Predicting Cardiovascular Events or Death in People with Dysglycemia Using Machine Learning Methods

by

Yuanhong Zhang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2025

© Yuanhong Zhang 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Cox regression is commonly used to analyze time-to-event for patients in tabular medical data. Hazard ratio can then be calculated to show how much riskier an event may occur for a patient in one group versus the other. Nonetheless, the hazard ratio and Cox regression rely on the proportional hazards assumption, which is not guaranteed. In this paper, we investigate the use of machine learning models to predict patients' outcomes and identify key factors that may influence the outcomes. We focused on the ORIGIN Trial dataset because it has undergone extensive analysis using Cox regression, allowing us to compare the machine learning model results with previous findings. Three outcomes, major adverse cardiovascular events (MACE), expanded composite outcome (COPRIM2), and all-cause death (ALLDTH), were analyzed in this thesis. The machine learning models we used are Neural Network (NN), Random Forest (RF) and Gradient Boosted Trees (GBT), which were trained with nested cross-validation to tune their hyperparameters. When testing the trained models for all three outcomes, we found that machine learning models had higher Area-Under-the-Curve scores (AUCs) than Cox regression (0.91-0.95 vs 0.63-0.65), and Random Forest and Gradient Boosted Trees had excellent recall scores (0.80 - 0.88). Subsequently, we used SHAP values, mean decrease in AUC, and partial dependency plots (PDPs) to further examine variable importance for RF and GBT. For MACE and COPRIM2, prior cardiovascular events (priorcv), cancer, and blood lipid measures are the most important variables, while for ALLDTH, cancer and kidney functions related measures are the most important variables. The PDPs are harder to analyze than hazard ratio due to having no assumptions and fewer restrictions, but it is useful to estimate the non-linear relation between an explanatory variable and the average probability of the outcome occurring to patients in the dataset.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Anita Layton, who supported me during my time as a Master student and provided her valuable guidance on this thesis.

In addition, I am extremely thankful for Professor Hertzel Gerstein, who was able to find time to help me with his experience and expertise on the ORIGIN Trial.

I am also immensely grateful to my committee mmembers, Assistant Professor Anamaria Crisan and Yuntian Deng, for their constructive feedbacks and suggestions.

I also really appreciate the staff at University of Waterloo and Population Health Research Institute, who helped me with technical problems and accommodated my requests.

Finally, I could not have completed my research without my family, fiancée, and friends, who gave me encouragement and mental strength.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

List of Figures vi

List of Tables vii

List of Abbreviations viii

1 Introduction 1

1.1 Dataset Background . . . 3

1.2 Previous Results . . . 3

1.3 Overview . . . 4

2 Overview of Machine Learning Models Used 5

2.1 Neural Networks . . . 5

2.2 Tree Based Models . . . 9

2.2.1 Gradient Boosted Trees . . . 13

2.2.2 Random Forest . . . 14

3 Method **16**
3.1 Data Preprocessing . . . 16
3.2 Model Building . . . 18
3.3 Model Training . . . 18
3.4 Tuned Model Architecture . . . 20
3.5 Model Interpretabilty . . . 20

4 Result **22**
4.1 Model Prediction Comparisons . . . 22
4.2 Variable Importance . . . 24
4.2.1 Quantifiable Analysis . . . 24
4.2.2 Visual Analysis . . . 26

5 Conclusion **34**

References **36**

APPENDICES **40**

A Variable Definition, Type, and Preprocessing **41**

B GradientBoostedTrees **44**
B.1 ALLDTH . . . 44
B.2 COPRIM2 . . . 48
B.3 MACE . . . 52

C RandomForest **59**
C.1 ALLDTH . . . 59
C.2 COPRIM2 . . . 64
C.3 MACE . . . 68

List of Figures

2.1 A diagram of a feed-forward neural network with an input layer of dimension 3 in red, two hidden layers, each with 6 neurons, in blue, and an output layer of dimension 2 in green 6

2.2 A diagram of a neuron 7

2.3 A diagram of a decision tree and the regions it splits into 11

4.1 Partial Dependence Plots of Random Forest Model for ALLDTH. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explanatory variable measure on x-axis: (a) survival time, (b) whether they have cancer, (c) age, (d) estimated glomerular filtration rate (eGFR), (e) average blood glucose (sugar) levels for the last two to three months (hba1c), (f) ankle-brachial blood pressure index (abi) 30

4.2 Partial Dependence Plots of Gradient Boosted Trees Model for ALLDTH. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explanatory variable measure on x-axis: (a) survival time, (b) whether they have cancer, (c) age, (d) screening visit creatine value (SVCREV), (e) maximum grip strength (max_grip), (f) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke) 31

4.3 Partial Dependence Plots of Random Forest Model for MACE. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explanatory variable measure on x-axis: (a) survival time, (b) whether they have cancer, (c) age, (d) screening visit triglyceride value (SVTRIV), (e) screening visit high-density lipoprotein cholesterol value (svhdlv), (f) prior cardiovascular events (priorcv) 32

4.4 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explanatory variable measure on x-axis: (a) survival time, (b) body-mass index (bmi), (c) arm systolic blood pressure (sbparm), (d) arm systolic blood pressure (sbparm), (e) whether the patient is male, (f) screening visit ligh-density lipoprotein cholesterol value (svldlv) 33

B.1 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer 44

B.2 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit creatine value (SVCREV), (b) age 45

B.3 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) maximum grip strength (max_grip), (b) whether the patient uses statin 45

B.4 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior cardiovascular events (priorcv), (b) prior diagnosis of diabetes 46

B.5 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke), (b) whether the patient has hypertention 46

B.6 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether th patient uses renin angiotensin system drugs (ACE-Inhibitors or Angiotensin 2 blocker) , (b) Whether patient had laser therapy for eye disease 47

B.7 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) Whether the patient uses ASA or antiplatelet, (b) whether the patient uses anticoagulant 47

B.8 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient uses anticoagulant, (b) Whether the patient uses other BP Drugs (excluding Beta Blocker and RAS drugs) 48

B.9 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer 48

B.10 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit creatine value (SVCREV), (b) Hip circumference (rhipc) 49

B.11 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior cardiovascular events (priorcv), (b) Diabetes Duration . . 49

B.12 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) average blood glucose (sugar) levels for the last two to three months (hba1c), (b) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke) 50

B.13 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whather the patient has gestational diabetes, (b) whether the patient uses statin 50

B.14 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient has previous preeclampsia or eclampsia, (b) stable angina 51

B.15 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) ankle-brachial index during follow-up 51

B.16 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer 52

B.17 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit creatine value (SVCREV), (b) screening visit fasting glucose values (svfgv) 52

B.18 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit total cholesterol value (svchov), (b) Waist circumference (rwaistc) 53

B.19 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) kidney disease index (kdi), (b) average blood glucose (sugar) levels for the last two to three months (hba1c) 53

B.20 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit triglyceride value (SVTRIV), (b) prior cardiovascular events (priorcv) 54

B.21 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior myocardial infarction (priormi), (b) Hip circumference (rhipc) 54

B.22 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) maximum grip strength (max_grip), (b) number of children . . . 55

B.23 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whather the patient has gestational diabetes, (b) body-mass index (bmi) 55

B.24 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) ankle-brachial index during follow-up, (b) whether the patient uses diuretic 56

B.25 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient uses anticoagulant, (b) whather the patient has previous preeclampsia or eclampsia 56

B.26 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whehter the patient is male, (b) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke) 57

B.27 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient has hypertention, (b) whehter the patient drinks alcohol 57

B.28 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) ORIGIN allocations (1=glargine;2=standard care) 58

C.1 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer 59

C.2 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) age, (b) estimated glomerular filtration rate (eGFR) 60

C.3 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) ankle-brachial blood pressure index (abi), (b) Diabetes Duration 60

C.4 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior atrial fibrillation (prioraf), (b) average blood glucose (sugar) levels for the last two to three months (hba1c) 61

C.5 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) number of children, (b) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke) 61

C.6 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient uses diuretic, (b) whether the patient uses aldosterone antagonist 62

C.7 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior stroke (priorstrk), (b) prior cardiovascular events (priorcv) 62

C.8 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whehter the patient is male, (b) whather the patient has gestational diabetes 63

C.9 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whather the patient has gestational diabetes, (b) prior myocardial infarction (priormi) 63

C.10 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer 64

C.11 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior cardiovascular events (priorcv), (b) screening visit ligh-density lipoprotein cholesterol value (svldlv) 64

C.12 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) arm systolic blood pressure (sbparm), (b) average blood glucose (sugar) levels for the last two to three months (hba1c) 65

C.13 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) body-mass index (bmi), (b) age 65

C.14 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whehter the patient is male, (b) Whether the patient has cataracts 66

C.15 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient uses anticoagulant, (b) whether the patient uses aldosterone antagonist 66

C.16 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke), (b) Diabetes Duration 67

C.17 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whather the patient has gestational diabetes, (b) whehter the patient drinks alcohol 67

C.18 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whehter the patient drinks alcohol, (b) whather the patient has previous preeclampsia or eclampsia 68

C.19 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer 68

C.20 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) arm diastolic blood pressure (dbparm), (b) arm systolic blood pressure (sbparm) 69

C.21 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit high-density lipoprotein cholesterol value (svhdlv), (b) age 69

C.22 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit triglyceride value (SVTRIV), (b) Waist circumference (rwaistc) 70

C.23 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior cardiovascular events (priorcv), (b) whether the patient uses aldosterone antagonist 70

C.24 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient uses diuretic, (b) prior myocardial infarction (priormi) 71

C.25 Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whather the patient has previous preeclampsia or eclampsia . . 71

List of Tables

2.1 Definitions of different activation functions 7

4.1 Model Prediction AUCs on Test Data for the Three Outcomes 23
4.2 Model Prediction Accuracy and Recall on Test Data for the Three Outcomes 23
4.3 Variable Importance Ranked by SHAP Value 25
4.4 Variable Importance Ranked by Mean Decrease in AUC 25

List of Abbreviations

abi ankle-brachial blood pressure index 24, 27, 30, 60

ALLDTH all-cause death iii, 3, 4, 23–28, 30–35, 44–71

AUC Area-Under-the-Curve score iii, 1, 2, 4, 16, 22–25, 34, 35

BCE Binary Cross Entropy 8, 19, 22

bmi body-mass index 24, 28, 33, 55, 65

COPRIM2 expanded composite outcome iii, 3, 4, 23–26, 28, 33–35, 44–71

CV Cardiovascular 2–4, 24, 34

dbparm arm diastolic blood pressure 24, 69

eGFR estimated glomerular filtration rate 24, 26–28, 30, 60

EHR Electronic Health Records 1

GBT Gradient Boosted Trees iii, 5, 13, 16, 20, 22, 23, 34

hba1c average blood glucose (sugar) levels for the last two to three months 24, 26–28, 30, 50, 53, 61, 65

HF heart failure 1

kdi kidney disease index 24, 26, 53

LeakyReLu leaky rectified linear unit 6, 7

MACE major adverse cardiovascular events iii, 3, 4, 23–26, 28, 32, 34, 35, 52, 68

max_grip maximum grip strength 24, 31, 45, 55

NN Neural Network iii, 5, 16, 20, 22, 23, 34

PDP partial dependency plot iii, 4, 16, 21, 27–29, 35

prioraf prior atrial fibrillation 24, 61

priorcv prior cardiovascular events iii, 24–26, 32, 35, 46, 49, 54, 62, 64, 70

priormi prior myocardial infarction 24, 54, 63, 71

priorstrk prior stroke 62

ReLu rectified linear unit 6, 7

RF Random Forest iii, 5, 16, 20, 22, 23, 27, 28, 34

rhipc Hip circumference 24, 49, 54

rsmoke smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) 24, 31, 46, 50, 57, 61, 67

rwaistc Waist circumference 24, 53, 70

sbparm arm systolic blood pressure 24, 28, 33, 65, 69

svchov screening visit total cholesterol value 24, 53

SVCREV screening visit creatine value 24, 26, 31, 45, 49, 52

svfgv screening visit fasting glucose values 24, 26, 52

svhdlv screening visit high-density lipoprotein cholesterol value 24, 28, 32, 69

svldlv screening visit ligh-density lipoprotein cholesterol value 24, 28, 33, 64

SVM Support Vector Machine 1

SVTRIV screening visit triglyceride value 24, 32, 54, 70

Chapter 1

Introduction

Recent advancements in machine learning techniques have enabled modeling for more complex relations in many areas of study. A significant opportunity lies in tabular data in clinical research. Clinical research typically employs logistic regression to predict different categorical outcomes and uses Cox regression to identify risk factors. This paper will explore the possibility of using machine learning models to predict outcomes and identify risk factors for a clinical research dataset.

In 2015, Panahiazar and Taslimitehrani et al. used machine learning methods for heart failure (HF) survival analysis[1]. The goal of their study is to propose a HF survival risk prediction model by adding new variables derived using the Electronic Health Records (EHR) data to improve prediction accuracy and performance of the model. The models calculate a heart failure survival risk score for patients so that high-risk patients can be identified. They trained the models on 1560 patients and tested on a dataset with 3484 patients in order to show model generalizability. They used five classification algorithms: random forest, logistic regression, Support Vector Machine (SVM) regression, decision tree, and ada boost. They tested variable importance by comparing the Area-Under-the-Curve scores (AUCs) of models using only baseline features versus using extended features. The result showed that the models generally performed better with the extended features, except for SVM. Their best results came from logistic regression and random forest with the extended feature set. The paper showed that the machine learning models have good generalizability to new unseen data and new columns.

In 2018, Wong and Young et al. used machine learning methods to predict incident delirium risk based on EHR of hospitalized patients and compared the result to the usual nurse-administered questionnaires, which is the AWOL System[2]. Their dataset consisted

of 14227 patients in the train data and 3996 patients in the test data, and 796 clinical variables selected by an expert panel as features. They compared the following machine learning algorithms: penalized logistic regression (GLMNET), gradient boosting machine (GBM), artificial neural network with a single hidden layer (NN), linear support vector machine (SVM), and random forest (RF) using R. On test data, the baseline AWOL system achieved an AUC of 0.678, while GBM, GLMNET, and RF achieved the best AUCs of 0.855, 0.854, and 0.848, respectively. The best model, GBM, had a sensitivity of 0.597 at the 90% specificity threshold. GBM, GLMNET, and RF selected 345, 114, and 588 variables from the initial 796 variables. They used the variable selection frequency as a metric for variable importance. This is a good way to determine which variables are useful to the model, but it cannot determine how heavily the variables influence the model prediction.

Weng and Reps et al. applied machine learning methods to improve Cardiovascular (CV) risk prediction accuracy in 2017[3]. The dataset used was a routine clinical data of 378256 patients from UK, and 6.6% of the patients had CV events (24970 patients). They used four machine learning algorithms (random forest, logistic regression, gradient boosting machines, neural networks) and compared them to an established algorithm from American College of Cardiology (ACC) guidelines to predict first CV event over 10 years. They assessed the model accuracy with AUC and sensitivity. Compared to the ACC guideline, the use of neural networks improved AUC from 0.728 to 0.764, and the other three types of model had small margin of improvement for AUC as well. In addition to AUC, the machine learning models improved prediction accuracy and sensitivity. The researchers did not apply cross-validation during training, which may have increased the the performance further by tuning and validating the model's hyperparameters in a systematic way.

Weng and Reps et al. also tried to identify new risk factors. They ranked the variable importance by weighting in neural networks, variable selection frequency in random forest and gradient boosted machines, and coefficient effect size in ACC and logistic regression model. All machine learning models selected SES: Townsend Deprivation Index[4], which was not included in the ACC guideline. This variable is a measure in material deprivation, which may lead to an increase in mental and physical stress[5] and thus link to increased CV risk[6]. This shows that machine learning models can identify new risk factors. However, neural networks used more uncommon factors that do not have a clear scientific explanation, such as whether the patient's BMI is missing and ethnicity. This is likely due to the data is biased and neural networks are using that bias to make predictions, which shows data preprocessing is crucial before training machine learning models.

1.1 Dataset Background

The dataset used in this thesis was gathered from the ORIGIN Trial, where the design and final results were published[7, 8]. The researchers randomly assigned 12537 people, with a mean age of 63.5 with Cardiovascular (CV) risk factors plus impaired fasting glucose, impaired glucose tolerance, or type 2 diabetes, to receive insulin glargine (with a target glucose level glucose $\leq$ 5.3 mmol/L) or standard care and receive n-3 fatty acids or placebo with the use of 2-by-2 factorial design. The median follow-up time was 6.2 years.

This paper focuses on three binary outcomes measured for each patient. The first is major adverse cardiovascular events (MACE) which includes the composite of whether the patient suffered from nonfatal myocardial infarction, nonfatal stroke, or Cardiovascular (CV) death; the second is an expanded composite outcome (COPRIM2) that includes whether the patient had MACE, heart failure hospitalization, or revascularization; the third outcome of interest is all-cause death (ALLDTH) of the patient. These three outcomes are all binary variables with class imbalances, where there are fewer patients with outcomes than without. Specifically, the percentage of patients who had MACE is 16%, who had COPRIM2 is 28%, and who had ALLDTH is 15%.

1.2 Previous Results

Previous research found that the rates of incident Cardiovascular (CV) outcomes were similar in the insulin glargine and standard-care groups, with no significant difference in cancer[9]. Insulin glargine has also been found to have a neutral effect on both initial and recurrent heart failure hospitalizations[10]. Other factors that were observed to have a neutral effect on all-cause mortality, CV mortality, and CV outcomes are insulin resistance, supplementation with n–3 fatty acids, and non-severe hypoglycaemia[11]. In contrast, weight loss and hypoglycaemia are associated with an increased risk of mortality and CV outcomes. Patients with lower BMI have a higher mortality, while patients with normal BMI have the lowest mortality[11].

In the paper "Identifying Novel Biomarkers for Cardiovascular Events or Death in People With Dysglycemia" by Gerstein et al. in 2015[12], they focused on the three outcome predictions with the ORIGIN Trial data and identified risk factors. They also conducted a complementary biomarker study, which measured more biomarkers for a subset of the ORIGIN Trial patients (8401 of 12537). They used logistic regression models to predict the three outcomes. When using the ORIGIN Trial data without biomarkers, the C-statistics

for the models (which is the same as AUC for binary predictions) are approximately 0.63-0.65 for all three outcomes. When including the biomarkers, AUC on test data increases to 0.69 for MACE, 0.68 for COPRIM2, and 0.73 for ALLDTH. Using Cox proportional hazards models, they found 8 clinical factors that have comparatively high hazard ratios for the 3 outcomes, which are previous CV event, Albuminuria, gender, age for men greater than 55 years old and women greater than 65 years old, LDL/HDL, current smoking, previous diabetes mellitus, and hypertension. They also found 8 biomarkers linked to MACE, 12 biomarkers linked to COPRIM2, and 10 biomarkers linked to ALLDTH that can increase prediction accuracy.

1.3 Overview

In this paper, we focus on using machine learning methods to predict the patients' outcomes and determine the most important factors to these outcomes. We will leverage the following three machine learning models: neural networks, random forest, and gradient boosted decision trees. Neural networks are models that are neuron-based and computationally expensive. They are very powerful since if they are well-structured, they can imitate any multivariate function[13, 14], but tuning them can be a lot more difficult, as they are more susceptible to overfitting. On the other hand, Random Forest and Gradient Boosted Trees are more traditional machine learning methods that build decision trees to make predictions. These models are greedy estimators and are robust during training[15, 16]. All of these models are very mature and have readily available packages that allow the implementation of training and testing. The interpretability of these models improved with the ability to determine variable importance using various methods such as SHAP values, mean decrease in Area-Under-the-Curve score (AUC), and partial dependency plot (PDP). By analyzing these variable importance metrics, we can identify factors which influence outcomes of the patients as well similar to Cox regression.

Chapter 2

Overview of Machine Learning Models Used

In this paper, we focused on applying the following three machine learning models to the ORIGIN dataset: Neural Network (NN), Gradient Boosted Trees (GBT), Random Forest (RF). They have their strengths and weaknesses. A well designed Neural Network (NN) can approximate any multivariate function[14, 13], but can suffer from this too in practice due to high variance. Gradient Boosted Trees (GBT) and Random Forest (RF) are both ensemble tree-based models, which are more robust, but the explainability may be lower. In this chapter, we provide the necessary details to understand each of the models.

2.1 Neural Networks

Neural networks is a powerful model nowadays that is capable of representing any multivariate function given 3 layers[14, 13]. This makes it a good candidate for modeling patients and outcomes for the ORIGIN Trial dataset.

A neural network, as the name indicates, is a network of neurons. More specifically, the basic form of it is made up of an input layer, one or more hidden layer, and an output layer. This basic version is usually called a feed-forward neural network. We show an example of a neural network in Figure 2.1. The input layer is just an layer that accepts the input, and will have the same dimension as the input vector. The output layer is used to output the results in the desired output dimension. The hidden layers contain neurons, and each neuron accepts all outputs of the previous layer as a vector. Then, the neuron

processes them via a weighted sum, and applies an activation function to that weighted sum to generate the output[17].

