

Exploratory Data Analysis for Loan Default Prediction CASE STUDY

Chand Rayee
Kusum Chanchala



CONTENTS OF THIS TEMPLATE

You can delete this slide when you're done editing the presentation

<u>1. Introduction</u>	Predict loan defaulting using binary classification with features like amount, rate, term, credit history, employment, and income.
<u>2. Data Cleaning and Preprocessing</u>	Address missing data, transform categories to numbers, manage outliers, and apply data scaling or normalization techniques.
<u>3. Univariate Analysis</u>	Conduct univariate analysis: examine feature distribution, calculate summary statistics, and detect skewness and outliers.
<u>4. Bivariate Analysis</u>	Investigate feature-target correlations, visualize with plots, calculate correlation coefficients, and test group differences using chi-square and ANOVA.
<u>5. Multivariate Analysis</u>	Explore feature interactions using multivariate analysis, apply PCA or t-SNE, and visualize with correlation or scatter plot matrices.
<u>6. Feature Selection and Importance</u>	Determine key predictors of loan default using methods like correlation, feature elimination, and tree importance; assess business impact and clarity.



TABLE OF CONTENTS

01

ABSTRACT

DISCUSSION

04

02

INTRODUCTION

CONCLUSION

05

03

**CASE
PRESENTATION**

ROADMAP

06



01

ABSTRACT





ABSTRACT

- **Loan Default Prediction Study**

- **Dataset Overview**

- Entries: 39,717
- Columns: 111
- Features: Loan amount, interest rate, term, credit history, employment, income

- **Objective**

- Predict loan default (charged-off vs. fully paid/current)

- **Data Preprocessing**

- Removed columns with >90% missing data
- Imputed values, converted categories, addressed outliers

- **Exploratory Data Analysis (EDA)**

- Univariate: Distributions, summary statistics
- Bivariate: Relationships with loan status
- Multivariate: High-dimensional visualization (PCA, t-SNE)

- **Feature Selection**

- Techniques: RFE, tree importance
- Outcome: Identified key predictors

- **Modeling**

- RandomForestClassifier
- Achieved high accuracy

- **Conclusion**

- Insights on factors influencing defaults
- Impact on risk assessment and lending decisions



INTRODUCTION

BACKGROUND

Analyze factors causing loan defaults, clean data, preprocess, conduct EDA, and develop a model to predict potential defaulters.

IMPORTANCE

Predictive modeling in loan analysis is vital for assessing borrower risk, guiding lending decisions, and preventing financial losses.



2. Data Cleaning and Preprocessing

Initial Dataset Overview

Initial Dataset Overview:

- **Entries:** 39,717
- **Columns:** 111 (many with significant missing values)

Steps Taken:

Remove Columns with >90% Missing Values:

- Reduced dataset to a more manageable size.
- Remaining columns still had some missing values.

