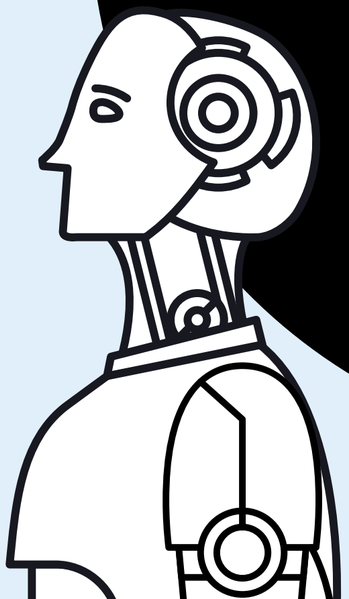
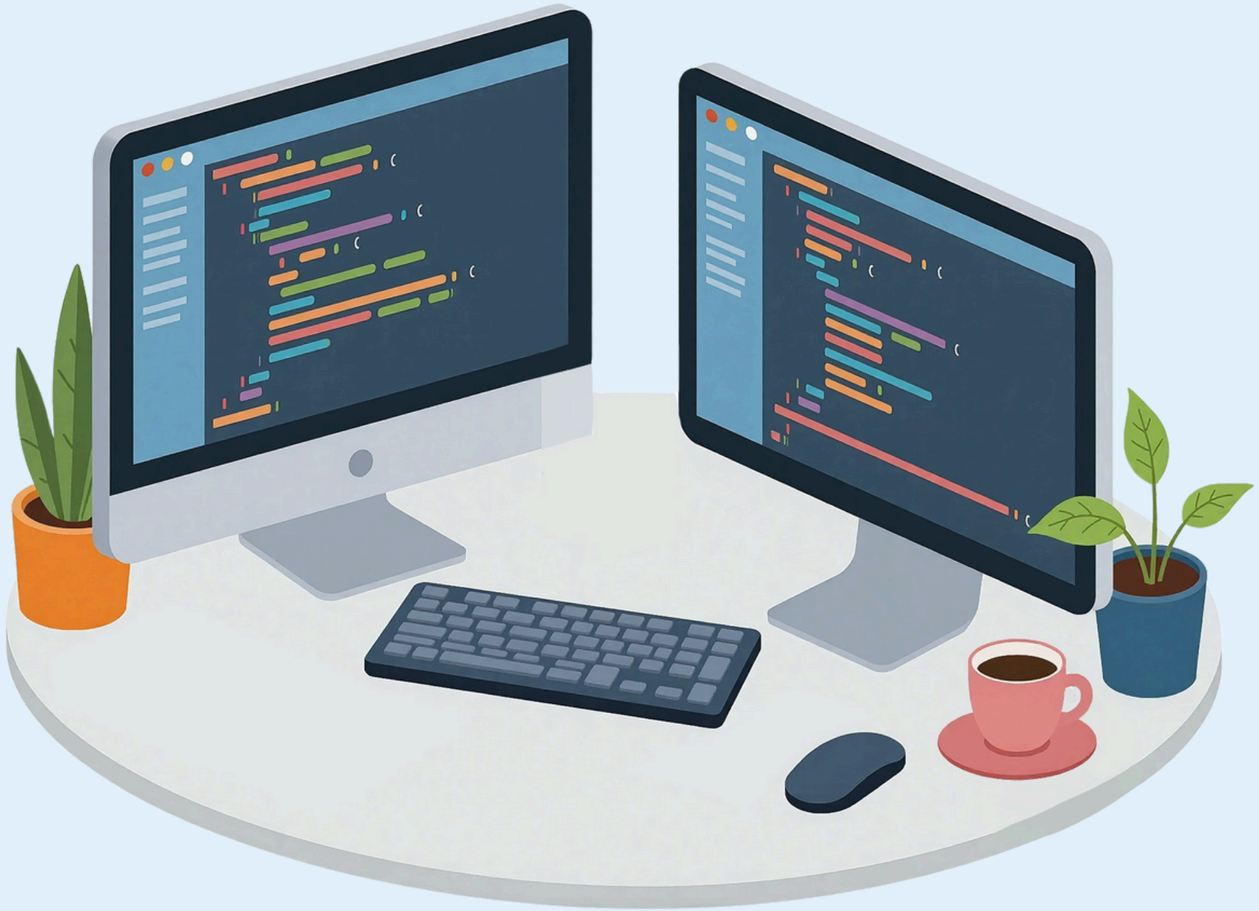
ENSEMBLE LEARNING IN MACHINE LEARNING