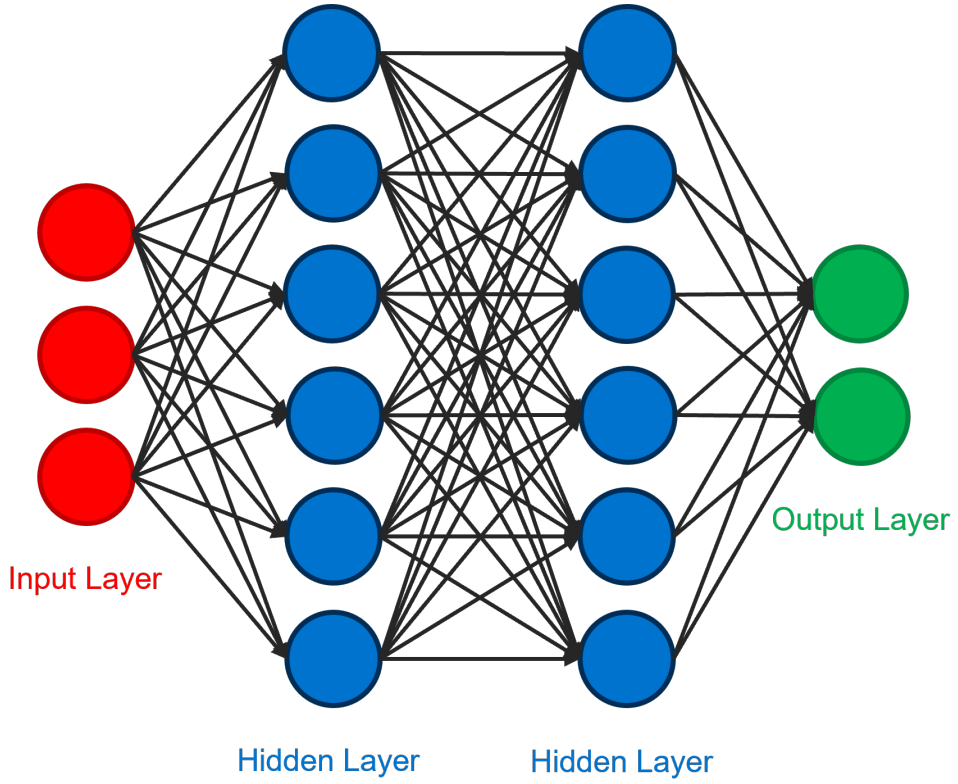


Figure 2.1: A diagram of a feed-forward neural network with an input layer of dimension 3 in red, two hidden layers, each with 6 neurons, in blue, and an output layer of dimension 2 in green

To show how each neuron in the hidden layer works, we include Figure 2.2 to better illustrate the process. w_0 is the bias, and w_i are the weights. a_0 is just a constant 1, and a_i are the outputs from the previous layer, which acts as the input of a neuron in the current layer. The function g is the activation function, which is applied to the sum and generates the final output. Formally, each neuron receives the input vector $\mathbf{a}$ and weights and bias vector $\mathbf{w}$, and it has activation function g, then its output is:

$$g(\mathbf{w}^T\mathbf{a}) \tag{2.1}$$

There are many activation functions for a neural network to choose from. Some common activation functions are: sigmoid, tanh, rectified linear unit (ReLu), and leaky rectified linear unit (LeakyReLu). We provide their function definitions in Table 2.1. In practice,

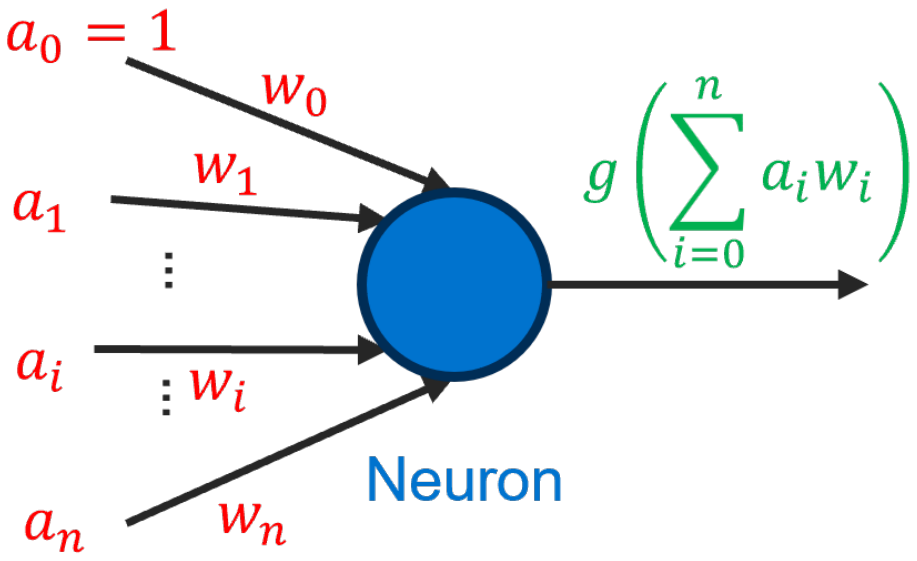


Figure 2.2: A diagram of a neuron

ReLu and LeakyReLu work the best since both of them do not suffer from the vanishing gradient problem. Moreover, LeakyReLu allows negative input values to "leak through" since it is slightly sloped in the negative domain. This prevents the dying neuron problem that ReLu suffers[18].

name	equation
sigmoid	$\sigma(x) = \frac{1}{1+e^x}$
tanh	$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
ReLU	$f(x) = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases}$
LeakyReLU	$f(x) = \begin{cases} \alpha x & x \leq 0 \\ x & x > 0 \end{cases}$ where α is a small positive value

Table 2.1: Definitions of different activation functions

The weights are usually randomly initialized at the creation of model, while the bias is set to 0 at the beginning. During training, we want to optimize a loss function, which is the objective of the model. A common loss function for binary classification problems is binary cross-entropy loss, while for regression problems it is the mean-squared error.

In this paper, we focus on classification, and our outcome is binary, so we used Binary Cross Entropy (BCE) loss[19]. Suppose we have n samples in a dataset X, each sample i is given by $\mathbf{x}_i, y_i$, where $\mathbf{x}_i$ is a input vector for the model and y_i is the expected output of either 1 or 0, and say we have a machine learning model $f(\mathbf{x}_i)$, then the BCE loss is given by:

$$l(X, f) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(f(\mathbf{x}_i)) + (1 - y_i)\log(1 - f(\mathbf{x}_i))\right) \tag{2.2}$$

The weights are updated via backward propagation[20]. The backward propagation algorithm is described in Algorithm 1 below, where l is the loss function, α is the learning rate, w is an arbitrary weight in the model f, and p is the number of epochs.

Algorithm 1 Backward propagation

Require: *data*
  $i \leftarrow 1$
  while $i <= p$ **do**
    $w \leftarrow w - \alpha \sum_{\text{data}} \frac{\partial l}{\partial w}$
    $i \leftarrow i + 1$ ▷ for each w in f
  end while

The algorithm uses the same idea as gradient descent, where we find the minimum of the loss function by taking a small step in the opposite direction of the gradient, as the gradient is the direction of the steepest ascent. This algorithm have many variants in implementation, because the amount of data may be big in which case it is difficult to calculate $\sum_{\text{data}} \frac{\partial l}{\partial w}$ with the entire dataset. A common solution is batching[21], and the batched backward propagation is described below in Algorithm 2, where additionally we have the batch size b, and let the number of samples in the dataset be n. The algorithm updates the weights based on the gradient of each batched data instead of the entire dataset.

Another improvement in practice is that instead of determining the epochs arbitrarily, we can add early stopping to prevent the model from overfitting to the data[22]. This can be done by splitting a validation set from the training data, and checking whether the loss on validation is decreasing. The algorithm is described below in Algorithm 3, suppose we want to use k number of samples as our validation set, and we want to stop model training if the validation hasn't improved in q epochs.

The early stopping variant is a good way to stop the model from overfitting to the training data, as it monitors the validation loss and restores the weights of the model

Algorithm 2 Batched Backward propagation

Require: $data$
  $data \leftarrow shuffle(data)$
  $i \leftarrow 1$
  while $i <= p$ **do**
    $j \leftarrow 0$
    while $b \times j < n$ **do**
      $w \leftarrow w - \alpha \sum_{data[jb:\max((j+1)b, n-1)]} \frac{\partial l}{\partial w}$ ▷ for each w in f
      $j \leftarrow j + 1$
    end while
    $i \leftarrow i + 1$
  end while

before the validation loss stops decreasing.

The difficulty of finding the best neural network lies in training the model and tuning its hyperparameters on the dataset. The hyperparameters for a neural network contain the following, but are not limited to: number of neurons in each layer, number of hidden layers, the activation functions for each layer, the learning rate, and the loss function. Hence, we usually use cross-validation before deciding the final model hyperparameters during training.

2.2 Tree Based Models

Tree-based models have been well-developed and studied. The most basic form is a single decision tree. A decision tree is used to partition the feature space into non-overlapping regions[23]. The terminal nodes of the tree would specify a region of the feature space. An internal node of a decision tree is a node in the decision tree that is not a leaf node. An example of a decision tree is shown below in Figure 2.3, where if we have two features A and B, and feature A may have values between [0, 1], feature B may have values between [0, 1], then the decision tree separates the feature space into three regions, Class1, Class2, Class3. Note that the region label does not need to be classes, it could also be a numeric or a vector, which means a decision tree can be applied to both classification or regression problems.

Formally, say the features are vectors $\mathbf{x}$, the feature space is $\mathbb{X}$, the decision tree partitions the feature space into J regions, where J is the number of terminal nodes. The

Algorithm 3 Batched Backward propagation With Early Stopping

Require: $data$
  $data \leftarrow shuffle(data)$
  $validation_data \leftarrow data[0:k]$
  $data \leftarrow data[k:]$
  $r \leftarrow 0$
  $i \leftarrow 1$
  $min_loss \leftarrow \infty$
  while $i <= p$ **do**
    $j \leftarrow 0$
    while $b * j < n$ **do**
      $w \leftarrow w - \alpha \sum_{data[jb:(j+1)b]} \frac{\partial l}{\partial w}$ ▷ for each w in f
      $j \leftarrow j + 1$
    end while
    $cur_loss \leftarrow l(X, f)$
    if $cur_loss \geq min_loss$ **then**
      if $r \geq q$ **then**
        break ▷ restore all w in f to values q epochs ago
      else
        $r \leftarrow r + 1$
      end if
    else
      $min_loss = cur_loss$
      $r \leftarrow 0$
    end if
    $i \leftarrow i + 1$
  end while

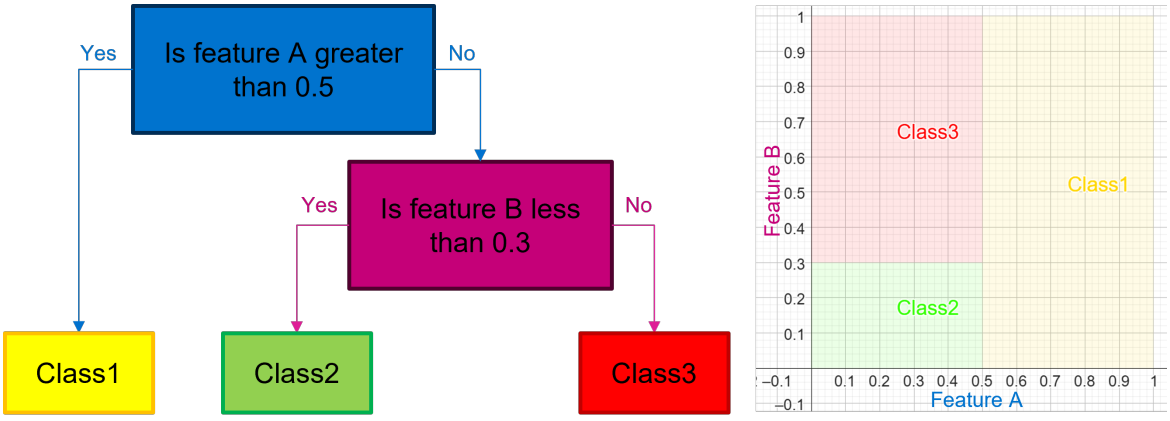


Figure 2.3: A diagram of a decision tree and the regions it splits into

regions R_i are such that $\mathbf{x} \in \bigcup_{j=1}^{J} R_j = \mathbb{X}$. Suppose we have K classes, then for a tree based model $\hat{\boldsymbol{\rho}}(\mathbf{x})$, the predicted probability vector for the input vector $\mathbf{x}$ would be:

$$\hat{\boldsymbol{\rho}}(\mathbf{x}) = (\widehat{\Pr}(y = 1|\mathbf{x}), \widehat{\Pr}(y = 2|\mathbf{x}), ..., \widehat{\Pr}(y = K|\mathbf{x})) \tag{2.3}$$

The predicted probability vector for each region R_j is constant, so we can rewrite $\hat{\boldsymbol{\rho}}(\mathbf{x})$ as:

$$\hat{\boldsymbol{\rho}}(\mathbf{x}) = \sum_{j=1}^{J} \mathbb{I}\{x \in R_j\} \hat{\boldsymbol{\rho}}_{R_j} \tag{2.4}$$

where $\hat{\boldsymbol{\rho}}_{R_j} = (\hat{\rho}_{R_{j,1}}, \hat{\rho}_{R_{j,2}}, ..., \hat{\rho}_{R_{j,K}})$, and $\hat{\rho}_{R_{j,k}}$ is the proportion of training data in region R_j belonging to class k:

$$\hat{\rho}_{R_{j,k}} = \frac{|\{i : \mathbf{x}_i \in R_j, y_i = k\}|}{|\{i : \mathbf{x}_i \in R_j\}|} \tag{2.5}$$

A decision tree can be created via recursive binary splitting[23]. The algorithm is described below in Algorithm 4. Let l be a desirable loss function, and $\mathbb{T}$ be the terminal condition:

The terminal condition for the decision tree building can be one of, but not limited to, the following:

1. No terminal nodes have more observations than a predefined maximum.
2. Tree depth reaches a predefined maximum.

Algorithm 4 Recursive Binary Splitting

$R_-(j,s) \leftarrow \{\mathbf{x}|x_j < s\}$
$R_+(j,s) \leftarrow \{\mathbf{x}|x_j \geq s\}$
$regions = \{\}$
while $\mathbb{T}$ is not True **do**
  if len($regions$) is 0 **then** ▷ First split

$$(j_1, s_1) = \arg\min_{(j,s)} \left[\sum_{i \in R_-(j,s)} l(y_i, \hat{\boldsymbol{\rho}}_{R_-(j,s)}) + \sum_{i \in R_-(j,s)} l(y_i, \hat{\boldsymbol{\rho}}_{R_+(j,s)}) \right]$$

    $regions.add(R_-(j_1, s_1))$
    $regions.add(R_+(j_1, s_1))$
  else
    Select a region such that $R_r(j_1, s_1) \in regions$
    $R_{r-}(j,s) \leftarrow \{\mathbf{x}|x_j < s\}$
    $R_{r+}(j,s) \leftarrow \{\mathbf{x}|x_j \geq s\}$

$$\begin{aligned}(j_2, s_2) = \arg\min_{(r,j,s)} & \sum_{i \in R_{r-}(j,s)} l(y_i, \hat{\boldsymbol{\rho}}_{R_{r-}(j,s)}) \\ & + \sum_{i \in R_{r+}(j,s)} l(y_i, \hat{\boldsymbol{\rho}}_{R_{r+}(j,s)}) \\ & + \sum_{R_q(j_q,s_q) \in regions} \sum_{i \in R_q(j_q,s_q)} l(y_i, \hat{\boldsymbol{\rho}}_{R_q(j_q,s_q)})\end{aligned} \tag{2.6}$$

    $regions.add(R_{r-}(j_2, s_2))$
    $regions.add(R_{r+}(j_2, s_2))$
    $regions.remove(R_r(j_1, s_1))$
  end if
end while
return $regions$

3. When the number of terminal nodes reaches a predefined maximum.

A single decision tree can be built relatively easily, but it often suffer from overfitting to the training data. A single decision tree may be pruned to improve performance, but a decision tree is not a very robust model, and changes in one sample in the training data can cause a large change in the tree created. Hence, a decision tree is considered a weak learner which has low bias but high variance. Instead of using a single decision tree, we often use an ensemble of decision trees in practice. This is also called the ensemble method, where we combine multiple weak learners to preserve low bias and reduce the variance. Two popular ensemble tree-based methods are Gradient Boosted Trees and Random Forest.

2.2.1 Gradient Boosted Trees

Gradient boosting is a greedy approximation technique that can be used to approximate any multivariate function[15]. Gradient Boosted Trees (GBT) are a common application of boosting techniques where the final strong machine learning model is built by adding a decision tree (weak machine learning model) at each training iteration to improve prediction accuracy.

GBT model works by generating new trees to learn the residuals of the previous tree. The general algorithm to create a gradient boosted tree model is as shown in Algorithm 5[23]. Let B be the number of trees we want to build in the gradient boosted tree model, and let λ be a shrinkage parameter, which is similar to the learning rate.

Algorithm 5 Gradient Boosting Trees

$\hat{\boldsymbol{\rho}}(\mathbf{x}) \leftarrow \vec{0}$
$r_i \leftarrow y_i$ ▷ for all i in the training data
$b = 1$
while $b \leq B$ **do**
Fit a tree $\hat{\boldsymbol{\rho}}^{(b)}(\mathbf{x})$ based on predetermined terminal conditions to the training data using the residuals r_i as the target probabilities for the input $\mathbf{x_i}$.
$\hat{\boldsymbol{\rho}}(\mathbf{x}) \leftarrow \hat{\boldsymbol{\rho}}(\mathbf{x}) + \lambda\hat{\boldsymbol{\rho}}^{(b)}(\mathbf{x})$
$r_i \leftarrow r_i - \lambda\hat{\boldsymbol{\rho}}^{(b)}(\mathbf{x}_i)$
end while
return $\hat{\boldsymbol{\rho}}(\mathbf{x}) = \sum_{b=1}^{B} \lambda\hat{\boldsymbol{\rho}}^{(b)}(\mathbf{x})$

2.2.2 Random Forest

Random forest models is an ensemble-based model that has comparable accuracy performance to AdaBoost but is more robust[16]. A random forest is made up of many decision trees where each tree can vote for the result, and the prediction is based on the voted results. It would not overtrain with more trees being added to the ensemble.

The algorithm to generate a random forest from training data is as described in Algorithm 6[23].

Algorithm 6 Random Forest

1. Generate datasets $\tilde{\mathcal{D}}^{(1)}, \tilde{\mathcal{D}}^{(2)}, ..., \tilde{\mathcal{D}}^{(B)}$ from non-parametric bootstrapping
2. Create a decision tree $\tilde{\mathcal{T}}^{(b)}$ for each $\tilde{\mathcal{D}}^{(b)}$ where $b \in \{1, 2, ..., B\}$ with the recursive binary split algorithm.
3. Note that for each split creation in the origin recursive binary split algorithm, instead of choosing the best splits in all p features, we can only choose a split from a randomly selected $\sqrt{p}$ number of features.
4. After all $\tilde{\mathcal{T}}^{(b)}, b \in \{1, 2, ..., B\}$ trees are created, say the class probability predictor associated with $\tilde{\mathcal{T}}^{(b)}$ is $\tilde{\boldsymbol{\rho}}^{(b)}(\mathbf{x})$, then return the final probability predictor:

$$\hat{\boldsymbol{\rho}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \tilde{\boldsymbol{\rho}}^{(b)}(\mathbf{x})$$

Randomly selecting a subset of features at the splitting step allows the trees to be less similar. Thus, the trees are less correlated which allows more effective variance reduction[16]. Hyperparameters available for a random forest model include the terminal condition for the recursive binary splitting algorithm and the number of random features to select at each split. The higher the number of trees, the better and more robust the random forest tree model.

Variable importance in random forests can be determined by permutation-based difference in out-of-bag errors, which can be used to perform feature selection for a random forest model[23]:

1. For each bootstrapped dataset $\tilde{\mathcal{D}}^{(b)}, b \in \{1, 2, ..., B\}$, we first calculate the prediction

error on out-of-bag observations using the trained tree $\tilde{\mathcal{T}}^{(b)}$:

$$\widetilde{\text{Err}^{(b)}} = \sum_{j=1}^{J} \sum_{\tilde{\mathbf{x}}_i \in R_j(\tilde{\mathcal{T}}^{(b)}) \wedge i \notin \tilde{\mathcal{D}}^{(b)}} l(\tilde{y}_i, \hat{\boldsymbol{\rho}}_{R_j(\tilde{\mathcal{T}}^{(b)})}(\tilde{\mathbf{x}}_\mathbf{i}))$$

2. Then, we permute the values for each feature $j = 1, 2, ..., p$ in the bootstrapped dataset and generate datasets with one column of data randomly permuted. We then train and calculate the out-of-bag error on the permuted bootstrap dataset like in step 1, and let $\widetilde{\text{Err}}_{(j)}^{(b)}$ denote this error.

3. Let $\Delta_{(j)}^{(b)} = |\widetilde{\text{Err}^{(b)}} - \widetilde{\text{Err}}_{(j)}^{(b)}|$, Then we can use the following average (scaled) increase in permuted out-of-bag error to measure variable importance:

$$\Delta\widetilde{\text{Err}}_{(j)} = \frac{\frac{1}{B}\sum_{b=1}^{B} \Delta_{(j)}^{(b)}}{\text{sd}(\Delta_{(j)}^{(1)}, \Delta_{(j)}^{(2)}, ..., \Delta_{(j)}^{(B)})}$$

where a larger $\Delta\widetilde{\text{Err}}_{(j)}$ indicates feature j has more predictive power (more important).

Chapter 3

Method

In this chapter, we focus on describing the process of training and testing the machine learning models on this dataset. The key steps are data preprocessing, model building, model training, tuning model architectures, and model interpretation. In data preprocessing, we removed 8% of the samples due to missing data and some columns due to missing data and bias. The removals did not impact the rate at which the three outcomes occur for the patients. We also applied scaling and one-hot encoding techniques when required. For model training, we used a nested cross-validation algorithm to tune the models. We selected the final model architectures for each of Neural Network (NN), Gradient Boosted Trees (GBT), and Random Forest (RF) in the tuning model architecture step. In the end, we tested the model and interpreted the models via SHAP ratio, average change in AUC after removing a variable, and partial dependency plot (PDP).

3.1 Data Preprocessing

The ORIGIN Trial dataset contains 79 columns of explanatory variables and 47 outcome variables, with a total of 12537 sample patients. Out of the 47 outcome variable, we are only interested in 6, they are: mace, survmace, coprim2, survcoprim2, alldth, and survalldth. The surv prefixed column is the survival time of the patient for that outcome, which is used as an independent variable for Cox regression. We use the survival time specific to each outcome when training the model to predict that outcome. We included a detailed table of variable names, their types, how we processed them, and their meanings in the Appendix A.

We then removed 9 variables with a lot of missing data. These variables are updated measures of the patients after the experiment has started. Thus, if the patient had died before the experiments ended, the data is missing. So using these data would make the response variable information indirectly available to the model and thus needs to be removed. There is also a follow-up days variable that accounts for how long the patients were tracked. It has a similar issue where patients were tracked for a shorter period, mostly because they died, so it needs to be hidden from the models.

In addition, we removed two columns without useful information, patient ID and date of randomization. These two columns are for record keeping and not meant for analysis. On the other hand, we created a column as the maximum grip strength. This is due to left grip and right grip strength being in separate columns, and for each patient, either left or right grip strength data is missing, so we merged them into one single column. For the next four columns (amount of physical exercise, previous preeclampsia or eclampsia, number of children, and Gestational diabetes), we did not find significance in the data being missing there, so we filled the missing data with the value 0.