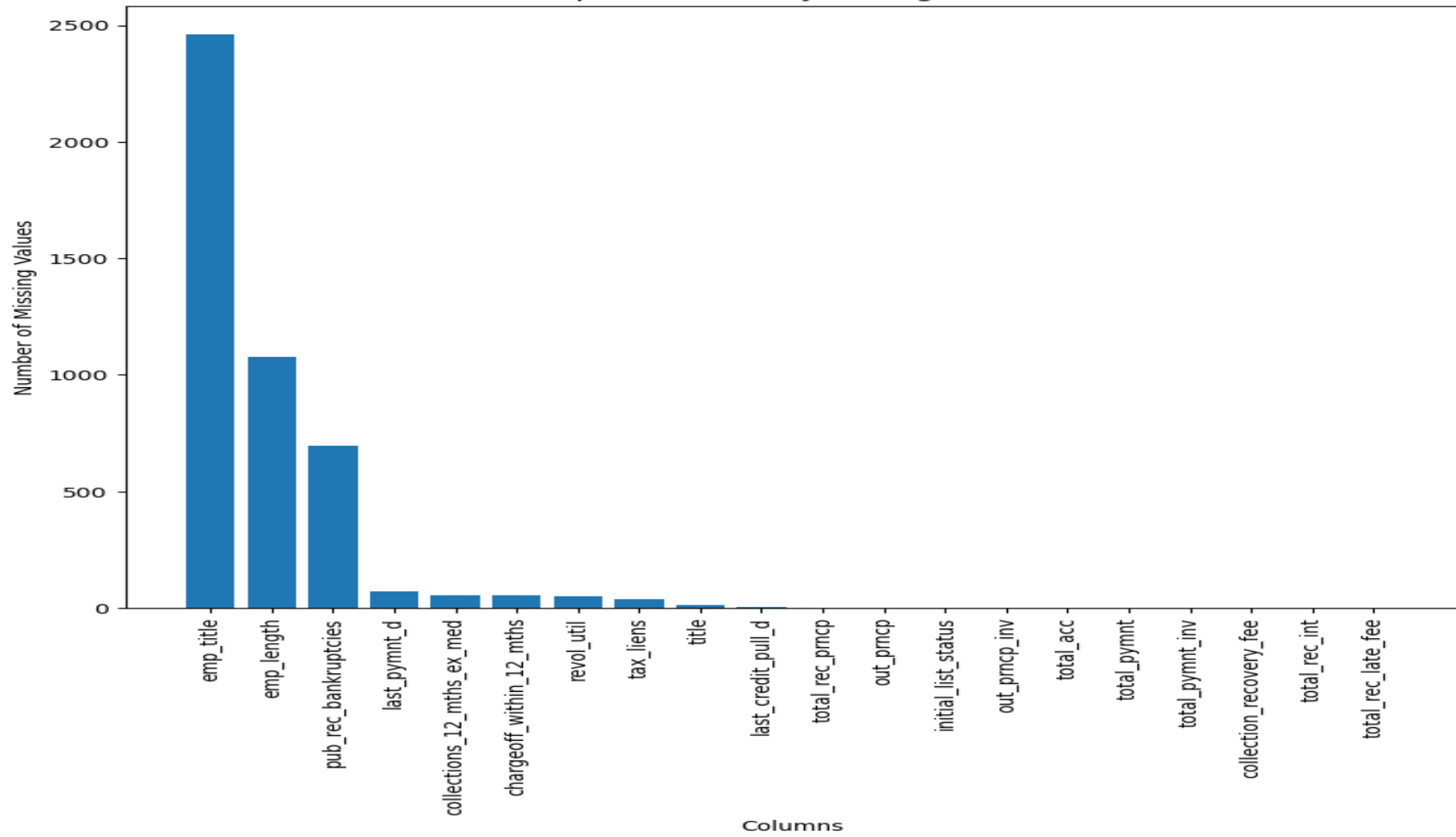
Handle Remaining Missing Values:

- **Numerical Columns:** Imputed with median values.
- **Categorical Columns:** Imputed with 'Unknown' and encoded using Label Encoder.

Outlier Detection and Removal:

- Used Interquartile Range (IQR) method to remove outliers from numerical columns.

Top 20 Columns by Missing Values





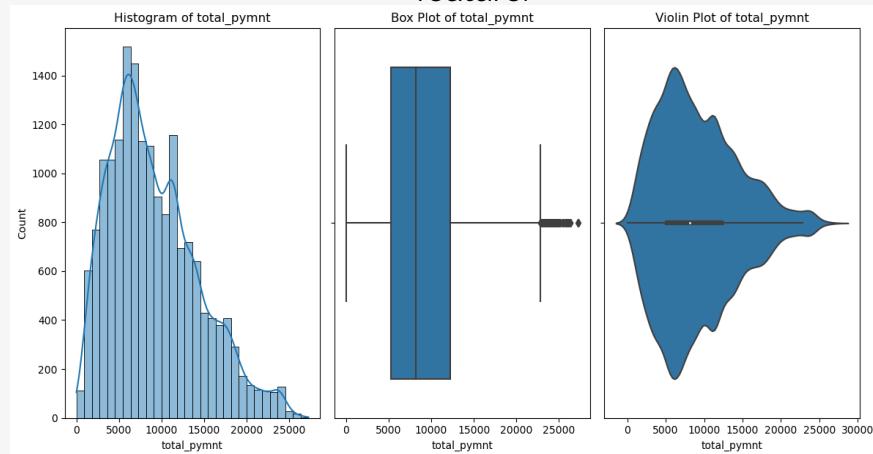
3. Univariate Analysis

Summary Statistics:

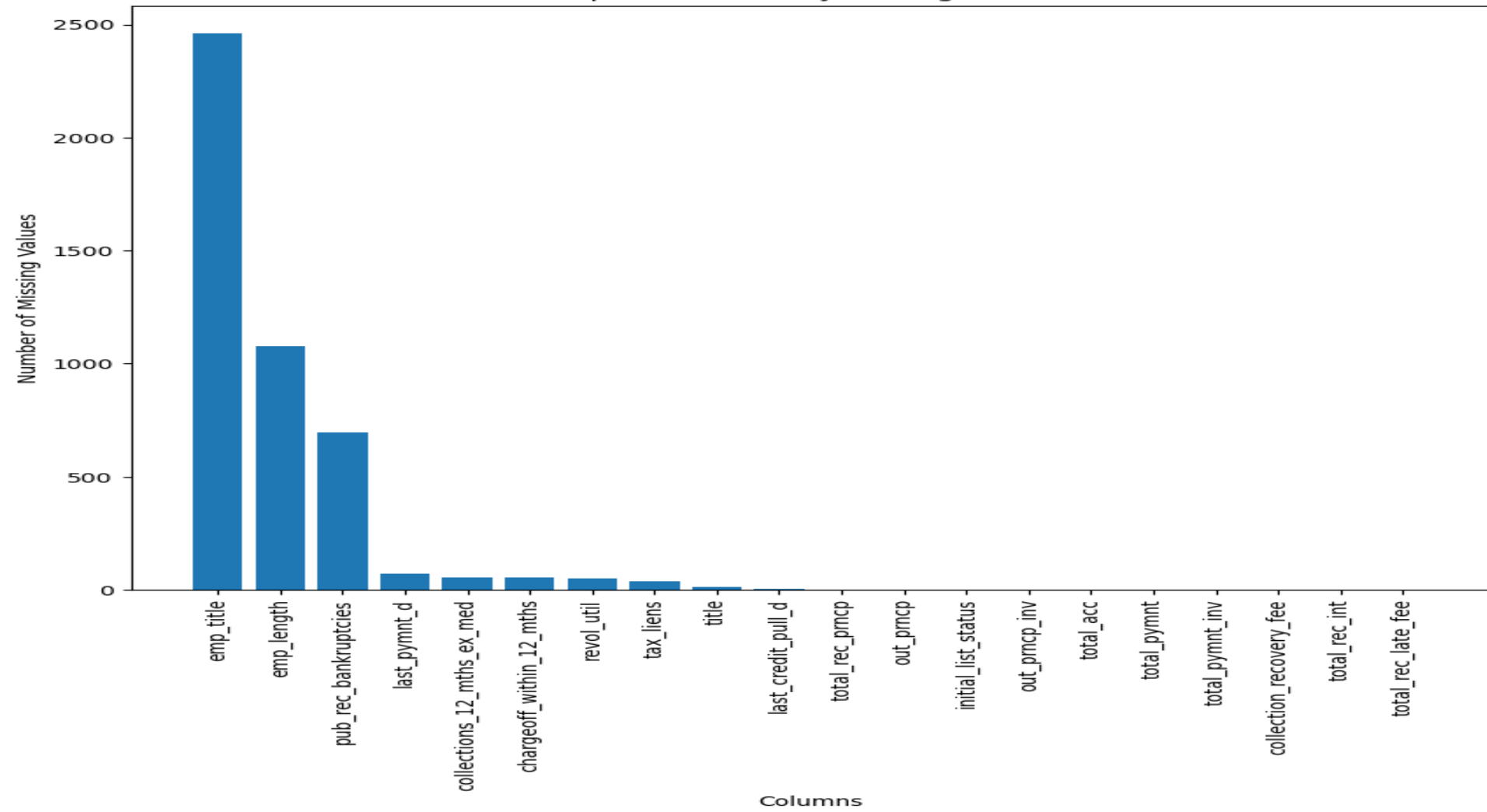
- Provided summary statistics for numerical features, including mean, median, mode, standard deviation, and skewness.
- Identified skewness and potential outliers in the data.

Visualizations:

Histograms, Box Plots, Violin Plots: Used to analyze the distribution of each numerical feature.



