CARDIOVASCULAR DISEASE DATA

Project Final Report

Group 6

MEMBER: Zijian Zhang, Zhouyuan Li, Yujin Song, Jimeng Lin, Yiyang Liu, Haiqiao Xiong

CONTENTS

1

INTRODUCTION

Background and Intention

PART ONE

The background of data

Cardiovascular disease

affects the heart or blood vessels

is chemic or hemorrhagic

happens in heart,cerebrum and systemic organization

results in tall lipidemia, sticky blood, atherosclerosis, hypertension.

STEP 1 INTRODUCTION - BACKGROUND

Cardiovascular disease

most common:
coronary artery disease

Heart attack, stroke, heart failure, arrhythmia, and heart valve problems

main causes:
smoking, high blood pressure, high cholesterol, glucagon level, drinking alcohol, lack of exercise, and obesity

serious disease: about 15 million people die of it each year

STEP 1 INTRODUCTION - INTENTION

1.predicts the probability of cardiovascular disease through variables interested

2.reminds people who have a high probability to timely medical examination

3.attracts customers to buy our company's health product and improves revenue

2

DIDA FRAMEWORK

PART TWO

STEP 2 DIDA FRAMEWORK

D

DATA

Age, Gender, Physical indicators (height, weight, Systolic blood pressure, Diastolic blood pressure, Cholesterol, and Glucose level), Habits (smoke, alcohol, and Physical activity)

I

INSIGHTS

The risk of developing cardiovascular disease (Probability)

D

DECISION

Recommend medical products to customers who have a high probability of developing cardiovascular disease.

A

ADVANTAGE

Customers get healthier and the company gains revenue.

dataset resource: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

STEP 2 DIDA FRAMEWORK

It can help the company manager to make more specific decisions.

Manager can be more intuitive to find out whether this person is a potential customer.

Hence, the manager can arrange ways to attract that customer to buy the company products

It can help the company to reduce costs.

Company can send advertisements to the particular customers who need the company’ s medical products.

We can pay attention to the factors that affect the final results and then adjust and enhance the company’ s further strategy.

3

DATA MINING PROCESS

Logistic Regression,Classification Tree,Nearest Neighbor,Neural Network

PART THREE

STEP 3 DATA MINING PROCESS - SET THE MODEL

We pick 4 models: logistic regression, classification tree, nearest neighbor, and neural network, to figure out the highest AUC and the best model that fit our goal.

Numerical variables: age, height, weight, systolic blood pressure, and diastolic blood pressure

Categorical variables: gender, cholesterol, glucose, smoking, alcohol intake, physical activity, and cardio

Then we create dummy variables for categorical variables and dropped redundant dummies.

Data partition: test size = 0.2

Dependent variable: cardio_1

Cross validation: k = 5

STEP 3 DATA MINING PROCESS - MODEL ANALYSIS

LogisticRegression	Penalty Level	AUC
	3.50446847	0.7893673867740303

We generated 1000 individual alphas from 0.001 to 100 to test the model.
The model gave us 3.50 as the penalty level, and 0.789 as the AUC

Nearest Neighbor	Optimal Number of Neighbors	AUC
	39	0.7337301258538164

We set max_k as 100. The model shows that the optimal number of neighbors is 39, and the AUC is 0.734.

Neural Network	AUC
	0.8035827730964054

We restrict the search of optimal model in 1 to 15 hidden layers. It returns the AUC of 0.804.

STEP 3 DATA MINING PROCESS - MODEL ANALYSIS

Classification Tree	Optimal Tree Depth	Number of Leaf Nodes	AUC
	7	121	0.7951785200984685

We set the level of depth between 1 and 10, and the analysis turned out that the best pruned tree has 7 levels of depth and 121 leaf nodes.

From the 121 leaf nodes, we choose the top 40% most effective nodes and the top 40% nodes with largest sample size, and found the overlap of these two groups.

So we decided to report the following 15 nodes.
Their nodes ID#: 121, 104, 102, 116, 95, 105, 68, 66, 35, 92, 118, 96, 119, 50, 115.
And the tree returns AUC of 0.795.