For the country column (a categorical variable with 40 categories), the approach is different depending on the machine learning models used. When training neural networks, it's unsensable to use one-hot encoding because that would be adding too many columns, so we mapped it to each country's estimated cardiovascular death rate[24] given their age groups and sex in the year 2012, which was when the experiment ended. This solves the one-hot encoding issue, and the new representation of countries is now value-based. When training tree-based machine learning models, they are capable of using categorical variables, so no changes are required for the country column. And because of this characteristic, we converted the ethnicity-related one-hot encoding back to categorical values.

A note on the dataset is that we found that the data sampling may be biased, as during feature selection, country, ethnicity and europe are frequently selected, and some factors, like age or smoking, are removed. We believe this could be because the patients sampled are not entirely random at each different location, which causes country, ethnicity, and europe to be correlated with multiple other factors. We decided to remove these three explanatory variables for the random forest and gradient-boosted tree model so they do not rely on them and provide better model interpretability.

After removing all regions and ethnicity related information, we found that the column 'mmse1' is often selected by the tree-based models. However, this column contains the test score for mini-mental state examination [25], which is a measure for cognitive imparement. There is no clear link or pattern between this score with any of the outcomes, so it could be that the data is biased, where the mmse score correlates with the outcomes in this dataset

but may not have strong predictive power in general. Therefore, we removed this column too to remove potential bias.

After preprocessing all the columns, we still had some records with missing data, but the distribution of outcomes for these records are the same as the rest of the dataset, so we decided to remove these records. We are left with 11596 sample patients in the end and only deleted about 7.5% of the data.

3.2 Model Building

We used Tensorflow[26] to build neural networks. The package allows for flexibility and ease when changing model configurations and hyperparameter tuning, with the benefit of preventing overfitting by using early stopping and a validation dataset during training.

For tree-based models, we built them with Yggdrasil Decision Forests[26]. The package has many tree-based models implemented and enables customizations of hyperparameters. It also provides additional model interpretability methods. We enabled feature selection for tree-based models as it will reduce correlation between columns and boost both the prediction results and model interpretation.

3.3 Model Training

All machine learning models are trained by using nested cross-validation techniques. We describe this algorithm in Algorithm 7, where we start with the entire dataset *data* and the number of samples in the dataset n, and let $train_model(train_data)$ to represent the model training procedure that will return a trained model, and let $evaluate_model(model, test_data)$ be an evaluation procedure to evaluate the trained model.

By using nested cross-validation, we can test different hyperparameters for each machine learning model within the inner cross-validation loop and then evaluate their effectiveness again in the outer cross-validation loop.

The neural networks were trialed by experimenting with different architectures and different weightings in the loss functions. The architecture of neural networks seems to have little impact on the model performance, given that there is a moderate number of neurons to learn the data. The loss functions (binary cross-entropy loss) can be weighted by class ratio to tune the recall and precision rate. When we assign a weight for the log loss

Algorithm 7 Nested Cross Validation

Require: $data, n$
  $data \leftarrow shuffle(data)$ ▷ Randomize data order
  $i \leftarrow 0, j \leftarrow 0$
  while $j < 5$ **do**
    $holdout_set \leftarrow data[j * n/5 : (j+1) * n/5]$
    $cv_set \leftarrow data \setminus data[j * n/5 : (j+1) * n/5]$
    while $i < 5$ **do**
      $test_set \leftarrow cv_set[i * n/5 : (i+1) * n/5]$
      $train_set \leftarrow data \setminus data[i * n/5 : (i+1) * n/5]$
      $model \leftarrow train_model(train_set)$
      $evaluate_model(model, test_data)$ ▷ Select the best hyperparameters
      $i \leftarrow i + 1$
    end while
    $model \leftarrow train_model(cv_set)$
    $evaluate_model(model, holdout_set)$
    $j \leftarrow j + 1$
  end while

close to the ratio of the number of positive samples to the number of negative samples, both the number of true positives and the number of false positives increases in the test dataset. The weight would penalize false negatives more, hence the decrease in false negatives. If we increase this weight to exceed the ratio, the number of false positives increases drastically, while decreasing this weight would decrease false positives but also decrease the number of true positives. This observation may be further tested with cross-validation and by using the same model architecture to compare TP, FP, and FN given different weights.

We tested many hyperparameters for gradient-boosted trees, and a lot of them did not result in significant differences. Some fields worth experimenting on are the max depth of trees, shrinkage, and focal loss. Usually for boosted trees, a low max_depth would be preferable, as the resulting model will be made of many weak learners. However, given the complexity of the dataset, a max depth of 4 seems to work best, as the conditions to distinguish positive cases can be difficult. Shrinkage is similar to the learning rate in neural networks, and a lower shrinkage seems to work better for slower but more consistent learning. The focal loss is almost the same as the weighted BCE loss for neural networks. The alpha term is the weight, and it has a gamma term that is the exponent for the difference between the predicted probability and the true label. The gamma term makes minor differences, and the alpha term behaves similarly to the weight of BCE.

For this dataset, random forest can be a good model because of its robustness, which can be used to contrast with gradient boosted trees. It also fits classification problems since the votes can be extrapolated to probabilities for each class. The random forest has many hyperparameters available to tune, but the results show minor differences. The max depth of trees has a minor effect on model performance.

3.4 Tuned Model Architecture

We tried multiple possible hyperparameters in the inner cross-validation loop, but the performance of the models of the same type is similar. The results differ by a small margin in terms of AUC in the inner cross-validation loop, all within the following range for all 3 outcomes: (NN$[0.89 - 0.95]$, GBT$[0.92 - 0.96]$, RF$[0.91 - 0.97]$). We present the final architecture of each model we chose for prediction comparison in this section.

The neural network model contains three hidden layers with 100 neurons in each layer. The activation function used on hidden layers is leaky-Relu with a slope of 0.01 for negative values. The output layer is activated with a sigmoid function to output probabilities. The model uses 20% of the training data for validation during training for early stopping the model if the F1 score does not increase after 50 epochs.

The gradient-boosted tree model is set to have a maximum of 4000 trees, and each tree has a max depth of 4. The model also has a shrinkage of 0.002 and used a backward feature selection method to select the best features before training the final model by computing permutation variable importance. Early stopping is also enabled by checking the loss increase, and the gradient-boosted tree model will stop adding new trees when the validation loss would not improve after adding another 100 trees. 10% of the data is used as validation for early stopping purposes.

The random forest model is set to have a maximum of 1000 trees, each tree has a max depth of 10. The model used a backward feature selection method to select the best features before training the final model by computing out-of-bag variable importance. Since the model output depends on the voting of the trees, we use the vote percentage as probability.

3.5 Model Interpretabilty

We need to understand which independent variable influenced the predicted outcomes of the models the most. For Cox regression, this can usually be done by calculating the hazard

ratio given the proportional hazards assumption, where a higher hazard ratio means the variable influences the outcome more. However, machine learning models do not require this assumption. Thus, machine learning models require other methods of determining the most influential variables.

There are a few ways to quantify variable importance. A common quantifier for machine learning models is SHAP values[27]. The values can be approximated for each input variable of the machine learning model, and the higher the Shap value, the higher the importance of the input variable to the model output. Some quantifiers specific to tree-based models are the sum of the split scores using a specific feature and the number of nodes using a specific feature. Larger scores indicate more important features.

Visualization can also help with model interpretation greatly, by visualizing high-dimensional functions like those represented by machine learning models is difficult[15]. One such good visualization method is a partial dependence plot (PDP), which shows the model's prediction marginalized based on each feature value[15, 28]. We provide the formal definition of partial dependence here. Given a machine learning model $\hat{f}$, say $\mathbf{x}$ is the input vector to the machine learning model, X is the random variable of the input vector, the set of features it uses is S, and K is a subset of features, the partial dependence of $\hat{f}$ on K, denoted as $\hat{f}_K(\mathbf{x}_K)$, is defined as[29]:

$$\hat{f}_K(\mathbf{x}_K) = \mathbb{E}_{\mathbf{x_K}}\left[\hat{f}_K(\mathbf{x}_K, X_{S\setminus K})\right] = \int \hat{f}_K(\mathbf{x}_K, X_{S\setminus K}) d\mathbb{P}(X_{S\setminus K}) \tag{3.1}$$

During model interpretation, the partial dependence, $\hat{f}_K$, can be estimated by calculating averages in the training data:

$$\hat{f}_K(\mathbf{x}_K) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_K(\mathbf{x}_K, \mathbf{x}^i_{S\setminus K}) \tag{3.2}$$

where i is the i-th sample in the training data and n is the size of the training data.

Chapter 4

Result

We gathered the results by following the proposed training and testing procedure in Chapter 3. To summarize, we compared the classification results by calculating the AUC of machine learning models with Cox regression, and found that machine learning models were able to achieve an AUC of above 0.9 while Cox Regression is worse than 0.8. We further investigated the performance of machine learning models, and found that although the accuracies are high for all of them, the recall rate for RF and GBT is above 0.8 for all three outcomes while NN's recall rate is lower at around 0.6. The tree-based models are better at distinguishing patients between their outcomes, and thus we computed the variable importance for the models and examined factors that are most correlated with the outcome.

4.1 Model Prediction Comparisons

4.1 and Table 4.2 below summarizes the results of the models. All models use BCE as a loss function during training.

The results in Table 4.1 shows the AUC (C-statistic) for all models. The Cox regression model results are from Gerstein et al., 2015[12]. We can see that Cox regression model, a logistic regression-based model, has the lowest AUCs on all three outcomes in the test set, around 0.63-0.65, compared to Neural Networks, Random Forest, and Gradient Boosted Trees with AUC above 0.90. When adding additional biomarkers features for Cox Regression, its AUCs improved to close to 0.70 for all three outcomes, but still lower than Neural Networks, Random Forest, and Gradient Boosted Trees without the additional information. Note that the results from Gerstein et al. is from a subdataset of the ORIGIN Trial

Model	MACE (95% CI)	COPRIM2 (95% CI)	ALLDTH (95% CI)
Cox Regression Model (Base Model)	0.65(0.63-0.68)	0.65(0.63-0.67)	0.63(0.61-0.66)
Cox Regression Model (With Bio Markers)	0.69(0.67-0.72)	0.68(0.66-0.70)	0.73(0.70-0.75)
NN	0.92(0.91-0.94)	0.93(0.92-0.95)	0.94(0.93-0.96)
GBT	0.94(0.93-0.95)	0.95(0.95-0.97)	0.95(0.94-0.96)
RF	0.94(0.93-0.95)	0.95(0.94-0.96)	0.95(0.94-0.97)

Table 4.1: Model Prediction AUCs on Test Data for the Three Outcomes

since only 8401 patients participated in the biomarker study, which partially explains the worse AUC score since there is less data during training. However, we can see that the machine learning models have a clear advantage when it comes to distinguishing whether a patient has a particular outcome or not, as the AUC score is higher by a large margin for NN, RF and GBT.

Since the dataset has slight class imbalances in the outcome variables, we also want to analyze the recall (sensitivity) score of each model, which will allow us to better understand whether the model can identify patients who actually have had an outcome. The recall scores and accuracies of each model is included in 4.2.

	MACE		COPRIM2		ALLDTH	
Model	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
NN	0.916	0.637	0.878	0.625	0.916	0.515
GBT	0.942	0.804	0.931	0.874	0.960	0.815
RF	0.941	0.819	0.930	0.878	0.963	0.818

Table 4.2: Model Prediction Accuracy and Recall on Test Data for the Three Outcomes

We can see that Random Forest and Gradient Boosted trees have similar recall and accuracy performance. Their accuraiesy on test data are all above 93%, while the recall scores are a bit lower, but still above 0.8 for all outcomes. The similarity is expected as RF and GBT are both tree-based machine learning models, and both are robust models that are less susceptible to overfitting. The same cannot be said to neural networks, where their accuracies are high as well for all 3 outcomes, all around 90% range, but the recall score is significantly worse than the tree-based models, with only 0.5-0.65 for all three outcomes, meaning that it does not identify the ones who had one of the outcomes well. We believe this is due to the NN models usually have high variance, and thus they suffer

from overfitting and are harder to tune in this case. This was evident during the training of the neural networks, where the validation loss stopped decreasing within a few epochs when trialed with multiple hyperparameters. Thus, we were unable to imporve the performance of nerual networks further.

4.2 Variable Importance

Since the Random Forest and Gradient Boosted Trees have better AUCs, accuracies, and recalls, we believe these models can be better alternatives than Cox regression when analyzing the dataset and identifying factors that affect outcomes the most. With the low recall rate for neural networks, the variable importance highlighted by it may not be a good indicator of whether the factor has a greater impact on the response variable. So we will focus on explaining variable importance from the tree-based models.

4.2.1 Quantifiable Analysis

Note that we exclude the survival time for ranking variable importance analysis since the survival time is the most important variable for all models, and it's used in Cox regression as the dependent variable for hazard functions. We also use the entire dataset to calculate variable importance quantifiers. Table 4.3 ranks variables by their SHAP value[27]. As we can see, the top two indicators for MACE and COPRIM2 is cancer and priorcv. While for ALLDTH, the top three indicators are cancer, age, and kidney functions measures (screening visit creatine value (SVCREV), estimated glomerular filtration rate (eGFR) and kidney disease index (kdi)). Some of the frequently appearing top factors for all three outcomes are age, kidney functions measures (SVCREV, eGFR and kdi), prior CV related events (priorcv, prior myocardial infarction (priormi), prior atrial fibrillation (prioraf)), male, blood pressure (arm diastolic blood pressure (dbparm), arm systolic blood pressure (sbparm) and ankle-brachial blood pressure index (abi)), blood sugar measures (average blood glucose (sugar) levels for the last two to three months (hba1c) and screening visit fasting glucose values (svfgv)), body physical condition related measures (body-mass index (bmi), Waist circumference (rwaistc), Hip circumference (rhipc), maximum grip strength (max_grip) and smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke)). The blood lipid related measures (screening visit high-density lipoprotein cholesterol value (svhdlv), screening visit total cholesterol value (svchov), screening visit ligh-density lipoprotein cholesterol value (svldlv), and screening visit triglyceride value (SVTRIV)) appear mostly for the outcome MACE and COPRIM2, but not for ALLDTH.

	MACE		COPRIM2		ALLDTH	
rank	Random Forest	GB Trees	Random Forest	GB Trees	Random Forest	GB Trees
1	cancer	cancer	priorcv	priorcv	cancer	cancer
2	priorcv	SVCREV	cancer	cancer	age	age
3	age	priorcv	male	SVCREV	egfr	SVCREV
4	sbparm	rwaistc	age	rsmoke	abi	priorcv
5	dbparm	hba1c	hba1c	hba1c	durDM	statin
6	svhdlv	svchov	sbparm	rhipc	hba1c	max_grip
7	priormi	diu	svldlv	durDM	diu	priordm
8	male	svfgv	rsmoke	gdiab	priorcv	rsmoke
9	diu	priormi	bmi	statin	nchild	ace_arb
10	rwaistc	max_grip	statin	preecl	rsmoke	hyperten

Table 4.3: Variable Importance Ranked by SHAP Value

	MACE		COPRIM2		ALLDTH	
rank	Random Forest	GB Trees	Random Forest	GB Trees	Random Forest	GB Trees
1	age	cancer	cancer	cancer	cancer	cancer
2	cancer	SVCREV	priorcv	priorcv	age	age
3	sbparm	priorcv	age	SVCREV	egfr	SVCREV
4	priorcv	svchov	hba1c	rsmoke	abi	priorcv
5	rwaistc	rwaistc	bmi	hba1c	hba1c	statin
6	dbparm	diu	sbparm	rhipc	durDM	max_grip
7	SVTRIV	hba1c	svldlv	durDM	priorcv	priordm
8	svhdlv	svfgv	male	gdiab	nchild	ace_arb
9	priormi	kdi	durDM	preecl	diu	anticog
10	male	priormi	rsmoke	statin	rsmoke	rsmoke

Table 4.4: Variable Importance Ranked by Mean Decrease in AUC

We also compare the mean decrease in AUC of models in Table 4.4, which shows how much the model depends on this variable to distinguish between patients with and without outcomes.

The rankings of the variables via mean decrease in AUC are very similar to the rankings via SHAP values. Although the order of the variables changes for some, we can still see cancer and priorcv are highly ranked as the most important predictors for MACE and

COPRIM2, and cancer, age and kidney function measures (eGFR, SVCREV, kdi) are highly ranked as the most important predictors for ALLDTH. hba1c is ranked relatively high consistently for COPRIM2 and ALLDTH as well.

In the paper "Identifying Novel Biomarkers for Cardiovascular Events or Death in People With Dysglycemia" by Gerstein et al. in 2015[12], they believed the most important clinical factors for MACE, COPRIM2, and ALLDTH were the following:

1. priorcv
2. Albuminuria
3. Whether the patient is men
4. Whether the age is greater than 55 for men and 65 for women
5. LDL/HDL ratio
6. Current smoking
7. Prior Diabetes
8. Hypertention

The results from machine learning models generally agree with these, as priorcv occurs frequently, and SVCREV and eGFR matches with a different type of kidney function measures. Albuminuria. Gender, age, smoking, blood lipids, blood pressure measures are shown as highly important for machine learning models too. Instead of prior diabetes, machine learning models chose to use more specific values such as svfgv and hba1c to differentiate the patients further.

4.2.2 Visual Analysis

We then examine the partial dependence plots of the tree-based machine learning models to gain more insight into these variables. The partial dependence plots are unrealistic when variables have strongly correlated features. However, it can be used to estimate how the average prediction in your dataset changes when a single feature is changed, assuming that none of the features are correlated[29].

Comparing the PDPs of Random Forests versus Gradient Boosted Trees when predicting ALLDTH, we see that the RF models have much smoother PDPs and thus the results may be more valueable, since we might not expect a significant change in a patient's outcome when the change in one factor is small from a medical perspective. The smoother PDPs can be linked back to the algorithm for RF, where only a smaller subset of features are considered for each split, meaning that the more useful features are not always used in the first few splits, and thus allowing the PDP for it to be more detailed as the more useful features are used in deeper leaves too. Deeper leaves have a smaller region associated with it, thus the change in partial dependence is less for Random Forest when the value of a feature changes.

Since the PDPs illustrate the average prediction value based on each explanatory variable, we can interpret the six PDPs of Random Forest Model for ALLDTH as follows:

(a) When survival time increases, the predicted probability of ALLDTH occurring for patients in this dataset would decrease.

(b) The average estimated predicted probability of ALLDTH occurring for patients in this dataset with cancer increases by more than 0.1.

(c) The average estimated predicted probability of ALLDTH occurring for patients in this dataset who are older than 75 is at least 0.02 higher than patients who are younger than 65.

(d) The average estimated predicted probability of ALLDTH occurring for patients in this dataset who has eGFR lower than 50 is at least 0.03 higher than patients who have eGFR about 90-100.

(e) The average estimated predicted probability of ALLDTH occurring for patients in this dataset increases as hba1c increases when hba1c is greater than 5.2.

(f) The average predicted probability of ALLDTH occurring for patients in this dataset who has abi higher than 1.2 is lower than patients who have abi lower than 0.8.

A detailed analysis and interpretation of these graphs can be a lot more difficult than the hazard ratio, especially for continuous variables, as one unit change of explanatory variable does not lead to a constant change in probability or risk of an outcome occurring for the patient. However, these graphs may represent the real probability change given a unit change in the response variable really well.

For example, eGFR is a measure of how well kidneys are functioning. A decreased eGFR under 70 can lead to a 68% higher risk of death from any cause[30]. Our graph indeed shows an increase in partial dependence of the ALLDTH outcome on eGFR starting to increase at below 90, and more drastically as the eGFR decreases further to below 70.

From an age perspective, the partial dependence increases more drastically when age is over 70, which suggests that the patients with age over 70 are more likely to be at risk for the ALLDTH outcome, which aligns with common expectations.

We also see a similar trend for hba1c, where the partial dependence for ALLDTH outcome is lowest around 5%-5.5%. The partial dependence elevates at around 5.7%, and further increases around 6.5%, and increases drastically when exceeding 6.5%. This aligns with the diabetes diagnosis, where the normal range of hba1c is below 5.7%, prediabetes range of hba1c is 5.7%-6.4%, and diabetes range of hba1c is higher than 6.4%[31]. The graph also indicates if the hba1c value is lower than 5, the ALLDTH outcome is more likely, and this confirms with the results from other studies which show that hypoglycaemia is linked to mortality[32].

We can also examine the RF model PDPs for the outcome MACE and COPRIM2. Interestingly, we see that for cancer, although it influences the two outcomes, the outcome average partial dependence is lower when with cancer. This is due to the dataset being mixed, and people with cancer are likely to have experienced the ALLDTH outcome without experiencing MACE and COPRIM2.

The blood lipids measure are related to the MACE outcome. We see that for triglyceride, the average partial dependence of MACE is the lowest around 0.8 mmol/L, and increases when the triglyceride measure elevates to 1.7 mmol/L, which is the upper end of the normal triglyceride measure range[33]. Then the risk continute to slowly elevates as the triglyceride value increases further. The average partial dependence of MACE increases for lower triglyceride measures too, but this may be due to the unrealistic data from setting a triglyceride value lower for patients with higher risks. We see a smoother trend for svhdlv, where the risk of MACE is higher when the value is below 1.5 mmol/L, while getting to the lowest risk at around 2.0 mmol/L.

For COPRIM2, we see that hba1c and svldlv both have the lowest estimated partial dependence at around normal ranges below 6% and 2 mmol/L [33] respectively, and the partial dependence increases as the hba1c and svldlv increases. Similar trend can be observed for sbparm and bmi, where the value around normal ranges have the lowest partial dependence, and the partial dependence increases when the sbparm and bmi are outside of the normal ranges.

We have included all PDPs from all tree-based models in Appendix B and C for future reference and reviews. These PDPs may not accurately capture the associated risks, but they can be good estimates of changes in predicted probability with respect to the explanatory variables. We may be able to find the underlying reasons of why a change in explanatory variable causes the response variable to change in the trajectory shown in the PDPs or create models that capture this relationship.

