Loan Default Prediction Using Machine Learning

Presented by Vincent Barchok Ngochoch

Data Science & ML Consultant

Email: vbarchok@gmail.com | LinkedIn: vincent-ngochoch-94095b64

Problem Statement

- Can we predict whether a customer will default on a loan based on historical data?
- What are the most influential factors leading to default?
- Which model performs better between Logistic Regression and Decision Tree?
- Objective: Reduce NPLs, improve risk pricing, and support early intervention.

Data Overview

- 10,000 customer records with financial features.
- Key Variables:
  - Employed (1/0)
  - Bank Balance (Numeric)
  - Annual Salary (Numeric)
  - Defaulted? (Target Variable)
- Binary classification: Predict 'Defaulted?'.

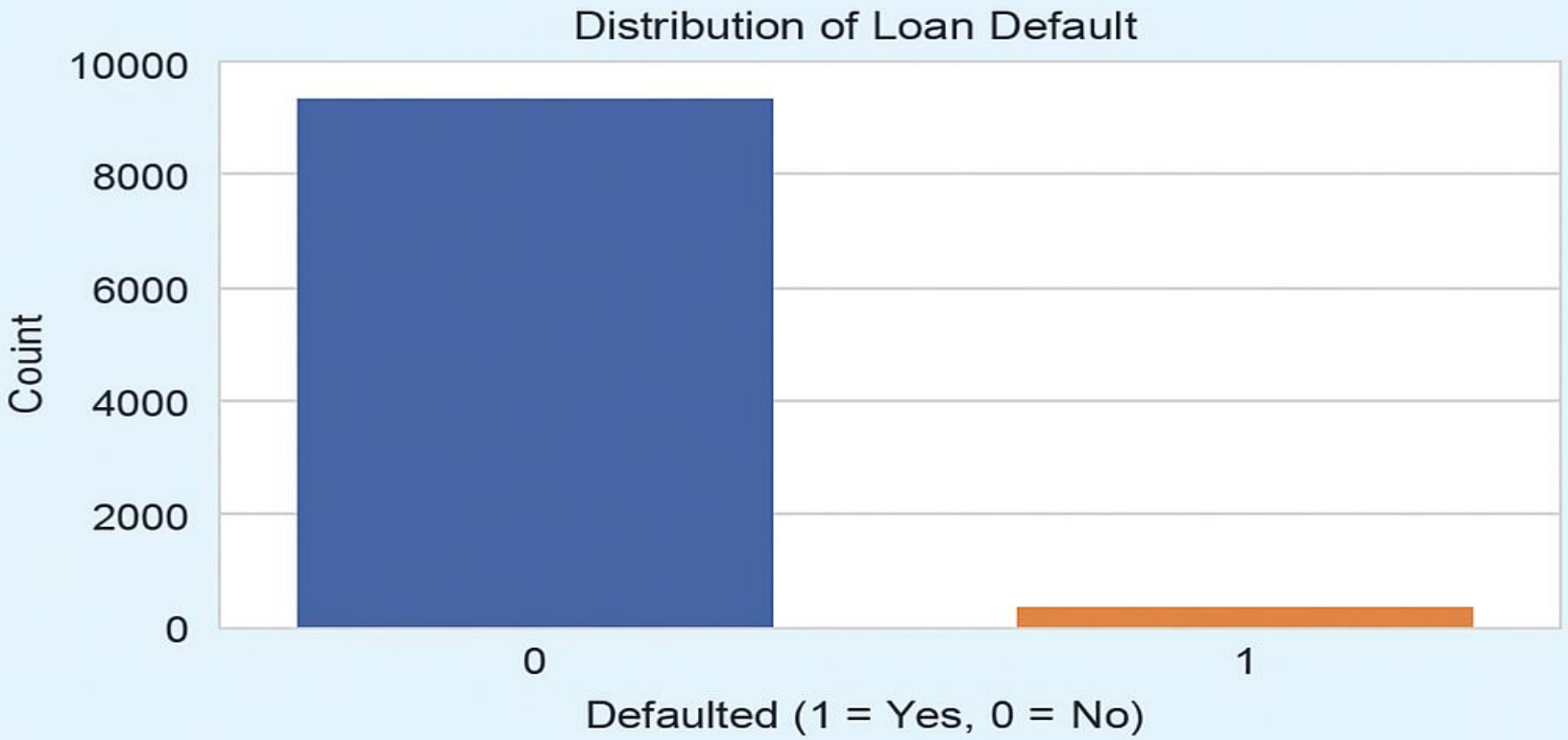
Exploratory Visualizations

- Include key plots:
- Default Rate by Employment Status
- Distribution of Bank Balance
- Correlation Heatmap
- Default Probability by Salary Bracket
- Brief annotations per plot to highlight key takeaways.

