Data driven Loan Approval: A Machine Learning Approach

Project phase : Methods, Findings and Recommendations

Date of Presentation: 03/25/2025
Akshaya Waddepally
aw93269n@pace.edu
Class Name: Practical Data Science
Program Name: MS in Data Science
Seidenberg School of Computer Science and Information Systems
Pace university

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modeling methods
- Findings
- Business recommendations and technical next steps
- Appendix

Executive Summary

- **Business Problem:**

Our company is facing challenges in efficiently assessing loan eligibility. The current process is manual and inconsistent, leading to delays, customer dissatisfaction, and potential financial risks due to incorrect approvals or rejections. We need a data-driven approach to enhance decision-making and reduce risk.

- **Proposed Solution:**

To address these challenges, we analyzed historical loan data to identify key factors influencing loan approvals. We then built and tuned a machine learning model that predicts the likelihood of a loan being approved based on applicant data. This data-driven approach improves the accuracy of loan approval decisions, reducing human error and financial risks. The model's performance was validated to ensure its reliability before deployment, enabling more efficient and fair decision-making for loan approvals.

Project plan recap

Deliverable	Due Date	Status
Data & EDA	03/25/2025	Complete
Methods, Findings and Recommendations	04/01/2025	Complete
Final Presentation	04/22/2025	In Progress

Data

Data

- **Data Source :** https://www.kaggle.com/code/johnpaulchikwe/loan-eligibility-prediction-using-machine-learning#Loan-Eligibility-Prediction-Using-Machine-Learning
- **Sample Size :** Approximately 614 loan applications.
- **Time Period :** Data spans across a defined period (Time period unknown).
- **Assumptions :**

-The dataset reflects real-world loan approval conditions, with consistent criteria for all applicants.

-The data is assumed to be representative of the general loan applicant population, without significant biases or underrepresentation

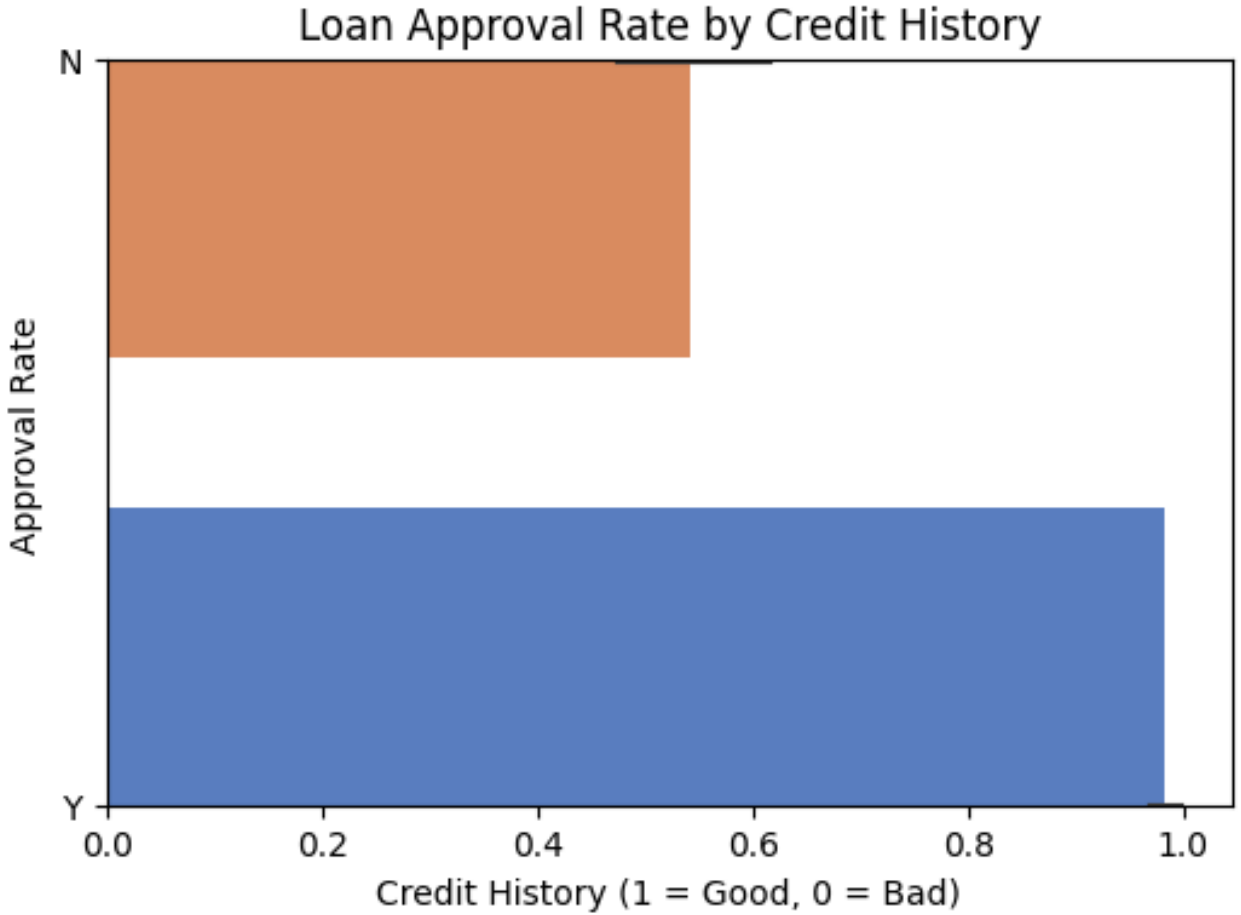
- **Key Inclusions & Exclusions:**

-Included: Applicant demographics, credit history, loan amount, property area, and employment details**.**

-Excluded: Personally identifiable information (PII) to maintain privacy.

