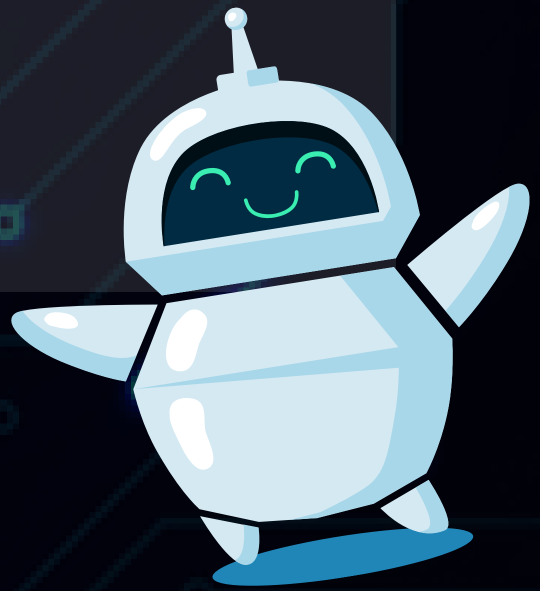
CREDIT RISK ASSESSMENT IN FINANCE INDUSTRY AND AI-ML SOLUTION

by: Group 3



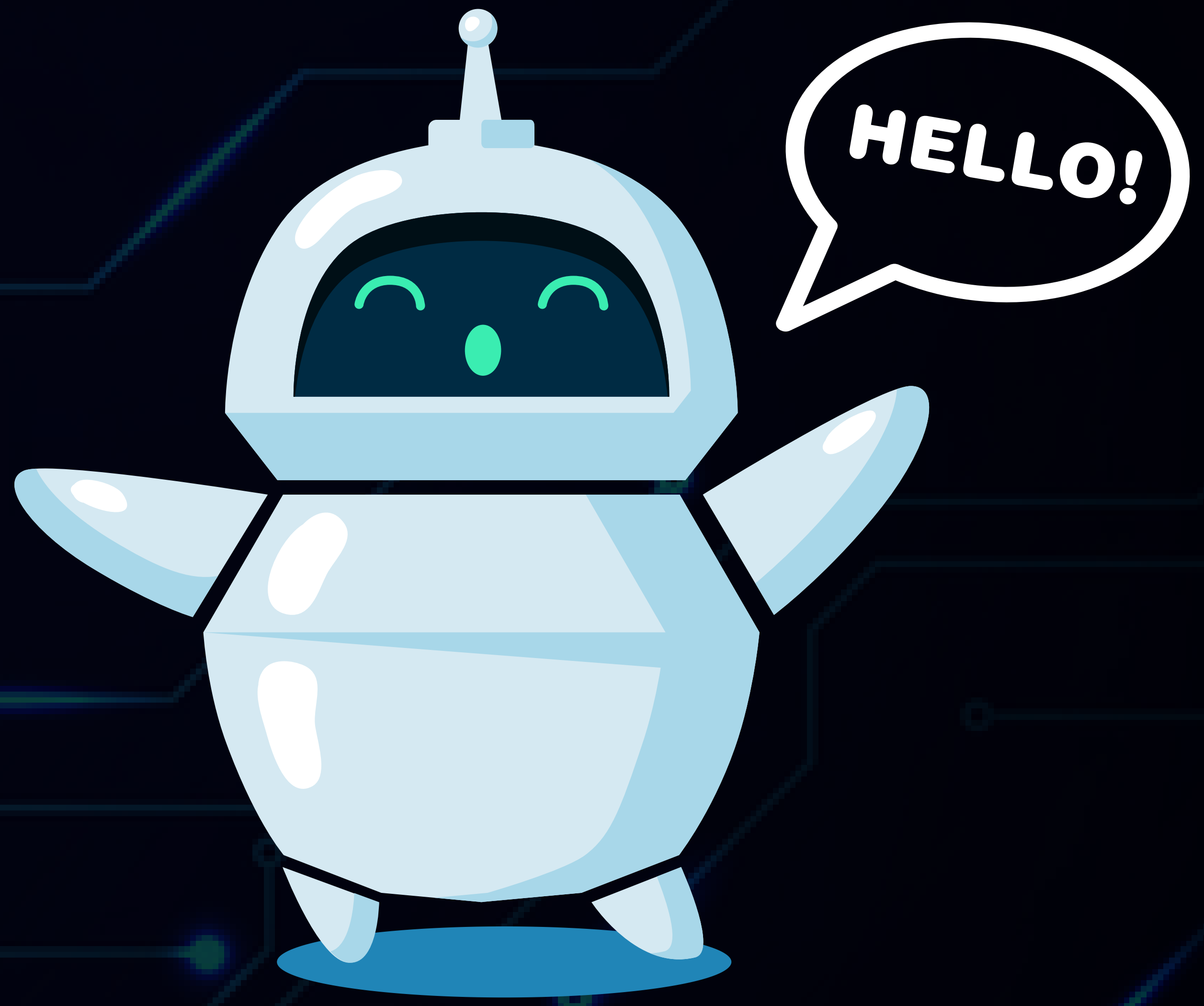
INTRODUCTION

- Core functions: Deposits, credit, investment, payment systems
- Banking = backbone of economic development
- Rising credit demand from retail, MSME, and digital borrowers
- Strong regulatory ecosystem (RBI, Basel norms). Shift toward automation & AI-driven risk models
- Growing reliance on digital lending increases credit risk exposure
- AI/ML enables early identification of high-risk borrowers

PROBLEM STATEMENT

How can financial institutions reduce loan defaults using ML while maintaining regulatory compliance?

- Credit datasets are imbalanced & small in size
- Traditional models miss complex borrower patterns
- Objective: Benchmark 8 ML algorithms to minimize misclassification risk



LITERATURE REVIEW

Author & Year	Models Compared	Key Finding
Yang et al. (2025)	XGBoost, LightGBM, CatBoost, LR, RF, MLP	Ensemble models performed best; strong robustness
Chawla et al. (2024)	DT, NN, KNN, LR	Decision Tree marginally outperformed NN
Ibrahim et al. (2024)	MLP + SOM	Hybrid SOM–MLP improved accuracy significantly
Chang et al. (2024)	SVM, LR	SVM achieved higher accuracy than LR
Xu (2024)	GBM, LR, DT, RF, SVM	GBM delivered highest AUC and accuracy
Shih et al. (2022)	Outlier Detection + ML	Outlier removal improved model stability & profitability
Chen et al. (2021)	DT, SVM, RF, KNN, LR, ANN	DT highest accuracy; SVM best overall performance
Laborda et al. (2021)	LR, SVM, KNN, RF	Feature selection boosted model performance
Silva et al. (2020)	Logistic Regression	LR correctly predicted ~90% defaults
Wang et al. (2015)	Lasso-LR vs DT, RF, LR	Lasso-LR best for large, imbalanced datasets

