2025-8-29

Random forest

—Based on Credit and Breast Cancer data

ZHANG JINGYI

Introduction

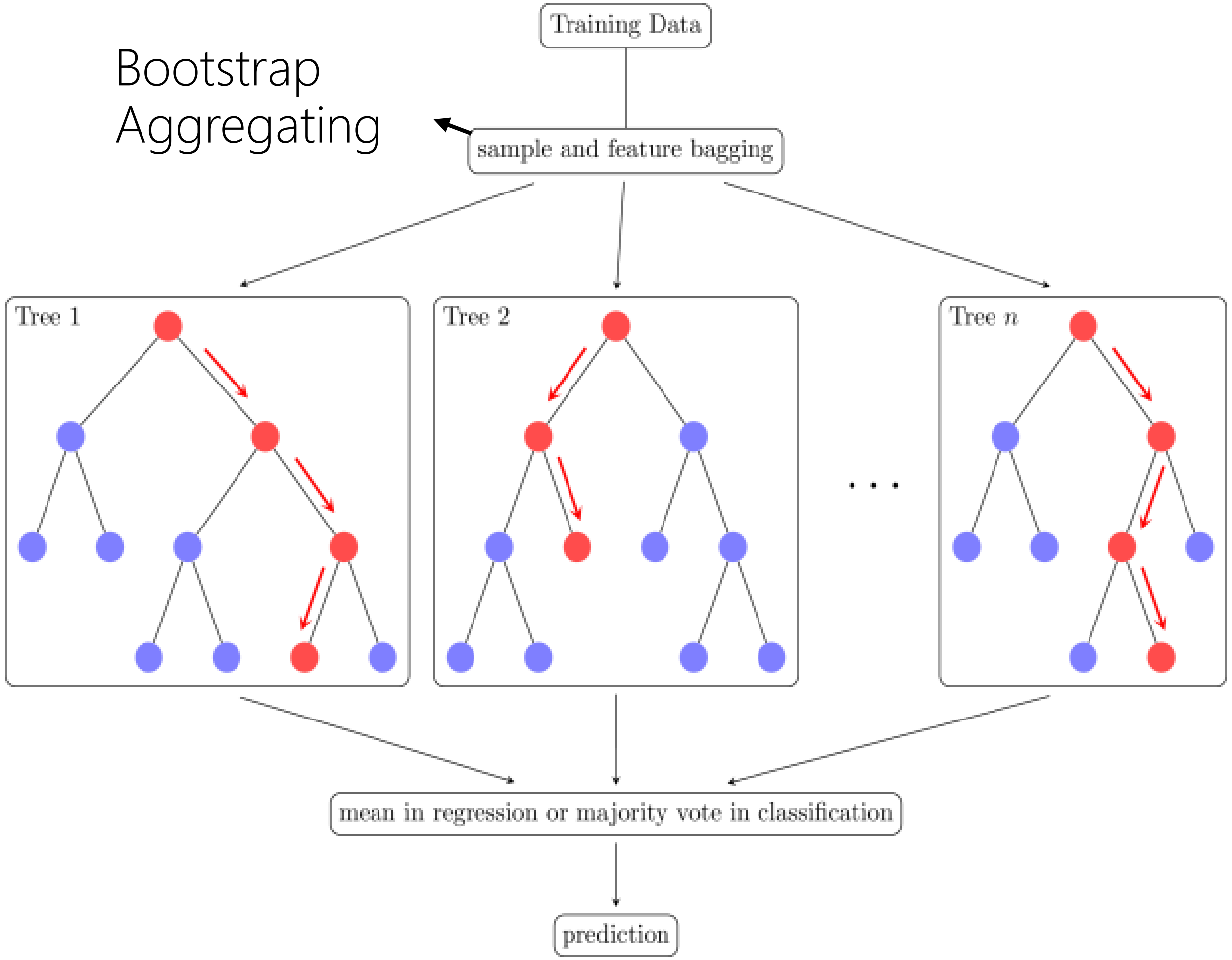
Dataset 1

Dataset 2

Discussion

Principle & Pros/Cons | Application Strategy

Bootstrap Aggregating



Advantages

- Strong resistance to overfitting
- Handles high-dimensional data well
- Can estimate feature importance