Presented by

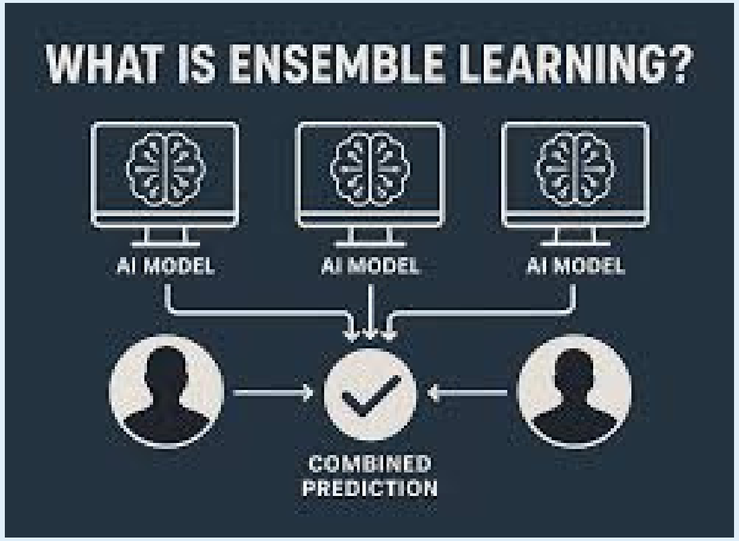
Pramod HV Pannaga Y P Srihari N Praneel Sanjay Biju

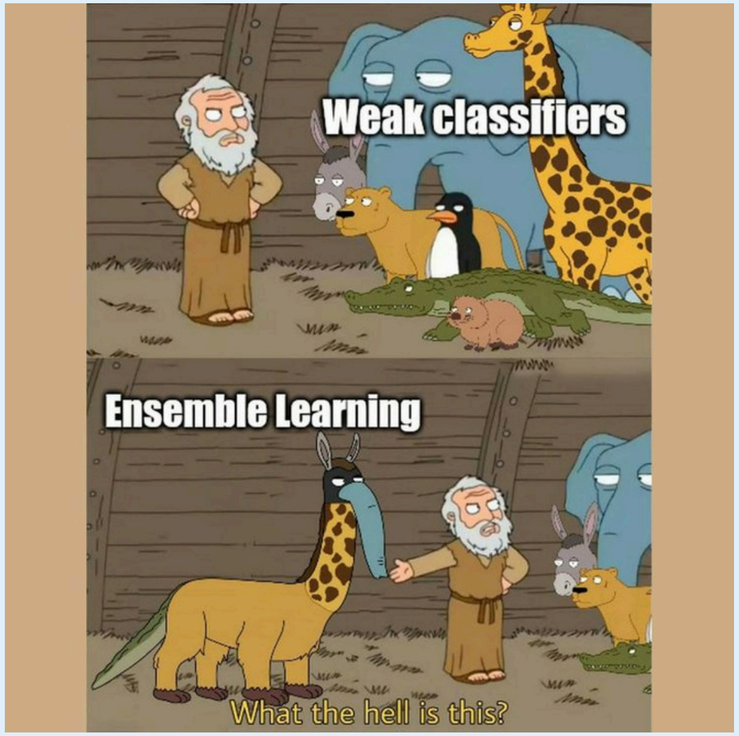
Contents



Introduction

- Ensemble learning combines multiple machine learning models to improve prediction performance
- Instead of relying on a single model, it aggregates several “weak” or “base” learners
- Common benefits:
  - Higher accuracy
  - Better generalization
  - Reduced overfitting
- Widely used in finance, healthcare, and recommendation systems

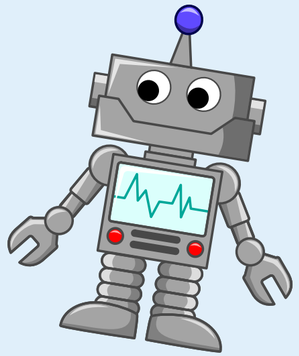




Problem Statement

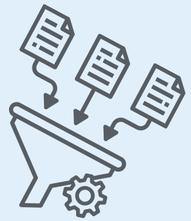
- Goal: Predict whether a loan applicant will default or not.
- Dataset includes features such as:
  - Income
  - Credit history
  - Has mortgage
  - Employment status
- Challenges:
  - Noisy and imbalanced data
  - Non-linear relationships between features
  - Risk of overfitting with single models
  - Need for robust and accurate prediction models

