Midterm Report

Beibei Cao

Introduction

Heart Diseases have led to 12 million deaths every year worldwide estimated by the World Health Organization. Deaths caused by heart diseases take up to half the deaths due to cardiovascular diseases in the United States. The identification of potential risk factors of heart diseases has become increasingly urgent as early prognosis could effectively promote improvement on problematic lifestyles, and further reducing the chance of the development of heart diseases. This analysis aims to prioritize the most relevant risk factors from available data of heart disease as well as predict the its risk using various classifiction method.

Data description:

Demographic:
- `male`: male 1 female 0 (Binary)
- `age`: age of the patient (Continuous)
- `education` education level (Treated as continuous)

Behavioral:
- `current_smoker`: whether or not the patient is a current smoker (Binary)
- `cigs_per_day`: the number of cigarettes smoked per day (Continuous)

Medical (history):
- `bp_meds`: blood pressure medication (Binary)
- `prevalent_stroke`: stroke history (Binary)
- `prevalent_hyp`: hypertensive history (Binary)
- `diabetes`: diabetes history (Binary)

Medical (current):
- `tot_chol`: total cholesterol level (Continuous)
- `sys_bp`: systolic blood pressure (Continuous)
- `dia_bp`: diastolic blood pressure (Continuous)
- `bmi`: Body Mass Index (Continuous)
- `heart_rate`: heart rate (Continuous)
- `glucose`: glucose level (Continuous)

Response variable:
- `ten_year_chd`: 10 year risk of coronary heart disease CHD (Binary)

The data set is available on the Kaggle website. The data originally contains 4,238 of 15 predictor variables and 1 response variable. Missing data takes a very small percentage less than 1%. By visualizing the missing data, we noticed that the majority of missing data is the measurements of glucose level. The correlation matrix in the later section indicates that glucose is highly correlated with the diabetes history. Thus, we imputed the missing values of glucose with the diabetes column using BagImpute. Other observations with missing values were omitted directly (0.4%). Categorical variables are formatted as factors and the response variable is converted to string indicators “yes” and “no” in order to facilitate downstream modeling.

Exploratory analysis/visualization

Table 1: Data summary

Name	heart
Number of rows	3987
Number of columns	16
Column type frequency:	
factor	7
numeric	9
Group variables	None