Disadvantages

- Hard to interpret
- Large models, slower to train
- High memory usage

Introduction

Dataset 1

Dataset 2

Discussion

Principle & Pros/Cons

Application Strategy

Workflow

Data preparation
↓
Feature selection
↓
Model building (randomForest)
↓
Model evaluation (accuracy / confusion matrix)
↓
Result interpretation

代码

```
# Load package
library(randomForest)

# Train model
model <- randomForest(
  Species ~ ., data = iris, ntree = 100,
  importance = TRUE
)

# View model results
print(model)
importance(model)

# Prediction
pred <- predict(model, iris)
table(pred, iris$Species)
```

Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

a) $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$

b) $Sensitivity = \frac{TP}{TP+FN}$

c) $Specificity = \frac{TN}{TN+FP}$

" How well does the Random Forest model perform in identifying high-risk customers? "

- **Why FN Matter** | ❌ **Bad customer → Predicted as good → Loan approved → Default → Loss**
- **Evaluation Focus** | ✅ **maximize sensitivity for the "bad" class (minimize false negatives)**

Source: How to avoid False Positives and False Negatives in Testing? | BrowserStack

Introduction
Dataset 1
Dataset 2
Discussion

status	age	job	purpose	amount	...	credit_risk
1	21	3	4	1049	...	1
2	36	3	4	2799	...	1
3	23	2	2	841	...	1
4	39	2	4	2122	...	1
5	38	2	4	2171	...	1
6	48	2	4	2241	...	1

status	age	job	purpose	amount	...	credit_risk
1	21	3	4	1049	...	good
2	36	3	4	2799	...	good
3	23	2	2	841	...	good
4	39	2	4	2122	...	good
5	38	2	4	2171	...	good
6	48	2	4	2241	...	good

Introduce-German Credit Dataset

- 1000 records;
- 20 predictors;
- Binary target variable (credit_risk: good/bad).

Preprocessing

- Converted codes to labeled factors;
- Set ordered factors for ordinal variables;
- Used conditional probabilities to assess feature impact.

visualizing

Introduction
Dataset 1
Dataset 2
Discussion

status	age	job	purpose	amount	...	credit_risk
1	21	3	4	1049	...	1
2	36	3	4	2799	...	1
3	23	2	2	841	...	1
4	39	2	4	2122	...	1
5	38	2	4	2171	...	1
6	48	2	4	2241	...	1

Account Status Category	Interpretation	Good (%)	Bad (%)
no checking account	no account or very poor credit	28	73
< 0 DM	overdrawn account	48	52
1 - < 200 DM	low balance	77	23
⩾ 200 DM or salary assignment	strong financial status	91	9