METHODOLOGY

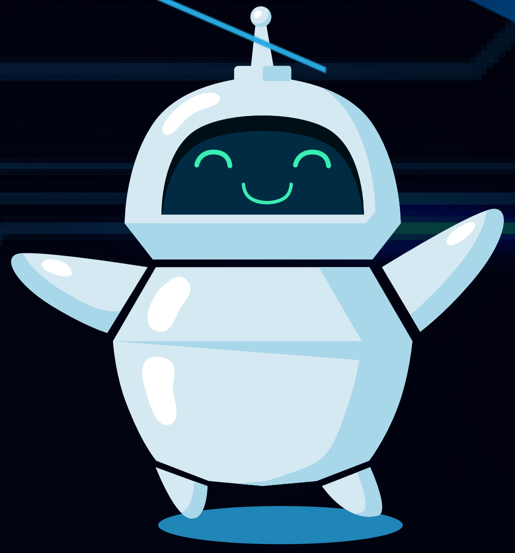
Data Preparation

- German Credit Dataset
- Train-test split: 80/20
- Outlier removal: Cook’s Distance
- Log transformation of skewed variables
- One-hot encoding → 61 features
- Z-score standardization

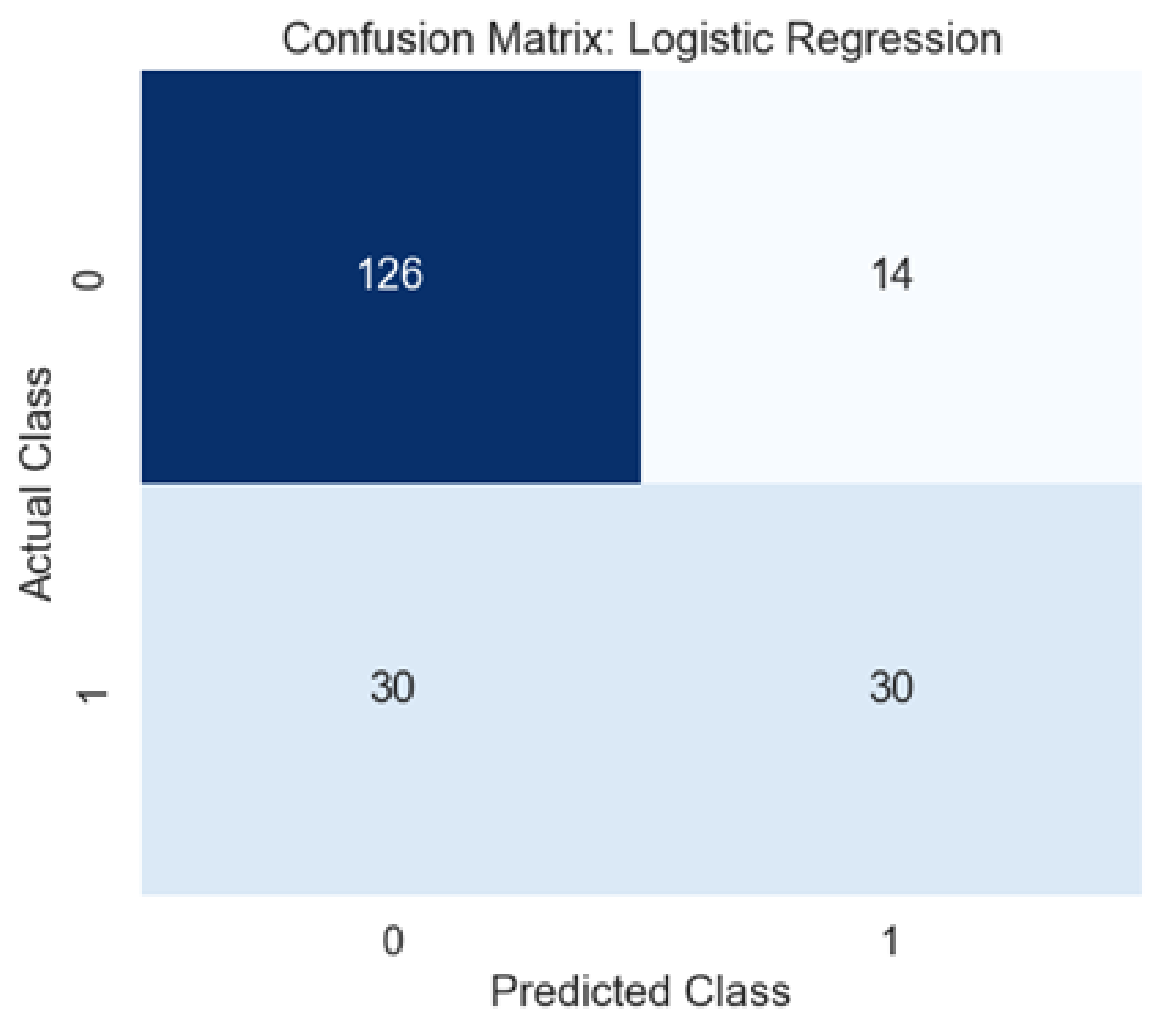
Models Evaluated:

LR, DT, RF, XGBoost, SVM, KNN, LightGBM, MLP

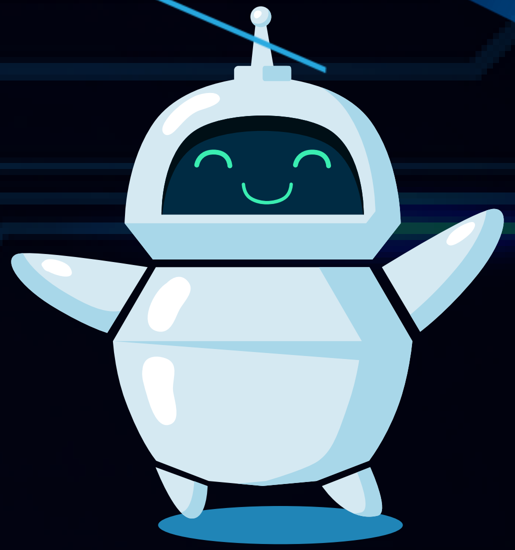
LOGISTIC REGRESSION:



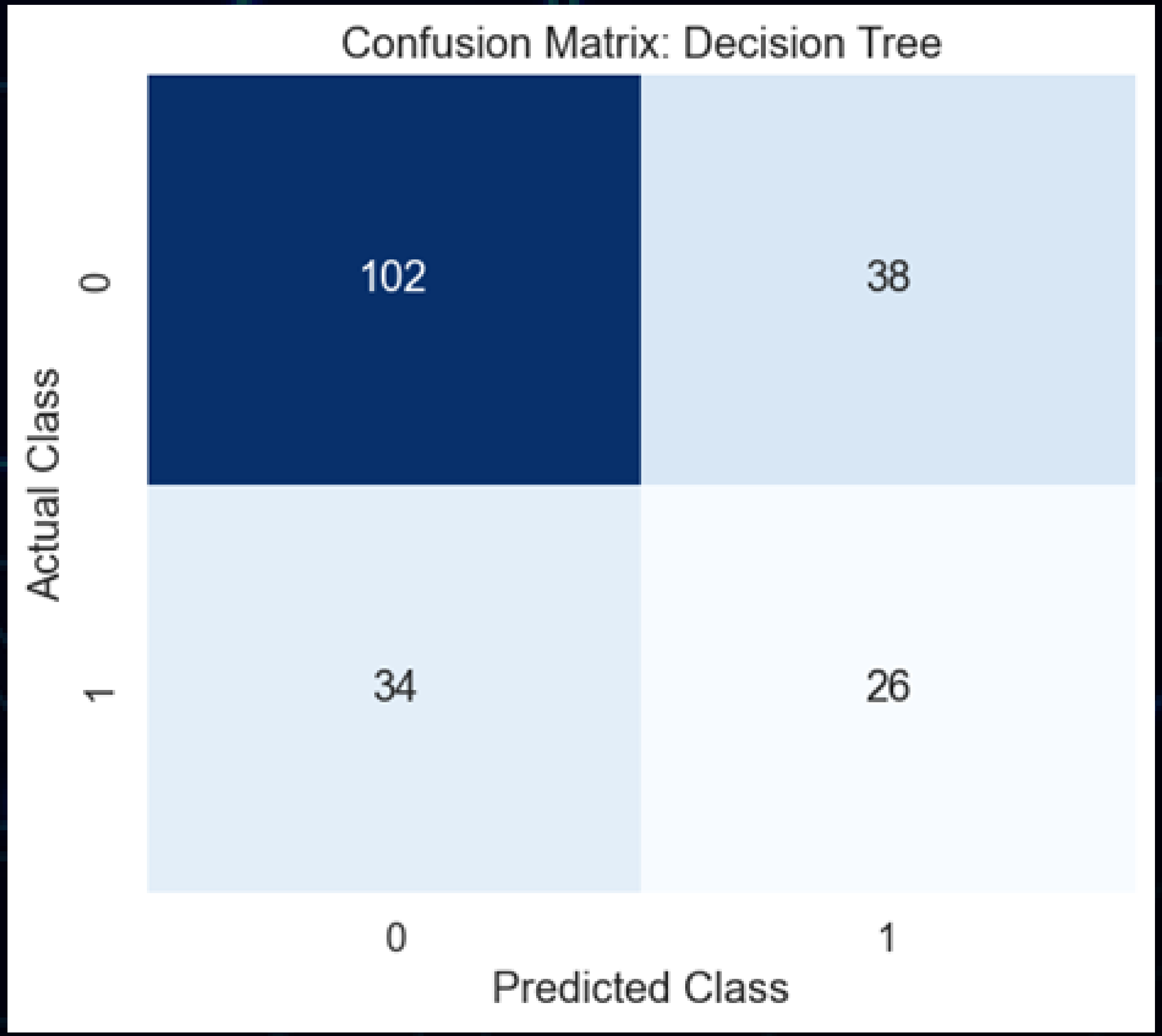
- Linear baseline model
- L2-regularized LR with ‘lbfgs’ solver
- Accuracy: 78%
- ROC–AUC: 0.7931
- Weakness: lower recall for defaulters (50%)
- Strength: Interpretability & regulatory friendliness



DECISION TREE:

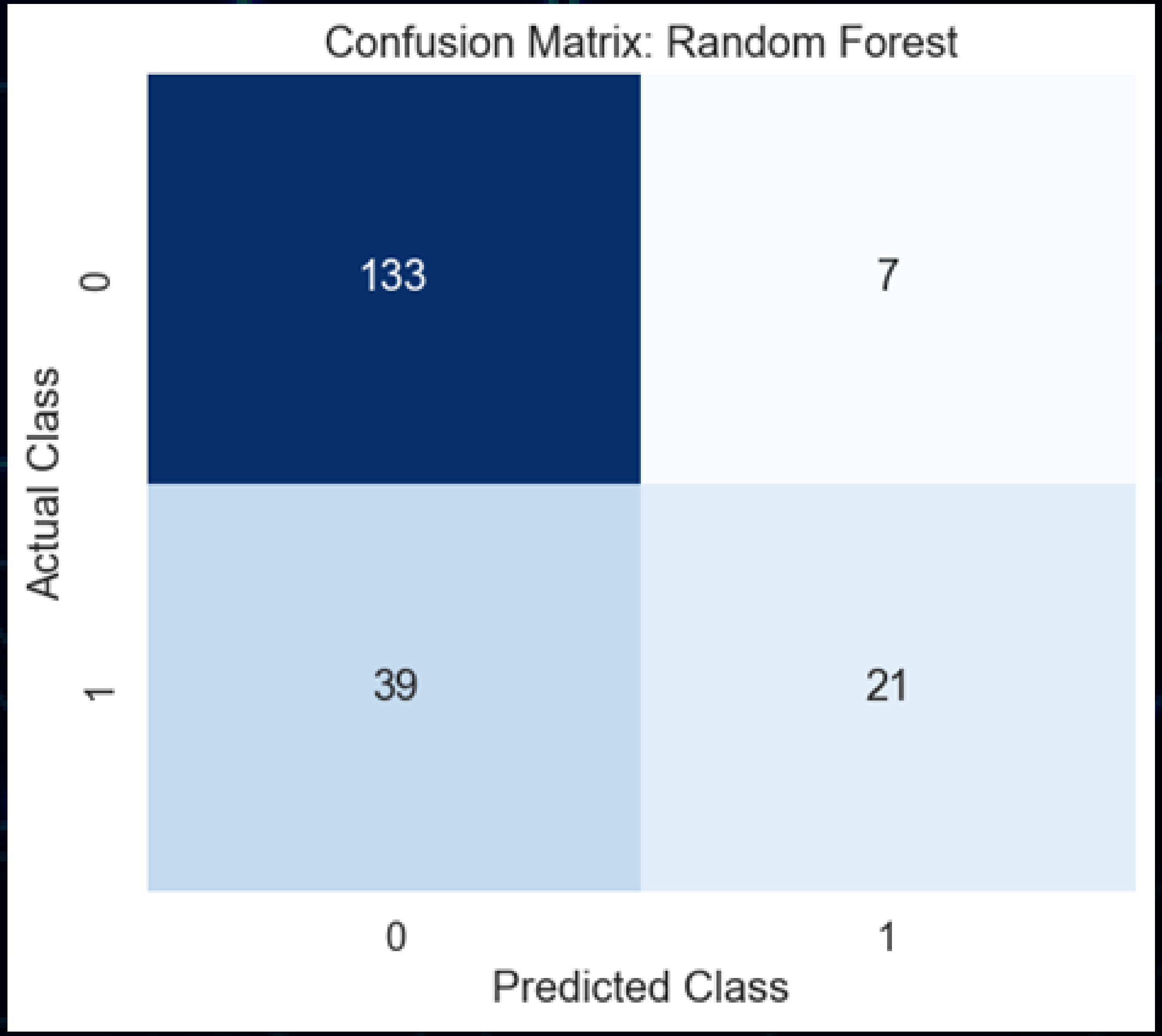


- Non-linear splits using Information Gain
- Max depth = 5 to avoid overfitting
- Accuracy: 64%
- ROC-AUC: 0.6205
- High False Negatives → weak risk detection