STEP 3 DATA MINING PROCESS - MODEL RESULTS

LogisticRegression	Penalty Level	AUC
	3.50446847	0.7893673867740303

Classification Tree	Optimal Tree Depth	Number of Leaf Nodes	AUC
	7	121	0.7951785200984685

Nearest Neighbor	Optimal Number of Neighbors	AUC
	39	0.7337301258538164

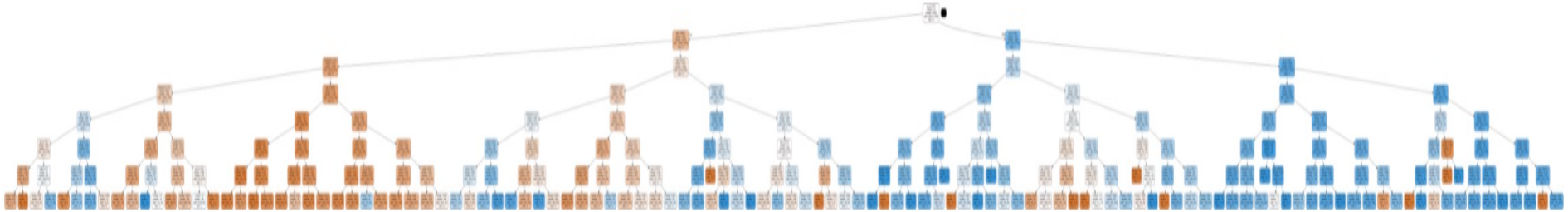
Neural Network	AUC
	0.803582730964054

4

ANALYSIS AND RESULTS

PART FOUR

STEP 4 ANALYSIS AND RESULTS - METHOD CHOICE



CHOICE

According to the AUC performance of these models, we can find out that the Neural Network model has the highest AUC then classification tree, logistic regression, and nearest neighbor.
However, considering the interpretable ability we select the **classification tree model** to be the final selected model.

REASON

Since the ultimate goal of this case is to predict potential cardiovascular disease, interpretability is very crucial. We can see how each predictors weight and affect every specific observation, and give customized advice (what is next step of treatment) to every patient.

STEP 4 ANALYSIS AND RESULTS - TREE NODES (15 IN ALL)

Leaf node ID = 121
Path = ['ap_hi > 129.5', 'ap_hi > 138.5', 'ap_hi > 149.5', 'ap_lo > 68.5', 'age > 18372.0', 'weight > 63.90999984741211', 'age > 18374.0']
sample = 5069
value = [752, 4317]
class = 1

Leaf node ID = 116
Path = ['ap_hi > 129.5', 'ap_hi > 138.5', 'ap_hi > 149.5', 'ap_lo > 68.5', 'age <= 18372.0', 'weight > 69.5', 'age <= 18334.5']
sample = 1286
value = [132, 1154]
class = 1

Leaf node ID = 104
Path = ['ap_hi > 129.5', 'ap_hi > 138.5', 'ap_hi <= 149.5', 'gluc_1 > 0.5', 'cholesterol_1 > 0.5', 'ap_hi <= 143.5', 'age <= 21983.5']
sample = 3456
value = [676, 2780]
class = 1

Leaf node ID = 95
Path = ['ap_hi > 129.5', 'ap_hi > 138.5', 'ap_hi <= 149.5', 'gluc_1 <= 0.5', 'age <= 23464.5', 'age > 17993.5', 'cholesterol_1 <= 0.5']
sample = 926
value = [196, 730]
class = 1

Leaf node ID = 102
Path = ['ap_hi > 129.5', 'ap_hi > 138.5', 'ap_hi <= 149.5', 'gluc_1 > 0.5', 'cholesterol_1 <= 0.5', 'height > 153.5', 'age <= 23471.0']
sample = 1713
value = [251, 1462]
class = 1

Leaf node ID = 105
Path = ['ap_hi > 129.5', 'ap_hi > 138.5', 'ap_hi <= 149.5', 'gluc_1 > 0.5', 'cholesterol_1 > 0.5', 'ap_hi <= 143.5', 'age > 21983.5']
sample = 828
value = [127, 701]
class = 1

STEP 4 ANALYSIS AND RESULTS - TREE NODES (15 IN ALL)

Leaf node ID = 68
Path = ['ap_hi > 129.5', 'ap_hi <= 138.5', 'cholesterol_1 <= 0.5', 'cholesterol_2 <= 0.5', 'gluc_1 > 0.5', 'height <= 180.5', 'height <= 178.5']
sample = 742
value = [123, 619]
class = 1