Introduce-German Credit Dataset

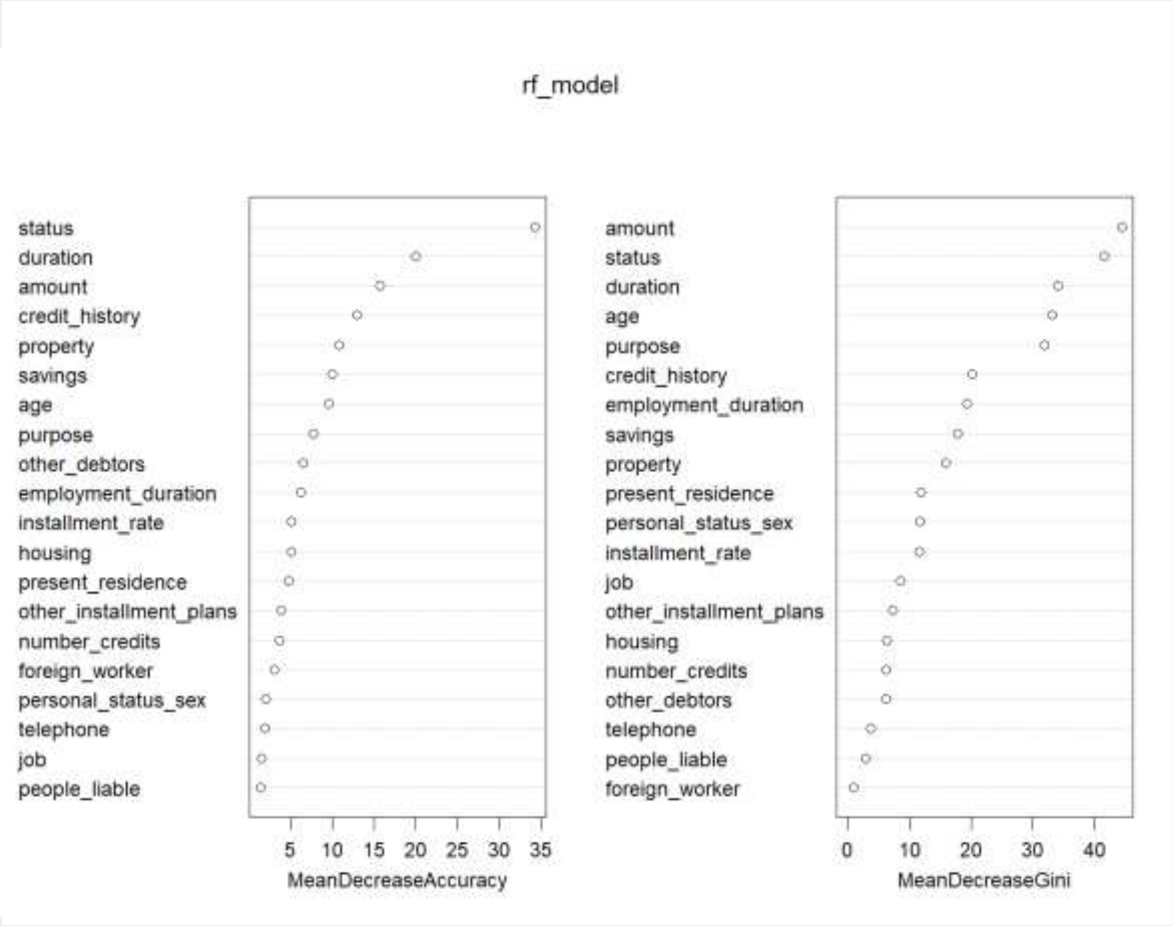
- 1000 records;
- 20 predictors;
- Binary target variable (credit_risk: good/bad).

Preprocessing

- Converted codes to labeled factors;
- Set ordered factors for ordinal variables;
- Used conditional probabilities to assess feature impact.

visualizing

- "Status" strongly predicts credit risk
- no account → bad risk;
- high balance → good risk.



Introduce-German Credit Dataset

- 1000 records;
- 20 predictors;
- Binary target variable (credit_risk: good/bad).