Exploratory Data Analysis

The Impact of Credit History on Loan Approval Rates

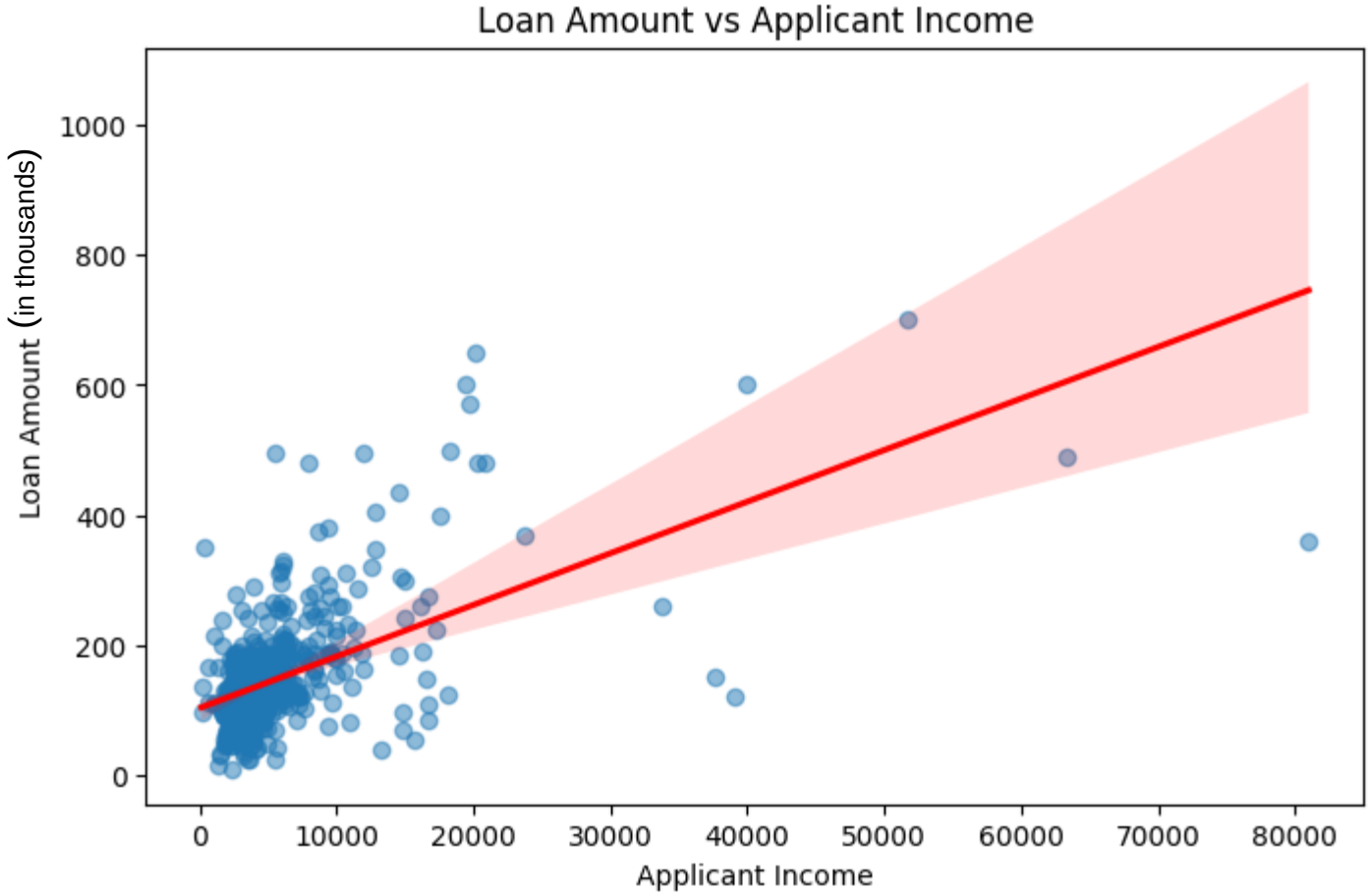


📌 Strong Correlation – Applicants with good credit history have significantly higher approval rates.

📌 Poor Credit = High Rejection – Most applicants with bad credit history are denied.

📌 Credit History is a Key Factor – Plays a crucial role in loan approval decisions

