone does not determine approval.
- Correlation-based feature removal improved performance.
- Statistical testing supported machine learning insights.

13. Conclusion

This project successfully integrated hypothesis testing with supervised machine learning to predict loan approval outcomes. Statistical analysis validated relationships between applicant attributes and loan decisions, while classification models automated prediction.

Logistic Regression achieved the best performance with 78.86% accuracy and a ROC-AUC score of 0.703. The findings demonstrate how data-driven systems can support financial institutions in reducing risk, improving efficiency, and enhancing decision-making accuracy.