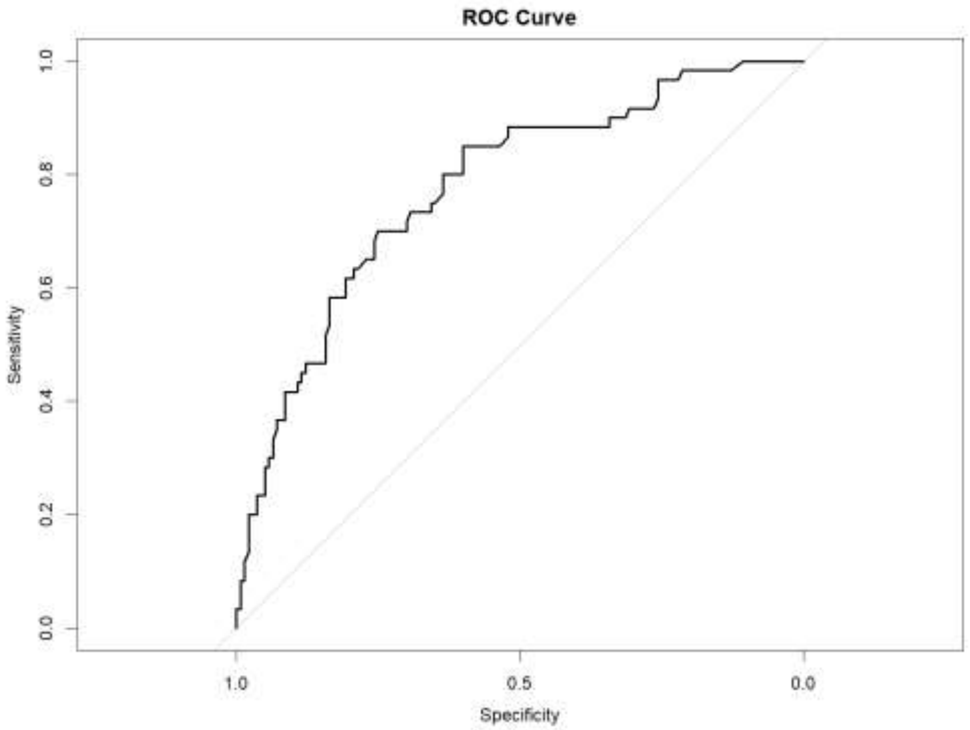
Preprocessing

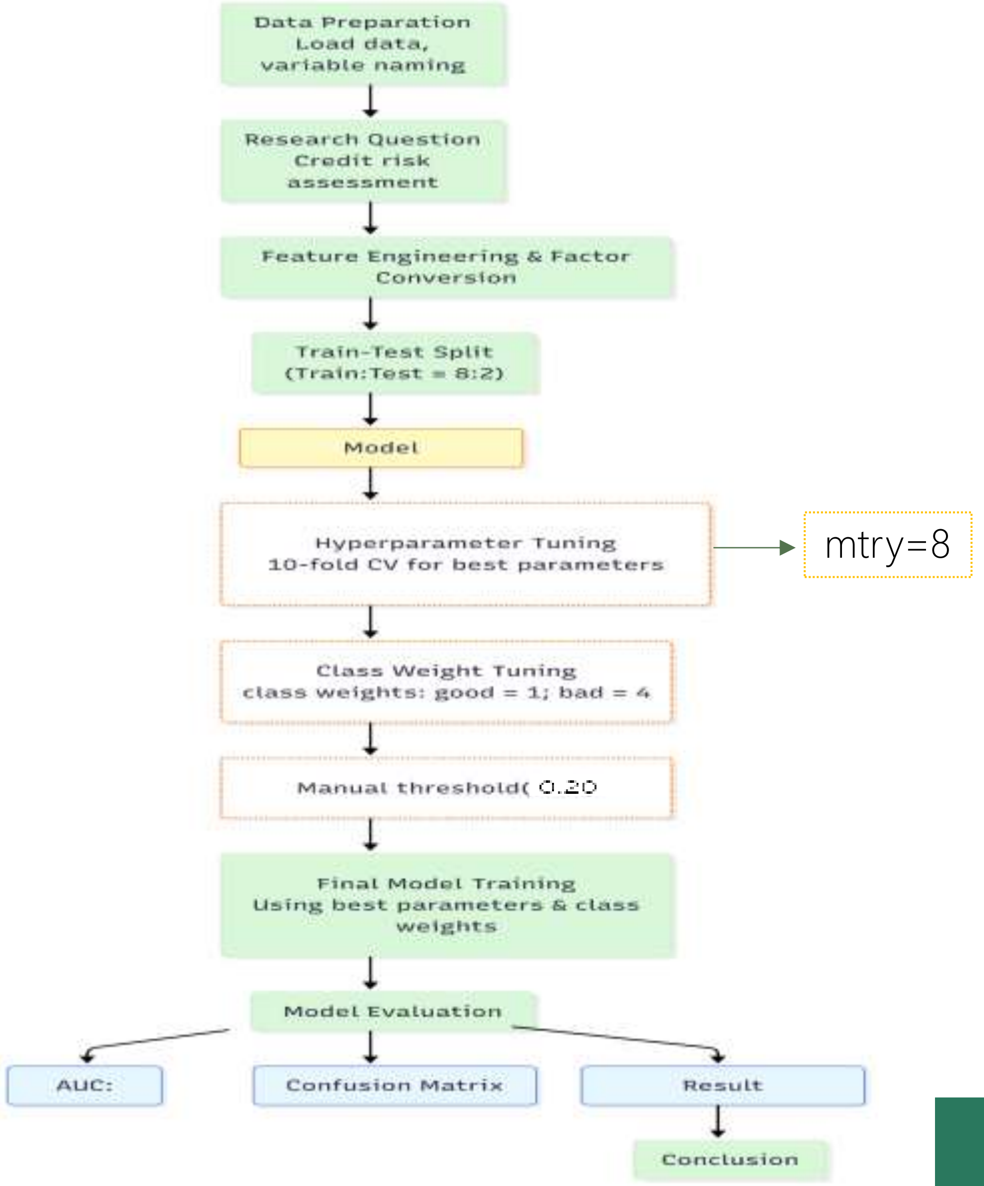
- Converted codes to labeled factors;
- Set ordered factors for ordinal variables;
- Used conditional probabilities to assess feature impact.