How Income Influences Loan Amount Decisions

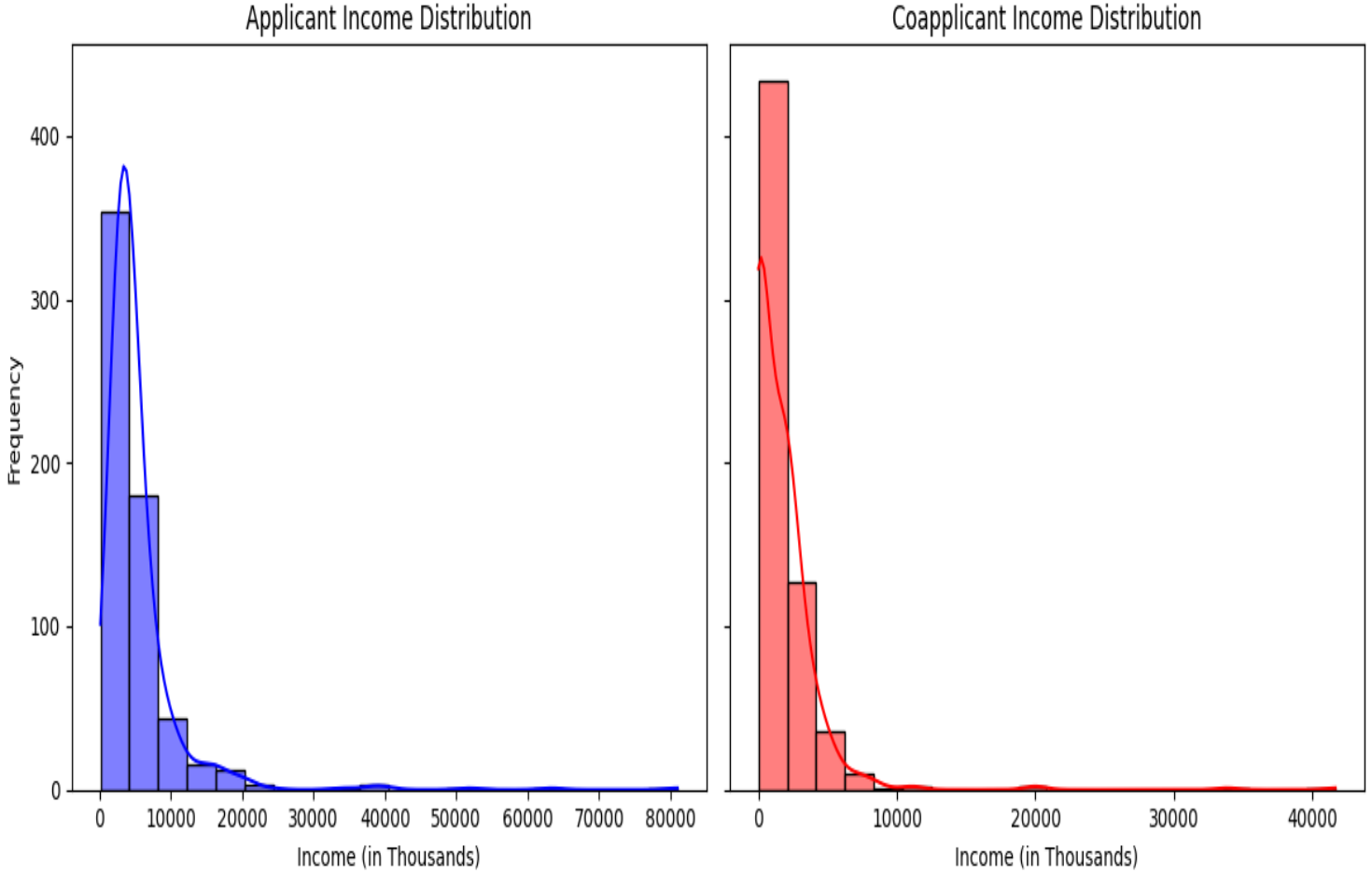


📌 **Positive Correlation**: Loan amounts tend to increase with higher applicant income, but the relationship is not perfectly linear.

📌 **Variance in Loan Approvals**: Some applicants with lower incomes receive higher loan amounts, indicating possible influence from other factors (co-applicant income, credit history, etc.).

📌 **Risk Identification:** Outliers suggest cases where high-income applicants receive relatively lower loans and vice versa.

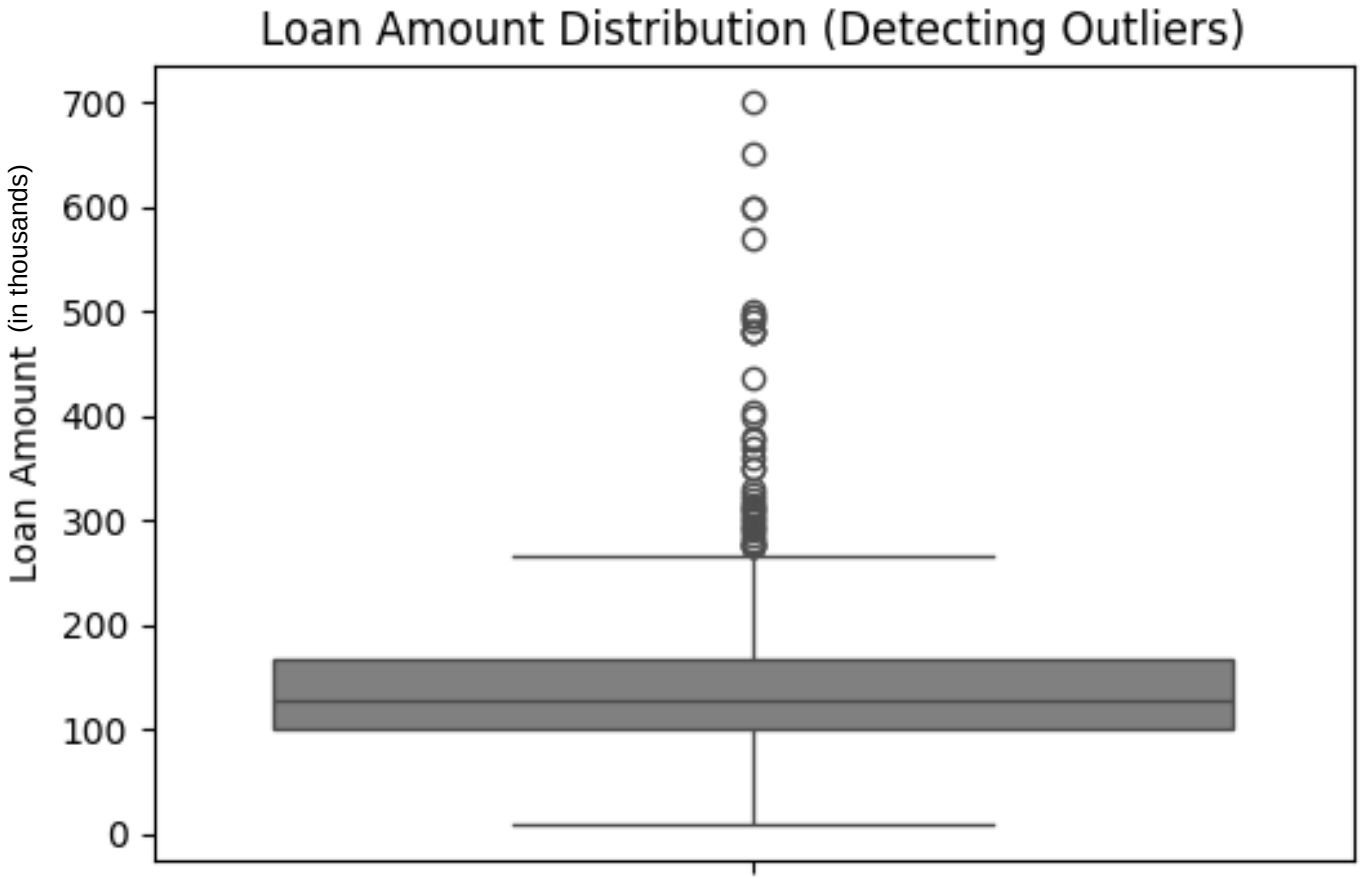
Distribution of **Applicant Income vs. Coapplicant Income**.