Top 20 Columns by Missing Values





4. Bivariate Analysis

Target Variable: Loan Status

•Numerical Features:

Visualized relationships using scatter plots and box plots.

•Categorical Features:

Visualized relationships using bar plots.

Correlation Analysis:

- Calculated correlation coefficients between numerical features and the target variable.
- Used scatter plots and correlation matrices to explore relationships.

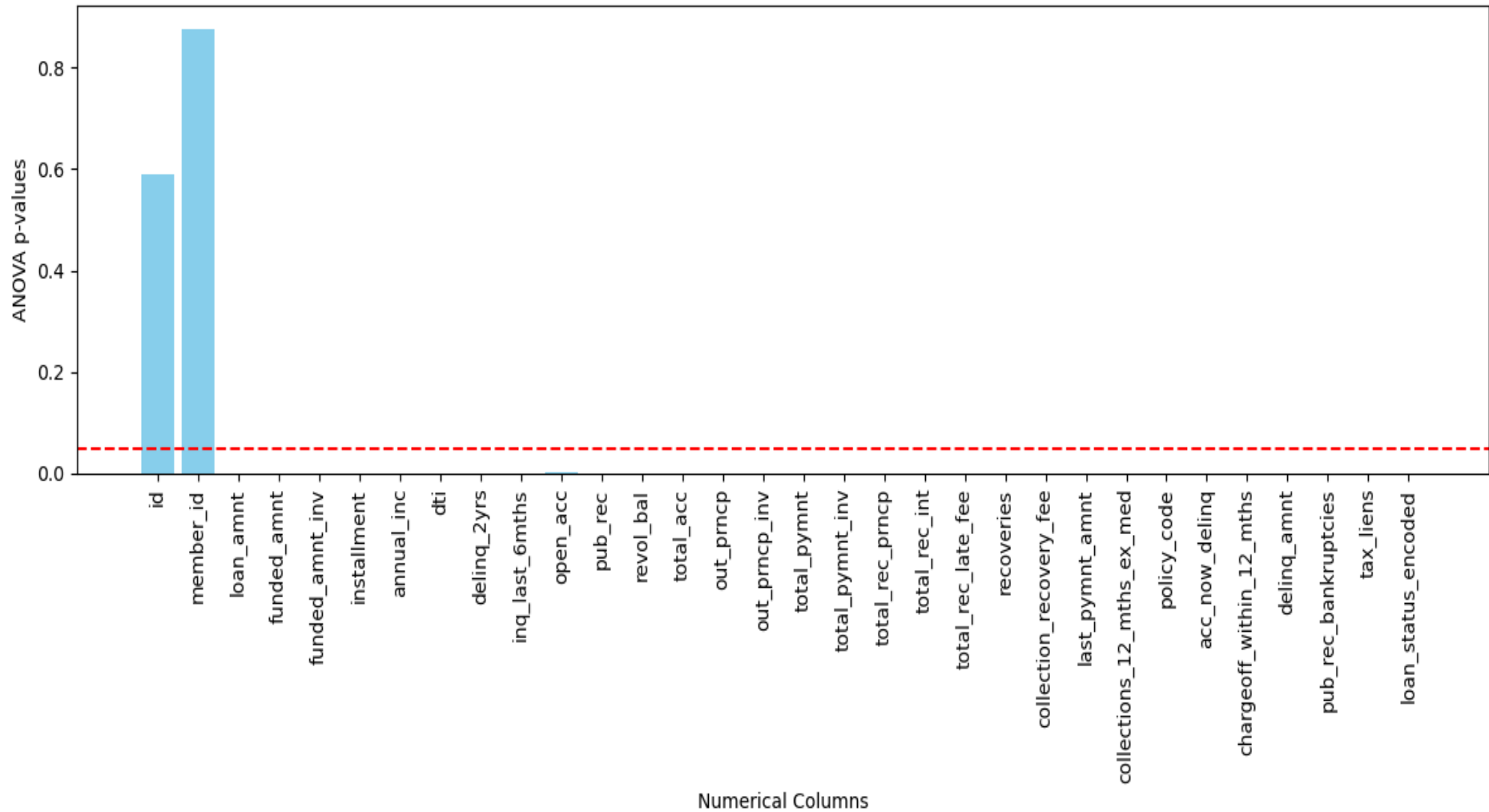
Statistical Tests

•Chi-square Tests:

Performed for categorical variables.

