

A close-up photograph of a person's hands in a dark suit with white cuffs. The left hand holds a black pen over a document, while the right hand holds a fan of US dollar bills. A laptop is partially visible on the right side of the frame.

LOAN DEFAULT PREDICTION ANALYSIS

BY CLAUDIA SAGINI



INTRODUCTION

In the fast-paced world of finance, managing risk and making informed lending decisions are critical to the success and stability of financial institutions. With the growing availability of data and advancements in machine learning, we have an unprecedented opportunity to leverage predictive analytics to address one of the most pressing issues in lending: loan defaults.

CONTENT

01

BUSINESS AND DATA
UNDERSTANDING

02

OBJECTIVES

03

DATA ANALYSIS

04

MODEL EVALUATION

05

INSIGHTS AND
RECOMMENDATIONS



BUSINESS UNDERSTANDING

We aim to help financial institutions manage loan portfolios and mitigate risks associated with loan defaults. Loan defaults can significantly impact the financial health and stability of lenders. Accurately predicting loan defaults is crucial for making informed decisions and safeguarding their interests.



Stakeholder:

Financial Institution: Banks, Credit Unions, etc.



Business Problem:

Loan defaults pose risks for lenders. Accurate prediction aids informed decisions, risk management, and portfolio optimization. Early identification of high-risk borrowers minimizes default exposure for stability. Thus, a reliable predictive model is crucial for risk management and operational sustainability.



Can Machine Learning Predict Loan Defaults?

Lenders face loan default challenges, but machine learning offers a solution. This project demonstrates how machine learning classifies loan applicants by default risk, improving approval decisions and cutting financial losses.



OBJECTIVES

- Develop a machine learning model to predict loan defaults, providing lenders with a powerful tool to forecast risk.
 - Identify key features that influence loan repayment behavior, offering insights into borrower characteristics and financial habits.
 - Evaluate the performance of the model using relevant metrics to ensure accuracy, reliability, and practical application.
 - Implement the model in the institution's loan approval process to reduce financial risk and enhance decision-making capabilities.
-



DATA UNDERSATANDING

To address the business problem of loan default prediction, we will leverage a comprehensive loan default prediction dataset.

This dataset contains a wealth of information about loan applicants, including their demographic details, financial characteristics, loan details, and repayment history.



Data Justification

We chose this dataset because it is a comprehensive representation of loan applicants and their repayment behavior, allowing us to build a robust machine learning model that can accurately predict loan defaults. The dataset also includes a range of features that can help us identify the most important factors influencing loan repayment behavior, such as credit score, debt-to-income ratio, and loan term.

What the Data Entails

Demographic Information: Age, education level, occupation, etc.

Financial Information: Credit score, debt-to-income ratio, income, etc.

Loan Details: Loan amount, loan term, interest rate, etc.

Target Variable: Loan default indicator (0 - paid in full, 1 - defaulted)

Limitations

The dataset is based on historical data and may not reflect current market trends or changes in borrower behavior.

The dataset may not include all relevant factors that influence loan repayment behavior, such as changes in interest rates or economic conditions.

The dataset may contain missing values or errors that could affect the accuracy of the model.



Data Preparation

These are some of the tasks performed here:

- Here are a few rows and columns of the the data.
- It contains 255347 rows and 18 columns.
- Data cleaned was performed.
- Encoding categorical variables
- Creating new features from existing ones
- Improve the quality of your data
- Reduce the risk of overfitting or underfitting

	LoanID	Age	Income	LoanAmount	CreditScore	MonthsEmployed	NumCreditLines	InterestRate	LoanTerm	DTIRatio	Education	EmploymentType	MaritalStatus	HasMortgage	HasDependents	LoanPurpose	HasCoSigner	Default
0	I38PQUQS96	56	85994	50587	520	80	4	15.23	36	0.44	Bachelor's	Full-time	Divorced	Yes	Yes	Other	Yes	0
1	HPSK72WA7R	69	50432	124440	458	15	1	4.81	60	0.68	Master's	Full-time	Married	No	No	Other	Yes	0
2	C1OZ6DPJ8Y	46	84208	129188	451	26	3	21.17	24	0.31	Master's	Unemployed	Divorced	Yes	Yes	Auto	No	1
3	V2KKSFM3UN	32	31713	44799	743	0	3	7.07	24	0.23	High School	Full-time	Married	No	No	Business	No	0
4	EY08JDHTZP	60	20437	9139	633	8	4	6.51	48	0.73	Bachelor's	Unemployed	Divorced	No	Yes	Auto	No	0