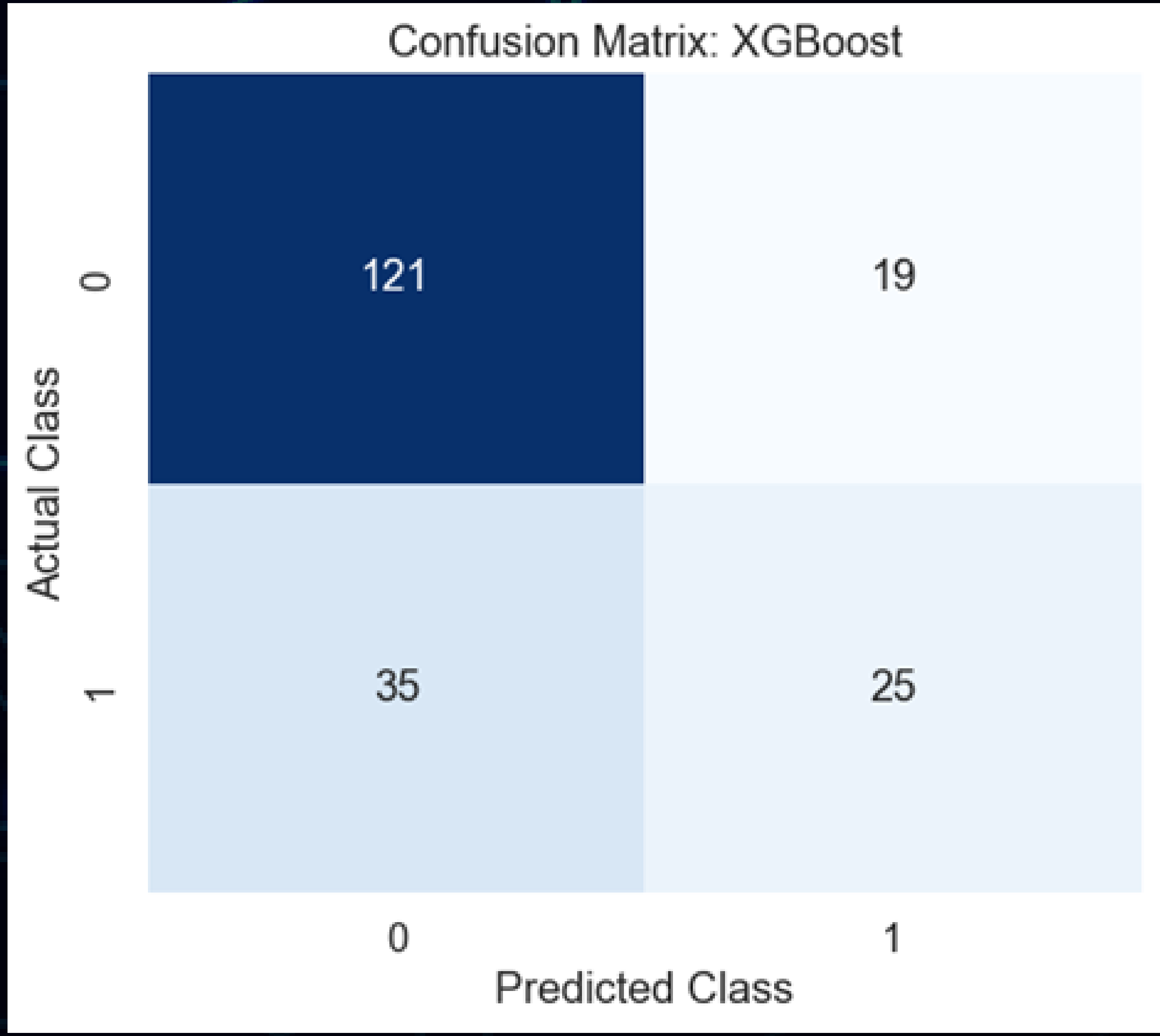
RANDOM FOREST:

- 100 trees, max depth 10
- Reduces variance vs. single tree
- Accuracy: 77%
- ROC-AUC: 0.7999 (Highest)
- Conservative model → 39 False Negatives

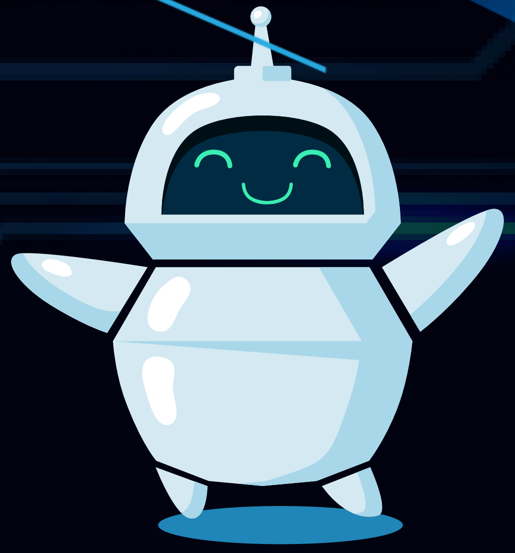


XGBOOST:

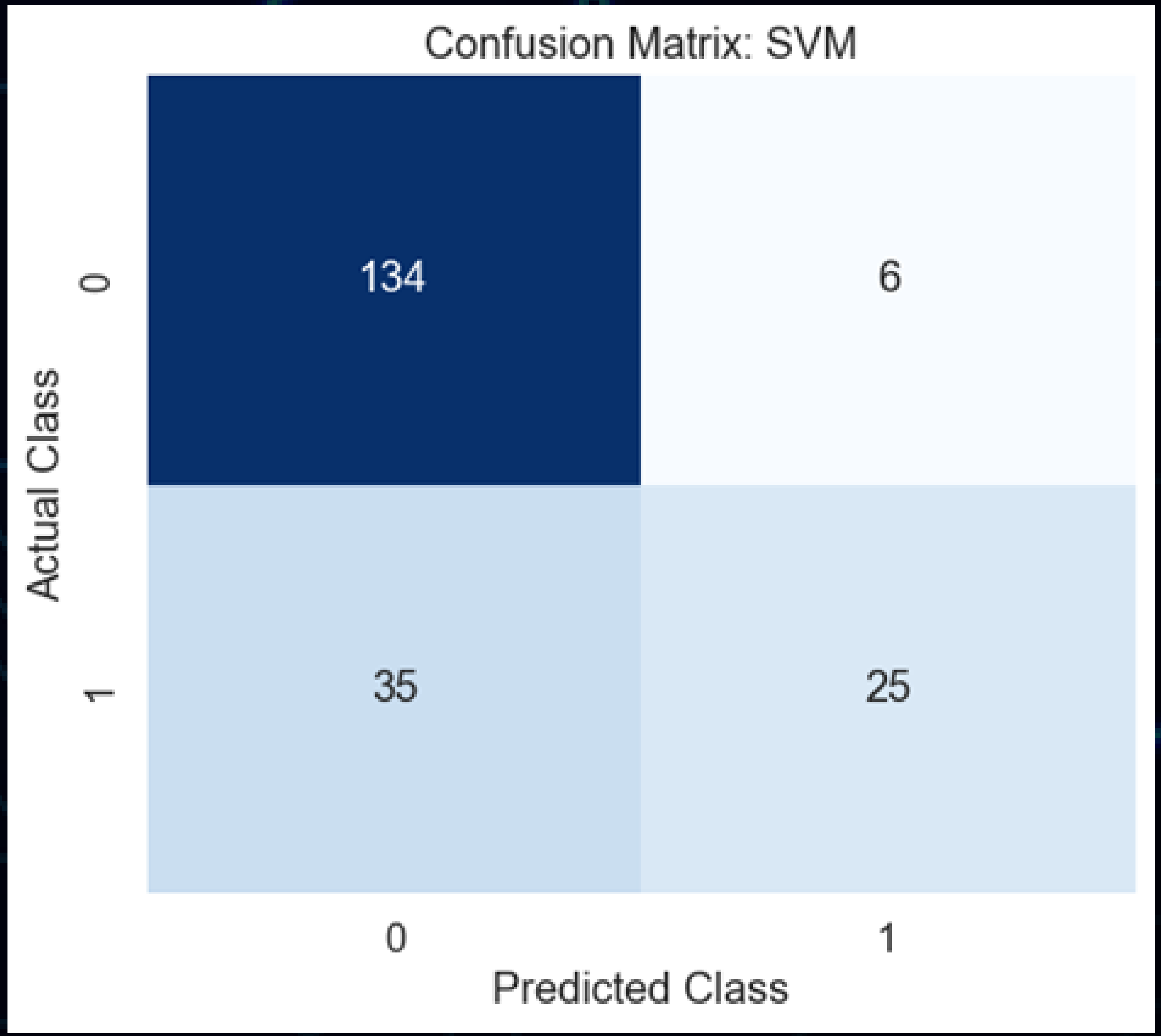
- Gradient boosting on sequential trees
- Accuracy: 73%
- ROC-AUC: 0.7189
- Overfits on small datasets (N < 1000)



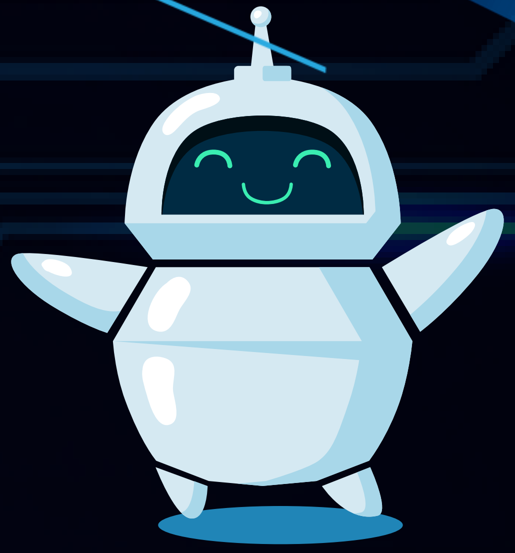
SUPPORT VECTOR MACHINE (SVM):



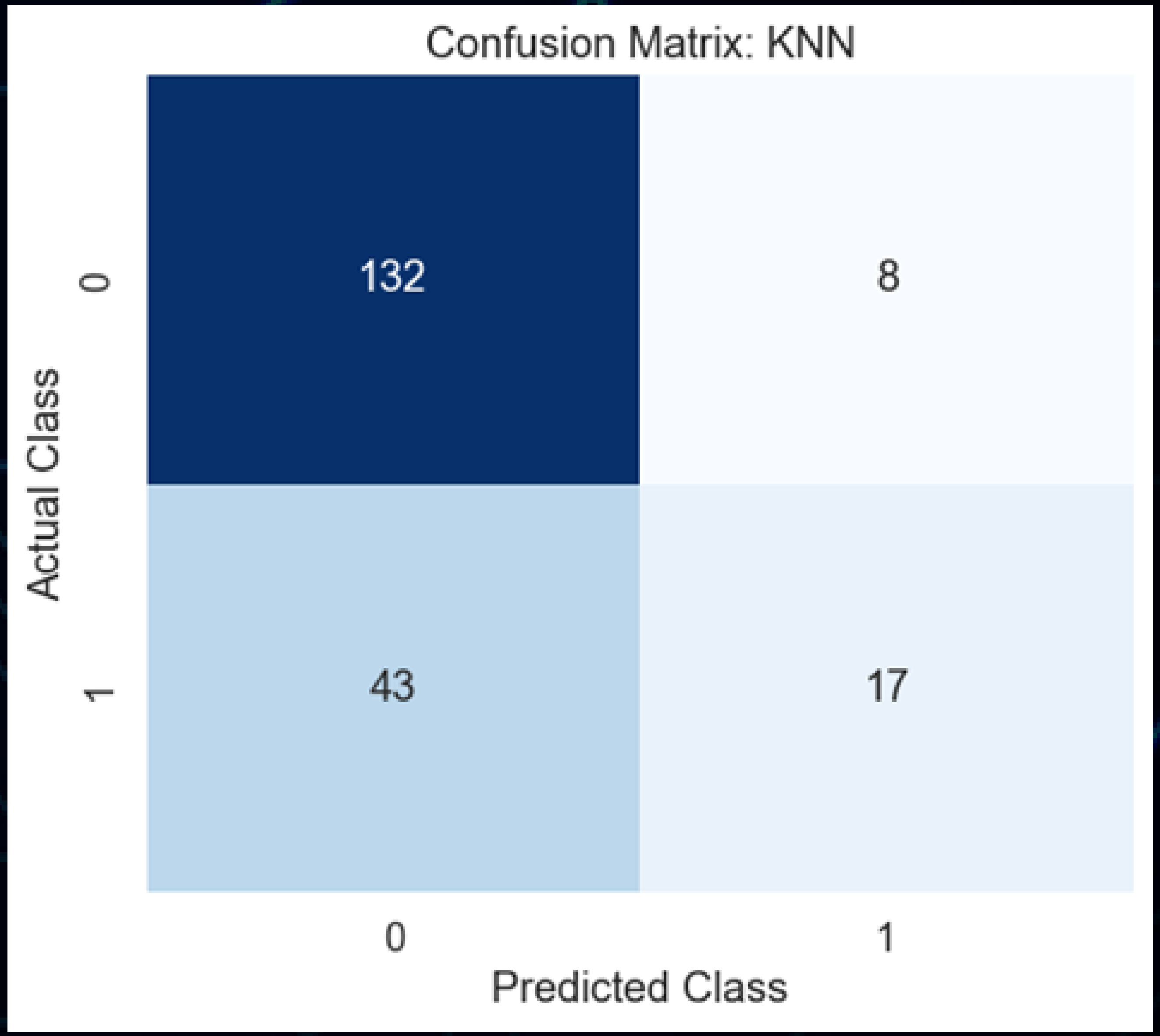
- RBF kernel for non-linear separation
- Accuracy: 79.5% (Highest overall)
- ROC–AUC: 0.7805
- Very low False Positives (best for customer friendliness)
- Still produces 35 False Negatives