Dataset Description

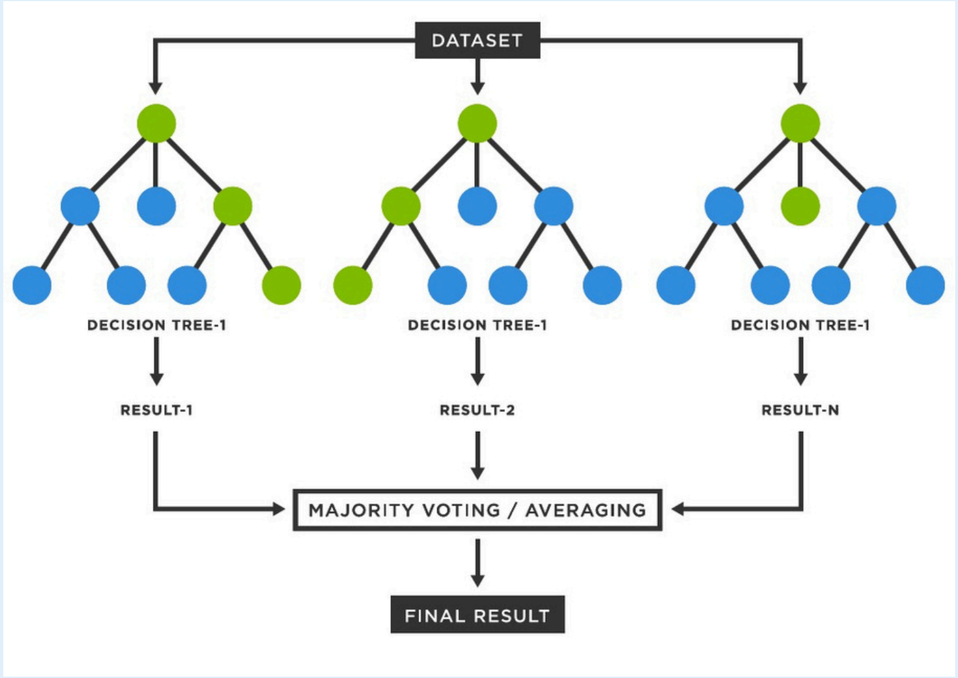


- Dataset used: Loan Default
- Dataset Size: 255,347 loan records with 18 features
- Objective: Predict whether a borrower will default on a loan
- Target Variable: Default (Binary: 1 = Default, 0 = No Default)
- Key Feature Categories:
  -Demographics: Age, Education, MaritalStatus, HasDependents
  -Financial Profile: Income, CreditScore, DTIRatio, NumCreditLines
  -Loan Details: LoanAmount, InterestRate, LoanTerm,LoanPurpose
  -Employment & Stability: EmploymentType, MonthsEmployed
  -Risk Mitigation: HasMortgage, HasCoSigner

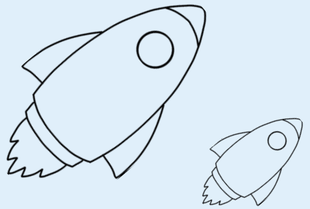
Bagging



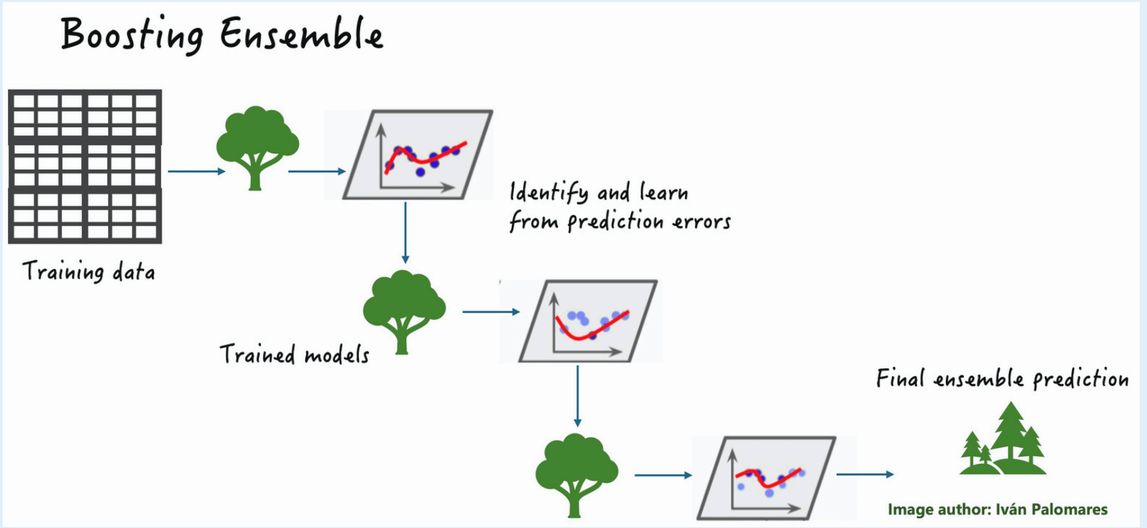
- Trains multiple models on different random subsets of the data
- Each subset is created using sampling with replacement
- Final prediction is made by:
  - Majority voting (classification)
  - Averaging (regression)
- Reduces variance and overfitting
- Example:
  - Random Forest using multiple decision trees for loan approval prediction



