Two-Stage Classification Model for Imbalanced Data

R26121065 吳思蒨

Dataset

Dataset Description

- **Features : Contains multiple features related to demographic, clinical, or behavioral data (e.g., age, cholesterol, blood pressure).**
- **Target Variable : cardio: Binary label indicating the presence (1) or absence (0) of cardiovascular disease.**

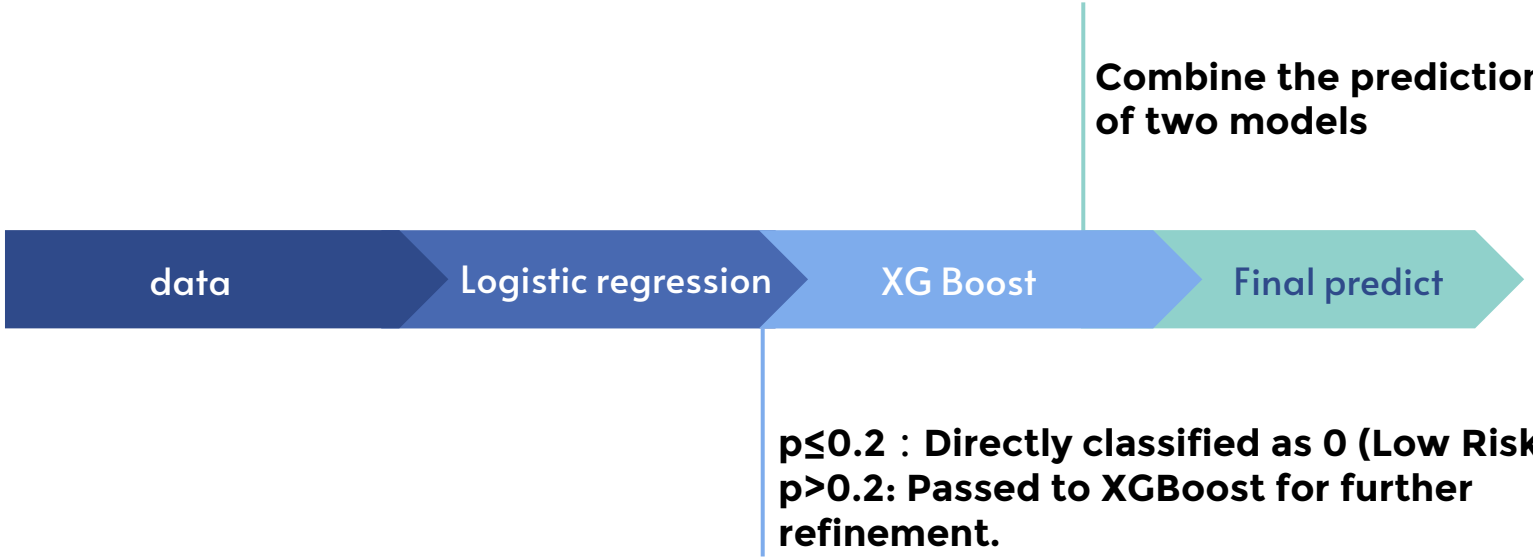
Imbalanced Data Problem

- **Total sample size : 70,000**
- **Final sample size : 38523 (after sampling)**
- **Class 0 (Negative) : 35021**
- **Class 1 (Positive) : 3502**

Cost Asymmetry

- **The cost of a False Negative (FN) is ten times higher than that of a False Positive (FP).**

Two-Stage Classification Process



Results

Stage1 (logistic regression)

	0	1
0	1418	4439
1	31	533

Precision : 0.11
Recall : 0.95

Stage2 (weighted XGBoost)

	0	1
0	3324	1115
1	175	358

Precision : 0.24
Recall : 0.67

Final prediction

	0	1
0	4742	1115
1	206	358

Precision : 0.24
Recall : 0.63

Threshold=0.5

Comparison

Normal XGBoost

	0	1
0	5762	95
1	512	52

Precision : 0.35
Recall : 0.09

Weighted XGBoost

	0	1
0	4900	957
1	258	306

Precision : 0.24
Recall : 0.54

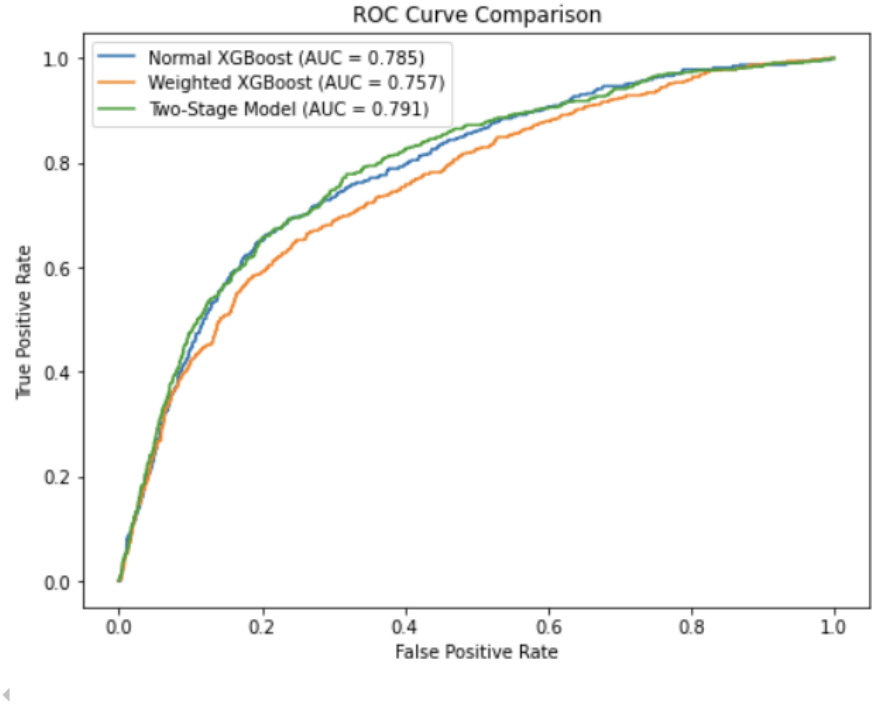
Two-Stage Model

	0	1
0	5094	763
1	261	303

Precision : 0.28
Recall : 0.54

Threshold=0.6

ROC curve


ROC Curve Comparison
Normal XGBoost (AUC = 0.785)
Weighted XGBoost (AUC = 0.757)
Two-Stage Model (AUC = 0.791)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

SUMMARY

Key Achievements

- **Proposed a Two-Stage Classification Model combining Logistic Regression and Weighted XGBoost.**
- **Improved recall for minority class (Class 1) to 0.54, while balancing precision.**

Future Directions

- **Further optimization of loss weights to improve precision.**

Thanks for your attention.