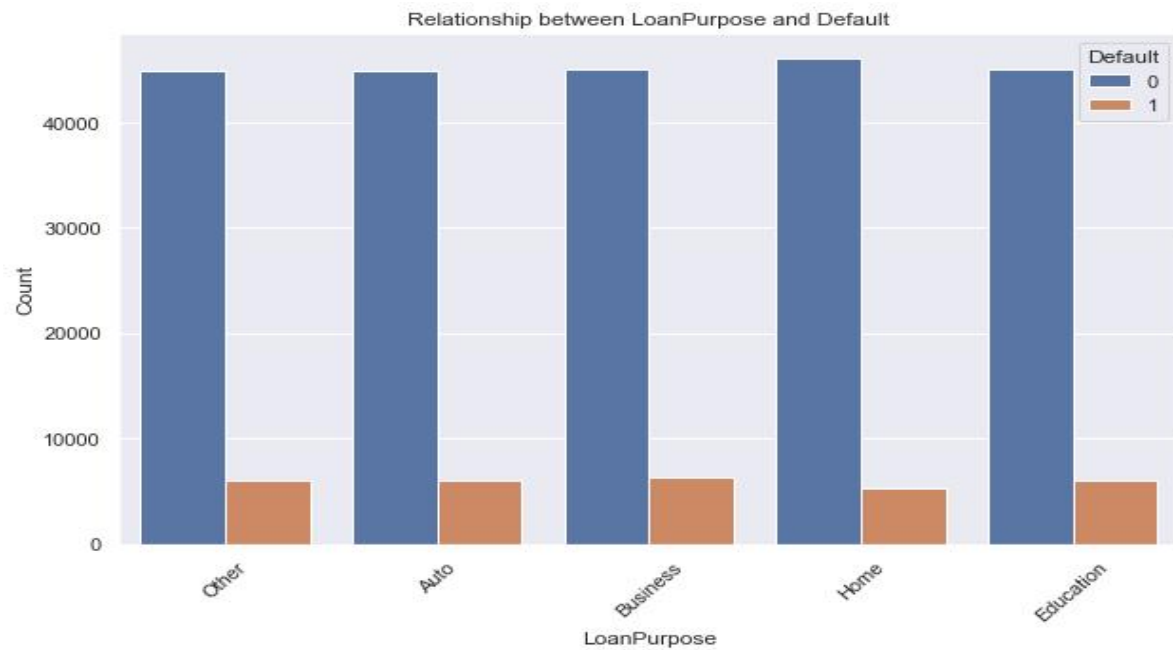
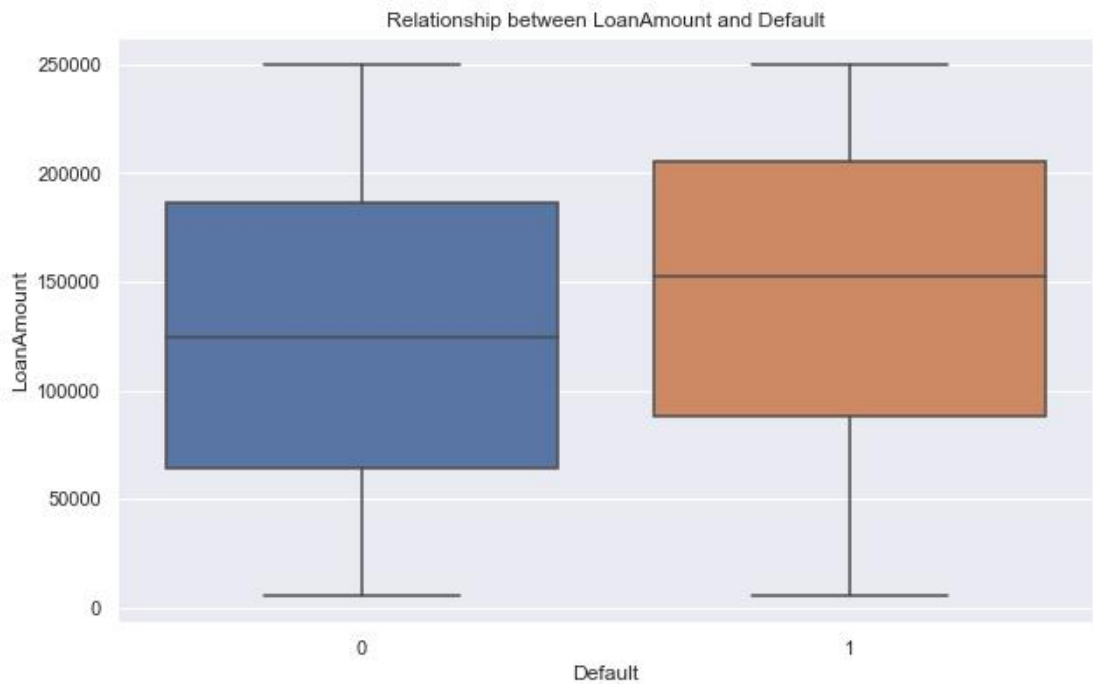