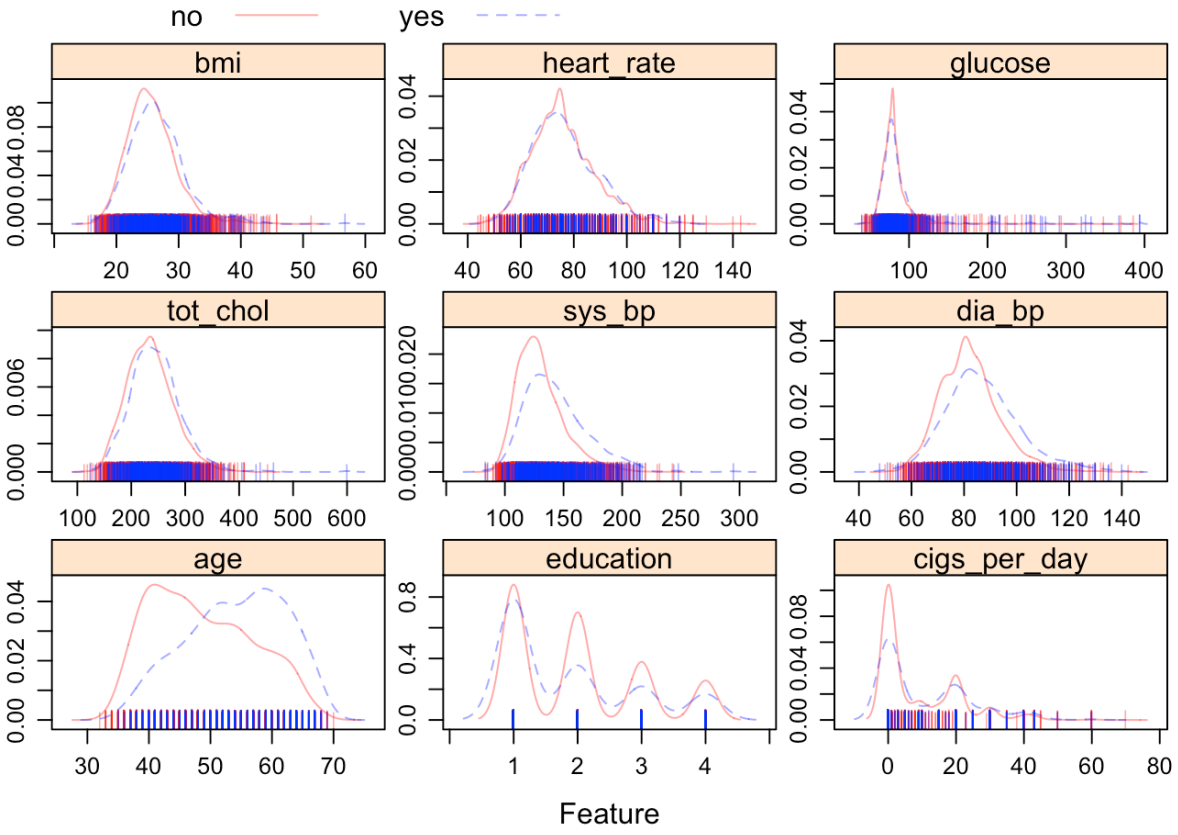
Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
male	0	1	FALSE	2	0: 2260, 1: 1727
current_smoker	0	1	FALSE	2	0: 2029, 1: 1958
bp_meds	0	1	FALSE	2	0: 3870, 1: 117
prevalent_stroke	0	1	FALSE	2	0: 3965, 1: 22
prevalent_hyp	0	1	FALSE	2	0: 2753, 1: 1234
diabetes	0	1	FALSE	2	0: 3886, 1: 101
ten_year_chd	0	1	FALSE	2	no: 3392, yes: 595

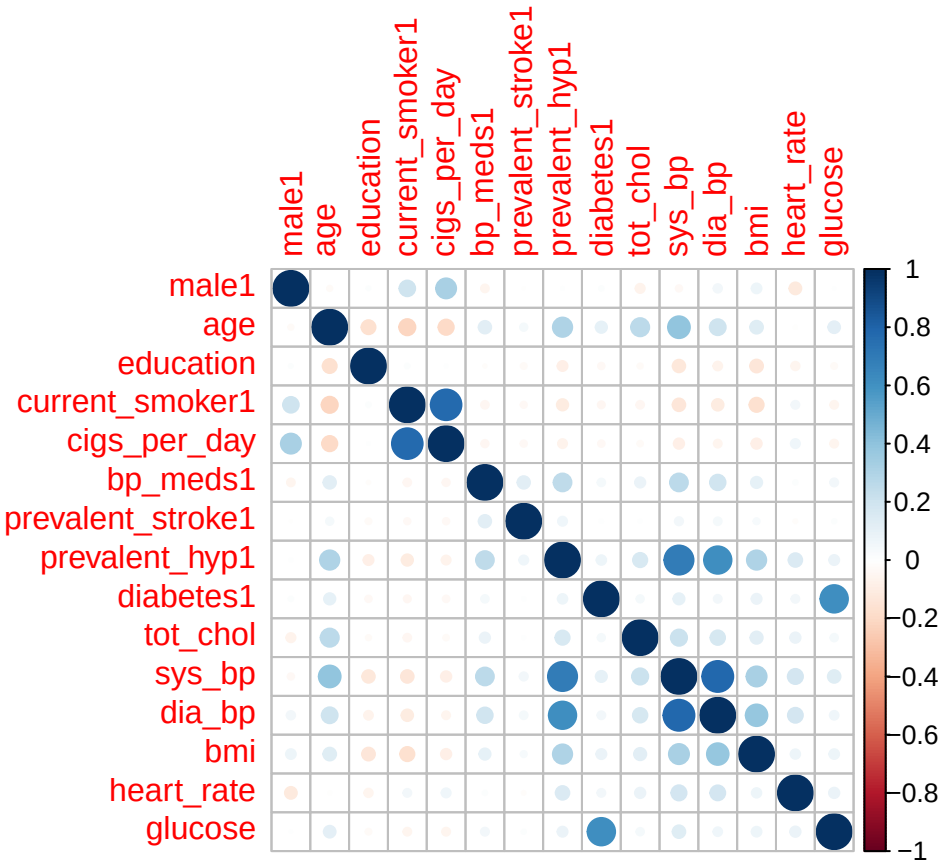
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
age	0	1	49.48	8.53	32.00	42.00	49.00	56.00	70.0
education	0	1	1.98	1.02	1.00	1.00	2.00	3.00	4.0
cigs_per_day	0	1	9.02	11.91	0.00	0.00	0.00	20.00	70.0
tot_chol	0	1	236.62	44.02	113.00	206.00	234.00	263.00	600.0
sys_bp	0	1	132.22	21.95	83.50	117.00	128.00	143.50	295.0
dia_bp	0	1	82.86	11.88	48.00	75.00	82.00	89.50	142.5
bmi	0	1	25.77	4.08	15.54	23.06	25.38	27.99	56.8
heart_rate	0	1	75.87	12.09	44.00	68.00	75.00	83.00	143.0
glucose	0	1	81.72	22.99	40.00	72.00	79.00	85.00	394.0

