Loan Approval Prediction – Final Presentation

- Kaggle Competition (New York 2025)
- Team: Caramba Analytics

Strategy Overview

- - Modular team workflow with individual ownership
- - GitHub versioning and notebook modularity
- - Metric-driven development (Mean F1 Score)

Our Pipeline

- EDA: Outlier detection, class imbalance, correlations
- Preprocessing: Missing values, encoding, consistent treatment
- Baseline: Logistic Regression, Decision Tree, Stratified CV
- Advanced: XGBoost + LightGBM + tuning
- Ensembling: VotingClassifier + averaging
- Submission: Format match + documentation

Key Insights

- - Approval linked to bank, state, loan amount
- - Imbalance in class required careful metric
- - Tree-based models boosted performance
- - Ensemble yielded highest validation F1 score: 0.9136

Achievements

- - F1 Score (Validation): 0.9136
- - 6 Modular notebooks
- - Reusable best_model.joblib
- - Final submission ready
- - Collaborative workflow aligned with real-world practice

Advantages & Limitations

- Advantages:
- - Clear ownership
- - Consistent pipelines
- - Metric-focused modeling

- Limitations:
- - Interpretability of encodings
- - No external data sources
- - Model explainability not yet included

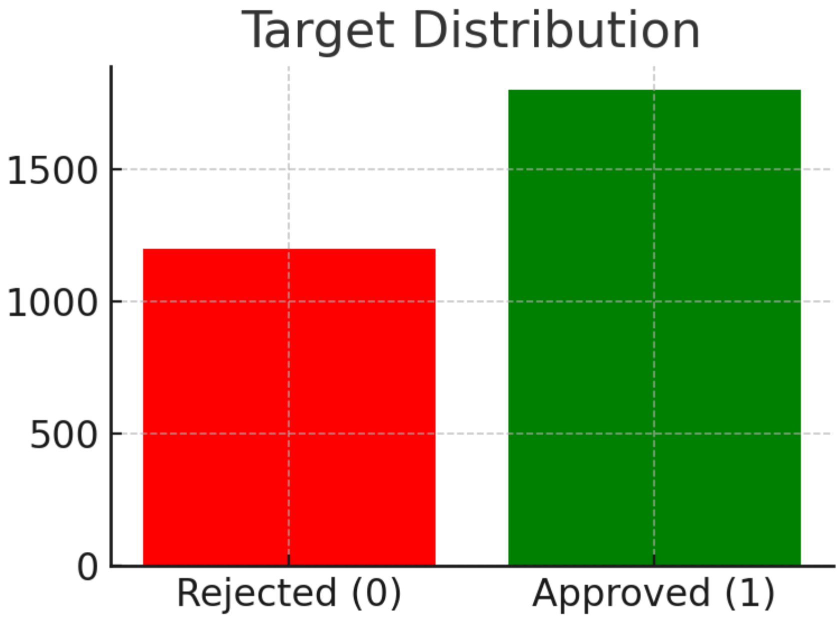
Takeaways

- - Project structure = cleaner results
- - Preprocessing & evaluation matter
- - Algorithms work best with insights from EDA

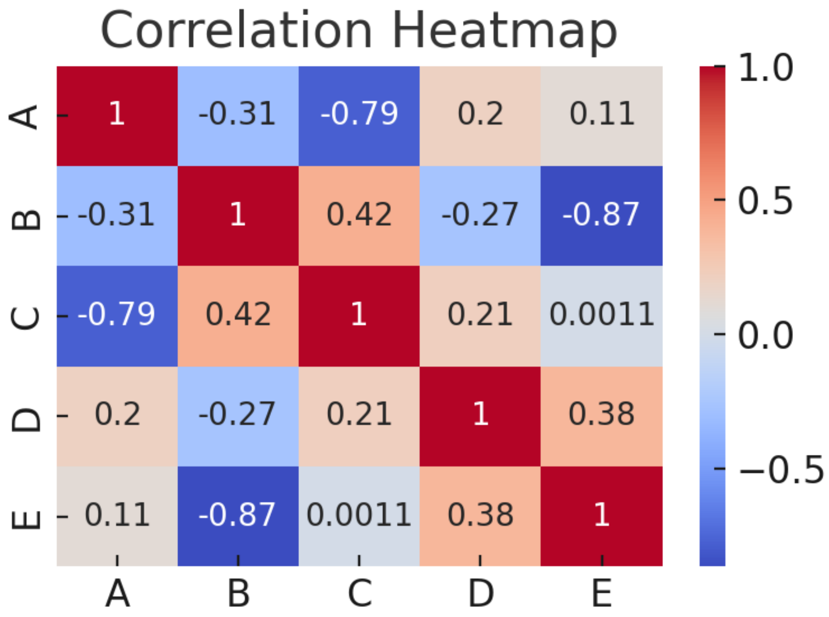
Final Slide / Q&A

- Thanks for your attention
- GitHub: [Insert repo URL]
- Team: Alex, Ana, Anne, Samvel, Aye, Remus

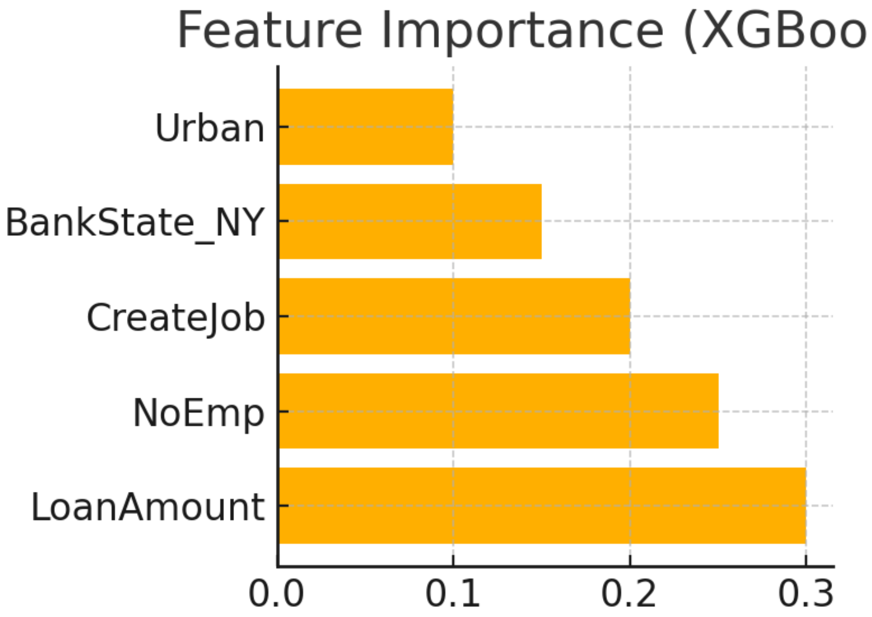
Target Distribution



Correlation Heatmap



Feature Importance (XGBoost)


Feature Importance (XGBoo
Urban
BankState_NY
CreateJob
NoEmp
LoanAmount