visualizing

- Top 5 features
  - ➢ Loan amount, status, duration, credit history, purpose
- Less feathures
  - ➢ foreign_worker,people_liable

- Model Evaluation: ROC & AUC





🔷 AUC Score: 0.777

- ROC curve rises toward top-left
- Model performs **better** than random guessing
- Indicates good classification ability

	Actual: good	Actual: bad
Pred: good	85 (TP)	**11 (FN)**
Pred: bad	**55 (FP)**	49 (TN)

Metric	Value	Business Meaning
Accuracy	0.670	Low, but not critical in imbalanced problems
Sensitivity (Recall)	0.817	Detecting good customers (positive class)
Specificity	0.607	Detecting bad customers (true negative rate)
FN Rate	**0.183**	**Low FN rate achieved**
FP Rate	**0.393**	**high rejection of good customers**

✅ **TP (True Positive)** = 85
→ Good customers correctly approved

✅ **TN (True Negative)** = 49
→ Bad customers correctly rejected

❌ **FP (False Positive)** = 55
→ Good customers wrongly rejected
(missed opportunity)

❌❌ **FN (False Negative)** = 11
→ Bad customers wrongly approved
⚠️ **leads to default risk**

‼️ **False Negatives** are the most critical, as they result in **actual financial loss**.

How should the Random Forest Model handle recurrence prediction in breast cancer?

False Negatives

- ✓ Missed recurrence delays treatment
- ✓ High medical and personal cost
- ✓ Must be minimized

False Positives

- ✓ May lead to unnecessary exams
- ✓ Mild anxiety, manageable consequences
- ✓ Less harmful

Therefore, reducing false negatives should be the top priority.

Supported by: "Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning" (2021)

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
no-recurrence-events	40-49	premeno	0-4	0-2	no	2	right	right_low	no
no-recurrence-events	60-69	ge40	15-19	0-2	no	2	left	left_low	no

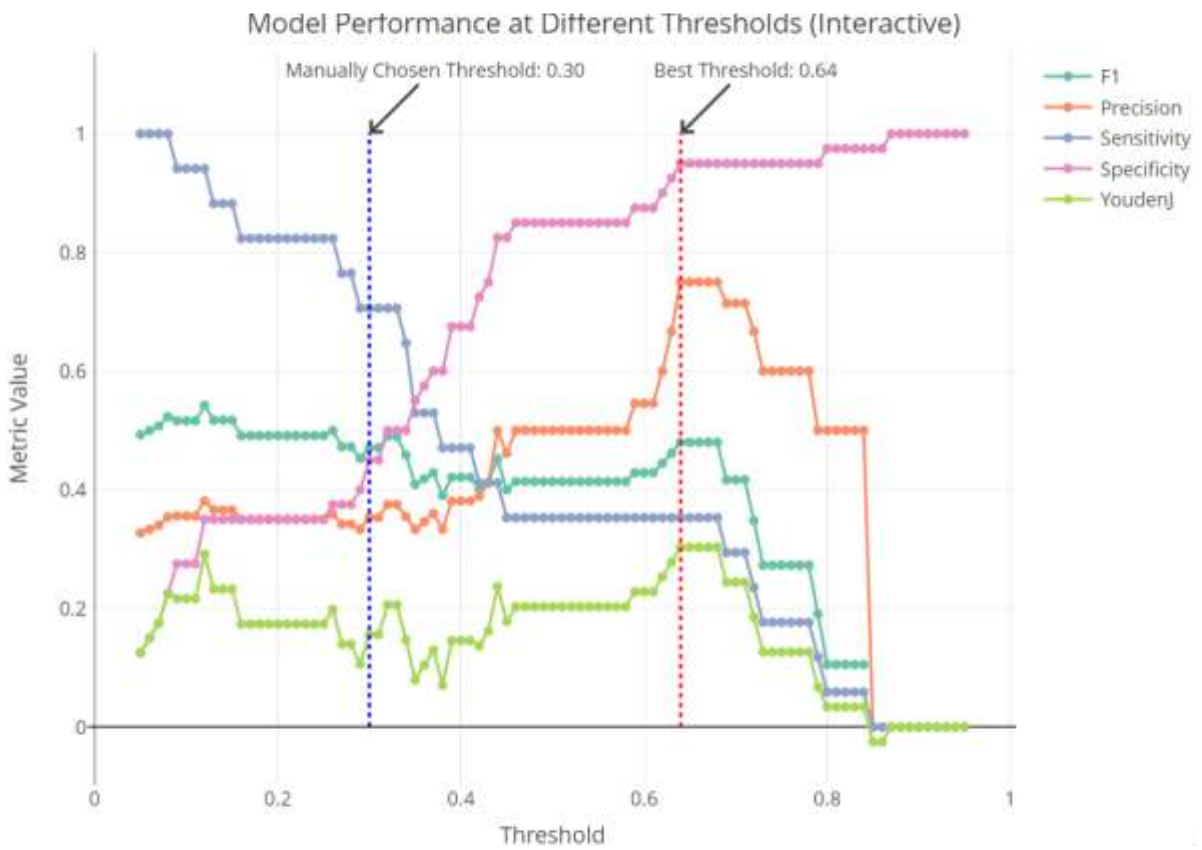
Recurrence	Age	Menopause	Tumor_Size	Inv_nodes	Node_caps	Deg_malig	Breast	Breast_quad	Irradiat
no_recurrence_events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
no_recurrence_events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
no_recurrence_events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
no_recurrence_events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
no_recurrence_events	40-49	premeno	0-4	0-2	no	2	right	right_low	no
no_recurrence_events	60-69	ge40	15-19	0-2	no	2	left	left_low	no

