PROJECT REPORT

GROUP 09

KAVYA PATI BANDLA

LIKHITHA GUTTAPALLI

SAI VITTALA AYYALASOMAYAJULA

SAI VENKATA SHIVA RAMA GANESH RAPETI

617-637-6694

857-204-7868

857-675-0628

857-398-7739

PATIBANDLA.KA@NORTHEASTERN.EDU

GUTTAPALLI.L@NORTHEASTERN.EDU

AYYALASOMAYAJULA.S@NORTHEASTERN.EDU

RAPETI.SA@NORTHEASTERN .EDU

Percentage of Effort Contributed by Student 1**: 25%**

Percentage of Effort Contributed by Student 2**: 25%**

Percentage of Effort Contributed by Student 3**: 25%**

Percentage of Effort Contributed by Student 4**: 25%**

Signature of Student 1**: kavya**

Signature of Student 2**: Likhitha**

Signature of Student 3**: Sai**

Signature of Student 4**: Ganesh**

Submission Date: 12-12-2024

Report

Loan Approval Prediction

1. Introduction

This project develops a machine learning model to predict loan approvals using applicant and loan data. We used a dataset from Kaggle for this purpose, focusing on exploring and preprocessing the data and then building models to predict loan approvals accurately. The main objective is to compare different models and choose the best one for a hypothetical real-world application.

2. Data Preparation

The dataset includes 45,000 entries with 14 different attributes related to applicants and loans, such as age, gender, education, income, and loan amount, among others. The steps taken in data preparation included:

- **Data Cleaning**: Checked and confirmed there were no missing values.
- **Feature Encoding**: Converted categorical data into numerical codes to make them suitable for modelling.
- **Feature Selection**: Selected important features based on their correlation with the target variable, 'loan_status'.
- **Dimensionality Reduction**: Used PCA to reduce the number of variables while retaining 95% of the data variance.
- **Normalization and Standardization**: Applied these techniques to key features to aid in model performance.

3. Exploratory Data Analysis

- **Univariate Analysis**: Found that the target variable 'loan_status' was imbalanced, with a higher proportion of rejected loans.
- **Multivariate Analysis**: Identified strong relationships between features like loan amount and credit score with the loan status.
- **Visualization**: Used scatter plots and heatmaps to better understand the relationships between various features.

4. Model Development

We developed and compared three different models:

- **Logistic Regression**: Good for binary outcomes and easy to interpret.
- **Decision Tree Classifier**: Useful for capturing complex patterns in data, but can overfit.
- **Random Forest Classifier**: Combines many trees to improve prediction accuracy and generalization.

Training and Testing:

- The data was split into 70% for training and 30% for testing.

- All models were trained using the pre-processed features.

5. Results and Evaluation

Performance metrics for the models were as follows:

Metric	Logistic Regression	Decision Tree	Random Forest
Train Accuracy	0.90	1.00	1.00
Test Accuracy	0.89	0.90	0.93
Test Precision	0.89	0.90	0.92
Test Recall	0.89	0.90	0.93
Test F1-Score	0.89	0.90	0.92

Model Comparison:

- Logistic Regression was straightforward but less effective with complex patterns.
- The Decision Tree showed perfect training results but tended to overfit.
- The Random Forest had the best overall performance, balancing accuracy and generalization.

6. Final Model Selection

The Random Forest Classifier was chosen for its high accuracy (93%), ability to handle complex and imbalanced data.

7. Conclusion and Recommendations

The Random Forest Classifier is recommended for deployment given its superior accuracy and robustness. It's well-suited for practical use in predicting loan approvals accurately. Going forward, we could investigate further feature engineering to enhance performance and explore strategies for integrating the model into operational loan approval processes.

From a business perspective, understanding the key features that influence loan decisions can help customize financial products to better meet customer needs. Additionally, further studies could focus on simplifying the model to maintain high accuracy while improving interpretability, which is crucial for real-world application and regulatory compliance.