•**ANOVA Tests:** Performed for numerical variables to check for significant differences between groups.

ANOVA Test P-Values for Numerical Variables





5. Multivariate Analysis

Techniques Used:

Correlation Matrix:

Visualized interactions among multiple features using a heatmap.

Scatter Plot Matrix:

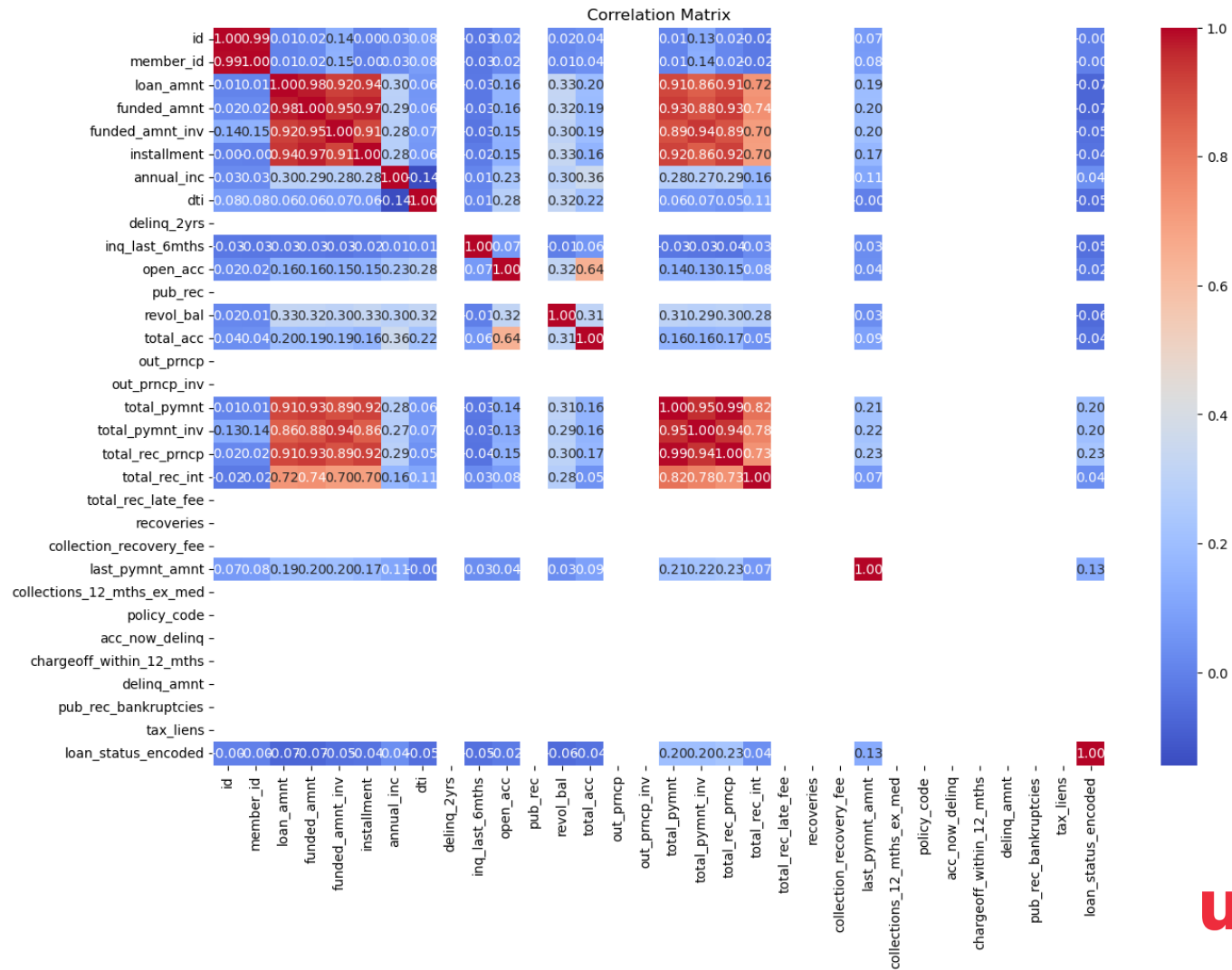
Used pair plots to explore relationships between features.