For nominal variables (binary), we generally have more people that do not have coronary heart disease in ten years than those who do for each medical history or condition, which is not very informative.



For continuous variables, an obvious observation is that those who have coronary heart disease in ten years tend to be of higher ages.

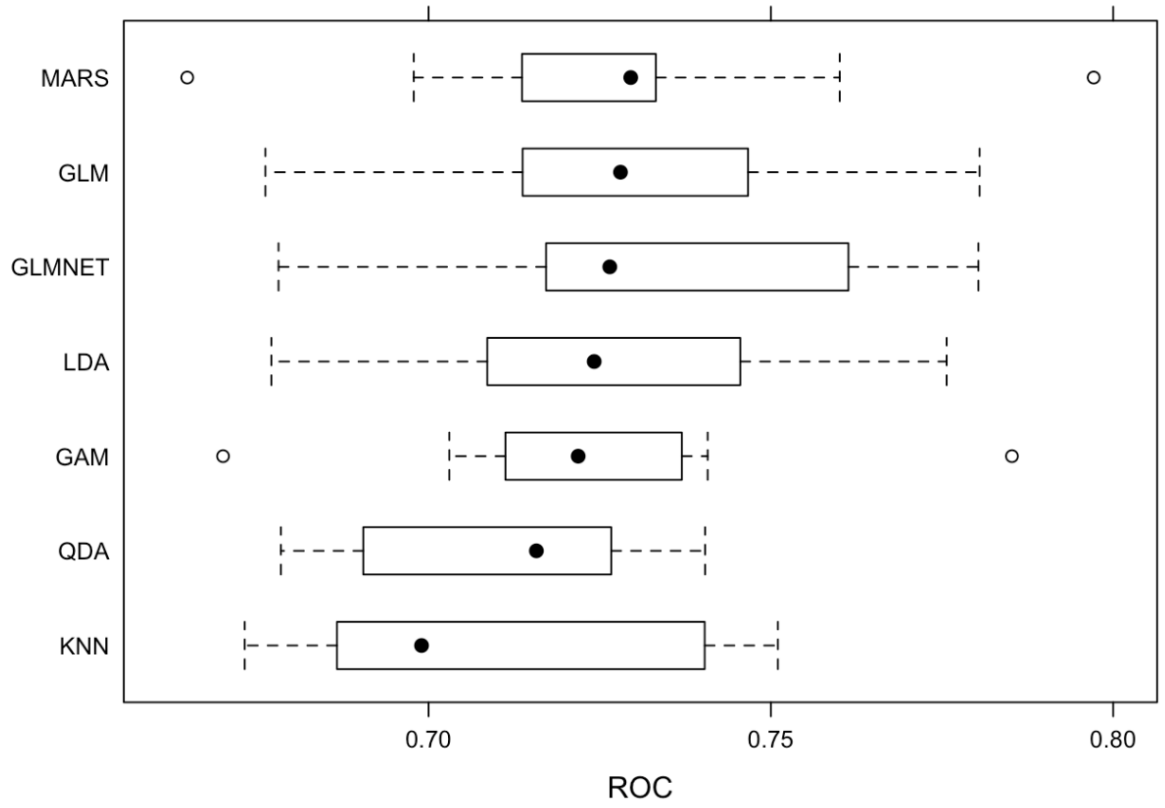


Some correlations are found: current smoker is positively correlated with the number of cigarettes smoked per day, systolic blood pressure/diastolic blood pressure is positively correlated with hypertensive history and glucose is positively correlated with diabetes history.