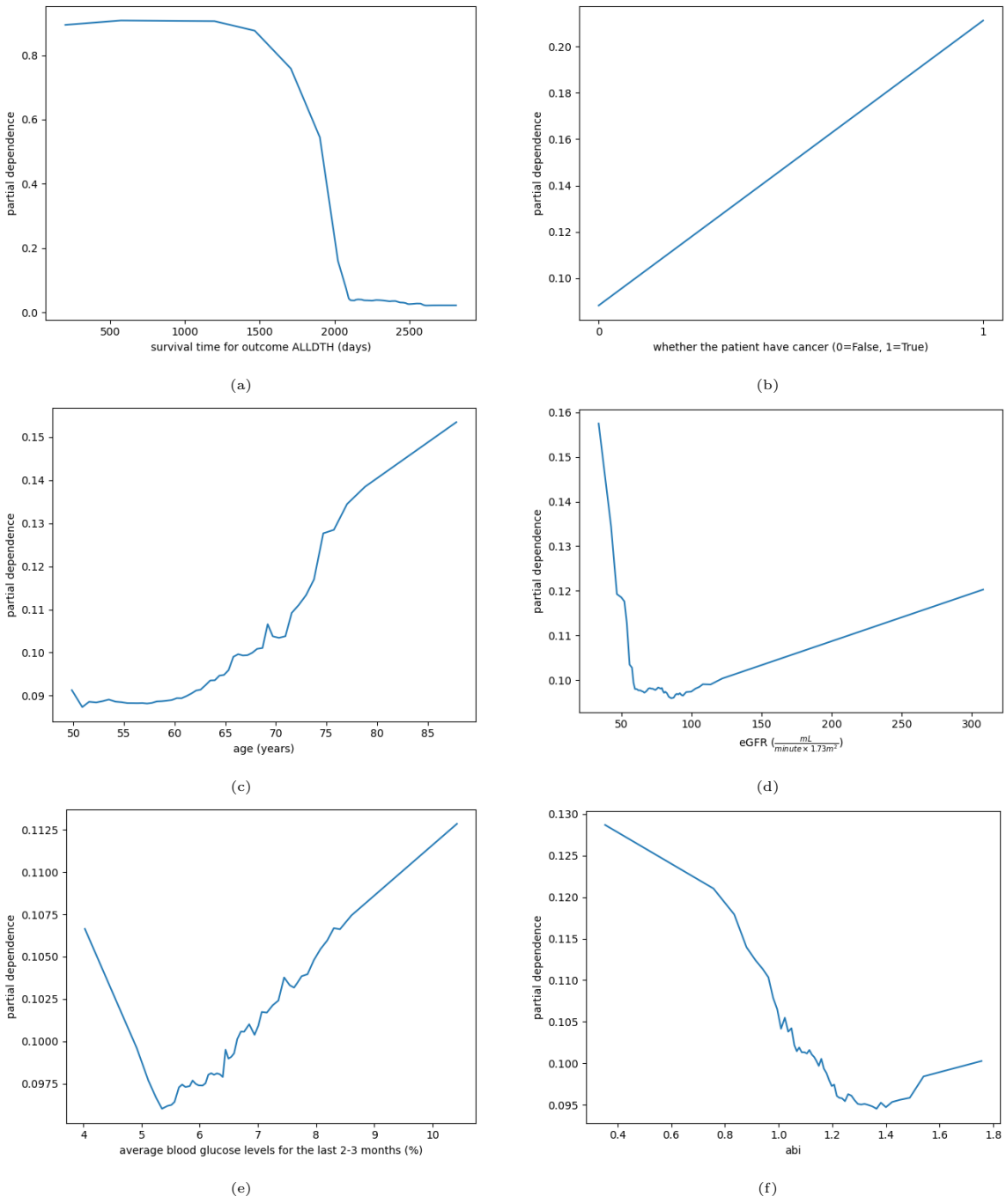


Figure 4.1: Partial Dependence Plots of Random Forest Model for ALLDTH. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explanatory variable measure on x-axis: (a) survival time, (b) whether they have cancer, (c) age, (d) estimated glomerular filtration rate (eGFR), (e) average blood glucose (sugar) levels for the last two to three months (hba1c), (f) ankle-brachial blood pressure index (abi)

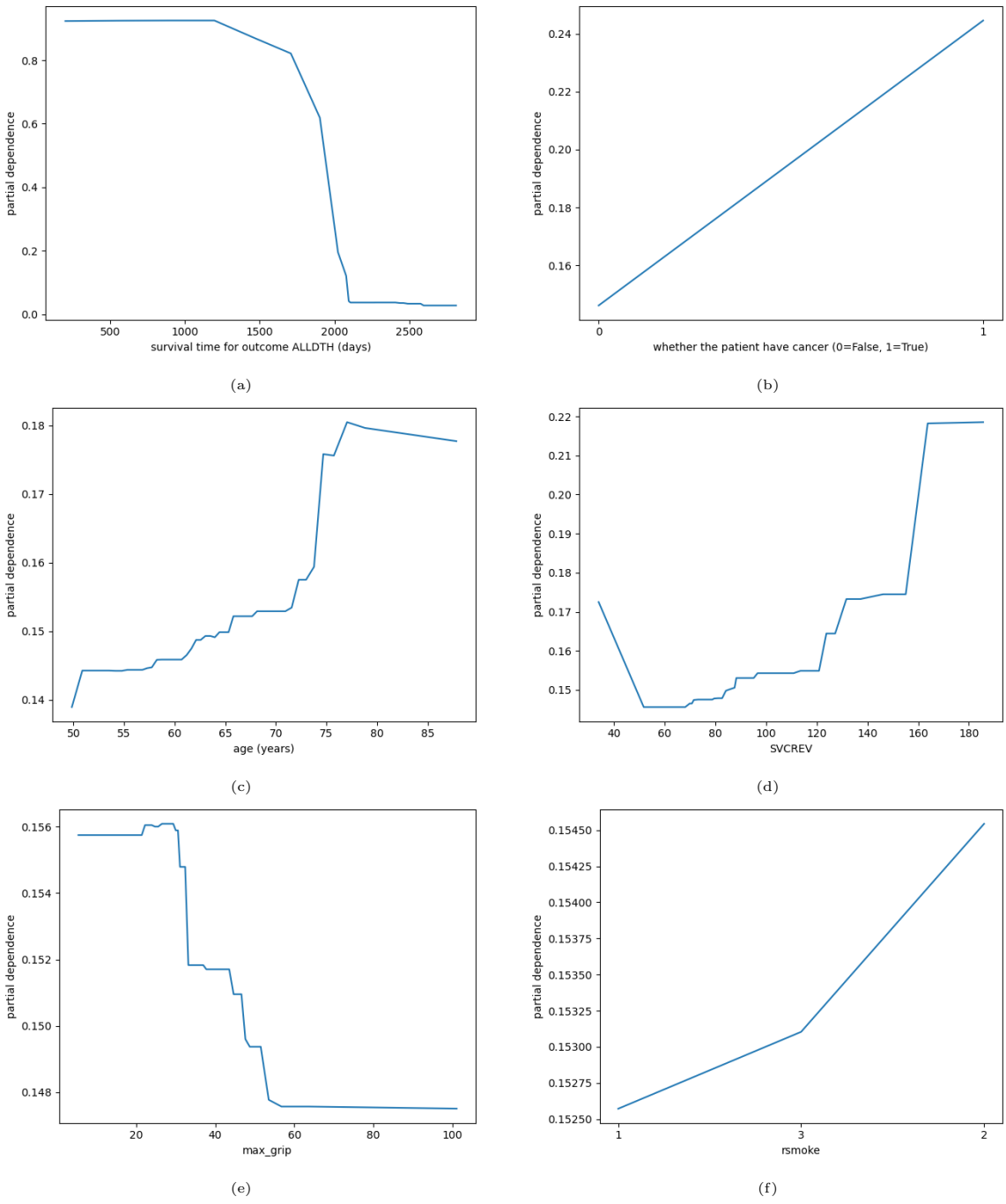


Figure 4.2: Partial Dependence Plots of Gradient Boosted Trees Model for ALLDTH. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explanatory variable measure on x-axis: (a) survival time, (b) whether they have cancer, (c) age, (d) screening visit creatine value (SVCREV), (e) maximum grip strength (max_grip), (f) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke)

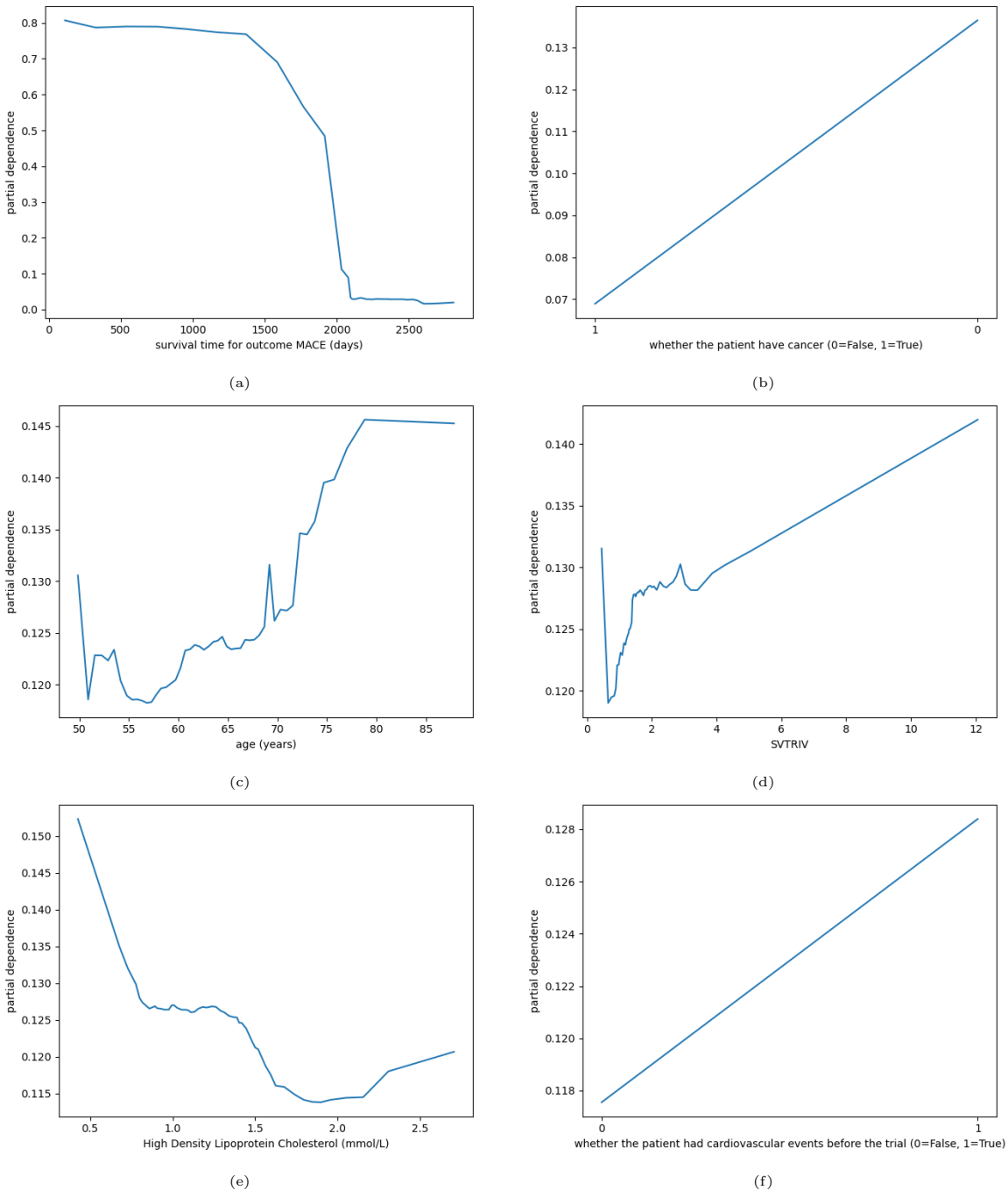


Figure 4.3: Partial Dependence Plots of Random Forest Model for MACE. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explanatory variable measure on x-axis: (a) survival time, (b) whether they have cancer, (c) age, (d) screening visit triglyceride value (SVTRIV), (e) screening visit high-density lipoprotein cholesterol value (svhdlv), (f) prior cardiovascular events (priorcv)

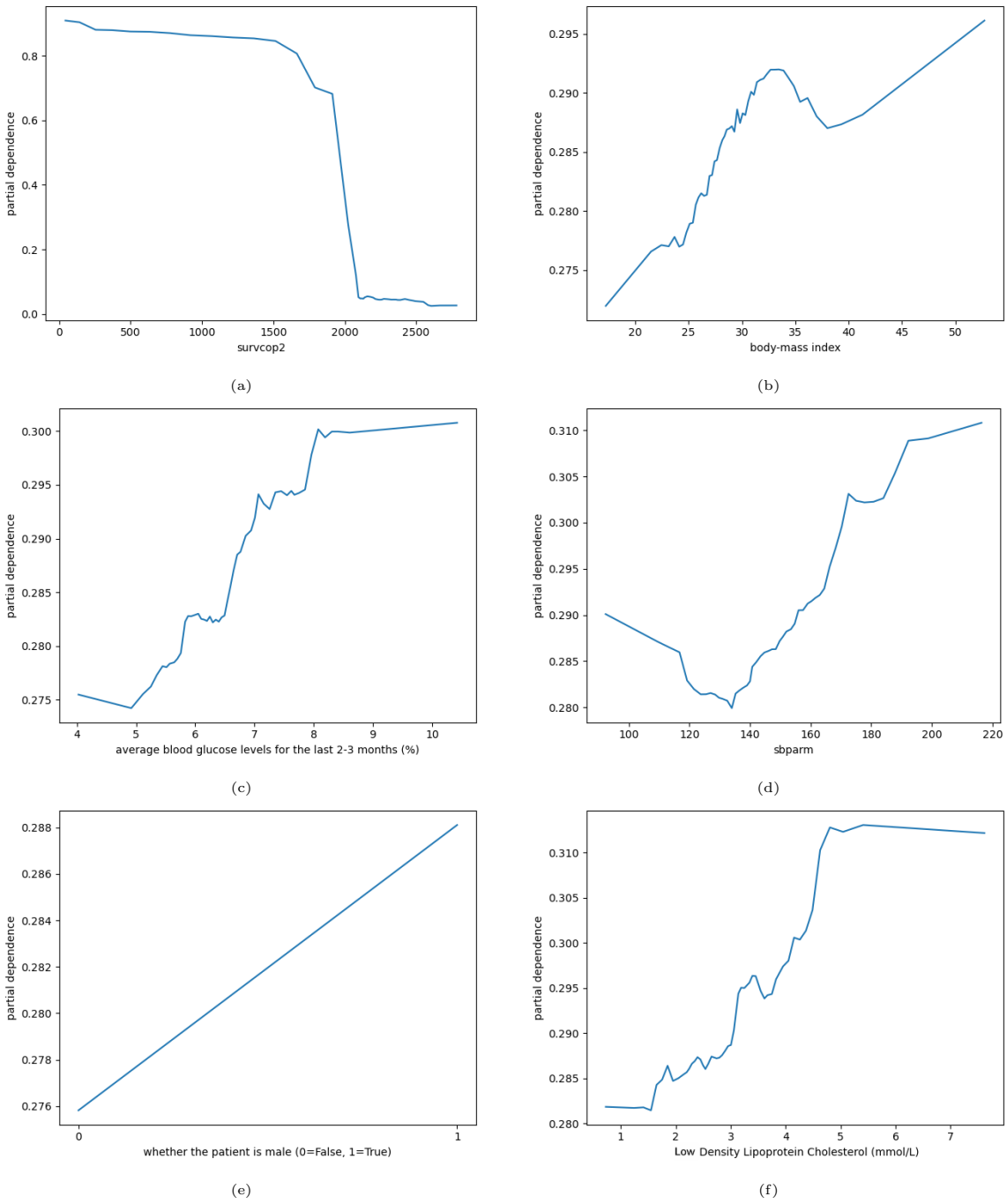


Figure 4.4: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explanatory variable measure on x-axis: (a) survival time, (b) body-mass index (bmi), (c) arm systolic blood pressure (sbparm), (d) arm systolic blood pressure (sbparm), (e) whether the patient is male, (f) screening visit ligh-density lipoprotein cholesterol value (svldlv)

Chapter 5

Conclusion

In this thesis, we employed the ORIGIN Trial dataset which contained data of 12537 individuals with with Cardiovascular (CV) risk factors plus impaired fasting glucose, impaired glucose tolerance, or type 2 diabetes. We focused on the three outcomes for these patients, which are major adverse cardiovascular events (MACE), expanded composite outcome (COPRIM2), and all-cause death (ALLDTH). We presented three types of machine learning models, Neural Network (NN), Random Forest (RF), and Gradient Boosted Trees (GBT), that can be trained on tabular data and predict outcomes. Each of these model required a slight different input format, which includes adjustments for categorical data and whether the data needs to be scaled. During data preprocessing, we removed columns and patients records that had a lot of NA values but do not change the outcome distribution.

When training the models, we applied nested cross-validation training algorithm to tune the hyperparameters and selected the best model hyperparameters to use on hold-out test sets. We compared the performance of the model on hold-out test sets with Cox regression models used in the previous paper. We found that the best performing models are random forest and gradient boosted trees. These machine learning models can outperform Cox regression model by a large margin in terms of AUC (C-statistic) on test data, where they had a score of 0.94-0.95 in all three outcomes compared to 0.63-0.65 for Cox regression model. We also examined model accuracy and recall for all three outcomes on test data since the dataset had outcome imbalances. We found that despite all models had excellent prediction accuracies of above 0.9, neural network models had suboptimal recall scores of lower than 0.65, while random forests and gradient boosted trees received recall scores of above 0.8.

After testing the deep learning models, we analyzed the variable importance in each

model. We focused on Random Forest and Gradient Boosted Trees since they are the best-performing models in terms of AUC and recall score; thus, they have more predictive power, and their variable importance would be more relevant. We used two different ways to analyze variable importance. Firstly, we quantified variable importance with SHAP values and mean decrease in AUCs, and noticed that for MACE and COPRIM2, the top two factors are priorcv and cancer, while for ALLDTH, the top three factors are cancer, age, and kidney functions related measures. We also found that while most risk factors such as age, blood sugar level, smoker categories, and body physical conditions are correlated with all three outcomes, the blood lipids measures are more correlated to the outcome MACE and COPRIM2. Secondly, we used PDPs to visualize the effect of each variable on prediction probability of the outcomes. We noticed that the graphs for continuous variables are harder to interpret due to non-linearity, but the non-linearity can match with the underlying connection between the variable and the outcome. We also identified that cancer is more correlated with ALLDTH, and negatively correlated with MACE and COPRIM2 since the patients who experienced ALLDTH may not have experienced MACE and COPRIM2. For continuous variables, the normal value range for each variable is often associated with lower partial dependence to the given outcome. We believe the PDPs can be very useful in identifying important factors to each outcomes and their relationships.

To summarize, machine learning models can be used to predict and identify variable importance for outcomes of patients in tabular medical data. They can achieve better prediction results than Cox regression. However, quantifying variable importance in machine learning models can be more difficult. The downside is that the average predicted probability of outcomes are non-linear with respect to each continuous variable, but the upside is that the non-linearity may estimate the actual effect of each factor and allow a closer examination of the reasons behind the relationship. Since the PDPs may not be smooth and thus theoretically difficult to justify, a potential future research direction may be modeling the relation between the underlying variables and the outcome probability mathematically with empirical research gathered constants given the estimated PDPs.

References

[1] M. Panahiazar, V. Taslimitehrani, N. Pereira, and J. Pathak, "Using ehrs and machine learning for heart failure survival analysis," *Studies in health technology and informatics*, vol. 216, p. 40–44, 2015.

[2] A. Wong, A. T. Young, A. S. Liang, R. Gonzales, V. C. Douglas, and D. Hadley, "Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment," *JAMA Network Open*, vol. 1, pp. e181018–e181018, 08 2018.

[3] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLOS ONE*, vol. 12, pp. 1–14, 04 2017.

[4] P. Townsend, P. Phillimore, and A. Beattie, *Health and Deprivation Inequality and the North (1st ed.)*. Routledge, 1988.

[5] S. Kley, "How material deprivation impacted economic stress across european countries during the great recession. a lesson on social comparisons," *Acta Sociologica*, vol. 65, no. 1, pp. 66–85, 2022.

[6] A. Steptoe and M. Kivimäki, "Stress and cardiovascular disease," *Nat Rev Cardiol*, vol. 9, pp. 360–370, 2012.

[7] The ORIGIN Trial Investigators, "Rationale, design, and baseline characteristics for a large international trial of cardiovascular disease prevention in people with dysglycemia: The origin trial (outcome reduction with an initial glargine intervention)," *American Heart Journal*, vol. 155, no. 1, pp. 26.e1–26.e13, 2008.

[8] T. O. T. Investigators, "n–3 fatty acids and cardiovascular outcomes in patients with dysglycemia," *New England Journal of Medicine*, vol. 367, no. 4, pp. 309–318, 2012.

[9] T. O. T. Investigators, "Basal insulin and cardiovascular and other outcomes in dysglycemia," *New England Journal of Medicine*, vol. 367, no. 4, pp. 319–328, 2012.

[10] H. C. Gerstein, H. Jung, L. Rydén, R. Diaz, R. E. Gilbert, and S. Yusuf, "Effect of basal insulin glargine on first and recurrent episodes of heart failure hospitalization," *Circulation*, vol. 137, no. 1, pp. 88–90, 2018.

[11] W. Doehner, H. C. Gerstein, J. Ried, H. Jung, C. Asbrand, S. Hess, and S. D. Anker, "Obesity and weight loss are inversely related to mortality and cardiovascular outcome in prediabetes and type 2 diabetes: data from the origin trial," *European Heart Journal*, vol. 41, pp. 2668–2677, 05 2020.

[12] H. C. Gerstein, G. Paré, M. J. McQueen, H. Haenel, S. F. Lee, J. Pogue, A. P. Maggioni, S. Yusuf, and S. Hess, "Identifying novel biomarkers for cardiovascular events or death in people with dysglycemia," *Circulation*, vol. 132, no. 24, pp. 2297–2304, 2015.

[13] V. E. Ismailov, "A three layer neural network can represent any multivariate function," *Journal of Mathematical Analysis and Applications*, vol. 523, no. 1, p. 127096, 2023.

[14] N. R. H., "Kolmogorov's mapping neural network existence theorem," *Proceedings of International Conference on Neural Networks, 1987*, vol. 3, pp. 11–13, 1987.

[15] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, p. 5–32, Oct. 2001.

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, *Deep Learning*, pp. 403–460. New York, NY: Springer US, 2021.

[18] T. Szandala, "Review and comparison of commonly used activation functions for deep neural networks," *CoRR*, vol. abs/2010.09458, 2020.

[19] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: theoretical analysis and applications," in *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, JMLR.org, 2023.

[20] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 10 1986.

[21] S. McCandlish, J. Kaplan, D. Amodei, and O. D. Team, "An empirical model of large-batch training," 2018.

[22] Y. Bai, E. Yang, B. Han, Y. Yang, J. Li, Y. Mao, G. Niu, and T. Liu, "Understanding and improving early stopping for learning with noisy labels," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, (Red Hook, NY, USA), Curran Associates Inc., 2021.