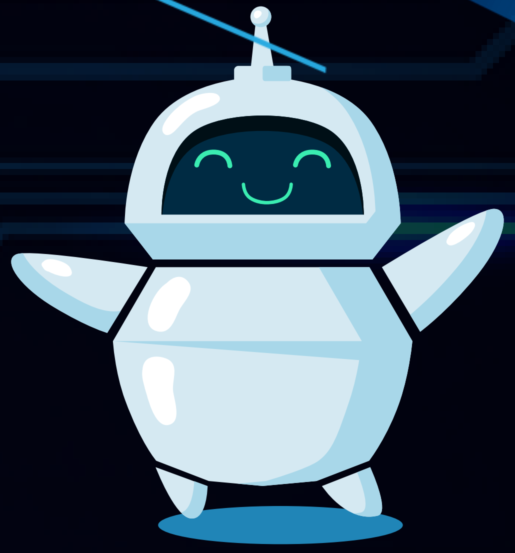
K-NEAREST NEIGHBORS (KNN):



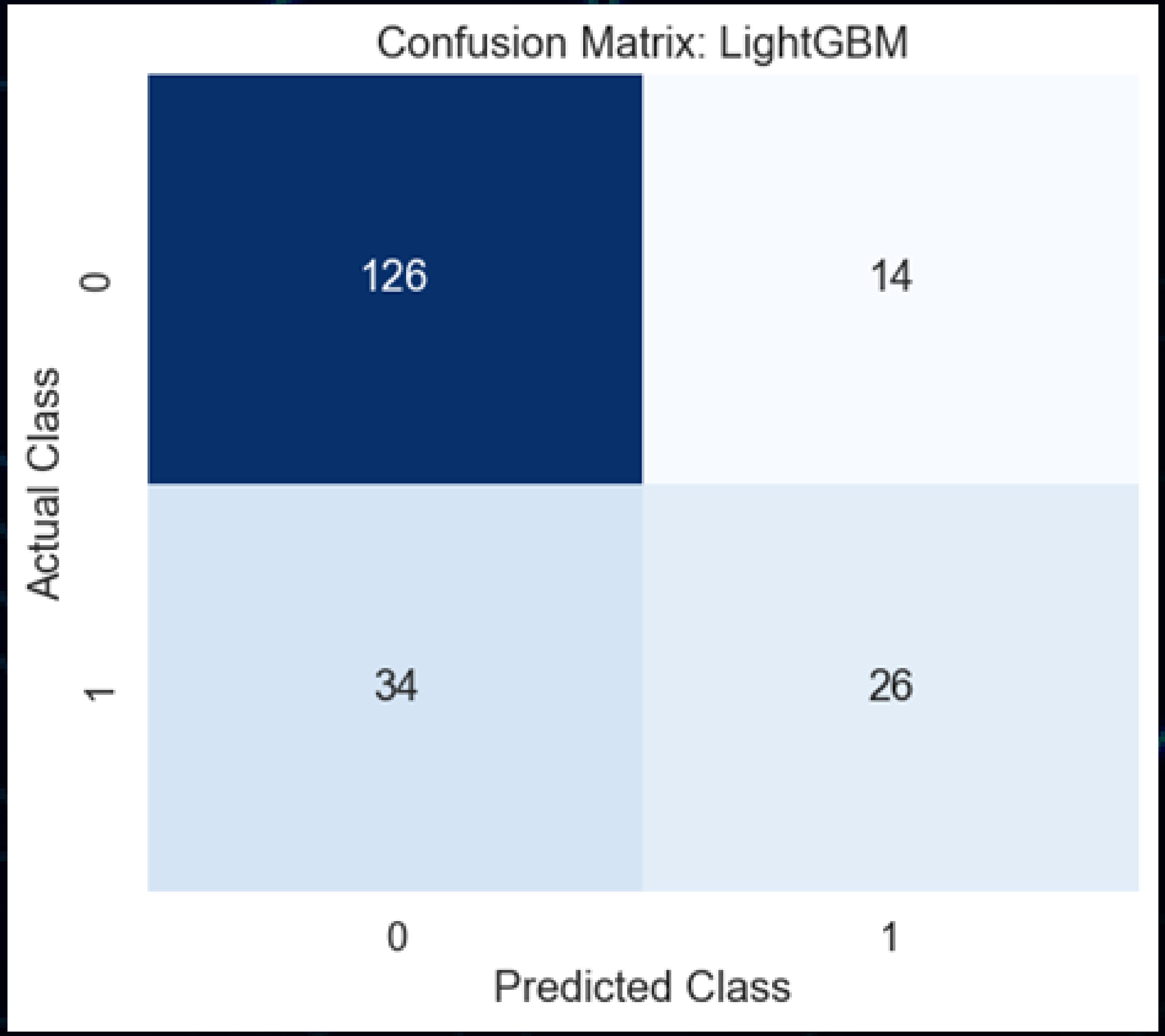
- k = 15 neighbours
- Accuracy: 74.5%
- ROC-AUC: 0.7090
- Struggles with high-dimensional data (61D)
- Highest False Negatives (43)