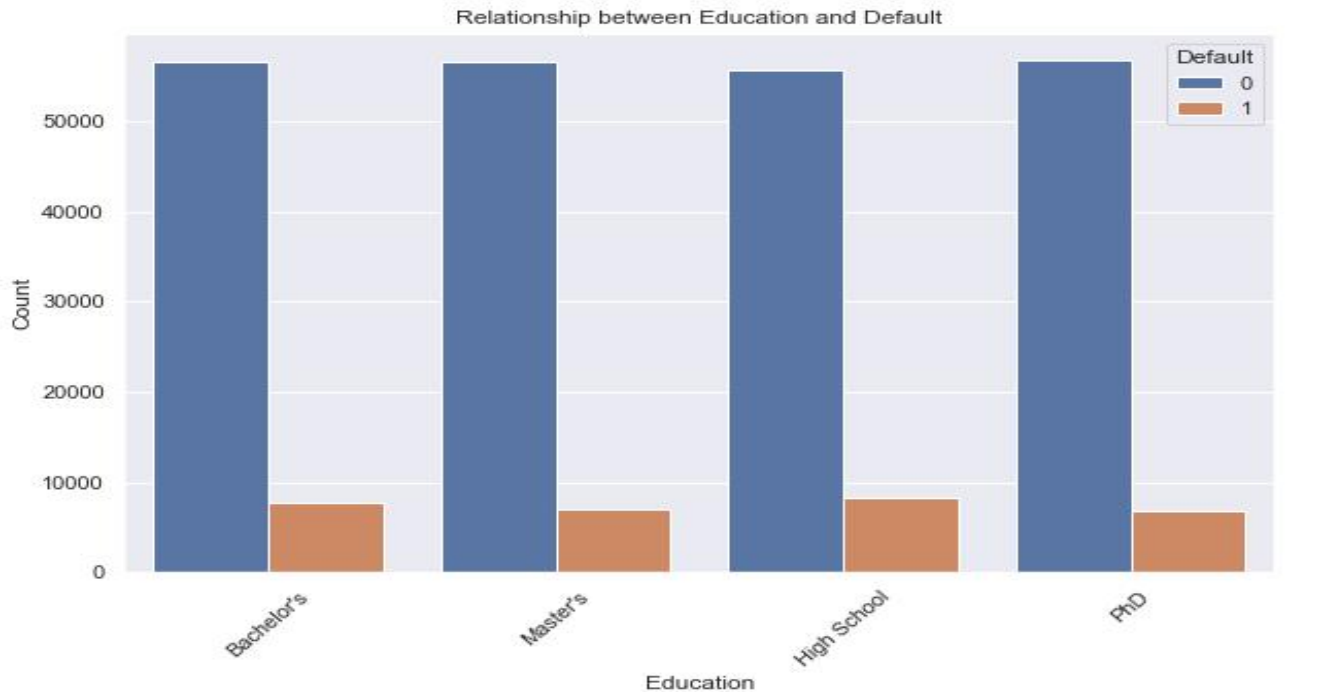
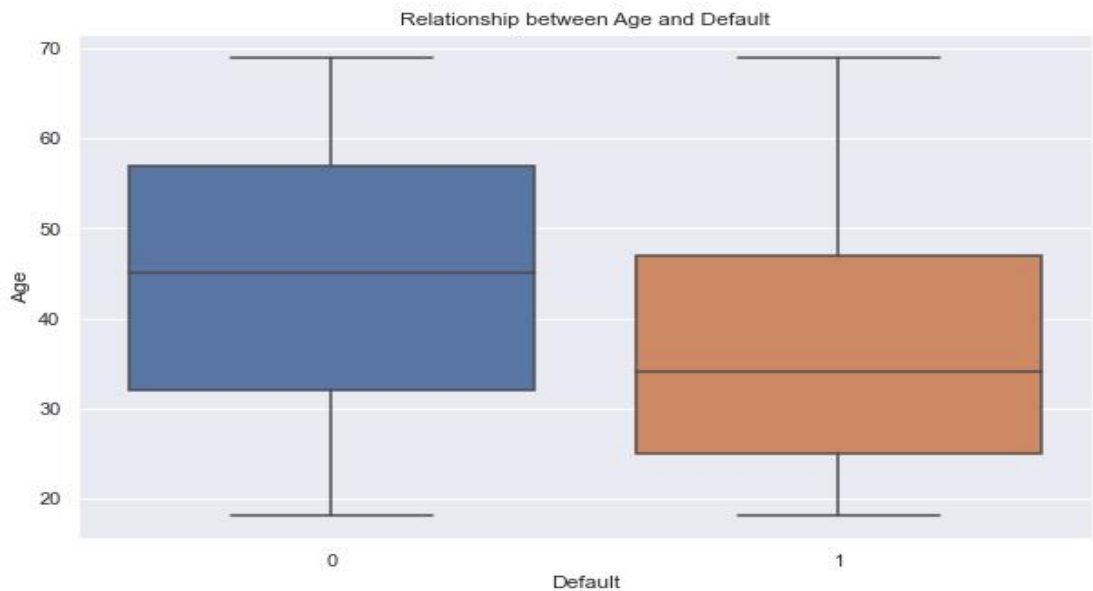