[23] G. James, D. Witten, T. Hastie, and R. Tibshirani, *Tree-Based Methods*, pp. 327–365. New York, NY: Springer US, 2021.

[24] G. B. of Disease Collaborative Network, L. National Heart, B. Institute, N. I. of Health (NIH), and A. C. of Cardiology, "Global burden of disease (gbd) cardiovascular burden estimates 1990-2022," 2023.

[25] R. M. Crum, J. C. Anthony, S. S. Bassett, and M. F. Folstein, "Population-based norms for the mini-mental state examination by age and educational level," *JAMA*, vol. 269, pp. 2386–2391, 05 1993.

[26] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[28] M. Guillame-Bert, S. Bruch, R. Stotz, and J. Pfeifer, "Yggdrasil decision forests: A fast and extensible decision forests library," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pp. 4068–4077, 2023.

[29] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 3 ed., 2025.

[30] A. S. Go, G. M. Chertow, D. Fan, C. E. McCulloch, and C. yuan Hsu, "Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization," *New England Journal of Medicine*, vol. 351, no. 13, pp. 1296–1305, 2004.

[31] S. I. Sherwani, H. A. Khan, A. Ekhzaimy, A. Masood, and M. K. Sakharkar, "Significance of hba1c test in diagnosis and prognosis of diabetic patients," *Biomarker Insights*, vol. 11, p. BMI.S38440, 2016. PMID: 27398023.

[32] P. E. Cryer, "Severe hypoglycemia predicts mortality in diabetes," *Diabetes Care*, vol. 35, pp. 1814–1816, 08 2012.

[33] "List of normal lab values," Dec 2020.

APPENDICES

Appendix A

Variable Definition, Type, and Preprocessing

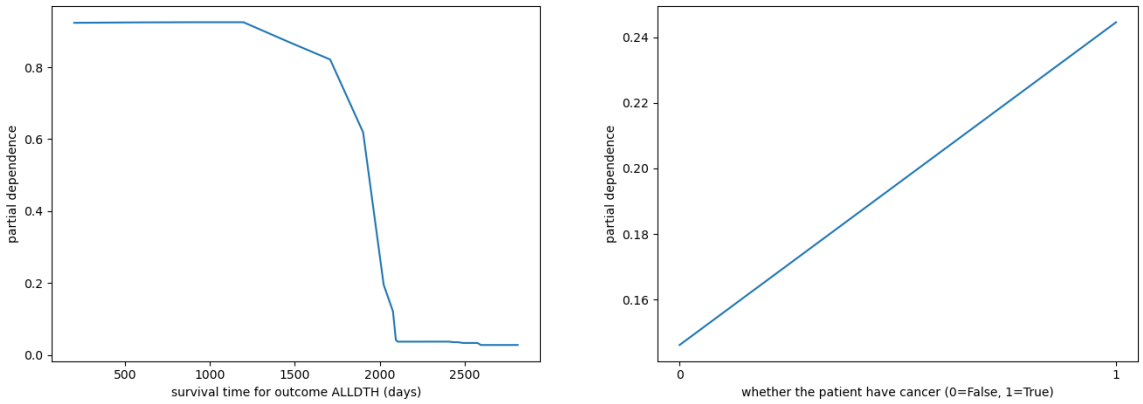
Variable	Meaning	Type	How we processed it
abi	ankle-brachial blood pressure index	continuous	kept
abi0		continuous	removed due to lots of missing data
abi09	ankle-brachial index during follow-up (certainly during the last visit according to the eCRF)	binary	kept
abi2		continuous	removed due to lots of missing data
abi6		continuous	removed due to lots of missing data
ace_arb	whether th patient uses renin angiotensin system drugs (ACE-Inhibitors or Angiotensin 2 blocker)	binary	kept
acr	Urine albumin to creatinine ratio	continuous	kept
age	age	continuous	kept
alcoh	whether the patient drinks alcohol	binary	kept
ald_ant	whether the patient uses aldosterone antagonist	binary	kept
alldth	whether the patient had all-cause death outcome	binary	kept
anticog	whether the patient uses anticoagulant	categorical	kept
asa_anti	Whether the patient uses ASA or antiplatelet	binary	kept
beta	whether th patient uses beta-blockers	binary	kept
bmi	Body-Mass Index	continuous	kept
calc_malbcreat	Calculated Albumin: Creatinine ration	continuous	kept
cancer	whether the patient have cancer	categorical	kept
catar	cataracts	categorical	kept
coprim2	whether the patient had expanded composite outcome	binary	kept
country	The country id of where the patient was located	categorical	removed, irrelevant
dbparm	Arm diastolic blood pressure	continuous	kept
dcct_a1c	Diabetes Control and Complications Trial A1C	continuous	kept
diu	whether the patient uses diuretic	binary	kept
durDM	Diabetes duration	continuous	kept
egfr	estimated glomerular filtration rate	continuous	kept
eth_black	Whether the patient is of Black ethnicity	binary	kept
eth_latin	Whether the patient is of Latin ethnicity	binary	kept
eth_oasian	whether the patient is other Asian	binary	kept
eth_other	Whether the patient is of other ethnicity	binary	kept
eth_sasian	whether the patient is south Asian	binary	Only kept for neural network
eth_white	Whether the patient is of White ethnicity	binary	kept
ethnic	the ethnicity of the patient	categorical	kept, but removed later due to bias for random forest and gradient boosted trees
EUFHIPC	End of Usual Follow-up Hip Circumference	continuous	removed due to lots of missing data
EUFWAISTC	End of Usual Follow-up Waist Circumference	continuous	removed due to lots of missing data
europe	Whether the patient is of European ethnicity	binary	kept, but removed later due to bias for random forest and gradient boosted trees
fishoil	whether the patient used fish oil	binary	kept
folup1	Duration of follow-up in days	continuous	removed, early ending follow-up leaks death information to models
gdiab	Gestational diabetes	categorical	kept, filled na with 0
glargranddt	Dates of randomization	categorical	removed, irrelevant
hba1c	average blood glucose (sugar) levels for the last two to three months	continuous	kept
hyperten	whether the patient has hypertension	binary	kept
kdi	kidney disease index	continuous	kept
laser	laser therapy for eye disease	categorical	kept

lgrip	left hand grip strength	continuous	kept, but merged with rgrip
M24HIPC	hip circumference after two years of follow-up	continuous	removed due to lots of missing data
M24WAISTC	waist circumference after two years of follow-up	continuous	removed due to lots of missing data
mace	whether the patient had major adverse cardiovascular events outcome	binary	kept
male	whether the patient is male	binary	kept
metf	whether the patient uses metf	binary	kept
mmse1	Mini-Mental State Examination (MMSE) score (baseline)	continuous	kept
nchild	Parity (number of children)	categorical	kept, filled na with 0
newdm	whether the patient has new onset diabetes	categorical	kept
othccb	Whether the patient uses other BP Drugs (excluding Beta Blocker & RAS drugs)	binary	kept
pid	Patient ID	discrete	removed, irrelevant
preecl	Previous preeclampsia or eclampsia	categorical	kept, filled na with 0
prioraf	whether the patient had atrial fibrillation before the trial	binary	kept
priorcv	whether the patient had cardiovascular events before the trial	binary	kept
priordm	prior diagnosis of diabetes	binary	kept
priormi	whether the patient had myocardial infarction before the trial	binary	kept
priorstrk	whether the patient had stroke before the trial	binary	kept
rgrip	right hand grip strength	continuous	kept, but merged with lgrip
rhipc	Hip circumference	continuous	kept
RMGLARG	ORIGIN allocations (1=glargine;2=standard care)	categorical	kept
rsmoke	1: no smoking; 2:current smoking; 3: former	categorical	kept
rwaistc	Waist circumference	continuous	kept
sbparm	Arm systolic blood pressure	continuous	kept
smoke	whether the patient is a current smoker	binary	kept
stang	stable angina	binary	kept
statin	whether the patient uses statin	binary	kept
strdctrynm	The country name of where the patient was located	categorical	kept, but removed later due to bias for random forest and gradient boosted trees
sulfon	whether the patient uses sulfon	binary	kept
survalldth	survival time for outcome all-cause death	continuous	kept
survcoprim2	survival time for expanded composite outcome	continuous	kept
survmace	survival time for major adverse cardiovascular events	continuous	kept
svchov	total cholesterol at the screening visit	continuous	kept
SVCREV	Creatinine	continuous	kept
svfgv	fasting glucose value	continuous	kept
svhdlv	High Density Lipoprotein Cholesterol	continuous	kept
svldlv	Low Density Lipoprotein Cholesterol	continuous	kept
SVTRIV	Triglyceride	continuous	kept
totalmets	Physical activity	continuous	kept, filled na with 0
unstang	unstable angina	binary	kept
WHR	waist circumference to hip circumference ratio	continuous	kept
WHR2y	waist to hip ratio after two years of follow-up	continuous	removed due to lots of missing data
WHREUF	End of Usual Follow-up Waist to Hip Ratio	continuous	removed due to lots of missing data

Appendix B

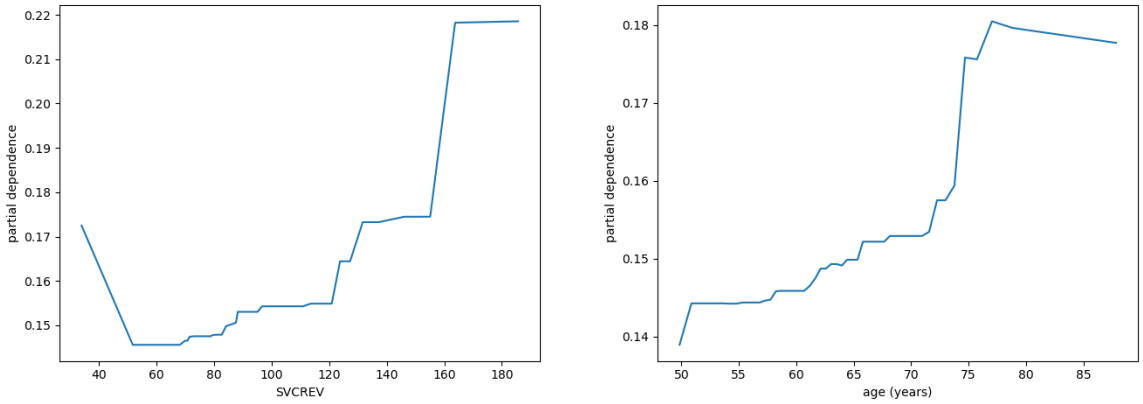
GradientBoostedTrees

B.1 ALLDTH



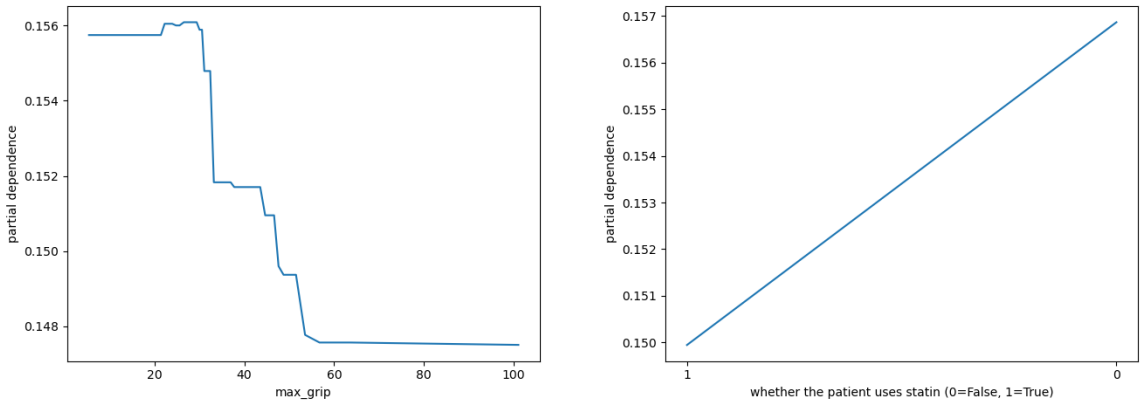
(a) (b)

Figure B.1: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer



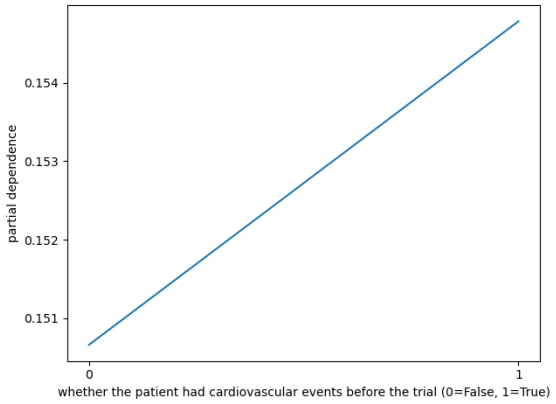
(a) (b)

Figure B.2: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit creatine value (SVCREV), (b) age

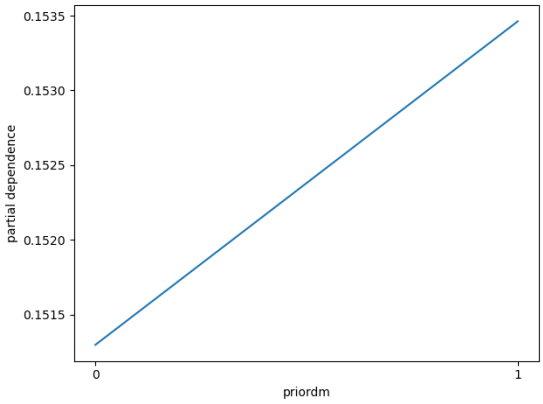


(a) (b)

Figure B.3: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) maximum grip strength (max_grip), (b) whether the patient uses statin

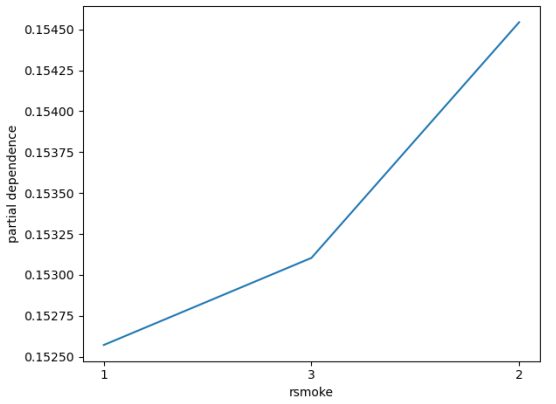


(a)

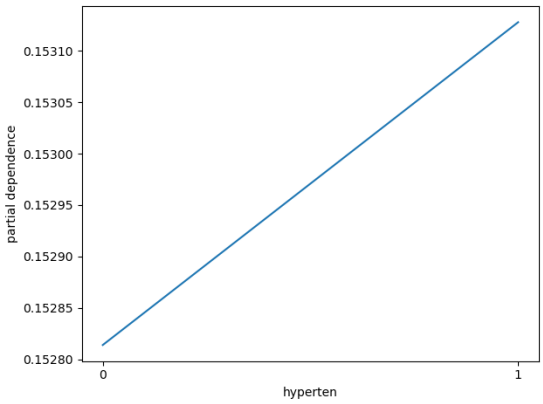


(b)

Figure B.4: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior cardiovascular events (priorcv), (b) prior diagnosis of diabetes

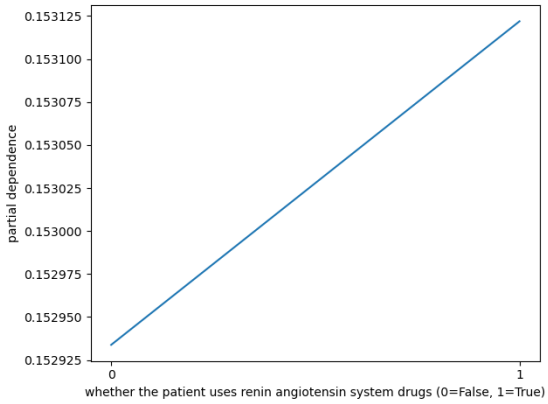


(a)

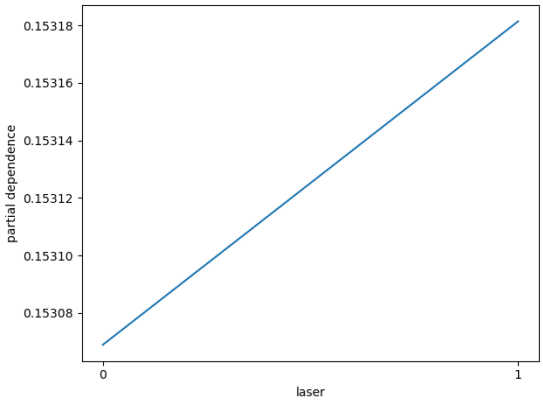


(b)

Figure B.5: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke), (b) whether the patient has hypertention

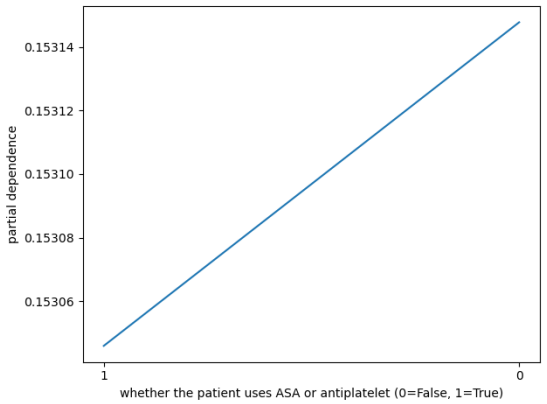


(a)

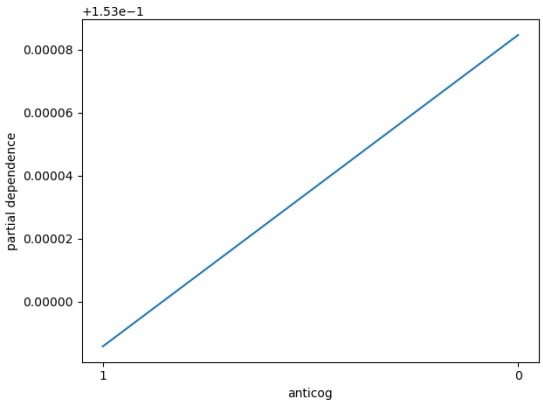


(b)

Figure B.6: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether th patient uses renin angiotensin system drugs (ACE-Inhibitors or Angiotensin 2 blocker) , (b) Whether patient had laser therapy for eye disease

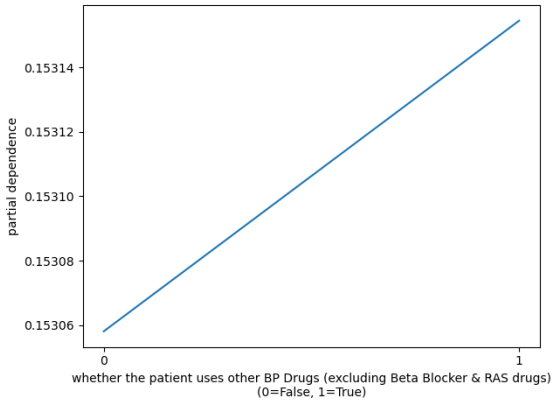


(a)



(b)

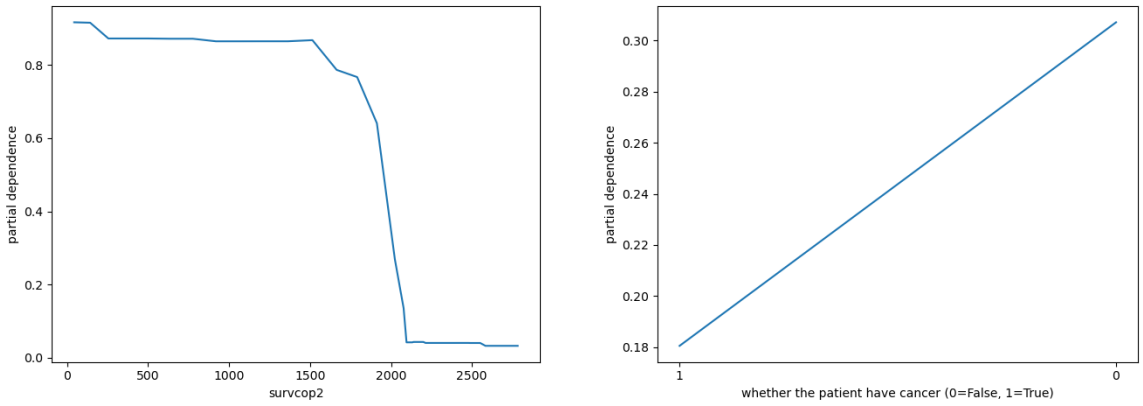
Figure B.7: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) Whether the patient uses ASA or antiplatelet, (b) whether the patient uses anticoagulant



(a)

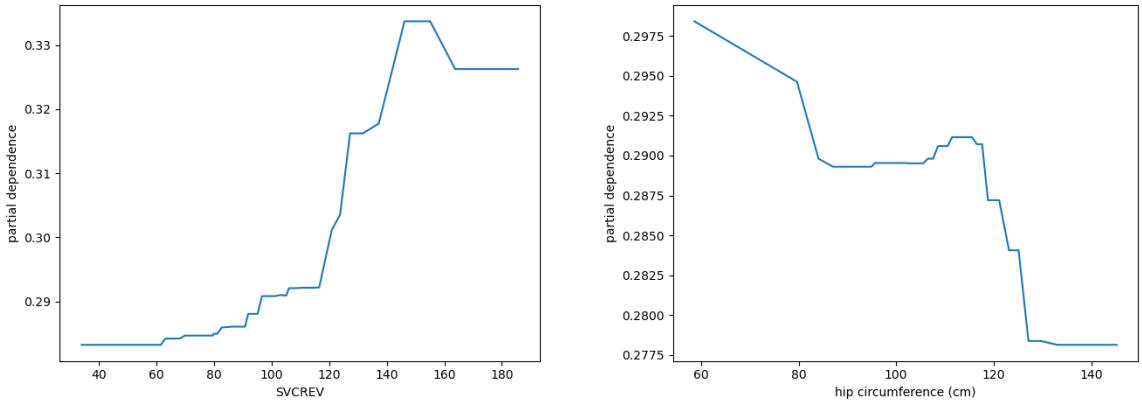
Figure B.8: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient uses anticoagulant, (b) Whether the patient uses other BP Drugs (excluding Beta Blocker and RAS drugs)