Boosting

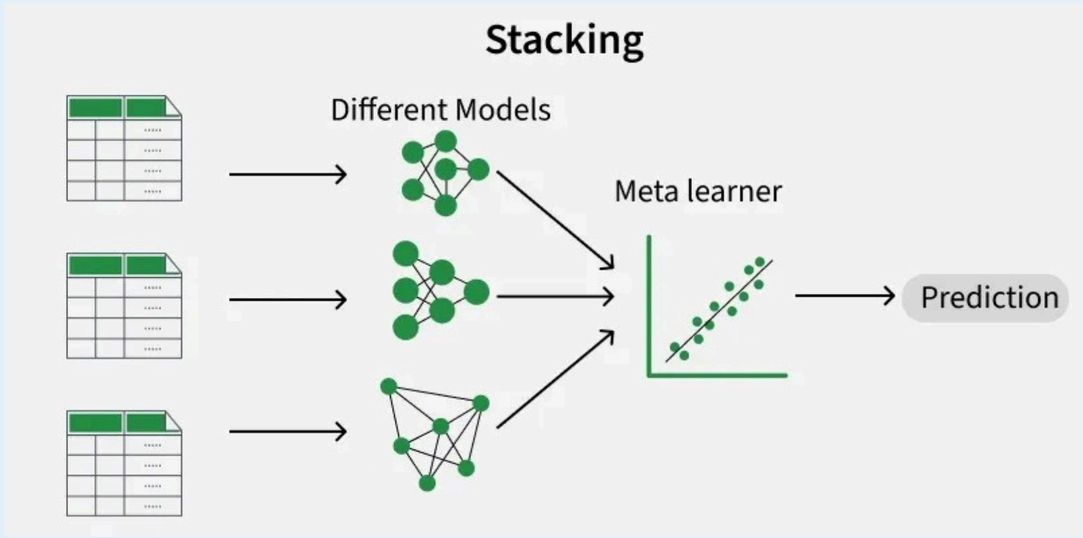


- Models are trained sequentially
- Each new model focuses on correcting errors made by previous models
- Assigns higher importance to misclassified loan applicants
- Improves model accuracy and reduces bias
- Popular boosting algorithms:
  - AdaBoost
  - Gradient Boosting
  - XGBoost



Stacking

- Combines predictions from multiple different models
- Uses a meta-model to learn how to best combine base model outputs
- Base models may include:
  - Logistic Regression
  - Decision Trees
  - Random Forest
- Meta-model learns optimal weighting for loan approval decisions
- Often achieves higher performance but is more complex



Comparisons

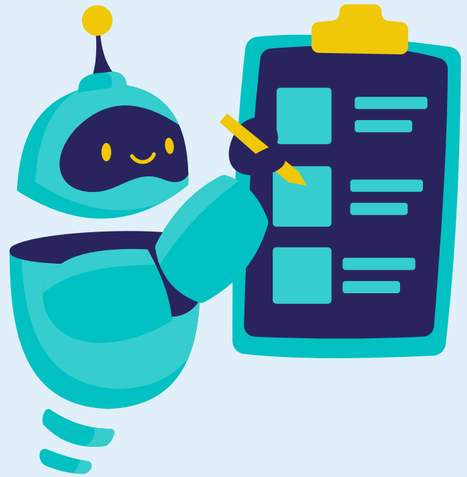
- Bagging
  - Best for reducing variance
  - Works well with unstable models
- Boosting
  - Best for improving accuracy
  - Sensitive to noisy data
- Stacking
  - Best overall performance
  - Computationally expensive
- Choice depends on dataset size, complexity, and performance needs