DATA ANALYSIS

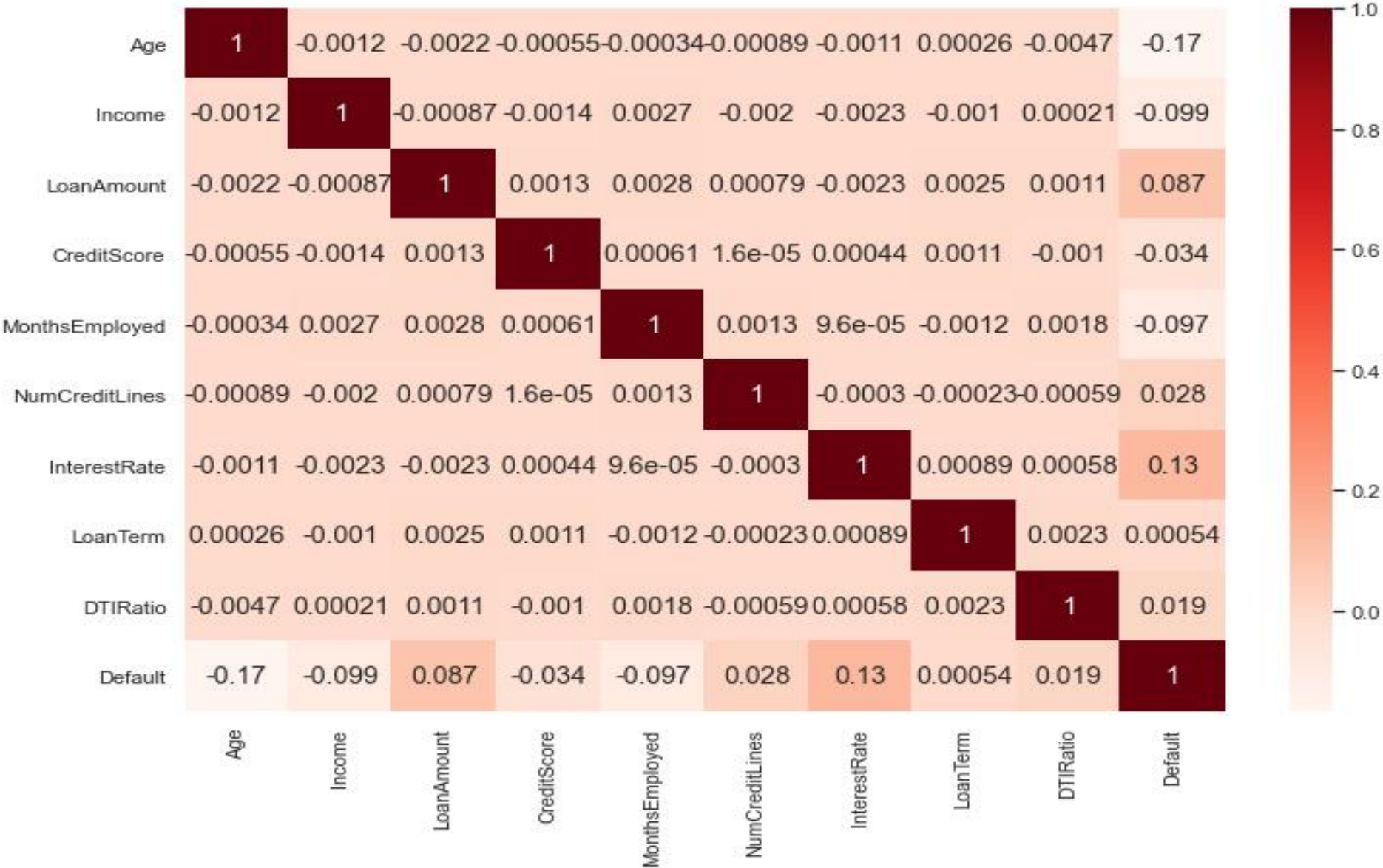
Here we did and Exploratory Data Analysis(EDA) on our data to see the distribution of various features and their relationship with the target variable, Default. Bivariate and Multivariate Analysis were performed.



Bivariate Analysis



Multivariate Analysis



From data analysis:

- Bivariate analysis: We can see that the boxplots show no outliers and we checked that the data had no missing values. The relation default and the other features were evenly distributed.
- Multivariate analysis: The correlation heatmap shows relationship between the numerical features. The feature with the most correlation with the target variable is InterestRate with 0.13



MODEL EVALUATION

Model evaluation is a crucial step in the machine learning process, where we assess the performance of a trained model on unseen data. This process helps us identify strengths and weaknesses, refine our model, and gain confidence in its ability to generalize to new data. By evaluating our model's performance, we can answer questions like how well it performs, whether it's overfitting or underfitting, and whether it's biased or error-prone. In this context, we'll be evaluating three machine learning models: Logistic Regression, Random Forest, and Decision Trees.

Why these models

1. Logistic Regression

A simple and interpretable model that can handle categorical outcomes. It's well-suited for problems with a small number of features and a small to moderate-sized sample size.

2. RandomForest

An ensemble model that combines multiple Decision Trees to improve predictive accuracy and reduce overfitting. Random Forest is particularly effective when dealing with high-dimensional data and can handle noisy or missing values.

3. Decision Trees

A popular model for classification problems, especially when dealing with high-dimensional data. Decision Trees can handle non-linear relationships and are relatively easy to interpret.

A few trade-offs

1. Logistic Regression

Simple and interpretable, but may not perform well on high-dimensional data or non-linear relationships.

2. RandomForest

Excellent for high-dimensional data, but may require more computational resources and may not be as interpretable as individual Decision Trees.

3. Decision Trees

Can handle non-linear relationships, but may suffer from overfitting or lack of interpretability.

Model Evaluation Metrics(Comparison between RandomForest and DecisionTrees after Hyperparameter Tuning)

Model	Precision	Recall	F1-Score	Accuracy	AUC Score
Random Forest	0.277	0.431	0.337	0.804	0.642
Decision Tree	0.184	0.454	0.262	0.704	0.596

The results show that Random Forest outperformed Decision Tree in all evaluation metrics.