- Most applicants and coapplicants have lower incomes (Right-skewed distribution)
- Applicant incomes are higher than coapplicants'.
- Coapplicants earn significantly less, making applicant income crucial for loan approval.

- **Suggestion**: Combined income should be considered in loan predictions.

Loan Amount Distribution: Identifying Outliers



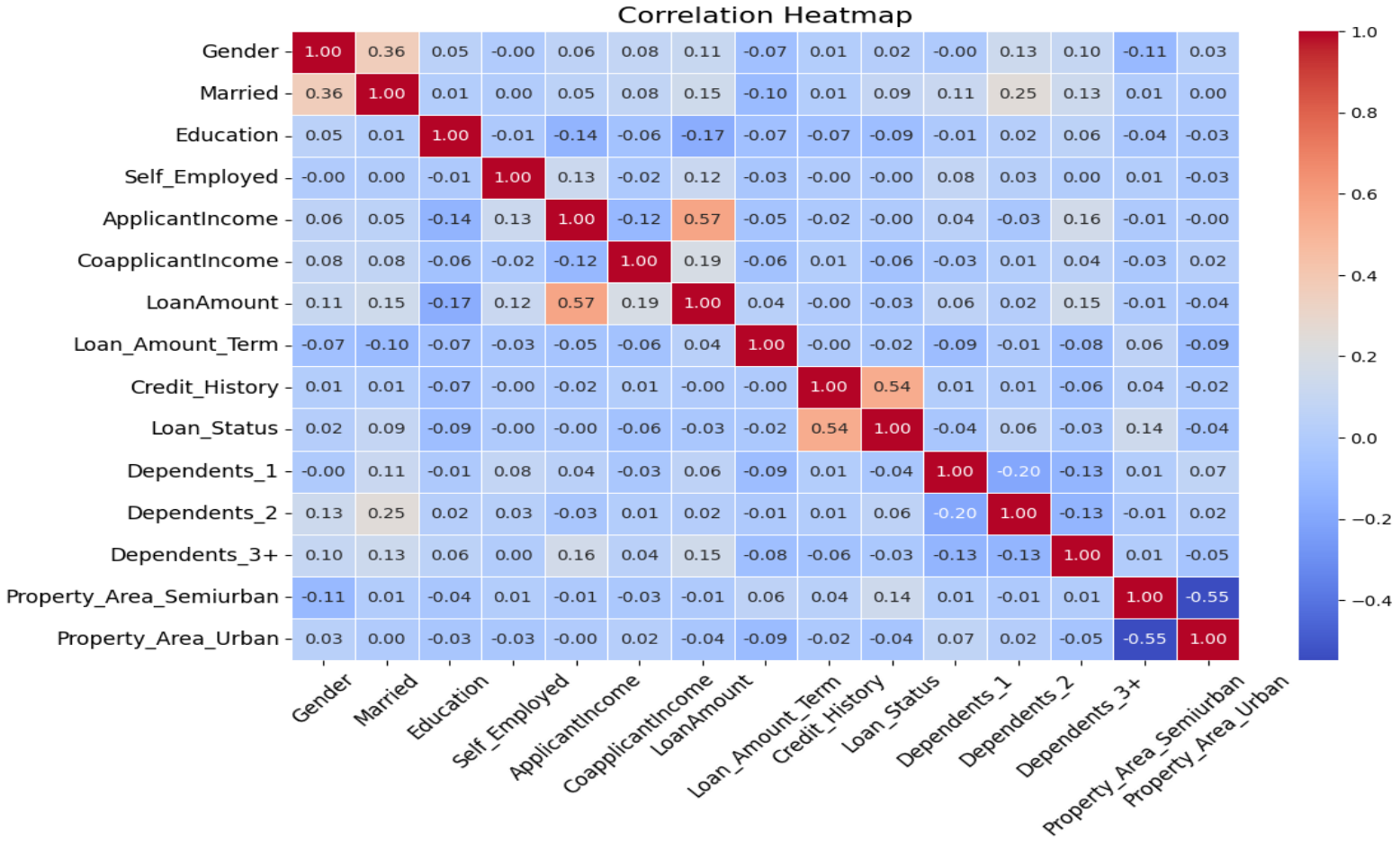
Key takeaways:

-Many high-value loan amounts are outliers, visible as dots above the box.

-The majority of loans fall within a reasonable range (~100-250 units).

-The distribution is right-skewed, meaning most loans are on the lower end, but some extreme values push the average up.

How Loan Variables Interact: A Heatmap Breakdown



-Weak correlations suggest some features may have less predictive power.

-Applicant Income & Loan Amount (0.57) show strong correlation, meaning **higher income leads to larger loan amounts.**

-Credit History & Loan Status (0.54) indicate **credit history significantly impacts loan approval.**

- Click here for some extra visualizations for a more insightful understanding of the data

Modeling methods

Predicting Loan Approvals Using Data

1) Outcome Variable (What We're Predicting):

- We are predicting **whether a loan application will be approved (Loan Status).**
- This is a binary classification problem, meaning the outcome is either Approved (Yes) or Not Approved (No).

Why this matters: Helps our financial institution automate decision-making and improve loan approval efficiency.

2) Features Used (Factors Considered for Prediction)

- **Applicant's Income & Co-applicant's Income** → Higher income increases approval chances
- **Loan Amount & Loan Term** → Larger loans may be riskier
- **Credit History** → Strong predictor of loan repayment behavior
- **Property Area & Dependents** → Captures risk factors for different applicants

3) Why Logistic Regression?

- Logistic Regression is best suited for this problem because:

-It's simple and interpretable – We can understand how each feature affects approval chances.
-It handles binary classification well – Since we're predicting Yes/No, it calculates the probability of approval.
-It works well with smaller datasets – Unlike complex models, it doesn't require huge amounts of data.

4) How the Model Works?

- Step 1: It looks at past loan applications and their outcomes.
- Step 2: It finds patterns between applicant details and approvals.
- Step 3: When a new application comes in, it calculates the probability of approval. If the probability is high (e.g., > 50%), the loan is approved; otherwise, it's rejected.