Leaf node ID = 66
Path = ['ap_hi > 129.5', 'ap_hi <= 138.5', 'cholesterol_1 <= 0.5', 'cholesterol_2 <= 0.5', 'gluc_1 <= 0.5', 'height > 150.5', 'weight <= 120.5']
sample = 577
value = [155, 422]
class = 1

Leaf node ID = 35
Path = ['ap_hi <= 129.5', 'age > 19960.5', 'age <= 22147.5', 'cholesterol_1 <= 0.5', 'cholesterol_2 <= 0.5', 'weight > 68.75', 'weight <= 99.5']
sample = 479
value = [147, 332]
class = 1

Leaf node ID = 92
Path = ['ap_hi > 129.5', 'ap_hi <= 138.5', 'cholesterol_1 > 0.5', 'age > 21731.0', 'smoke_0 > 0.5', 'ap_lo > 77.0', 'ap_lo > 89.5']
sample = 428
value = [138, 290]
class = 1

Leaf node ID = 118
Path = ['ap_hi > 129.5', 'ap_hi > 138.5', 'ap_hi > 149.5', 'ap_lo > 68.5', 'age > 18372.0', 'weight <= 63.90999984741211', 'weight <= 59.5']
sample = 423
value = [61, 362]
class = 1

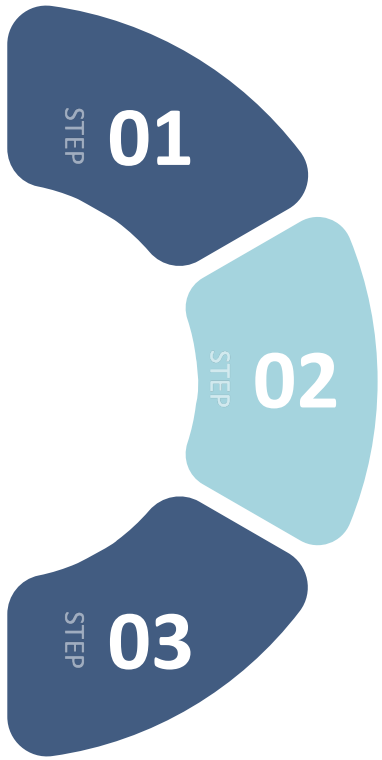
Leaf node ID = 96
Path = ['ap_hi > 129.5', 'ap_hi > 138.5', 'ap_hi <= 149.5', 'gluc_1 <= 0.5', 'age <= 23464.5', 'age > 17993.5', 'cholesterol_1 > 0.5']
sample = 373
value = [105, 268]
class = 1

Leaf node ID = 119
Path = ['ap_hi > 129.5', 'ap_hi > 138.5', 'ap_hi > 149.5', 'ap_lo > 68.5', 'age > 18372.0', 'weight <= 63.90999984741211', 'weight > 59.5']
sample = 355
value = [23, 332]
class = 1

Leaf node ID = 50
Path = ['ap_hi <= 129.5', 'age > 19960.5', 'age > 22147.5', 'cholesterol_1 <= 0.5', 'cholesterol_2 <= 0.5', 'height <= 186.0', 'weight > 65.5']
sample = 322
value = [58, 264]
class = 1

Leaf node ID = 115
Path = ['ap_hi > 129.5', 'ap_hi > 138.5', 'ap_hi > 149.5', 'ap_lo > 68.5', 'age <= 18372.0', 'weight <= 69.5', 'gender_1 > 0.5']
sample = 289
value = [39, 250]
class = 1

STEP 4 ANALYSIS AND RESULTS - FINAL RESULTS



We decided to give out customized messages to different patient group.
If the patient has a probability higher than 80%, then we will ask him to find cardiologist and start a treatment.

If the probability is between 80% and 65%, we will ask the patient to take a thorough examination to figure out whether he has the disease.

If the probability is between 65% and 50%, we will ask the patient to keep a healthy lifestyle and have close monitor on his heart condition

5

CONCLUSIONS

PART FIVE

KEY TAKEAWAYS

- The most impactful attribute is systolic blood pressure
- Classification Tree yields the best AUC, 0.795
- Customized suggestion and treatment for patient in different level of probability

Thank you for listening