Dataset Overview: Breast cancer

- Source: the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.
- Donors: M. Zwitter and M. Soklic.
- Size: 286 instances (201 no recurrence, 85 recurrence)
- Features: 9 attributes + 1 target (Recurrence)
- **Target labels: no_recurrence_events / recurrence_events**

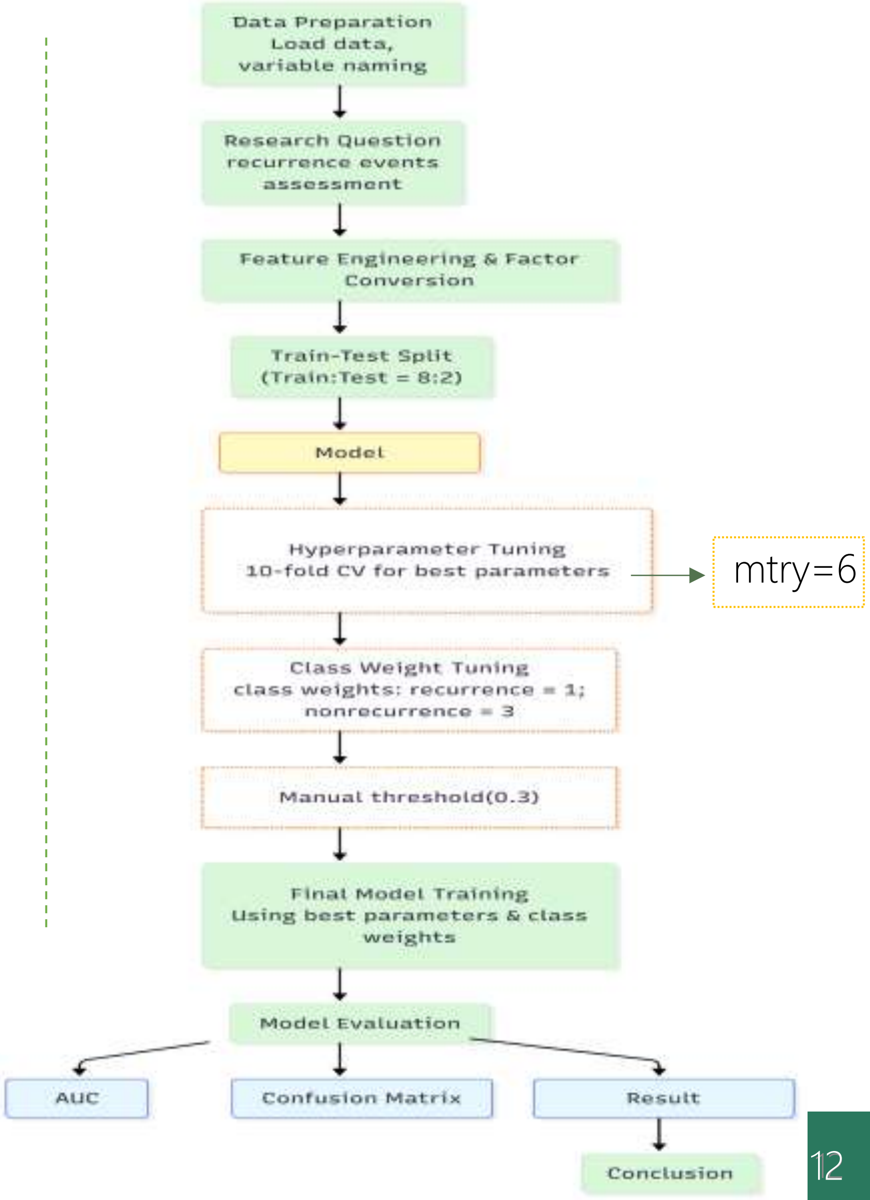
Data Preprocessing

- Renamed columns;
- Converted variables to proper types (factor/integer)



Healthcare Priority: Minimize False Negatives

- Threshold 0.3 **balances** sensitivity and specificity to prioritize minimizing false negatives
- Although **specificity drops**, extra checks (FP) are manageable
- **Better safe than sorry** in cancer recurrence prediction



	Actual: no_recurrence	**Actual: recurrence**
Pred: no_recurrence	18	5
Pred: recurrence	22	12

Metric	Value	Business Meaning
Accuracy	0.526	Low, but acceptable in medical classification
Sensitivity (Recall)	0.706	✅ High: most recurrence cases detected (true positive rate)
Specificity	0.450	Low: many non-recurrence predicted as recurrence