For a detailed technical explanation of the model:

- Click here

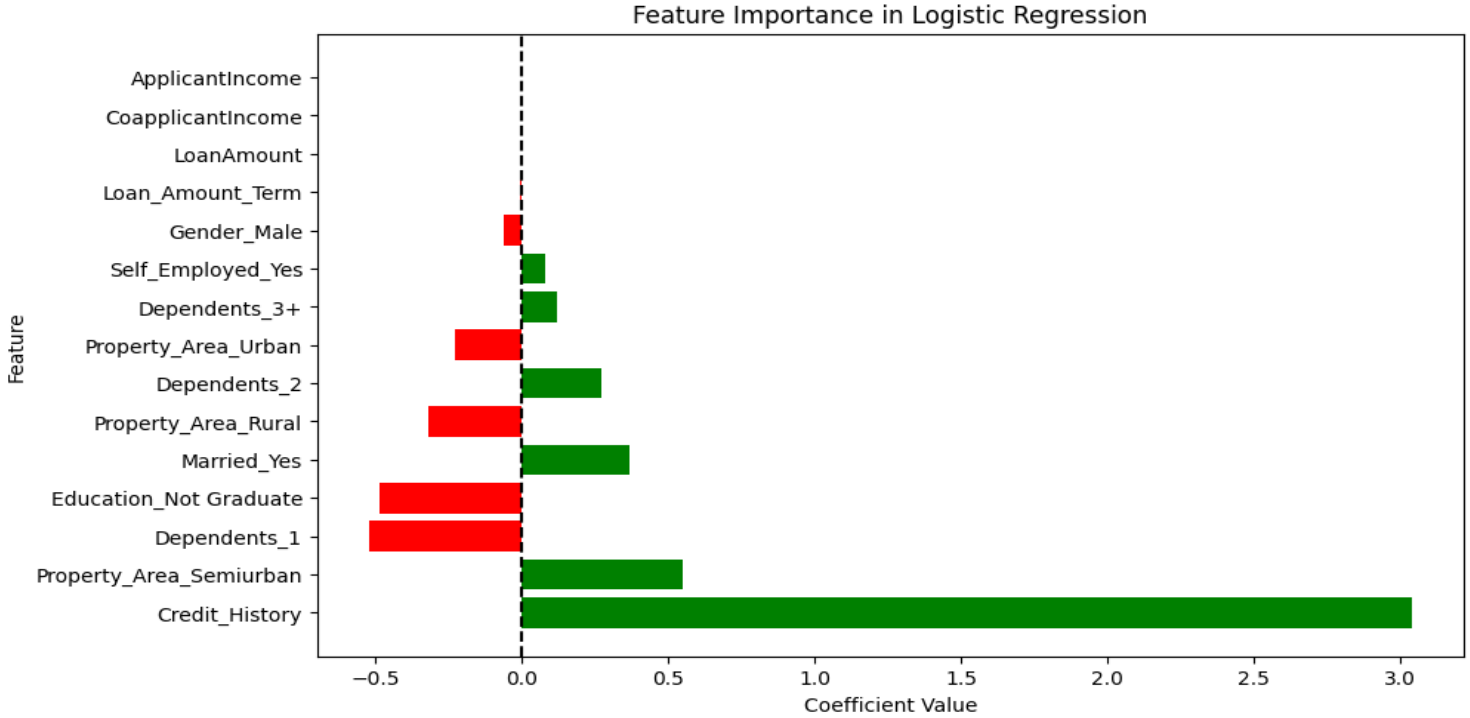
Findings

Key Modeling Findings & Business Insights

1) Key Results & Business Impact:

- Model Performance Overview: The logistic regression model achieved 79% accuracy, with a strong balance between precision (0.83) and recall (0.79), indicating reliable predictive performance for loan approval decisions.

2) Feature Importance Chart:



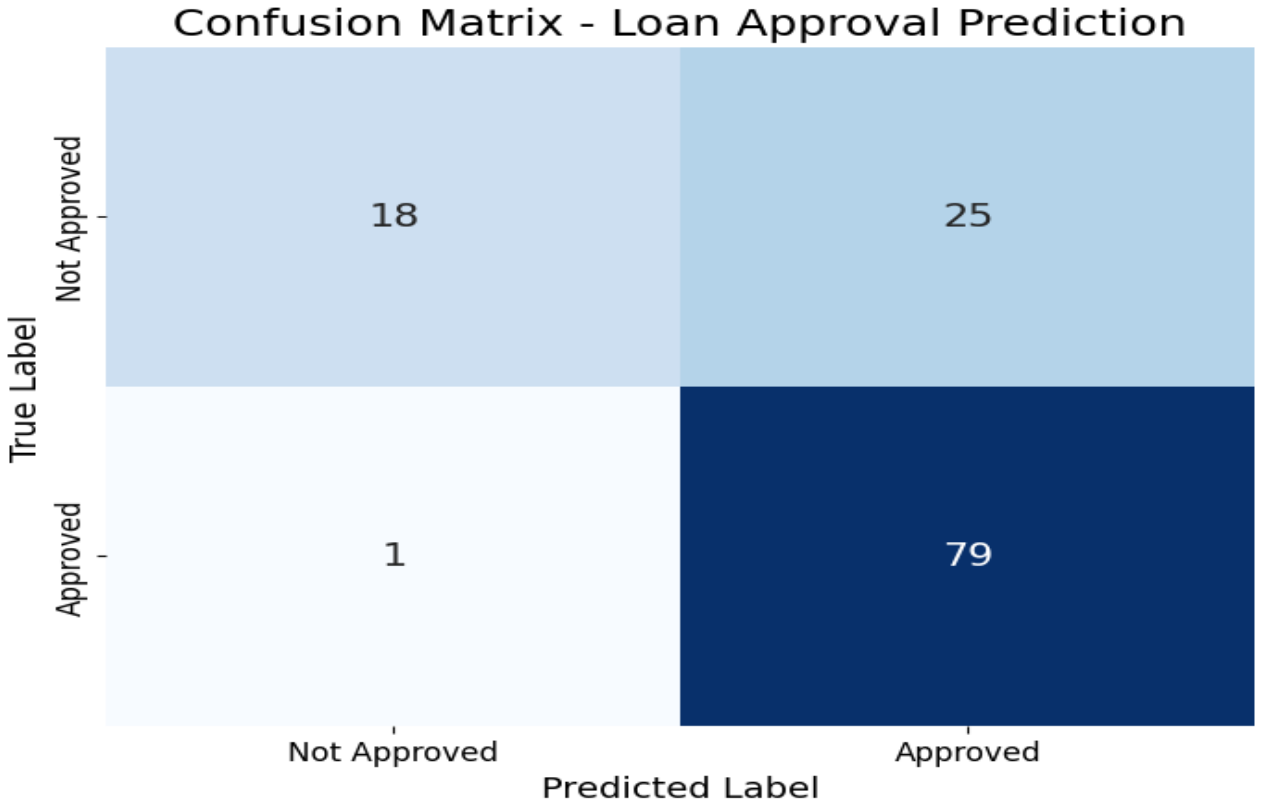
Key Takeaways:

- **Credit History Dominates** - By far the strongest predictor of loan approval (highest positive coefficient).
- **Income Matters** - Both applicant and co-applicant income significantly impact decisions.
- **Urban vs. Rural Bias** - Urban applicants have a slight edge over semiurban/rural.
- **Surprise Negatives** - Being self-employed or non-graduate shows negative effects.
- **Married Advantage** - Married applicants get a modest boost in approval odds.

Key Modeling Findings & Business Insights

Key Confusion Matrix:

- Shows the number of correct and incorrect predictions, demonstrating the model's effectiveness in classifying approvals vs. rejections.:



- The model achieved high accuracy (79.8%) with a strong recall (99%), effectively identifying approved loans.
- Precision (76%) indicates a moderate rate of false positives, which means some risky loans may still be approved.
- Suitable for scenarios prioritizing loan approval maximization but may need adjustments if minimizing risk is the goal.

Final Verdict: Deployable?

✅ Recommended for Production with Caution

The model shows **promising accuracy (79.8%)** and **high recall (99%)**, effectively capturing most approved loans. However, the **moderate false positive rate (25 cases)** indicates a risk of approving some non-eligible loans.

•Suitable for scenarios focused on **maximizing loan approvals**.

•Potential risk due to **misclassification of non-eligible loans**.

•Further improvements like **threshold tuning or using ensemble methods** can help balance accuracy and precision.

While the model is **deployable with caution**, evaluating the **financial impact of false positives** is recommended before production use.

Business Recommendations & Technical Next Steps

Business Recommendations

- **Strengthen Risk Assessment:** Since false positives pose financial risks, introduce an additional validation layer for high-risk applicants (e.g., manual review for borderline cases).
- **Enhance Data Collection:** Gather more granular financial data (like debt-to-income ratio) to improve prediction accuracy.
- **Monitor Model Performance:** Set up continuous performance monitoring to detect model drift or unexpected trends that could impact loan approval accuracy.

Data Science Next Steps

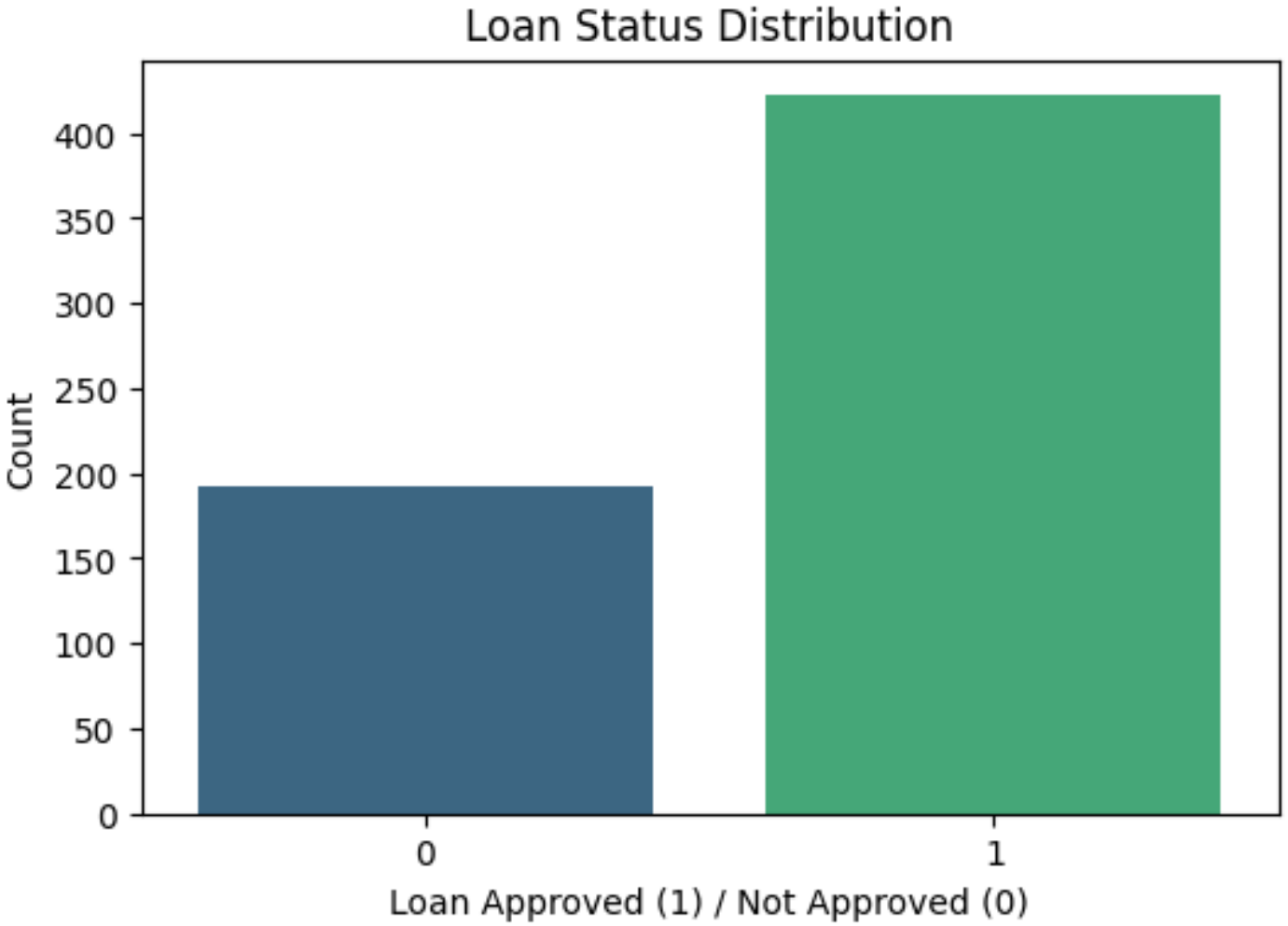
- **Model Optimization**: Experiment with ensemble methods (e.g., Random Forest, XGBoost) to reduce false positives while maintaining recall.
- **Feature Engineering**: Incorporate new features such as employment stability metrics or additional financial indicators.
- **Deploy with Caution:** Start with a pilot deployment to gauge real-world performance and refine the model before full-scale rollout