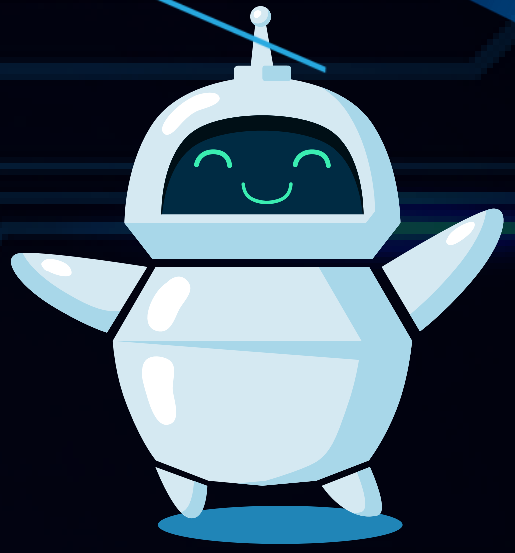
LIGHTGBM:



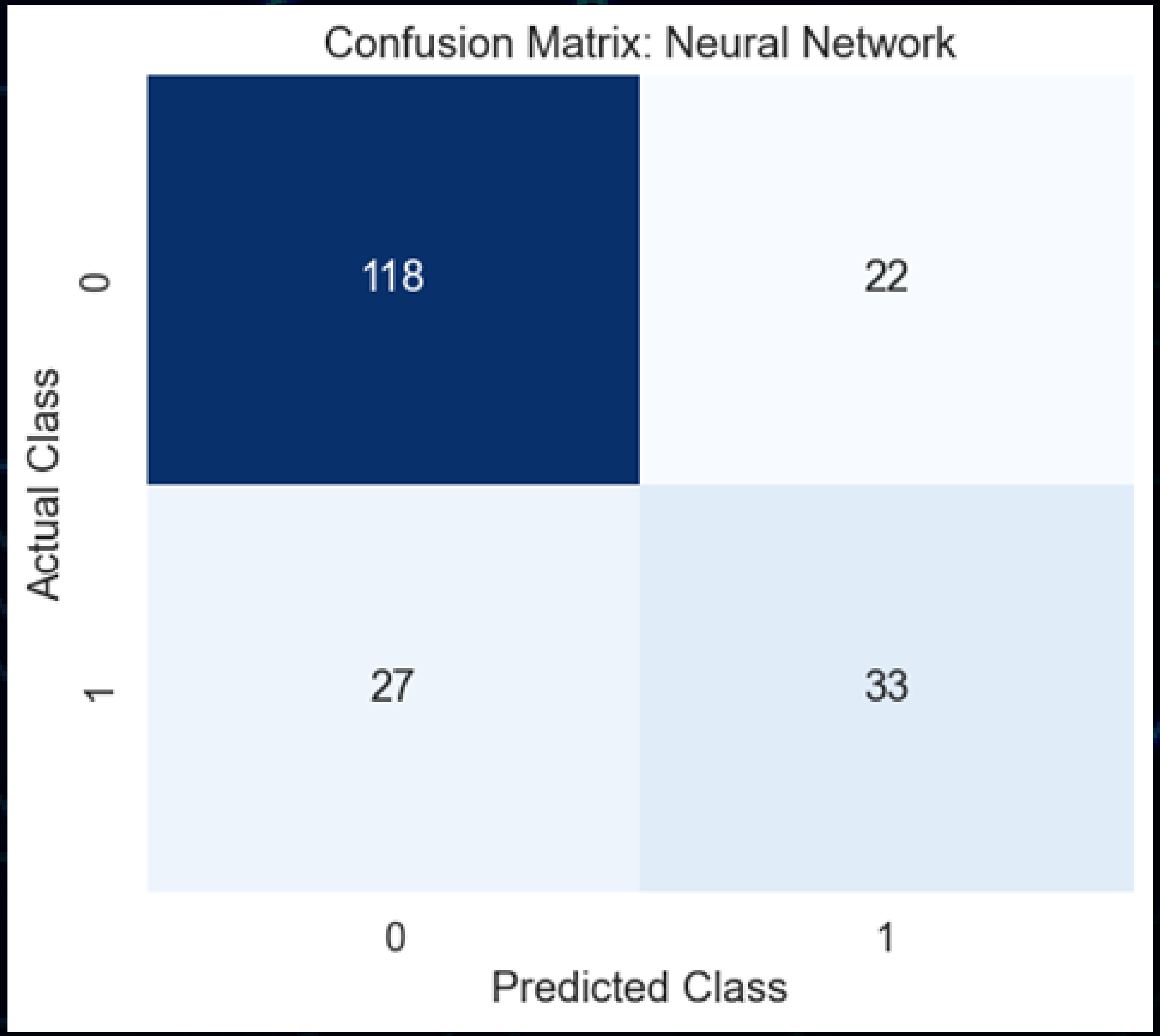
- Leaf-wise boosting with depth constraints
- Accuracy: 76%
- ROC-AUC: 0.7426
- 34 False Negatives
- Better than XGBoost but worse than RF



MULTI-LAYER PERCEPTRON (MLP):



- Two hidden layers (64–32) with ReLU
- Accuracy: 75.5%
- ROC–AUC: 0.7711
- Best Recall → only 27 False Negatives
- Best at catching bad borrowers