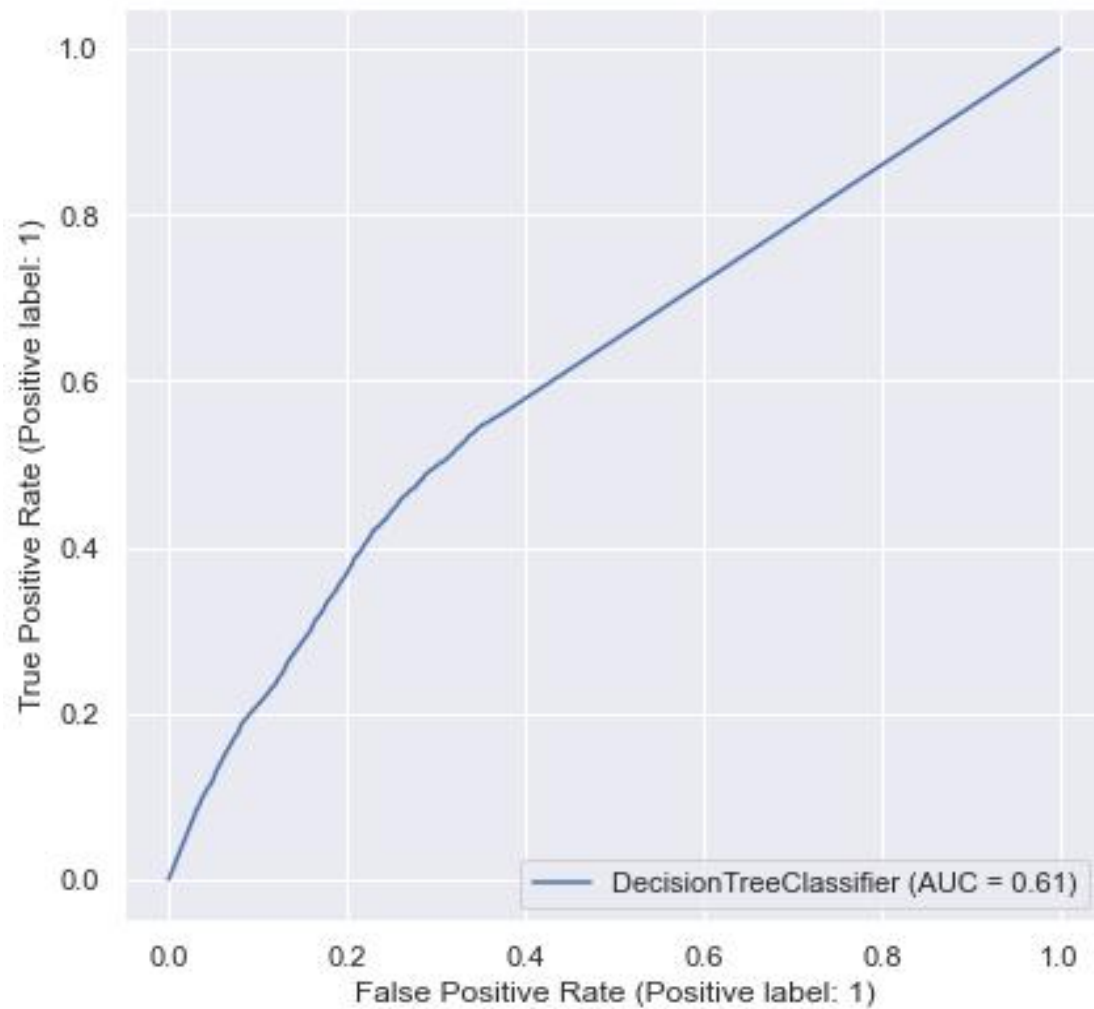
Random Forest demonstrated a more balanced performance, with a better trade-off between precision and recall.

It achieved higher accuracy and a better overall correctness of predictions.

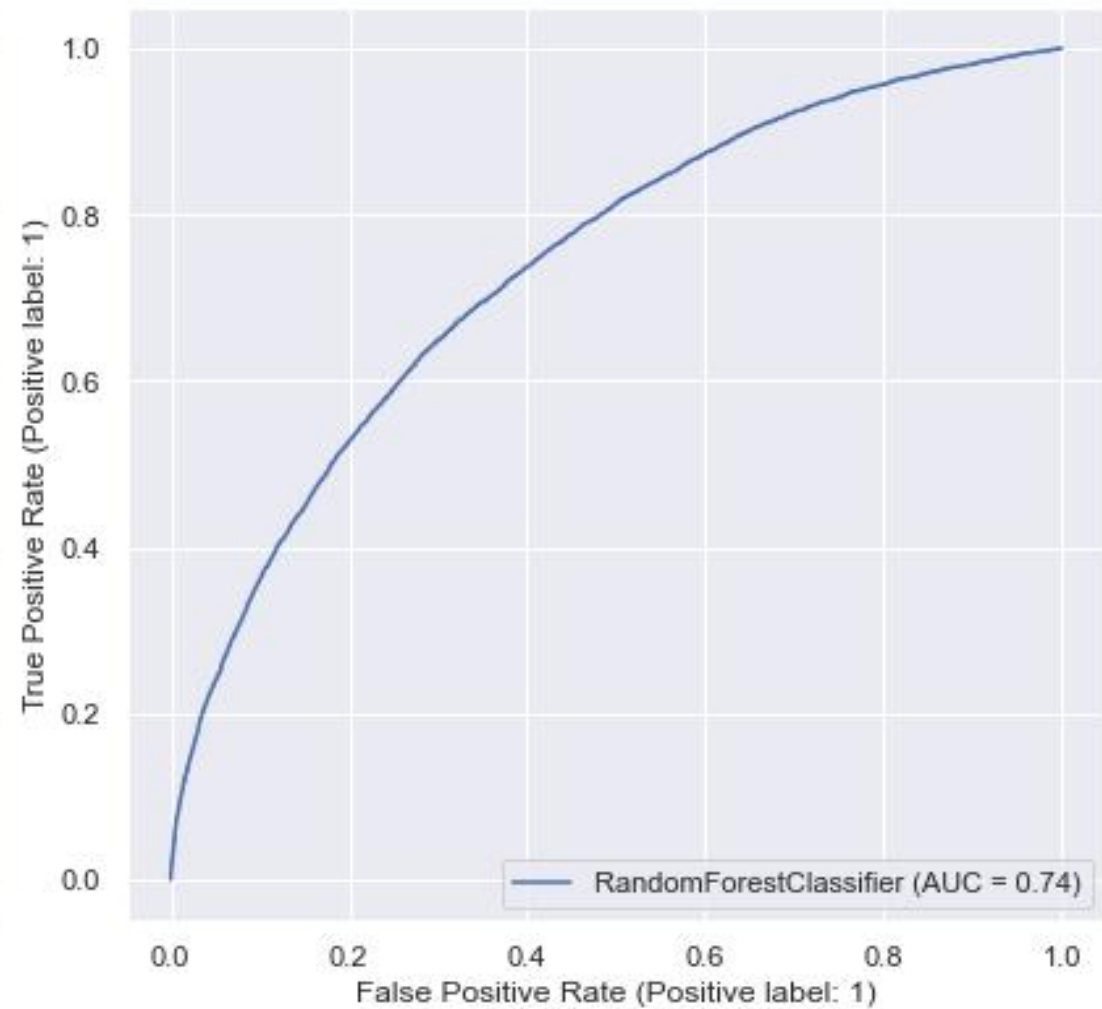
The model also showed better discrimination ability between positive and negative instances.