B.2 COPRIM2



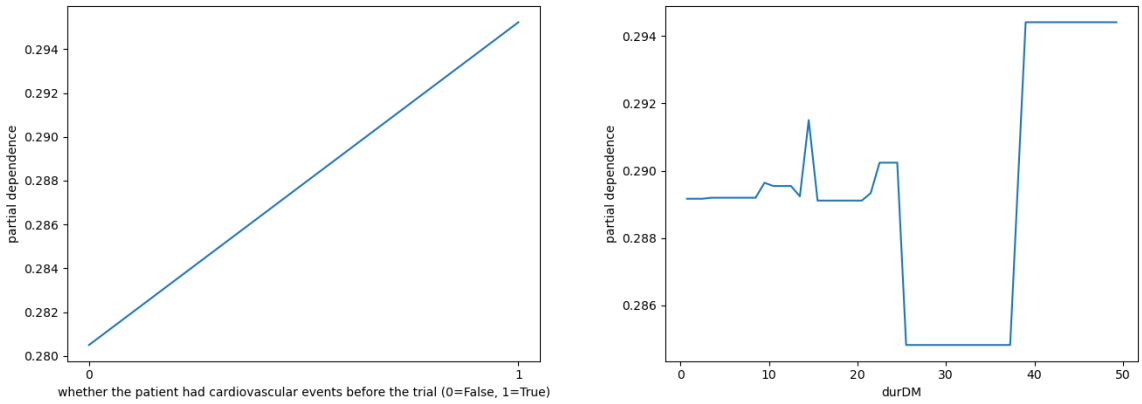
(a) (b)

Figure B.9: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer



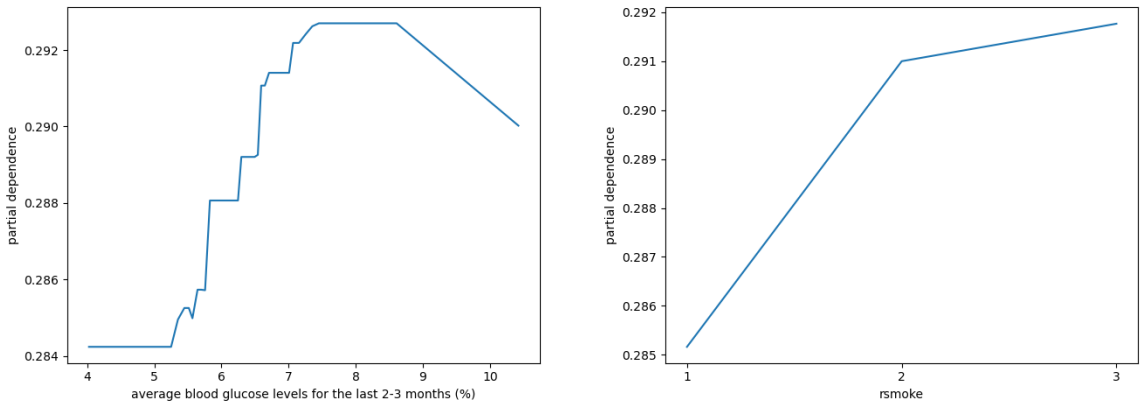
(a) (b)

Figure B.10: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit creatine value (SVCREV), (b) Hip circumference (rhipc)



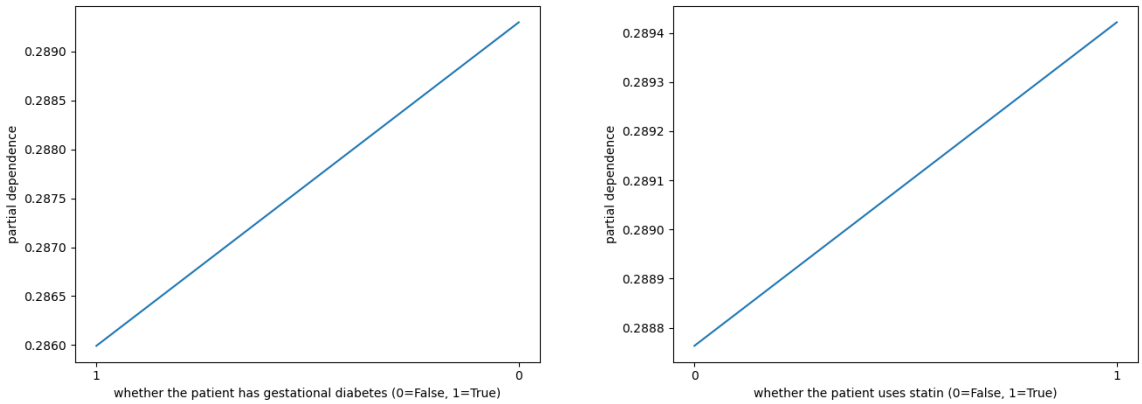
(a) (b)

Figure B.11: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior cardiovascular events (priorcv), (b) Diabetes Duration



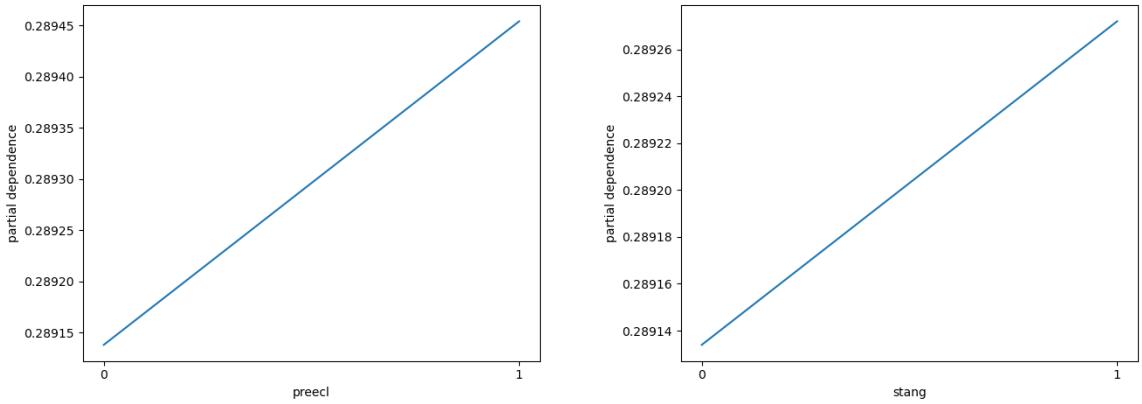
(a) (b)

Figure B.12: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) average blood glucose (sugar) levels for the last two to three months (hba1c), (b) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke)



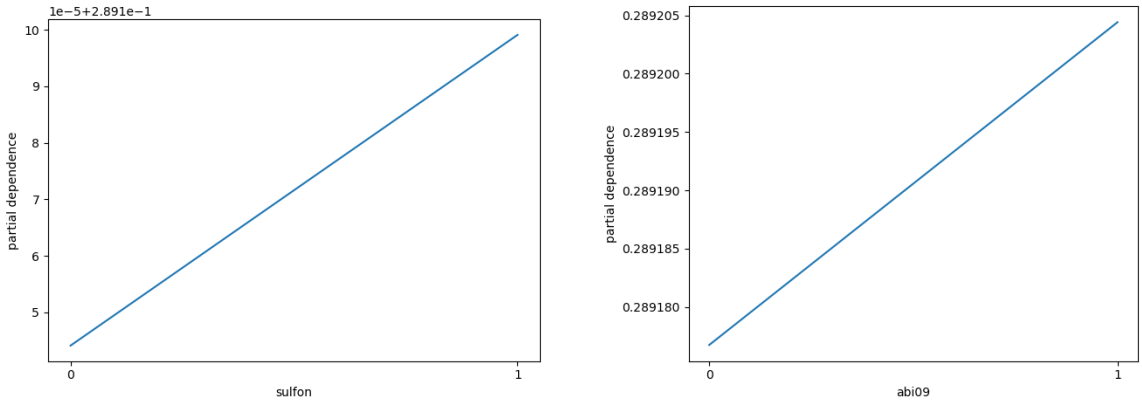
(a) (b)

Figure B.13: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whather the patient has gestational diabetes, (b) whether the patient uses statin



(a) (b)

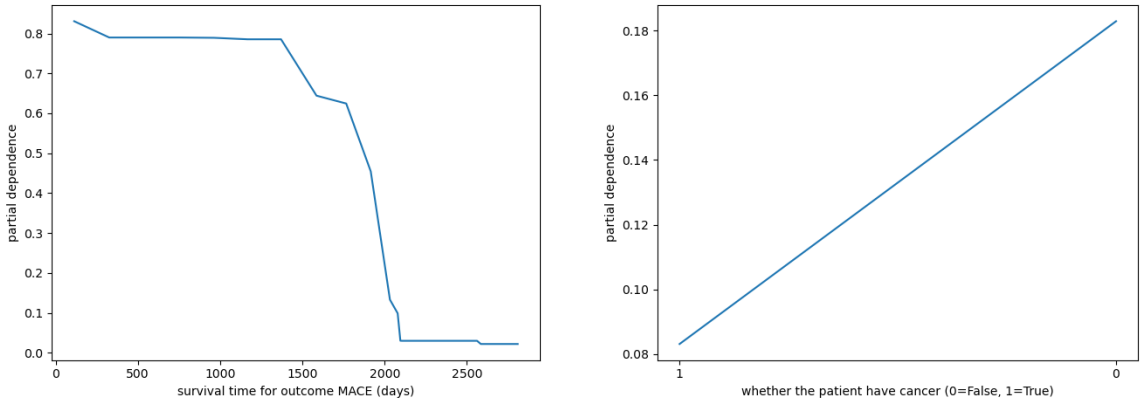
Figure B.14: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient has previous preeclampsia or eclampsia, (b) stable angina



(a) (b)

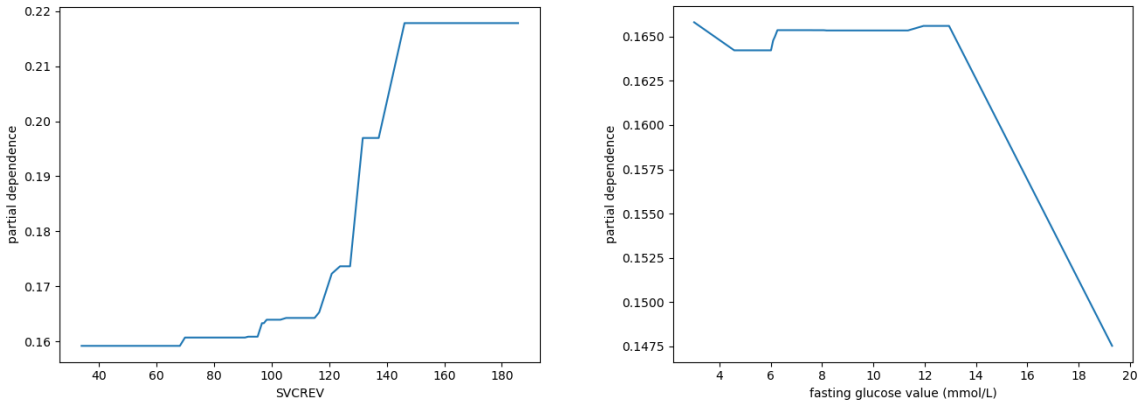
Figure B.15: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) ankle-brachial index during follow-up

B.3 MACE



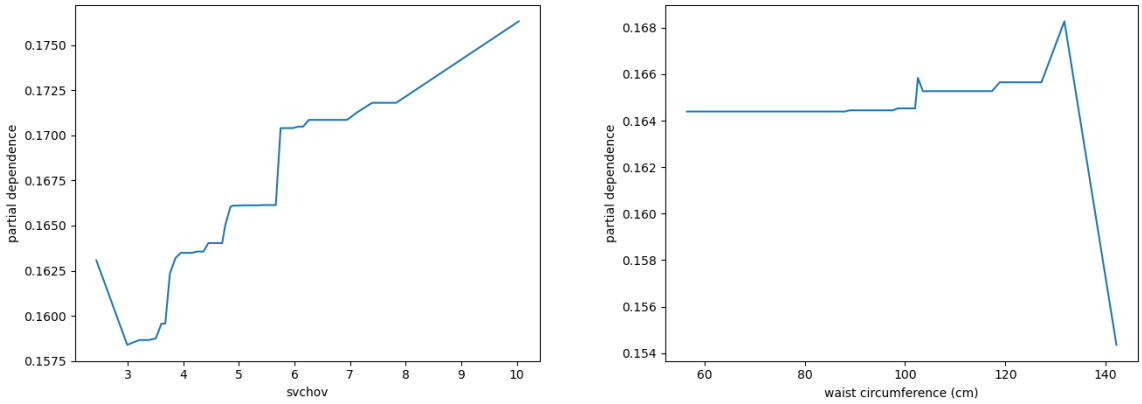
(a) (b)

Figure B.16: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer



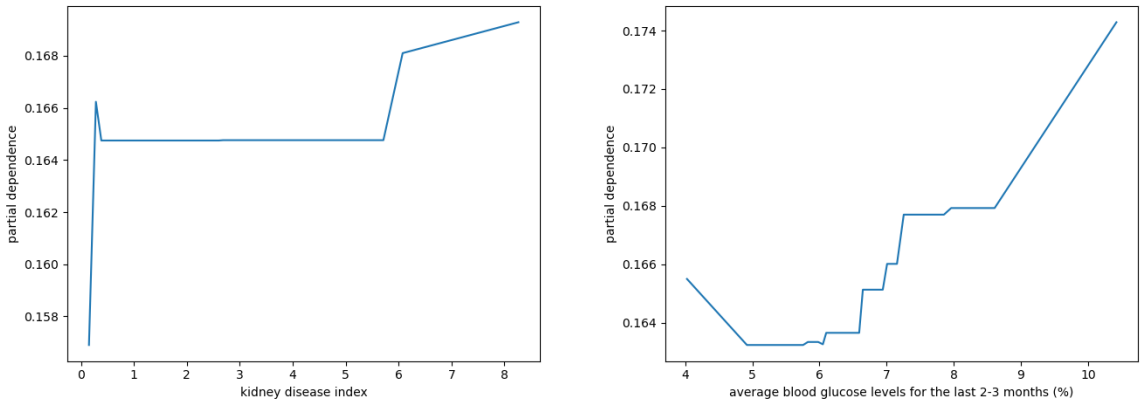
(a) (b)

Figure B.17: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit creatine value (SVCREV), (b) screening visit fasting glucose values (svfgv)



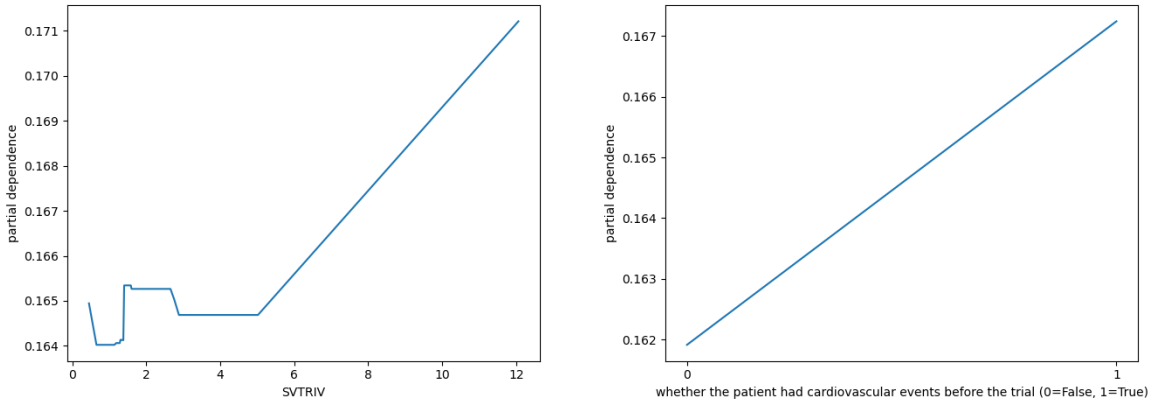
(a) (b)

Figure B.18: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit total cholesterol value (svchov), (b) Waist circumference (rwaistc)



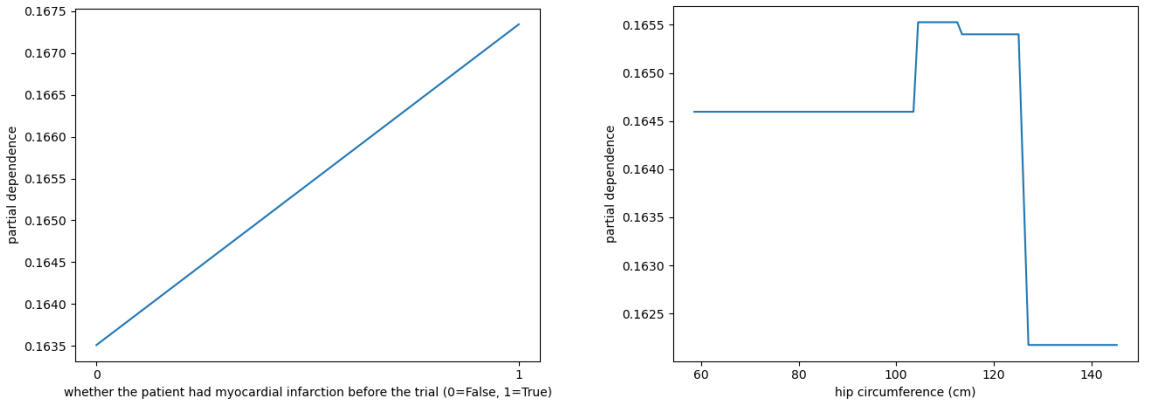
(a) (b)

Figure B.19: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) kidney disease index (kdi), (b) average blood glucose (sugar) levels for the last two to three months (hba1c)



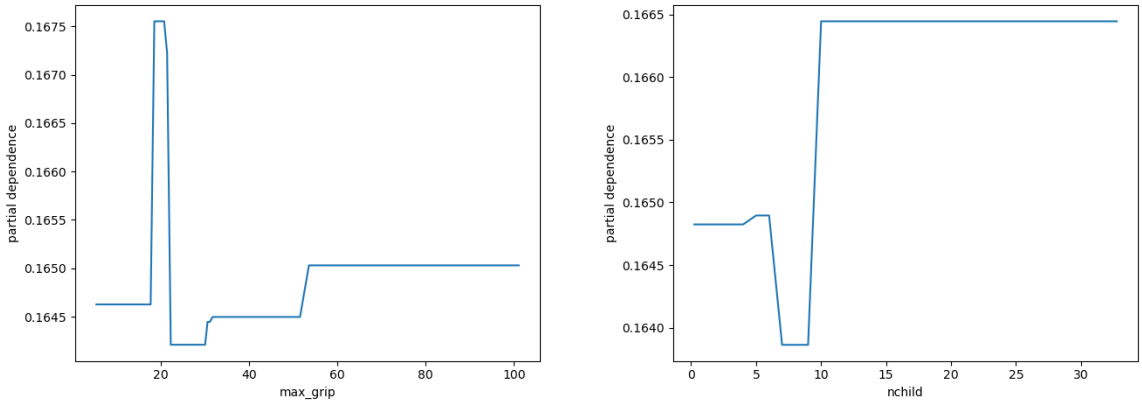
(a) (b)

Figure B.20: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit triglyceride value (SVTRIV), (b) prior cardiovascular events (priorcv)



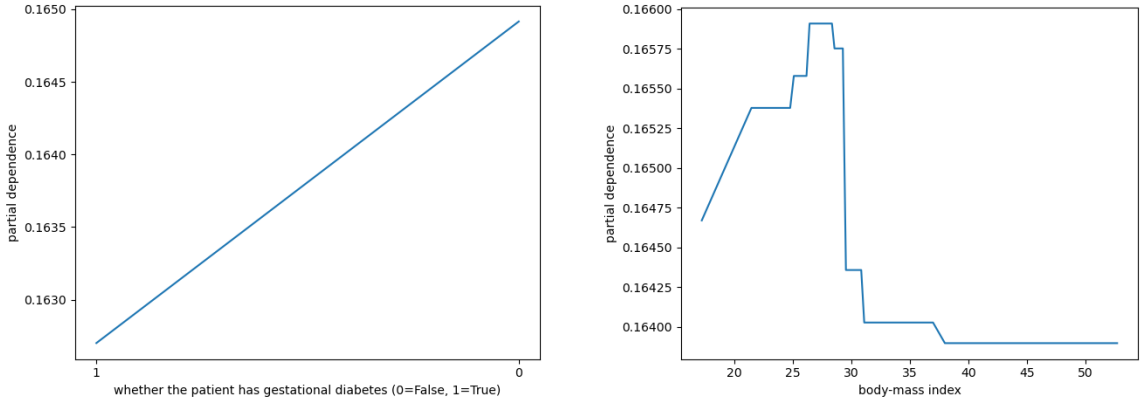
(a) (b)

Figure B.21: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior myocardial infarction (priormi), (b) Hip circumference (rhipc)



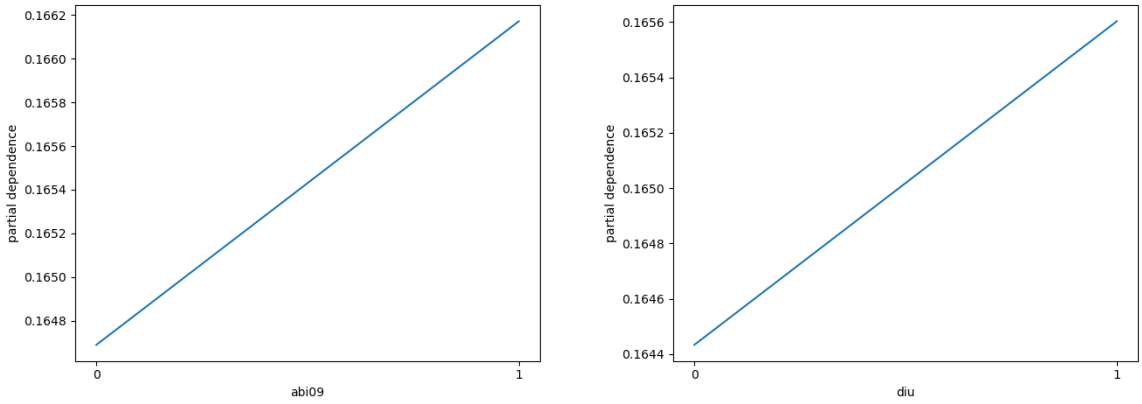
(a) (b)

Figure B.22: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) maximum grip strength (max_grip), (b) number of children