FN Rate	**0.294**	✅ **Low FN rate — very few recurrence patients missed**
FP Rate	**0.550**	**High: over-diagnosis is tolerable compared to under-diagnosis**

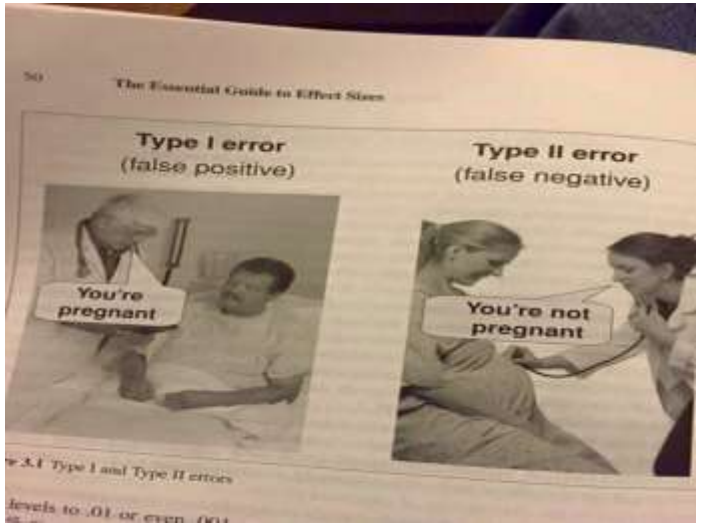
"In medical scenarios, missing a recurrence is riskier than over-diagnosing — thus FN ↓ is prioritized over FP ↓"

Introduction

Dataset 1

Dataset 2

Discussion



Summary

- Used Random Forest to classify "good" and "bad" credit risk.
- Ensemble trees provided strong and stable performance.

Current challenge:

- Type I (false positive) vs. Type II (false negative)
- hard to balance, especially when false negatives are high-risk.

📌 **What I've learned:**

"In real-world deployment, we should prioritize models that ensure high recall — even at the cost of some precision — to avoid missing high-risk events like credit defaults or breast cancer diagnoses. This experience made me realize that model optimization is not just about metrics, but about aligning with real-world consequences."