Parallel Coordinates Plot:

Visualized multidimensional data to identify patterns.

Dimensionality Reduction:

- **PCA (Principal Component Analysis):** Reduced dimensions and visualized data in 2D.
- **t-SNE (t-distributed Stochastic Neighbor Embedding):** Visualized high-dimensional data in 2D.





6. Feature Selection and Importance

Techniques Used:

Correlation Analysis:

Identified highly correlated features and visualized feature correlation matrix.

Recursive Feature Elimination (RFE):

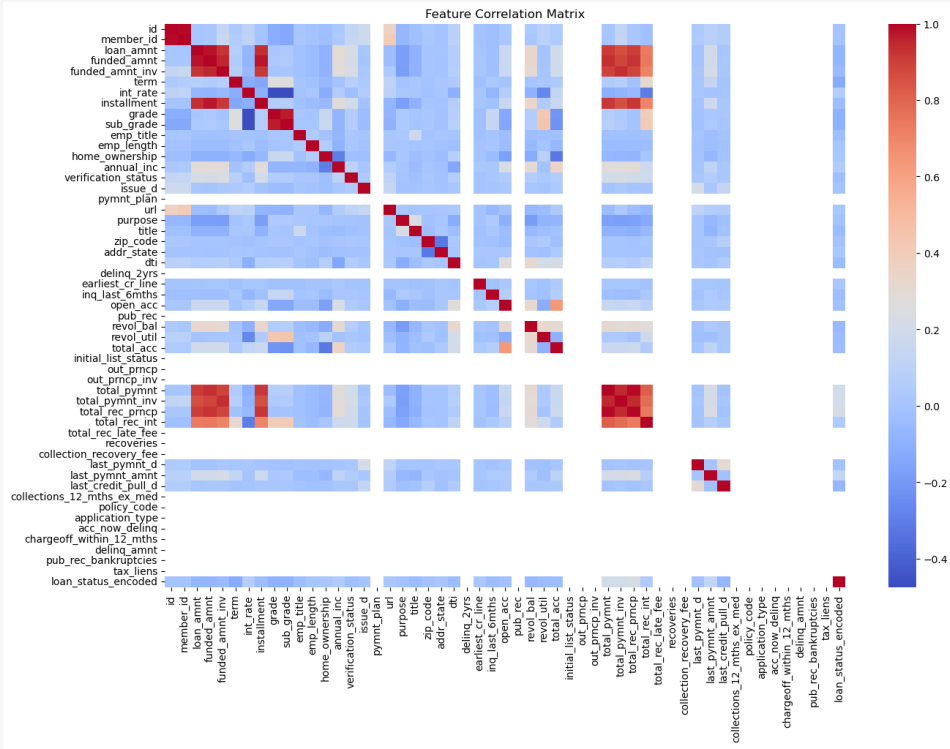
- Selected top 10 features using RFE with a RandomForestClassifier.
- Selected features: [Feature1, Feature2, ..., Feature10]

Tree-based Feature Importance:

Ranked features based on importance scores from a RandomForest model.

Feature Importance Plot:

Visualized feature importances with a bar plot.



Model Evaluation:

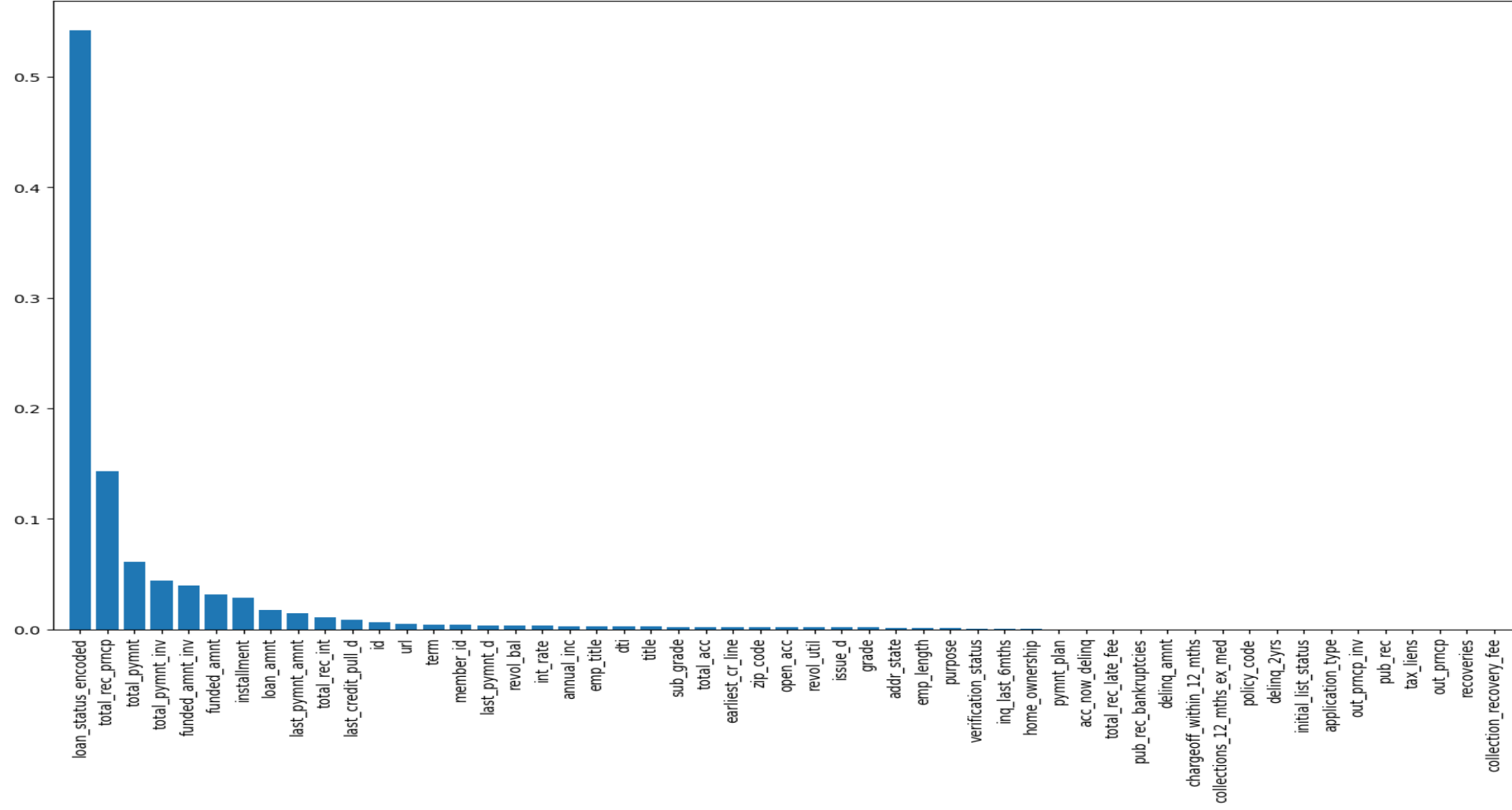
- Train-Test Split:** Split data into training (70%) and testing (30%) sets.
- Model Training and Testing:** Used RandomForestClassifier with top selected features.
- Model Accuracy:** Achieved an accuracy of 0.82 with top 10 selected features.



Key Findings:

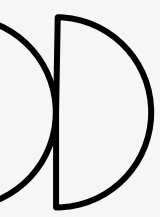
- Significant Features:** Identified the most relevant features influencing loan default, such as loan amount, interest rate, and borrower's credit history.
- Relationships:** Found significant relationships between loan status and various features, both numerical and categorical.
- Visual Patterns:** Used visualizations to highlight patterns and trends in the data.
- Dimensionality Reduction:** Successfully reduced data dimensions for better visualization and understanding.

Feature Importances by RandomForest



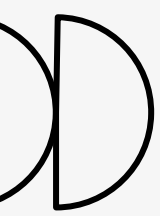
Business Implications:

- The insights gained can help the company make more informed lending decisions.
- Identifying key factors affecting loan defaults can improve risk assessment and potentially reduce credit losses.



CONCLUSION

The EDA provided a comprehensive understanding of the dataset, identified important features, and laid the groundwork for building an effective predictive model. This analysis can guide the company in mitigating risks and making data-driven decisions in lending practices.



THANKS

DO YOU HAVE ANY QUESTIONS?

Chandrayee.cse@gmail.com

chanchalakusum52@gmail.com