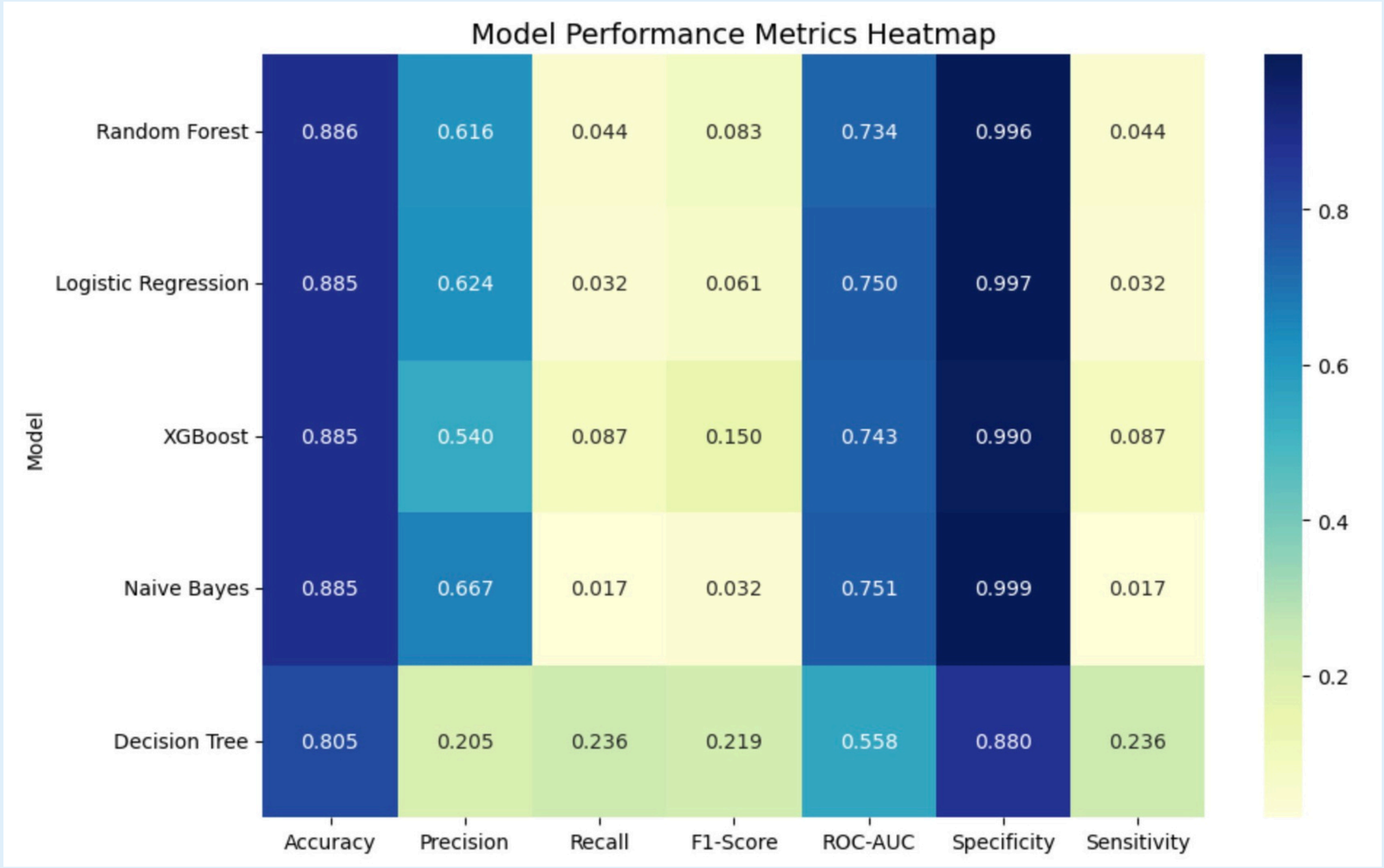
Other Applications

- Credit risk assessment
- Fraud detection
- Medical diagnosis
- Customer churn prediction
- Recommendation systems
- Stock market and financial forecasting

Conclusion

- Ensemble learning improves prediction accuracy and robustness
- Bagging reduces overfitting, boosting improves learning from mistakes
- Stacking leverages multiple models for optimal performance
- Highly effective for loan prediction problems
- Helps financial institutions make better, data-driven decisions




Model Performance Metrics Heatmap
Model
Random Forest
Logistic Regression
XGBoost
Naive Bayes
Decision Tree
0.886 0.616 0.044 0.083 0.734 0.996 0.044
0.885 0.624 0.032 0.061 0.750 0.997 0.032
0.885 0.540 0.087 0.150 0.743 0.990 0.087
0.885 0.667 0.017 0.032 0.751 0.999 0.017
0.805 0.205 0.236 0.219 0.558 0.880 0.236
Accuracy
Precision
Recall
F1-Score
ROC-AUC
Specificity
Sensitivity
0.8
0.6
0.4
0.2

THANK YOU....