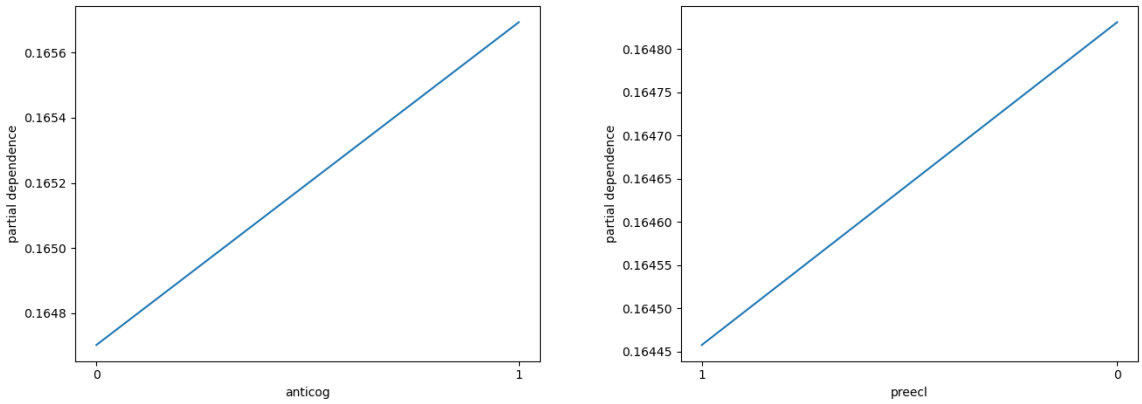
(a) (b)

Figure B.23: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whather the patient has gestational diabetes, (b) body-mass index (bmi)



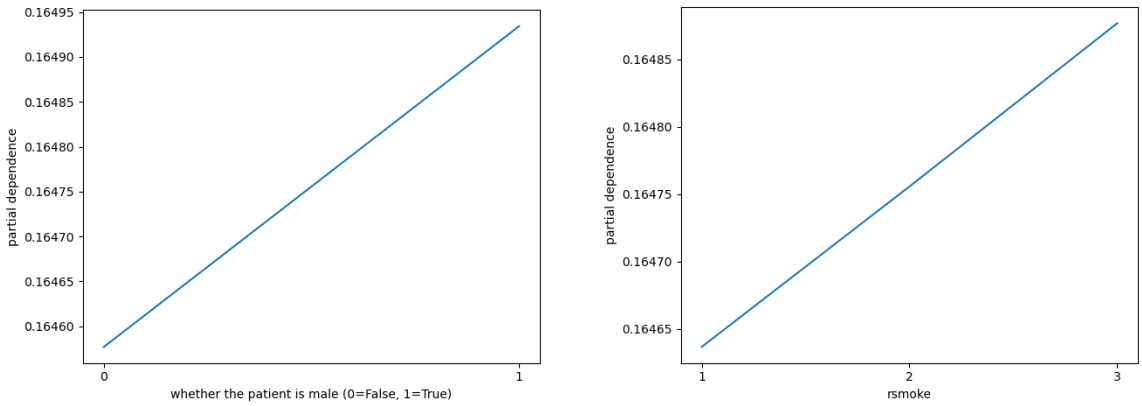
(a) (b)

Figure B.24: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) ankle-brachial index during follow-up, (b) whether the patient uses diuretic



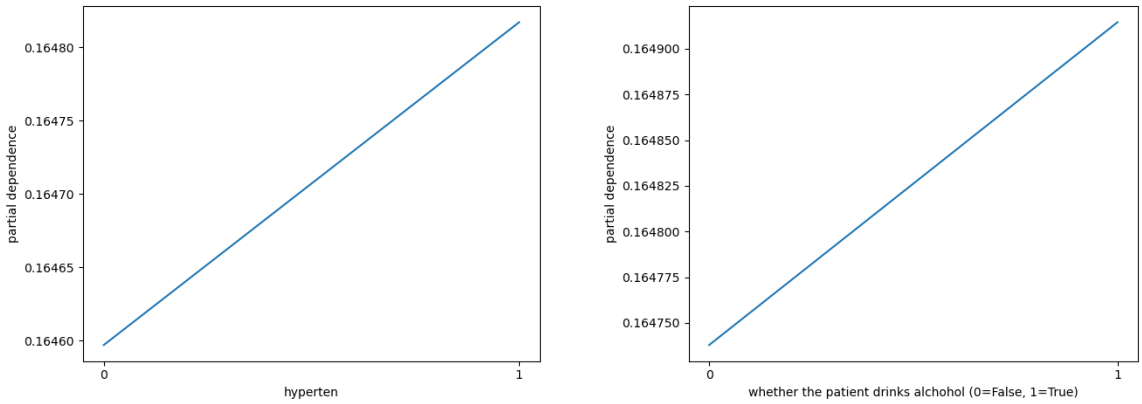
(a) (b)

Figure B.25: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient uses anticoagulant, (b) whather the patient has previous preeclampsia or eclampsia



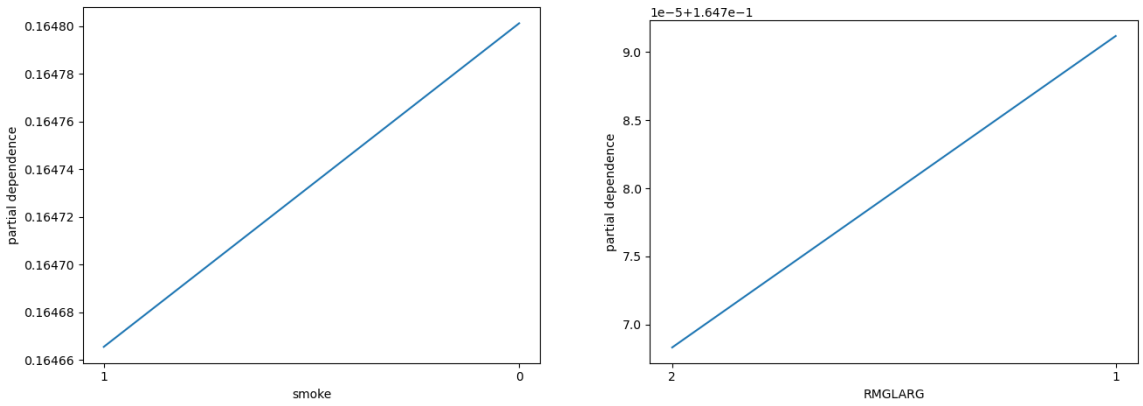
(a) (b)

Figure B.26: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whehter the patient is male, (b) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke)



(a) (b)

Figure B.27: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient has hypertention, (b) whehter the patient drinks alcohol



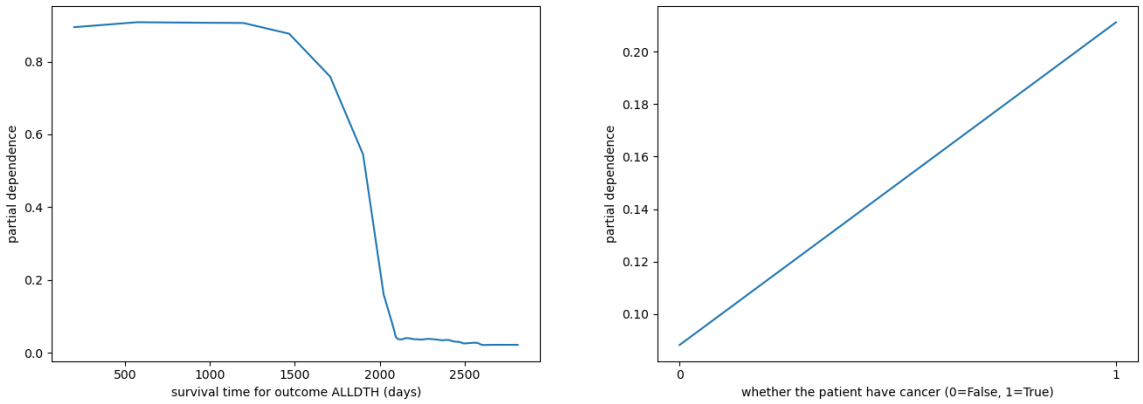
(a) (b)

Figure B.28: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) ORIGIN allocations (1=glargine;2=standard care)

Appendix C

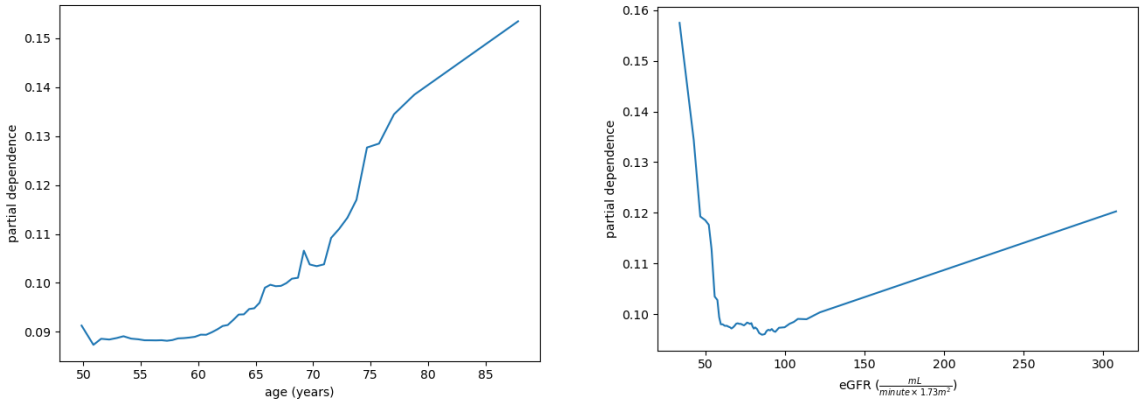
RandomForest

C.1 ALLDTH



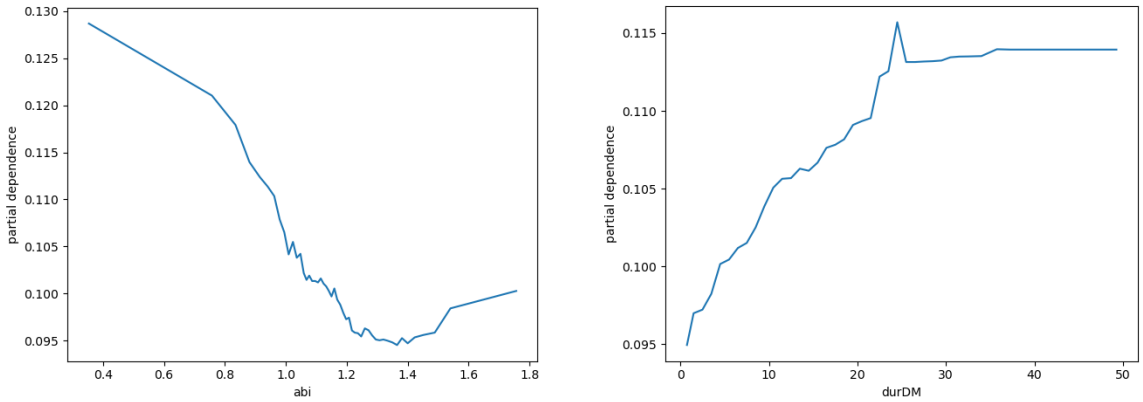
(a) (b)

Figure C.1: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer



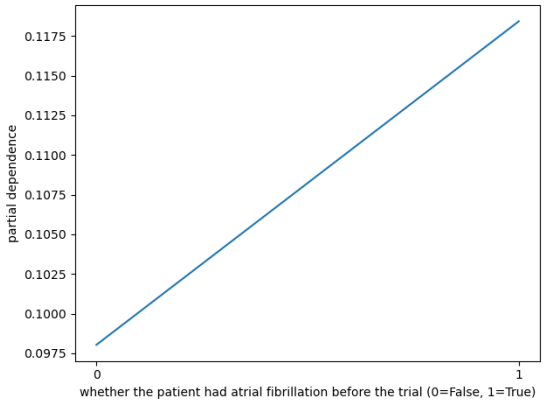
(a) (b)

Figure C.2: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) age, (b) estimated glomerular filtration rate (eGFR)

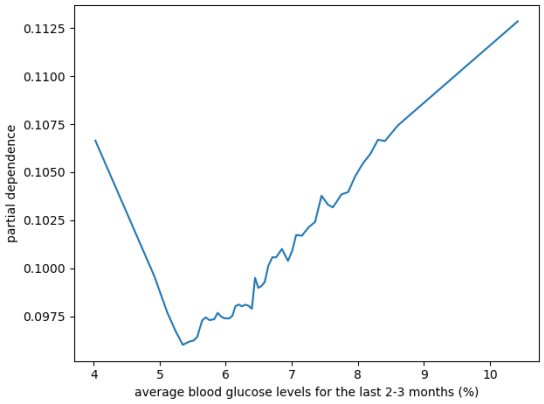


(a) (b)

Figure C.3: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) ankle-brachial blood pressure index (abi), (b) Diabetes Duration

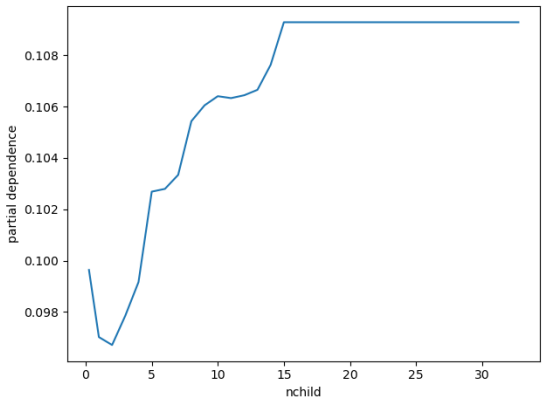


(a)

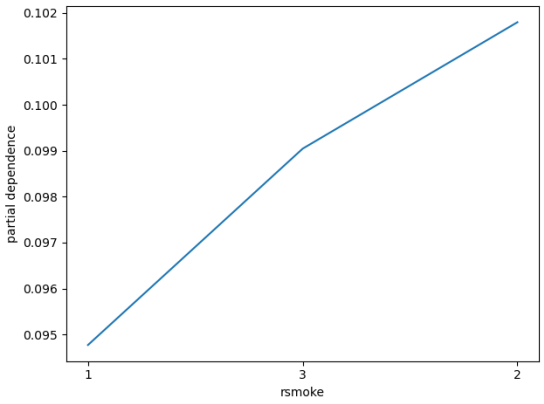


(b)

Figure C.4: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior atrial fibrillation (prioraf), (b) average blood glucose (sugar) levels for the last two to three months (hba1c)

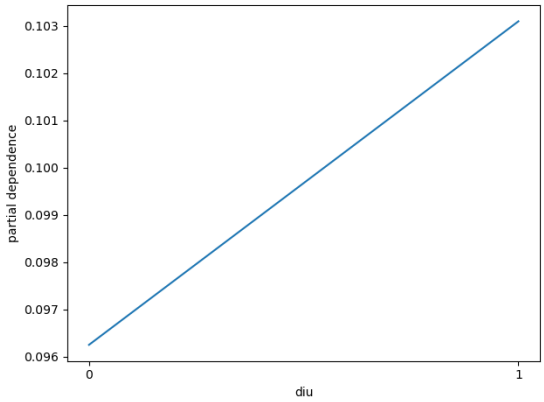


(a)

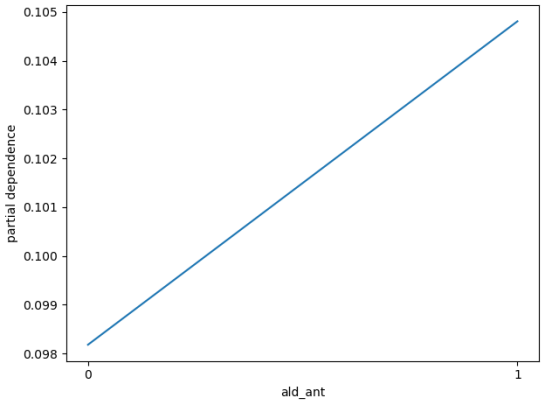


(b)

Figure C.5: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) number of children, (b) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke)

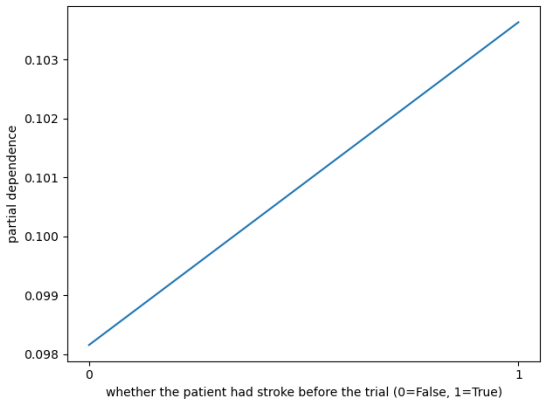


(a)

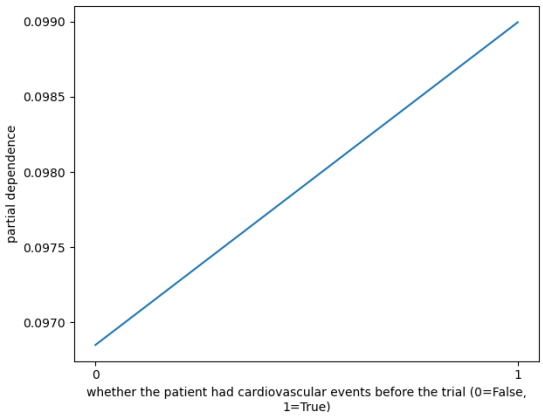


(b)

Figure C.6: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient uses diuretic, (b) whether the patient uses aldosterone antagonist

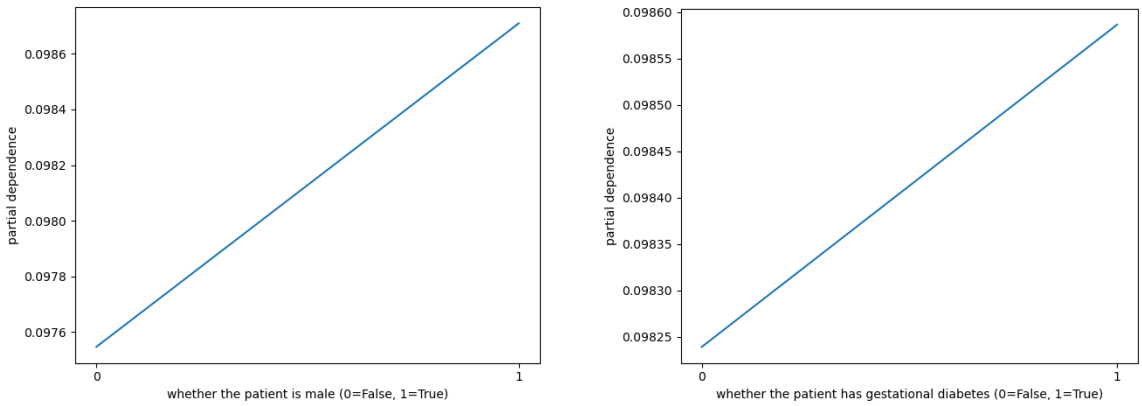


(a)



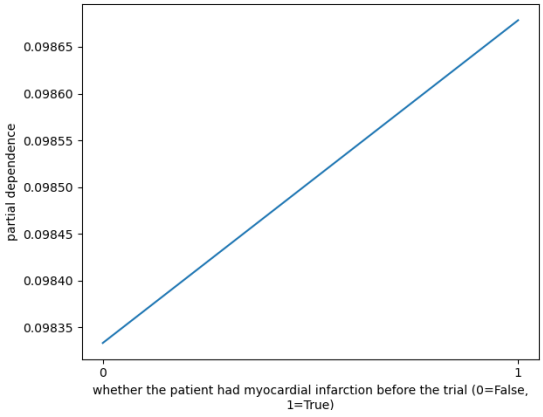
(b)

Figure C.7: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior stroke (priorstrk), (b) prior cardiovascular events (priorcv)



(a) (b)

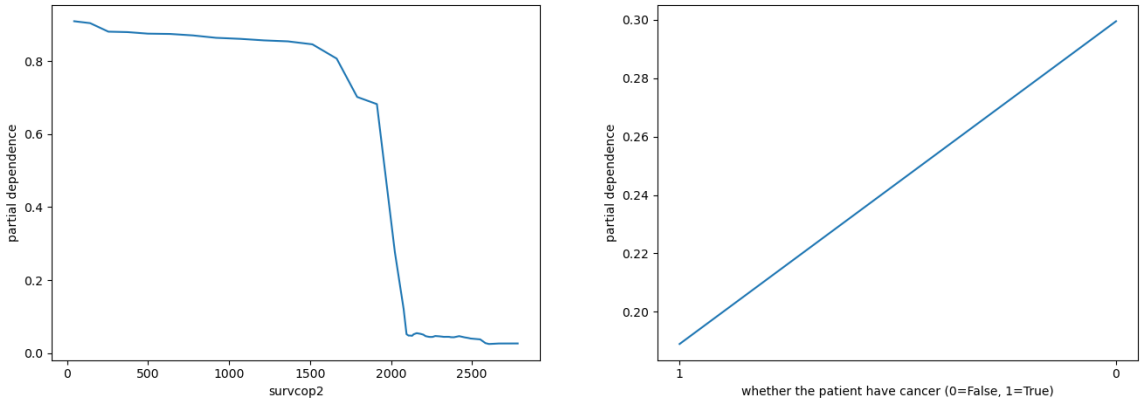
Figure C.8: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whehter the patient is male, (b) whather the patient has gestational diabetes



(a)

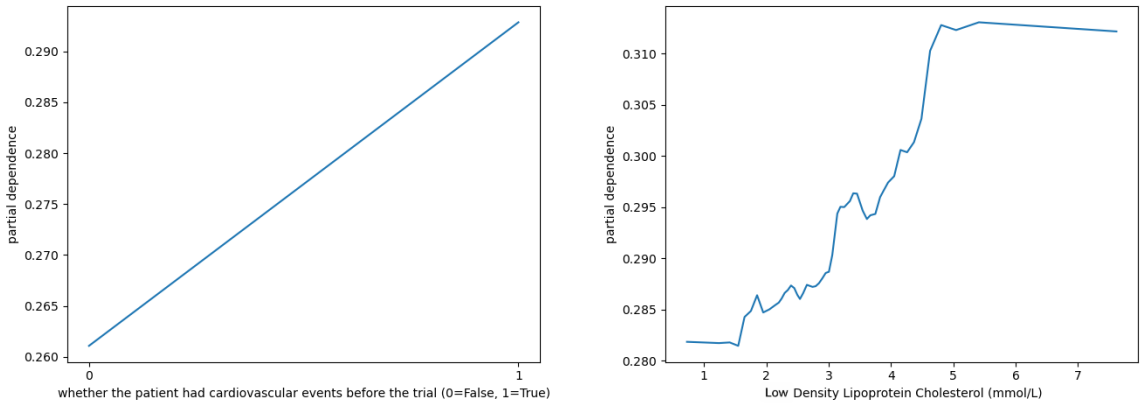
Figure C.9: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whather the patient has gestational diabetes, (b) prior myocardial infarction (priormi)