ROC-Curves

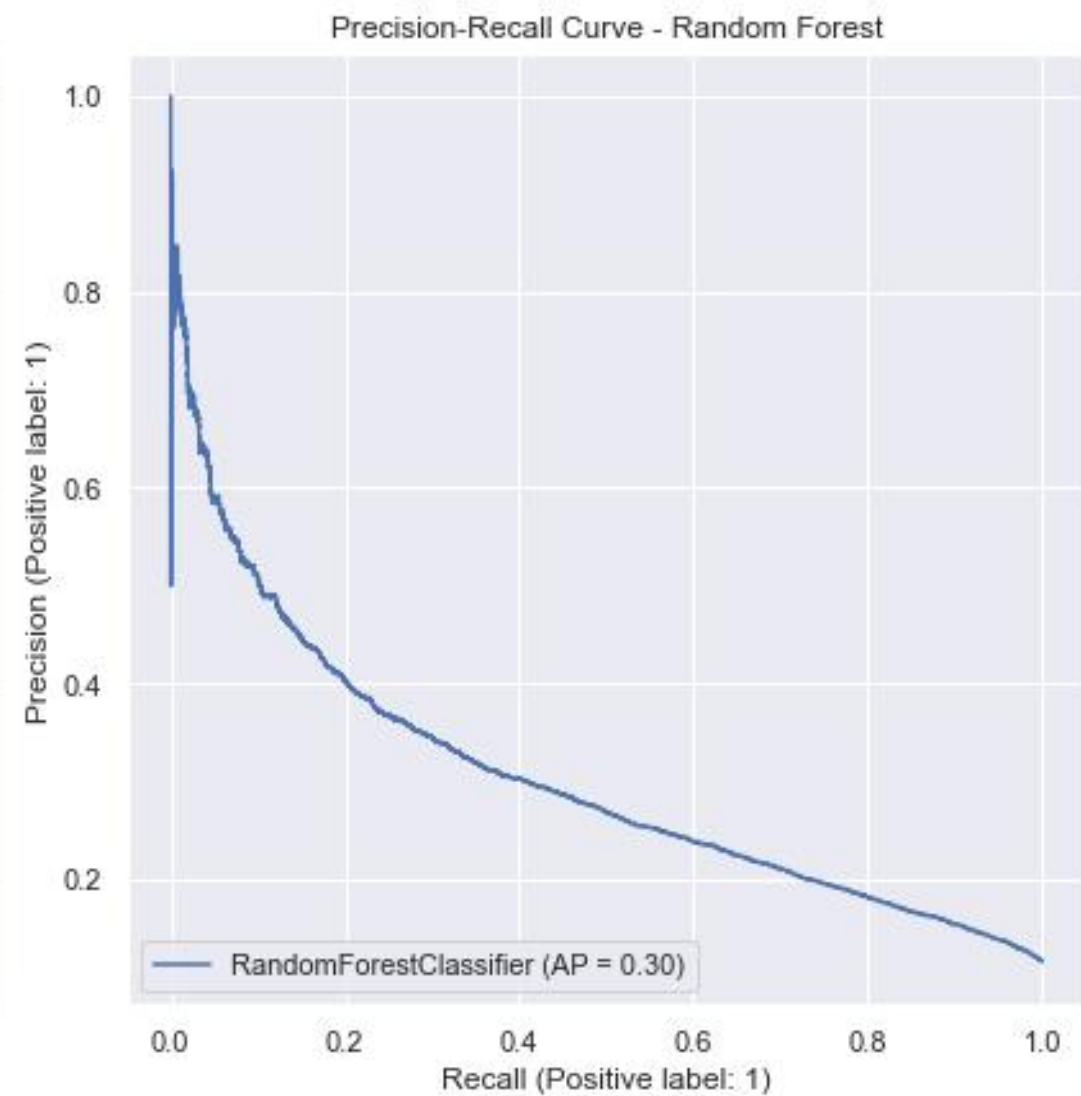
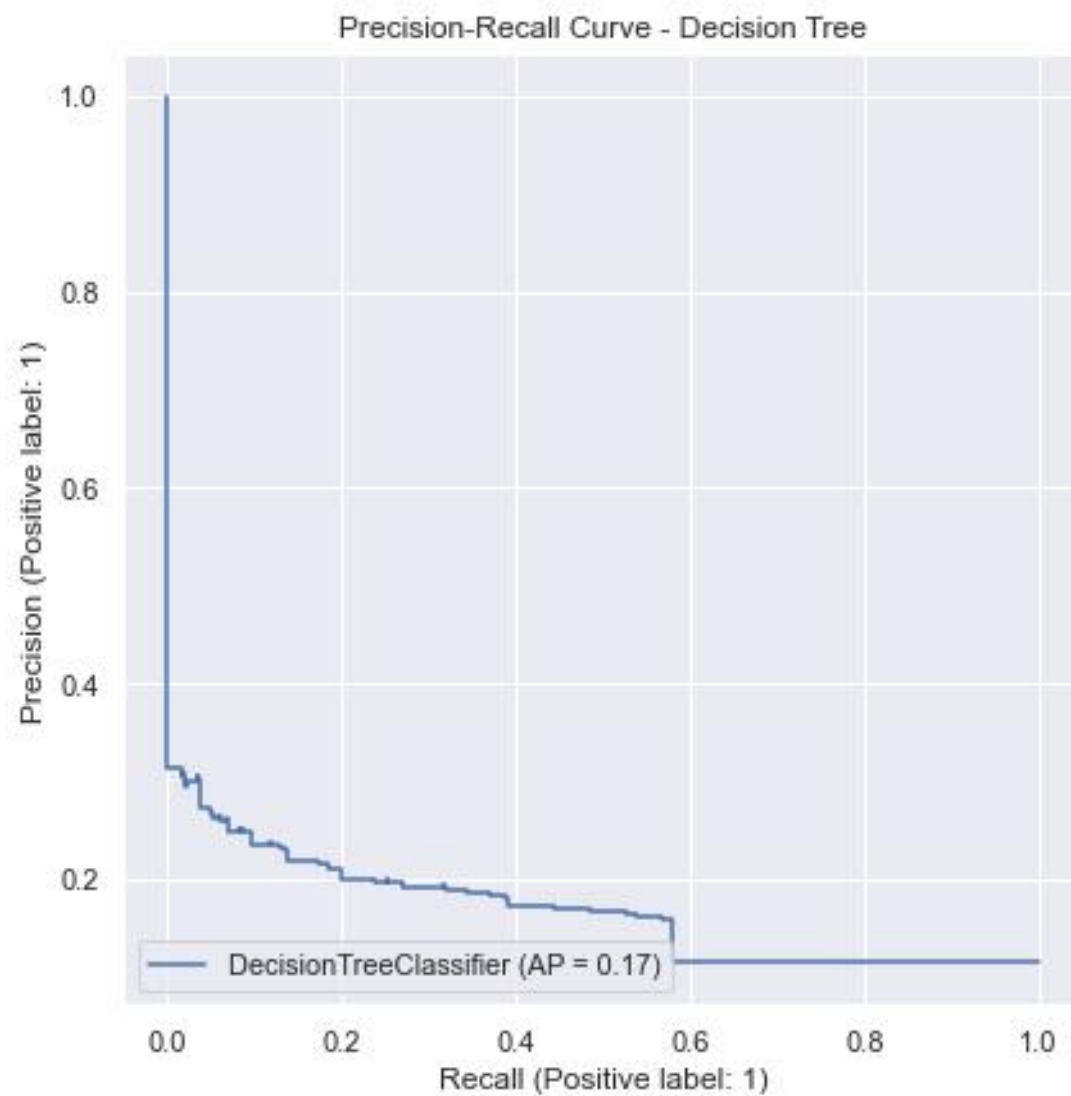
ROC Curve - Decision Tree