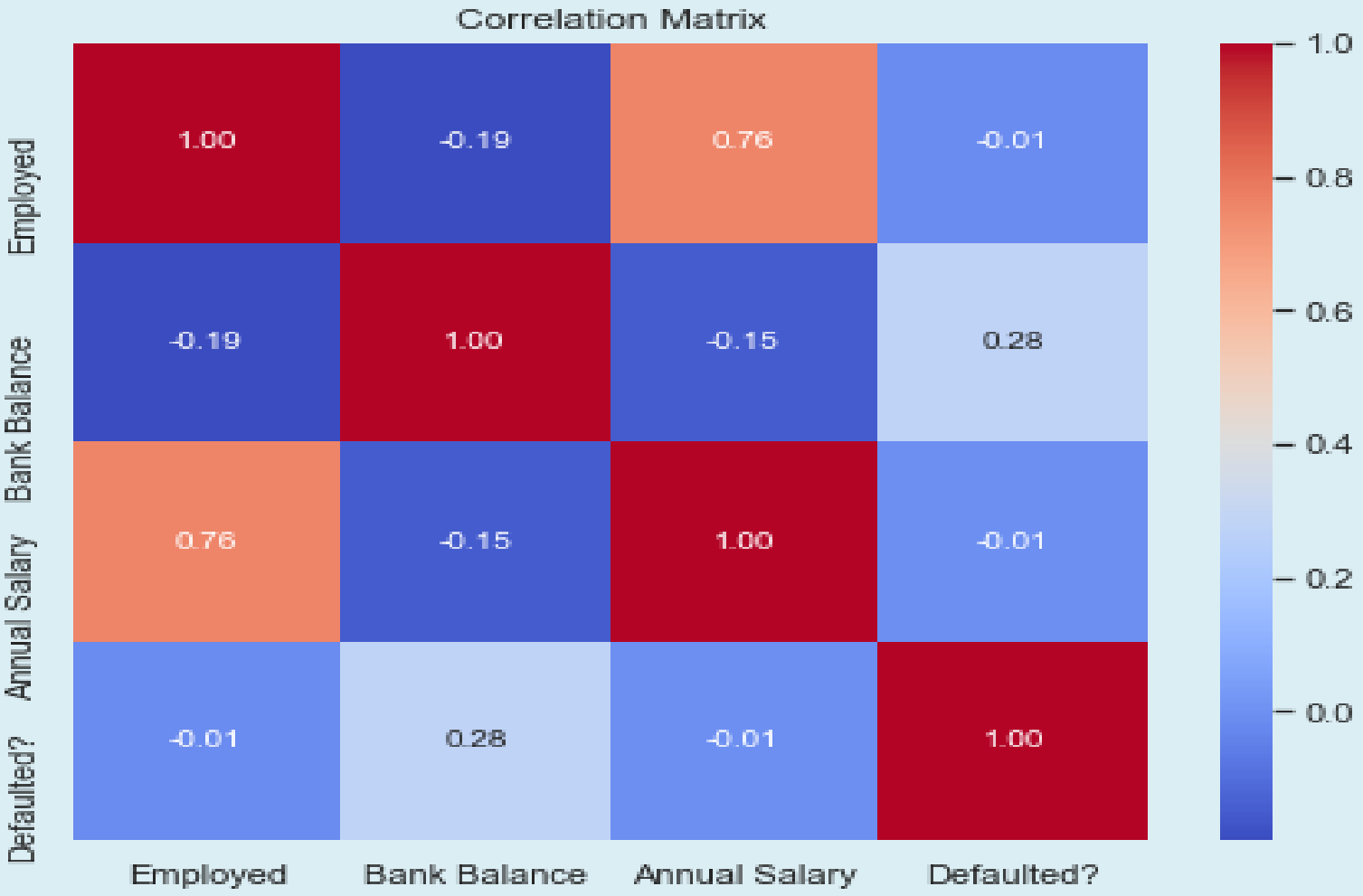
Key Drivers of Loan Default

- Top contributing features:
  - Bank Balance
  - Annual Salary
  - Employment Status
- Interpretation:
- Lower balances and unemployment significantly increase default risk.