Appendix

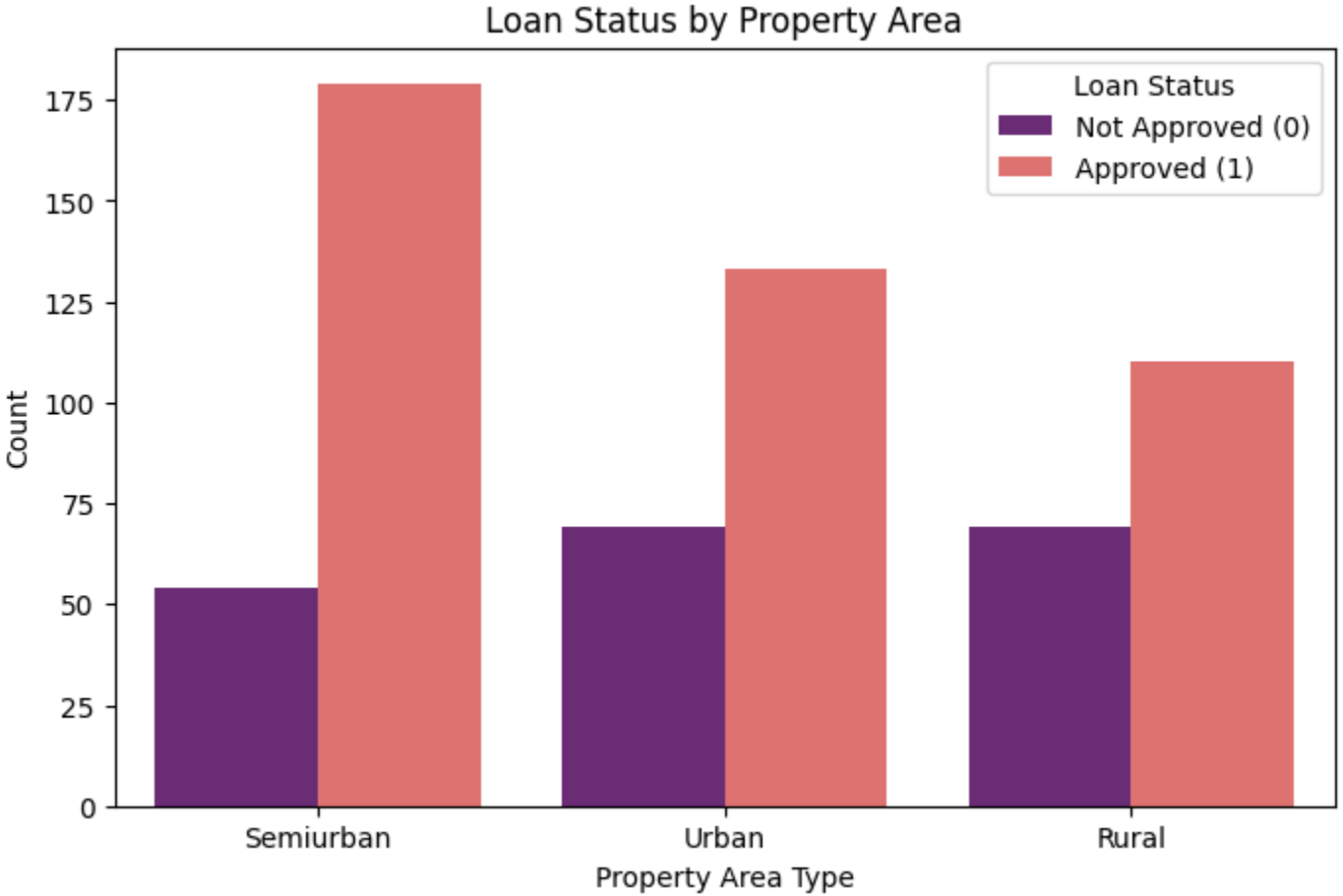
Data

📌 **Important Notes :**

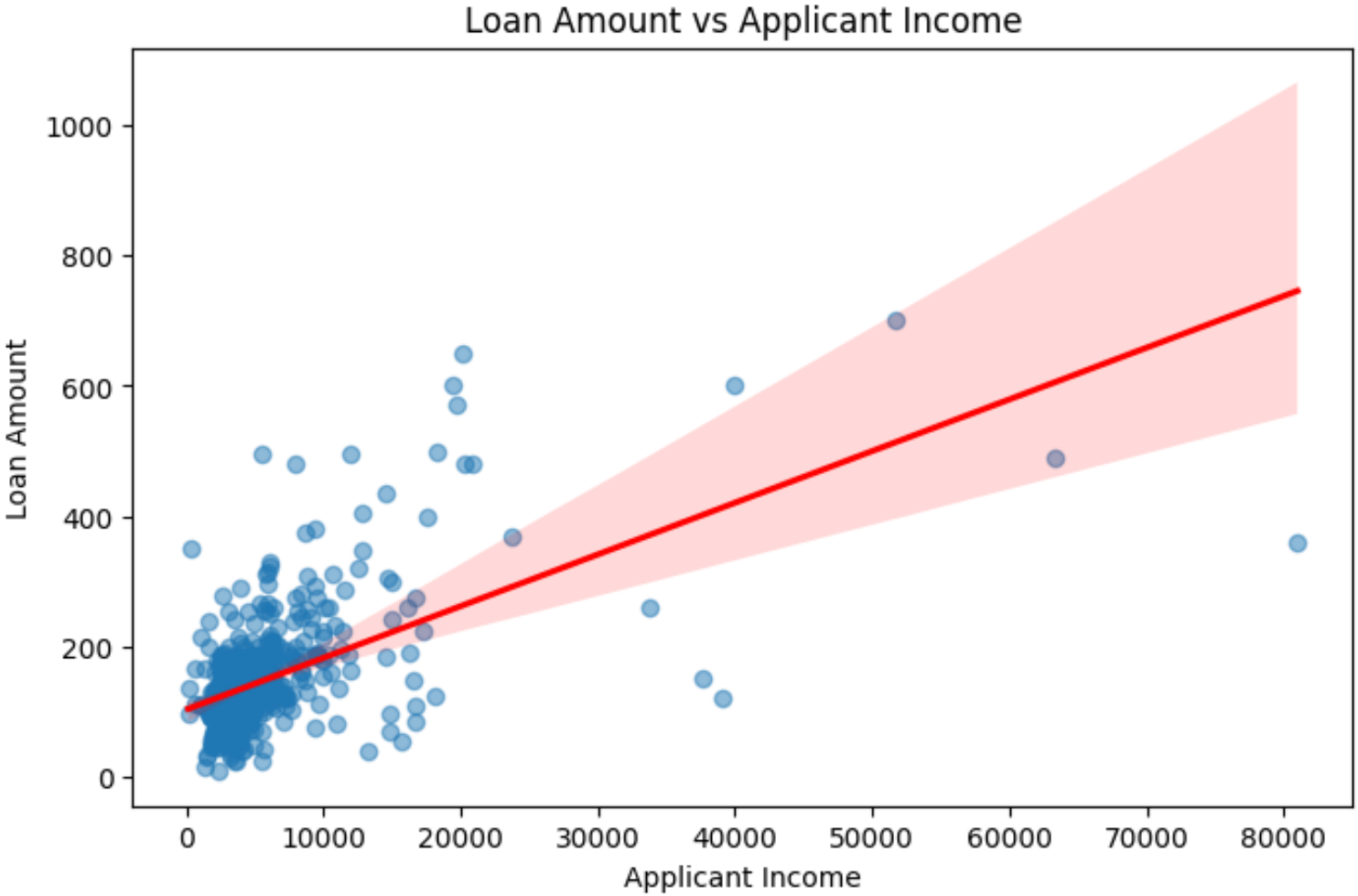
- **Data was preprocessed to handle missing values and inconsistencies.**
- **Imbalanced loan approval distribution was addressed to improve model fairness.**



- **More loans are approved than rejected**, indicating a positive approval trend.
- **Significant difference in approval vs. rejection rates**, suggesting influencing factors.
- **Loan approval bias or key factors are to be analyzed further** to understand decision patterns.



- **Highest approvals in semiurban areas**, suggesting better loan approval rates.
- **Urban & rural areas have lower approval counts**, but similar rejection numbers.
- **Loan approval trends vary by location**, indicating geographic influence on approvals.
- **Impact**: Property area type should be considered in loan predictions.



📌 Loan Amount Increases with Income but with significant variance.

📌 Outliers Exist – some high-income applicants receive smaller loans and vice versa.

📌 Income Alone Isn't Enough – other factors (credit history, DTI, co-applicant income) likely influence loan decisions.

Logistic Regression: Technical Overview

1) Why Logistic Regression?

- The business problem requires a binary classification of loan approval (Approved/Not Approved), making Logistic Regression a natural choice.
- It provides probabilistic outputs, which allow us to assess confidence in predictions, unlike tree-based models that produce hard classifications.
- Given our dataset size (~614 samples), simpler models like Logistic Regression generalize better, avoiding overfitting compared to complex models like Random Forest.

2) Feature Importance & Model Interpretability

- Key Features Impacting Loan Approval (based on logistic regression coefficients):