ROC Curve - Random Forest



Precision recall curves



Comparison of the Two Models:

ROC Analysis: The Random Forest classifier has a higher AUC (0.74) compared to the Decision Tree classifier (0.61), indicating better overall discriminative ability.

Precision-Recall Analysis: The Random Forest classifier also has a higher Average Precision (AP) score (0.30) compared to the Decision Tree classifier (0.17), indicating better performance in terms of precision and recall, especially useful in imbalanced datasets.

The Random Forest classifier proves once again to be the superior model for this classification task based on both the ROC and Precision-Recall curve evaluations as well.



RECOMMENDATIONS

Based on the evaluations, the Random Forest model emerges as the superior model. Although the model performs well, it is crucial to analyze feature importance and their coefficients through logistic regression to provide financial institutions with actionable insights. This enables better prediction and understanding of the factors influencing loan defaults.

Key Features Affecting Loan Defaults and Solutions (Based on Logistic Regression Coefficients)

Credit Score: A low credit score is a significant risk factor for defaults. Implement stricter credit score requirements for loan approval and provide financial literacy programs to educate borrowers on improving their credit scores.

Income Level: Lower income levels signify a higher likelihood of default. Verify income through reliable documentation and set income thresholds for different loan products. Offer lower loan amounts or different terms to lower-income applicants to ensure repayments are manageable.