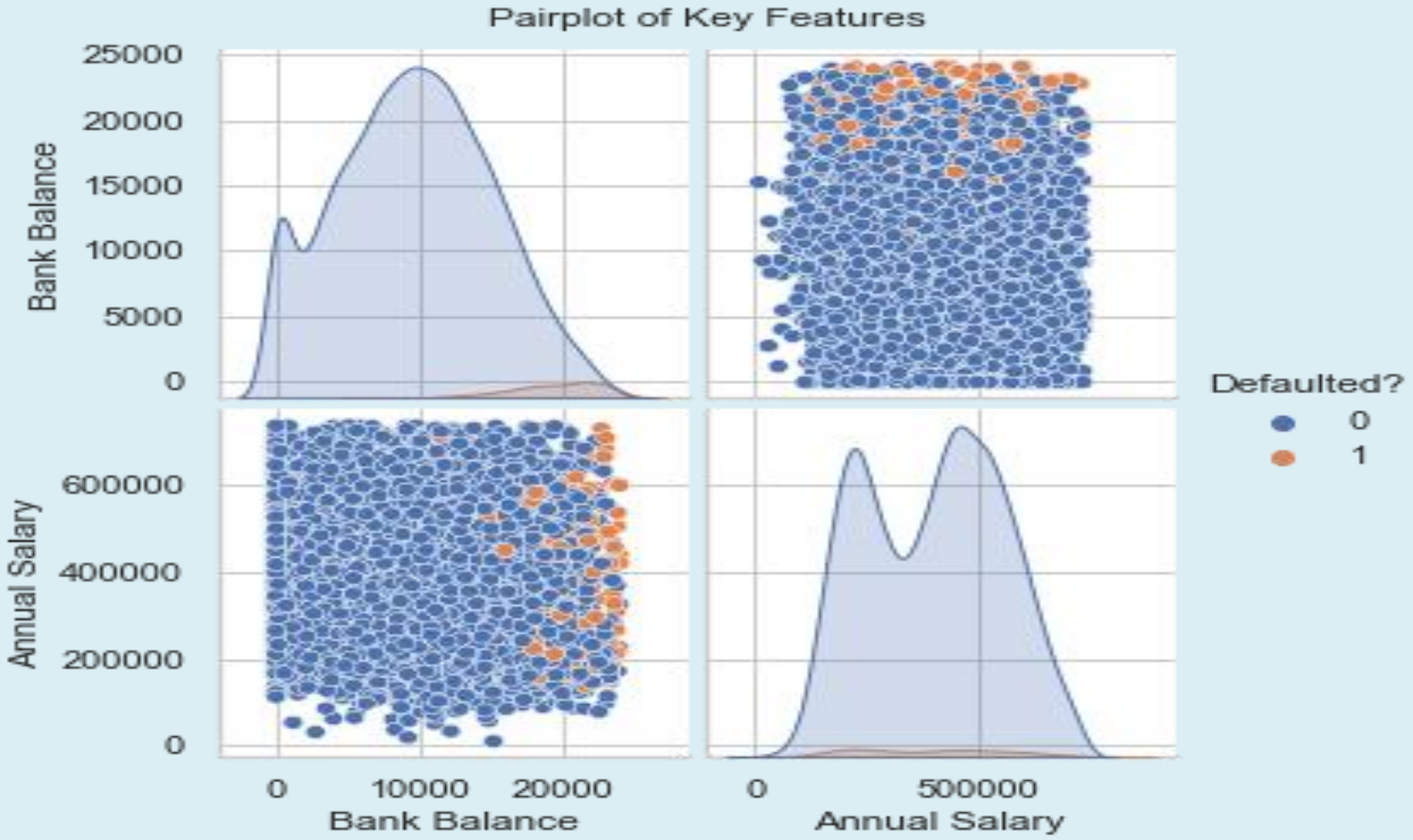
Visualizations



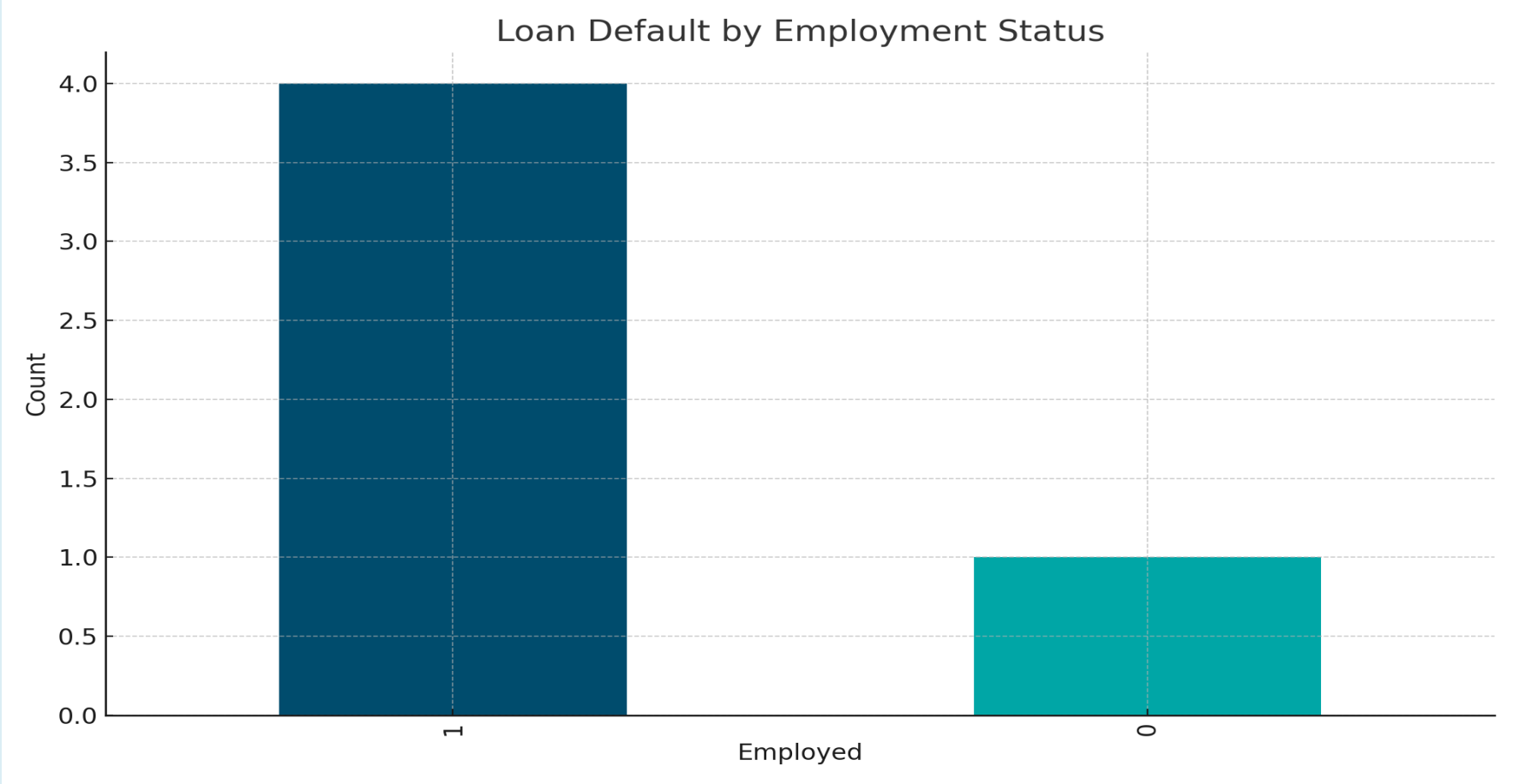
The dataset is highly imbalanced, with a large majoiṫy of customers not defaulting on their loans. This class imbalance (few defaults) highlights the need for careful model evaluation beyond accuracy, especially using metrics like recall and precision.



“Correlation Matrix: Annual Salary is strongly correlated with Employment Status (0.76), suggesting employed customers earn more. Bank Balance shows a mild positive correlation (0.28) with default, which is unexpected and warrants deeper analysis



"Pairplot of Key Features: Most defaulters (orange) cluster around higher bank balances and lower-to-mid salary ranges. The plot reveals potential interaction effects between features and default behavior, supporting the use of non-linear models like decision trees."



“Employed individuals (labelled '1') account for the majority of defaults, likely due to their higher representation in the dataset. However, the unemployed group (‘0’) shows a proportionally significant number of defaults relative to its size, highlighting employment status as a key factor in risk profiling.”

Model Performance Summary

- Baseline: Logistic Regression
- Final Model: Decision Tree Classifier
- Evaluation Metrics:
  - Accuracy, Precision, Recall, F1-Score, ROC AUC
- Decision Tree showed higher recall and better business applicability.

Strategic Recommendations

- Implement Decision Tree model in the credit scoring workflow.
- Use insights to segment borrowers and adjust pricing models.
- Conduct early outreach to high-risk profiles identified.
- Periodically retrain model with updated data.
- Goal: Smarter lending decisions, lower defaults, improved profitability.

Next Steps:

- Integrate the Decision Tree model into the bank’s credit scoring system.
- Monitor model performance regularly and assess its business impact.
- Retrain the model periodically using updated customer data.
- Consider testing more advanced models like Random Forest or XGBoost.
- Work with credit and collections teams to act on high-risk segments.

Thank You – Questions?

- Vincent Barchok Ngochoch.
- Email: vbarchok@gmail.com.
- LinkedIn: www.linkedin.com/in/vincent-ngochoch-94095b64

- Happy to take any questions.