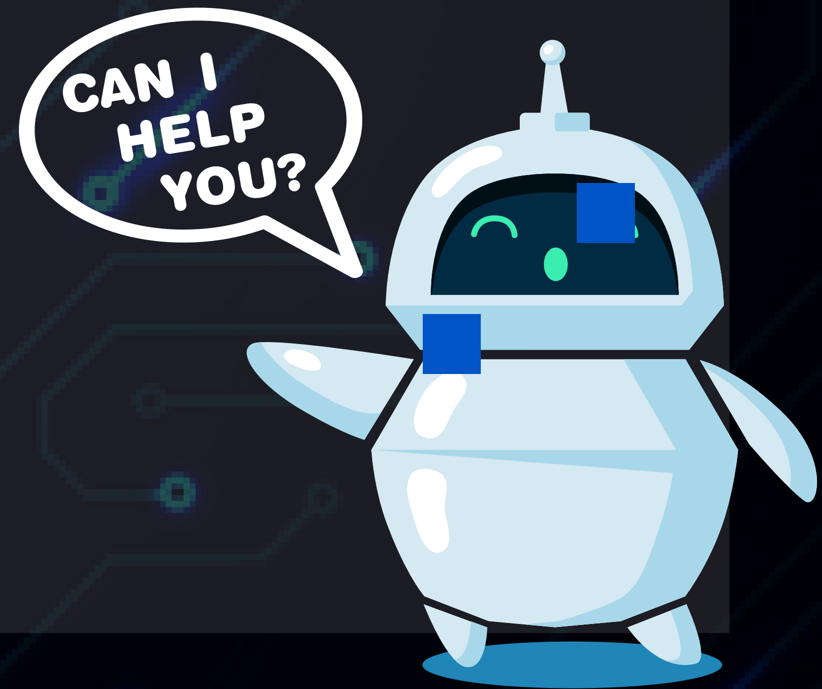


CONCLUSION & IMPLICATION

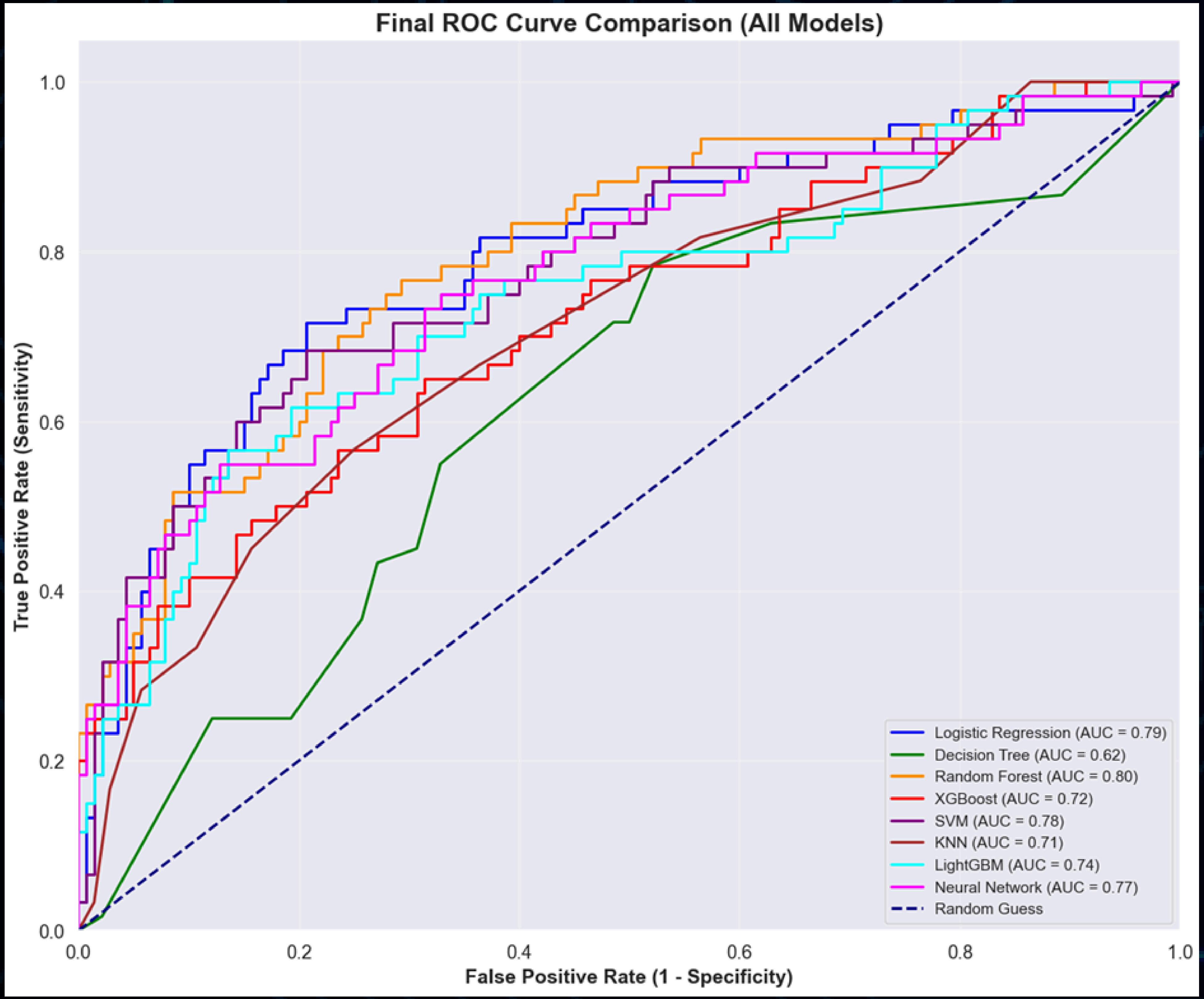
Rank	Model	Accuracy	ROC-AUC	False Negatives (Risk)
1	**Random Forest**	77.00%	**0.7999**	39
2	**Logistic Regression**	78.00%	0.7931	30
3	**SVM**	**79.50%**	0.7805	35
4	**Neural Network**	75.50%	0.7711	**27 (Best)**
5	**LightGBM**	76.00%	0.7426	34
6	**XGBoost**	73.00%	0.7189	35
7	**KNN**	74.50%	0.709	43
8	**Decision Tree**	64.00%	0.6205	34

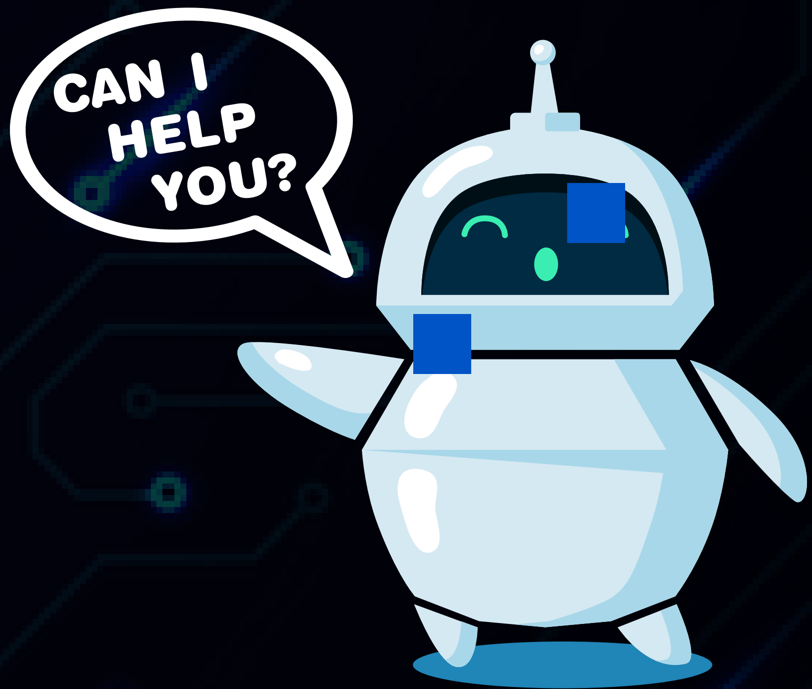
Random Forest = Best AUC → strongest for risk ranking
SVM = Best Accuracy → most correct decisions overall
MLP = Best Recall → lowest missed defaulters
Logistic Regression = Best practical model
- Balanced accuracy
- Low FN
- Most interpretable (RBI requirement)

Model complexity ≠ better performance
Boosting models underperformed due to small dataset size



CONCLUSION & IMPLICATION





THANK YOU!