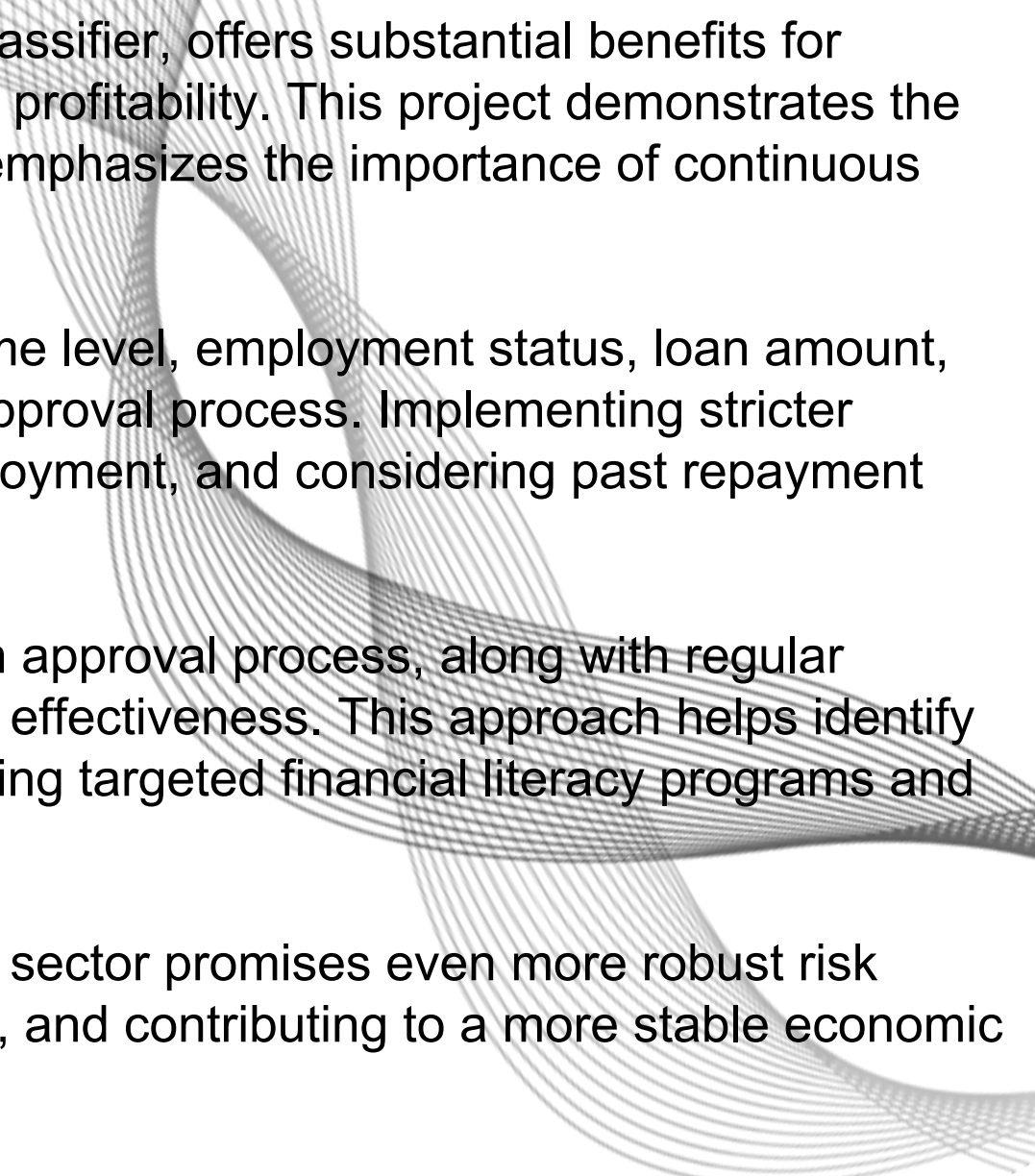
Employment Status: Unstable employment or unemployment increases default risk. Prioritize lending to individuals with stable employment and require employment verification. For those with unstable employment, offer loans with flexible repayment terms or smaller amounts.

Loan Amount: Higher loan amounts correlate with higher default probabilities. Conduct thorough affordability assessments to ensure the loan amount aligns with the borrower's financial capacity. Implement tiered lending where higher amounts are only available to lower-risk borrowers.

Previous Loan History: Previous defaults or late payments strongly predict future defaults. Utilize previous loan repayment behavior as a critical factor in the approval process. Offer financial counseling to borrowers with a troubled repayment history to help them improve their financial habits before approving

CONCLUSION





Integrating machine learning, especially the Random Forest classifier, offers substantial benefits for financial institutions, such as lower default rates and improved profitability. This project demonstrates the effectiveness of these models in predicting loan defaults and emphasizes the importance of continuous model improvement and data quality.

Identifying and addressing key features like credit score, income level, employment status, loan amount, and previous loan history can significantly enhance the loan approval process. Implementing stricter requirements, verifying documentation, prioritizing stable employment, and considering past repayment behavior are effective risk mitigation strategies.

The integration of these machine learning models into the loan approval process, along with regular updates and feedback loops, ensures sustained accuracy and effectiveness. This approach helps identify high-risk borrowers and provides valuable insights for developing targeted financial literacy programs and personalized loan products.

Further refinement of machine learning models in the financial sector promises even more robust risk management solutions, benefiting both lenders and borrowers, and contributing to a more stable economic environment.

THANKS

As we continue to advance in the field of machine learning and data science, the potential to revolutionize financial services grows exponentially. By leveraging these technologies, financial institutions can not only enhance their risk assessment capabilities but also offer more personalized and fair loan products to their customers. The ongoing commitment to innovation and data-driven decision-making will pave the way for a more resilient and inclusive financial ecosystem.