Feature	Coefficient	Interpretation
• Credit History	• +3.0384	The strongest predictor—applicants with a good credit history are much more likely to be approved.
• Property Area Semiurban	• +0.5492	Applicants from semi-urban areas have a slightly higher chance of approval compared to rural/urban areas.
• Married Yes	• +0.3688	Being married slightly increases approval chances.

Logistic Regression: Technical Overview

3) Model Performance & Justification (Comparison with Other Models)

Model	Accuracy	Presicion	Recall	F-1 score
Logistic Regression	0.79	0.83	0.79	0.76

Performance Metrics:

- Accuracy (0.79): The model correctly predicts loan approvals and rejections 79% of the time. This indicates a strong overall predictive capability.
- Precision (0.83): Of all the loans the model predicted as approved, 83% were actually approved. This is important because we want to minimize false positives (incorrectly approving a loan).
- Recall (0.79): The model captures 79% of all actual loan approvals. A higher recall ensures that we don't miss too many eligible loan applicants.
- F1-Score (0.76): A balance between precision and recall, ensuring that both false positives and false negatives are minimized effectively.

Justification for Logistic Regression:

- Interpretability: The model provides clear insights into feature importance, helping stakeholders understand why a loan was approved or denied.
- Performance: A strong balance between precision and recall makes this model reliable for making loan approval decisions.
- Scalability: Logistic regression is computationally efficient and scales well with large datasets, making it ideal for financial institutions handling numerous applications.

GO BACK TO THE FIRST SLIDE

- Click here to head to the first/main slide of the PPT