Technical Report for Supervisor

Amir A. Seid

2026-02-11

Purpose

This report documents the internal validation of a prediction model using Kaggle UCI Heart Disease data set. It aims to check the validity for both binary and 5-level ordinal classification.

Brief Summary

This analysis predicts heart disease (CAD) using the Kaggle UCI Heart Disease data set. Key steps performed include:

- Data cleaning and mice imputation We used Predictive Mean Matching for numerical variables and random forest for categorical variables after cross comparing the imputation results with logistic regression.
- Testing parallel assumption test (violated) We used Brant's test and visualizations to test the assumptions.
- Built a model using a Partial Proportional Odds logistic. Regression.
- Used Rubin's Rule to get pooled estimates for our model
- Built a binary and 5-level ordinal model to classify sick and healthy.
- Added a weighted factor (w)

  $w = \frac{1}{1.3 \cdot p^{1/3}}$

  where p is the class proportion
- to the predictive value matrix to increase the weight of higher class values.

Results

Our model failed to identify disease severity. Although the mean absolute error was below 0.5 (0.16), our multinomial model failed to identify class 4 and poorly performed on class 2 due to insufficient training samples. Class 4 had three values

Class Counts for our Target Variable in Test and Training data sets

Class	Observations-train	Observations-test
0	353	58
1	235	30
2	96	13
3	91	16
4	25	3

Confusion matrix for our weighted model (Multinomial prediction)

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
Class: 0	0.83	0.84	0.83	0.84	0.83	0.83	0.83	0.48	0.40	0.48	0.83
Class: 1	0.38	0.81	0.50	0.73	0.50	0.38	0.43	0.32	0.12	0.25	0.60
Class: 2	0.08	0.89	0.08	0.90	0.08	0.08	0.08	0.10	0.01	0.11	0.49
Class: 3	0.18	0.87	0.12	0.91	0.12	0.18	0.15	0.09	0.02	0.13	0.53

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
Class: 4	NA	0.98	NA	NA	0.00	NA	NA	0.00	0.00	0.03	NA

Confusion matrix for our weighted model Binomial Prediction(Healthy and Sick)

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
0.83	0.84	0.83	0.84	0.83	0.83	0.83	0.48	0.4	0.48	0.83

References

1. Heymans MW, Eekhout I. (2019). *Applied Missing Data Analysis*. Amsterdam: VU University.
2. Kaggle. (2019). UCI Heart Disease Dataset. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
3. Arı E. (2014). Parallel Lines Assumption in Ordinal Logistic Regression and Analysis Approaches. *International Interdisciplinary Journal of Scientific Research*, 1:8-23.