CS-433: Machine Learning Project 1

Peh Jin Yang, Celest Angela Tjong and Lee Ee Cheer

October 30, 2023

1 Introduction

Cardiovascular disease is a leading cause of death globally. This report outlines the steps taken in building a model to identify individuals with cardiovascular disease based on their lifestyle and medical background, possible future improvements and key findings.

2 Exploratory Data Analysis (EDA)

Out of **321** original columns, we selected **12** categorical features to construct our design matrix. While *pandas* is used, it is used solely for EDA and visualisation purposes. The Exploratory Data Analysis process is shown below.

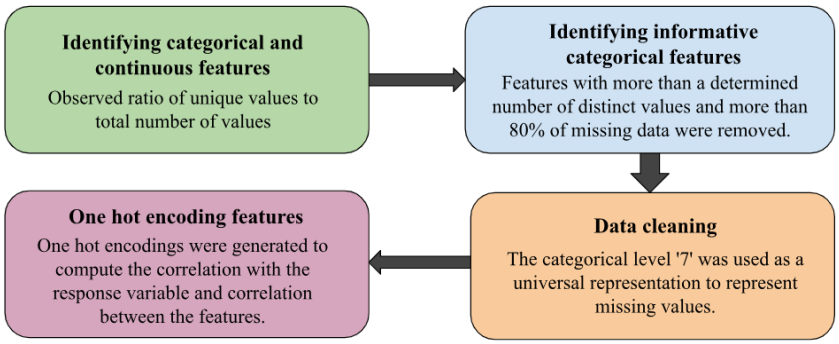

Figure 1: Data Preparation Process

3 Models and Method

3.1 Formulation of design matrix and data loading

A full-rank design matrix is prepared using one hot encoding and suppressing the categorical level '7' of each variable. Additionally, an intercept variable β_0 is added for model robustness.

3.2 Model training and evaluation

K-fold cross validation is performed with multiple epochs to generate random samples of the evaluation metrics on the train and validation data sets. This allows for the derivation of summary statistics which enables us to perform qualitative bias-variance decomposition analysis. Through an iterative model building process, it led up to the eventual optimal model, Model 5.

Model 1: **Mean Squared Error with original dataset**

Model 2: **Logistic Regression with original dataset**
These models have similar performance as a random predictor. Model 2 is preferable because of the higher recall and F1 score which suggests it can discriminate between positive and

negative samples better. The range of output values between 0 and 1 is also more desirable.

Model 3: **Logistic Regression with down-sampled dataset**
The model is able to fit the data better, as observed by the increase in F1 score. While the model managed to achieve low bias and high variance on the train-validation data set, there is a sharp decrease in performance as compared to the test data set. This suggests overfitting on the train-validation data set.

Model 4: **Logistic Regression with down-sampled, feature reduced dataset**
Feature reduction is performed to reduce noise in hopes of reducing the occurrence of False Positives and Negatives. For example, this is achieved by removing age-specific variables (e.g. '_FLSHOT6').

Model 5: **Regularised Logistic Regression with down-sampled, feature reduced dataset**
Given the degree of discrepancy between validation F1 and test F1 score, severe overfitting is present. Hence, a λ value of 8.25 is chosen to penalise the model sufficiently from overfitting but not too much which could result in underfitting.

4 Results

	Accuracy	Accuracy (Test)	Precision	Recall	F1	F1 (Test)
Model 1	0.912 ± 0.002	-	0.819 ± 0.150	0.003 ± 0.001	0.006 ± 0.002	-
Model 2	0.912 ± 0.002	-	0.769 ± 0.060	0.007 ± 0.003	0.014 ± 0.005	-
Model 3	0.750 ± 0.005	0.808	0.748 ± 0.010	0.754 ± 0.020	0.751 ± 0.008	0.376
Model 4	0.750 ± 0.008	0.751	0.748 ± 0.012	0.755 ± 0.011	0.752 ± 0.008	0.353
Model 5	0.717 ± 0.014	0.838	0.801 ± 0.022	0.579 ± 0.050	0.671 ± 0.026	0.387

Table 1: Model Results

Note: Columns not labelled with (Test) refers to results derived from the validation data set. Model 5 is chosen as the optimal model, trained with 100 epochs, a step-size γ of 0.1 and a regularisation term λ of 8.25.

5 Discussion

Given the F1 score of 0.387, further improvements can be made. Possible co-linearity between variables can be explored and exploited. For example, interaction terms could be added to the design matrix. Our team acknowledges that when we place more emphasis on recall, precision will be compromised. Thus, we have decided that weighted F1 scores could be utilised to give greater weight to False Negatives which is important in medical diagnosis.

6 Summary

In this report, a trained model is proposed to predict the occurrence of 'MICHD' given a certain clinical and lifestyle situation. We found that variables indicating commodities are more effective in identifying positive samples than variables related to a person's access to healthcare (e.g. vaccines or healthcare coverage).