C.2 COPRIM2



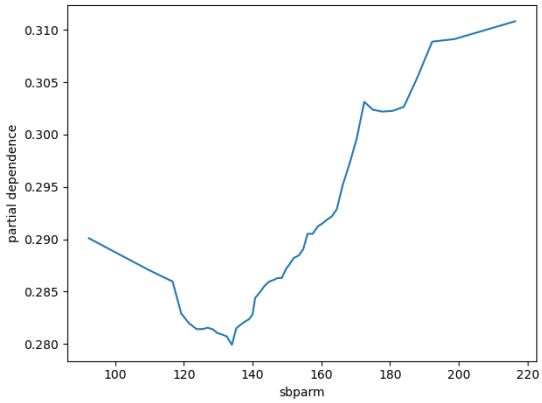
(a) (b)

Figure C.10: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer

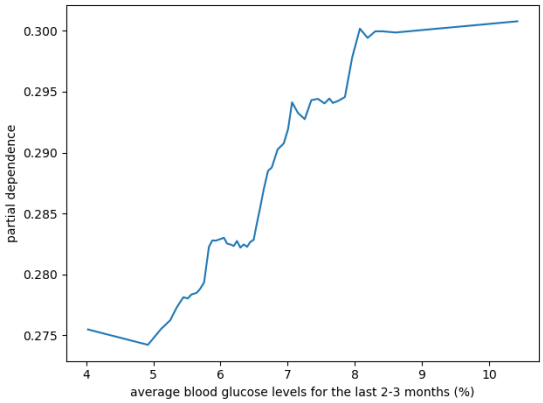


(a) (b)

Figure C.11: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior cardiovascular events (priorcv), (b) screening visit ligh-density lipoprotein cholesterol value (svldlv)

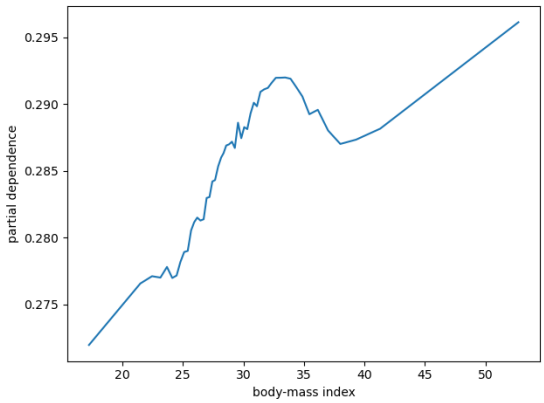


(a)

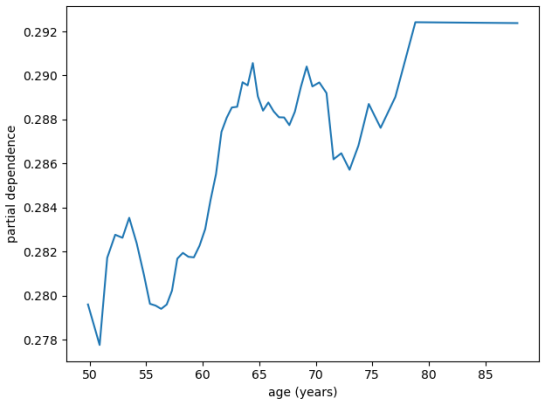


(b)

Figure C.12: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) arm systolic blood pressure (sbparm), (b) average blood glucose (sugar) levels for the last two to three months (hba1c)

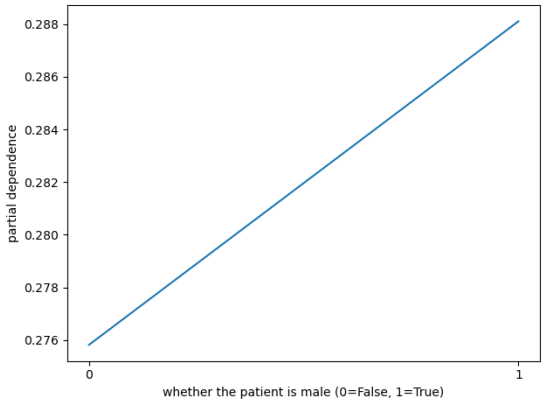


(a)

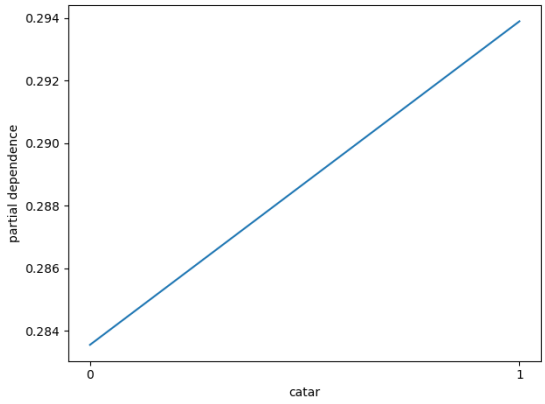


(b)

Figure C.13: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) body-mass index (bmi), (b) age

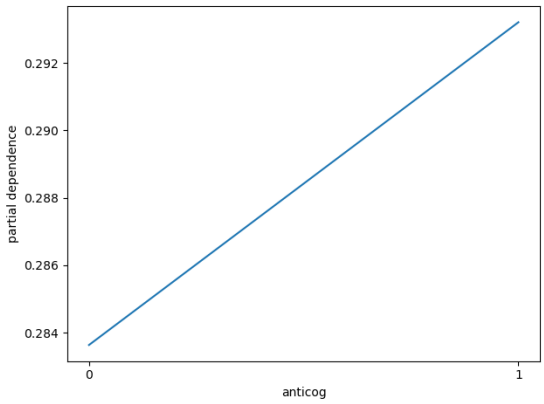


(a)

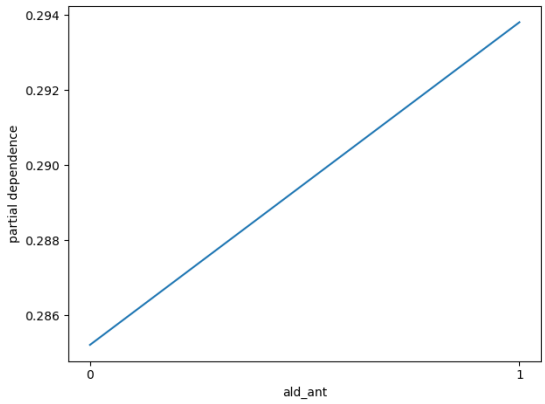


(b)

Figure C.14: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whehter the patient is male, (b) Whether the patient has cataracts

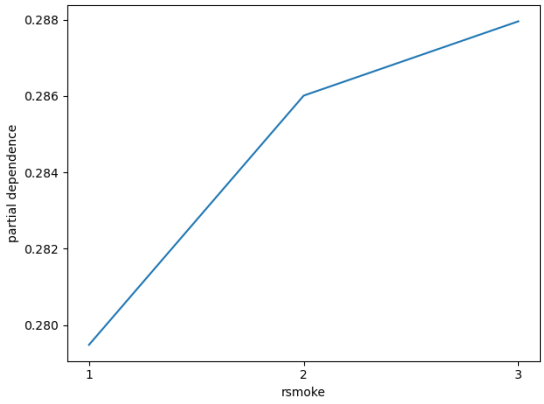


(a)

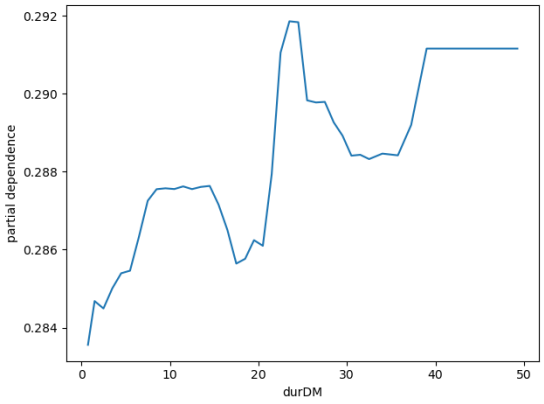


(b)

Figure C.15: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient uses anticoagulant, (b) whether the patient uses aldosterone antagonist

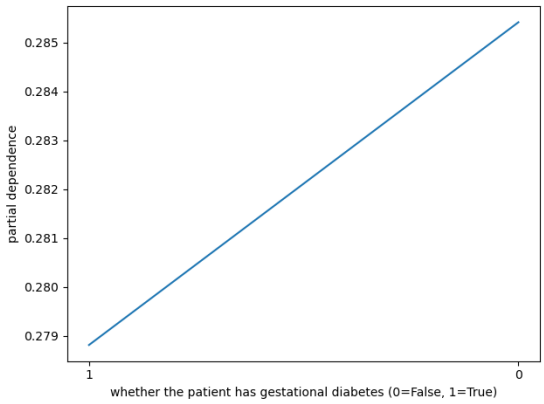


(a)

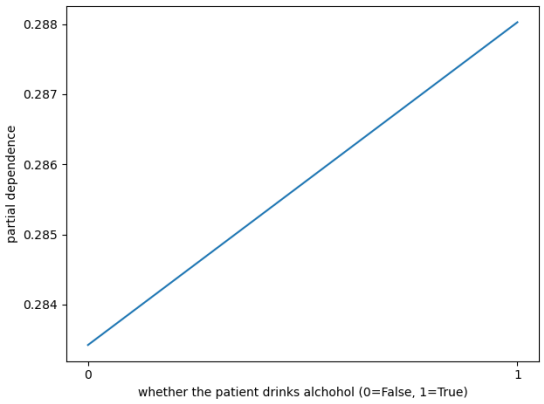


(b)

Figure C.16: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) smoker categories (1 - no smoking; 2 - current smoker; 3 - former smoker) (rsmoke), (b) Diabetes Duration

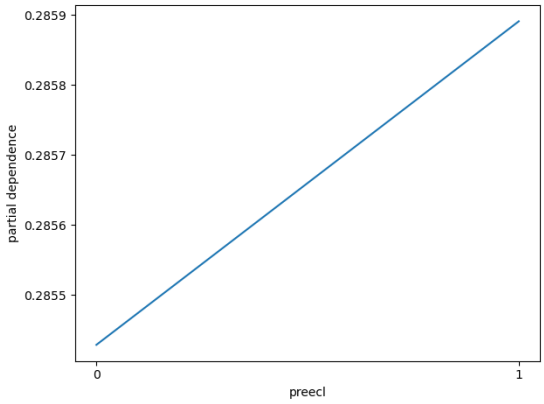


(a)



(b)

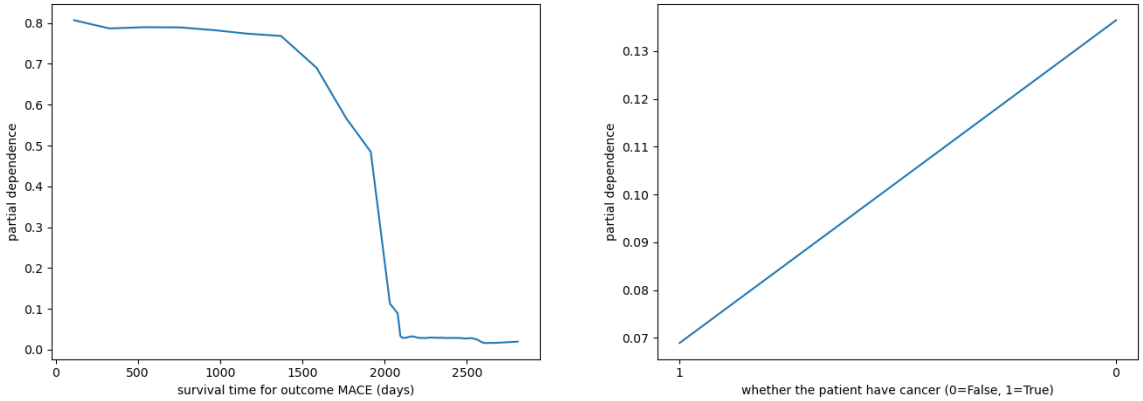
Figure C.17: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whather the patient has gestational diabetes, (b) whehter the patient drinks alcohol



(a)

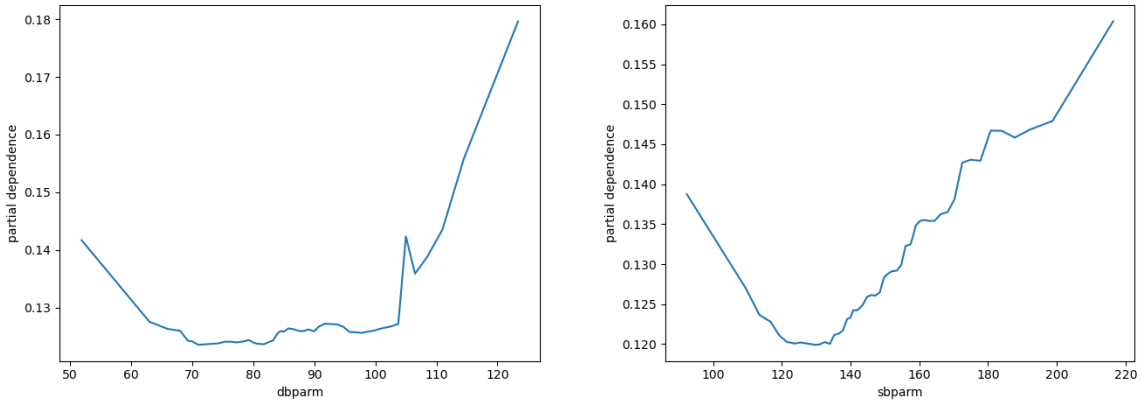
Figure C.18: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whehter the patient drinks alcohol, (b) whather the patient has previous preeclampsia or eclampsia

C.3 MACE



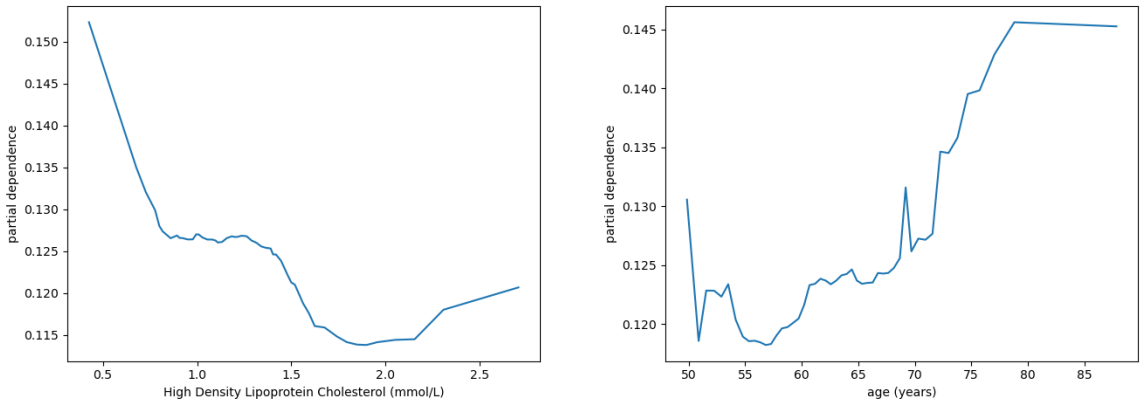
(a) (b)

Figure C.19: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) survival time, (b) whether they have cancer



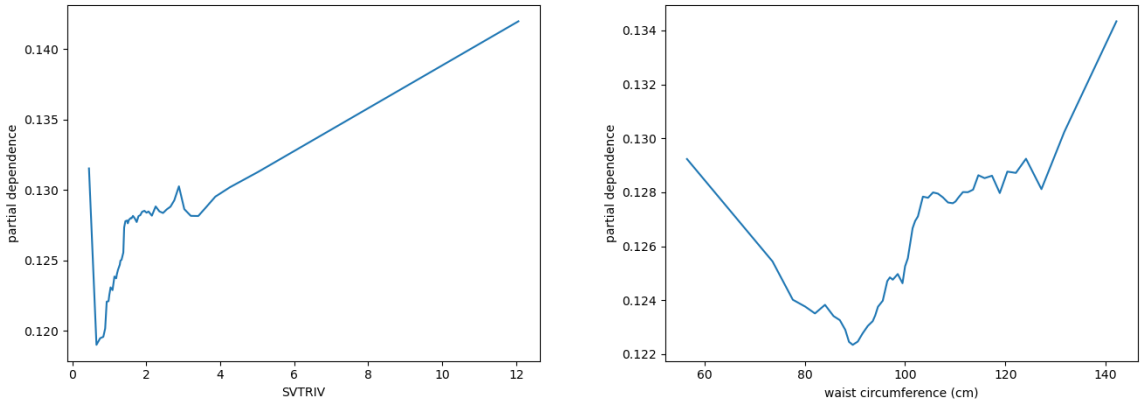
(a) (b)

Figure C.20: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) arm diastolic blood pressure (dbparm), (b) arm systolic blood pressure (sbparm)



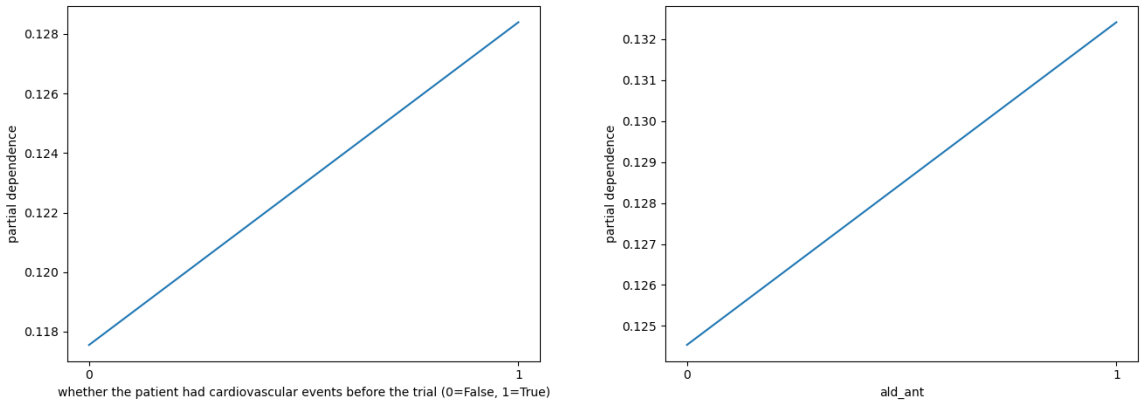
(a) (b)

Figure C.21: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit high-density lipoprotein cholesterol value (svhdlv), (b) age



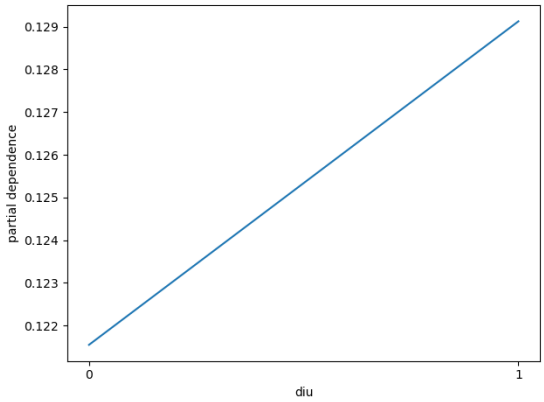
(a) (b)

Figure C.22: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) screening visit triglyceride value (SVTRIV), (b) Waist circumference (rwaistc)

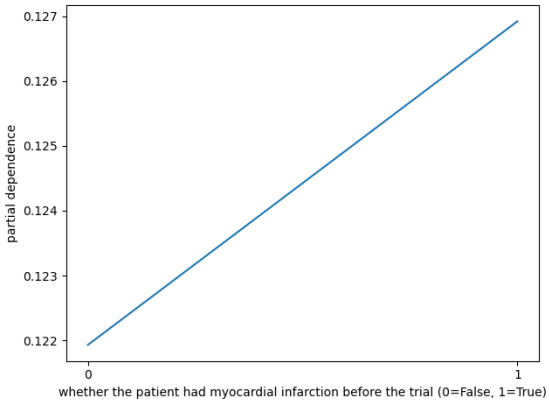


(a) (b)

Figure C.23: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) prior cardiovascular events (priorcv), (b) whether the patient uses aldosterone antagonist

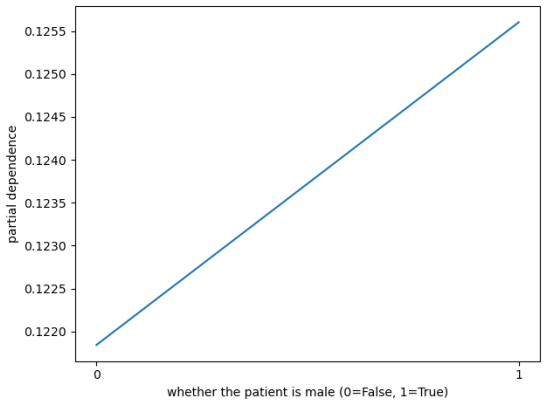


(a)

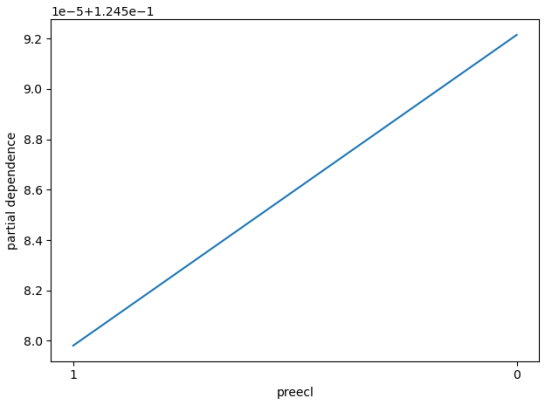


(b)

Figure C.24: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whether the patient uses diuretic, (b) prior myocardial infarction (priormi)



(a)



(b)

Figure C.25: Partial Dependence Plots of Gradient Boosted Trees for COPRIM2. Estimated average predicted probability of patients in dataset with the outcome ALLDTH is on y-axis given the following explantory variable measure on x-axis: (a) whather the patient has previous preeclampsia or eclampsia