Models

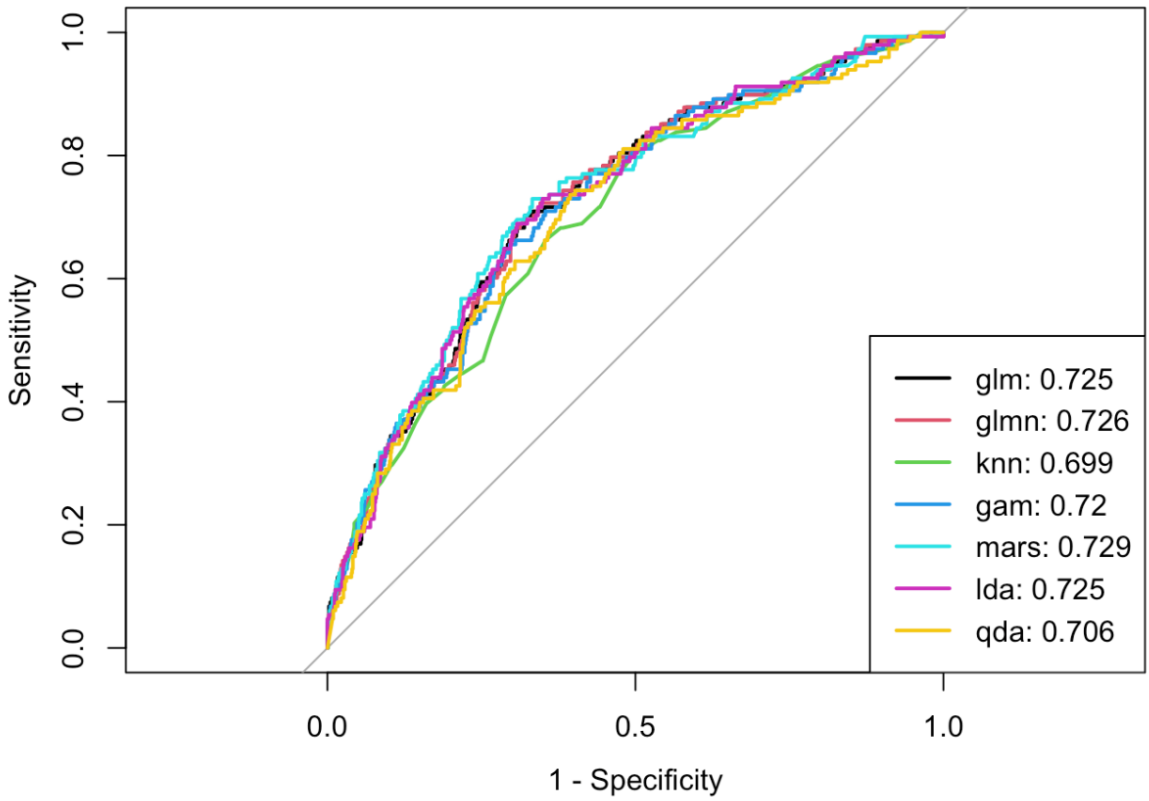
Models including Logistic Regression, Logistic Regression with Penalization, K-nearest Neighbors (KNN), Generalized Additive Model (GAM), Multivariate Adaptive Regression Splines (MARS), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are employed. All 15 predictor variables described in the introduction were used and the 10 year risk of coronary heart disease is the response variable. The `caret` package was employed for parameter tuning, model fitting and cross-validation all models in a consistent manner. The function `resamp()` was used for comparison of the seven models.

Training data performance:

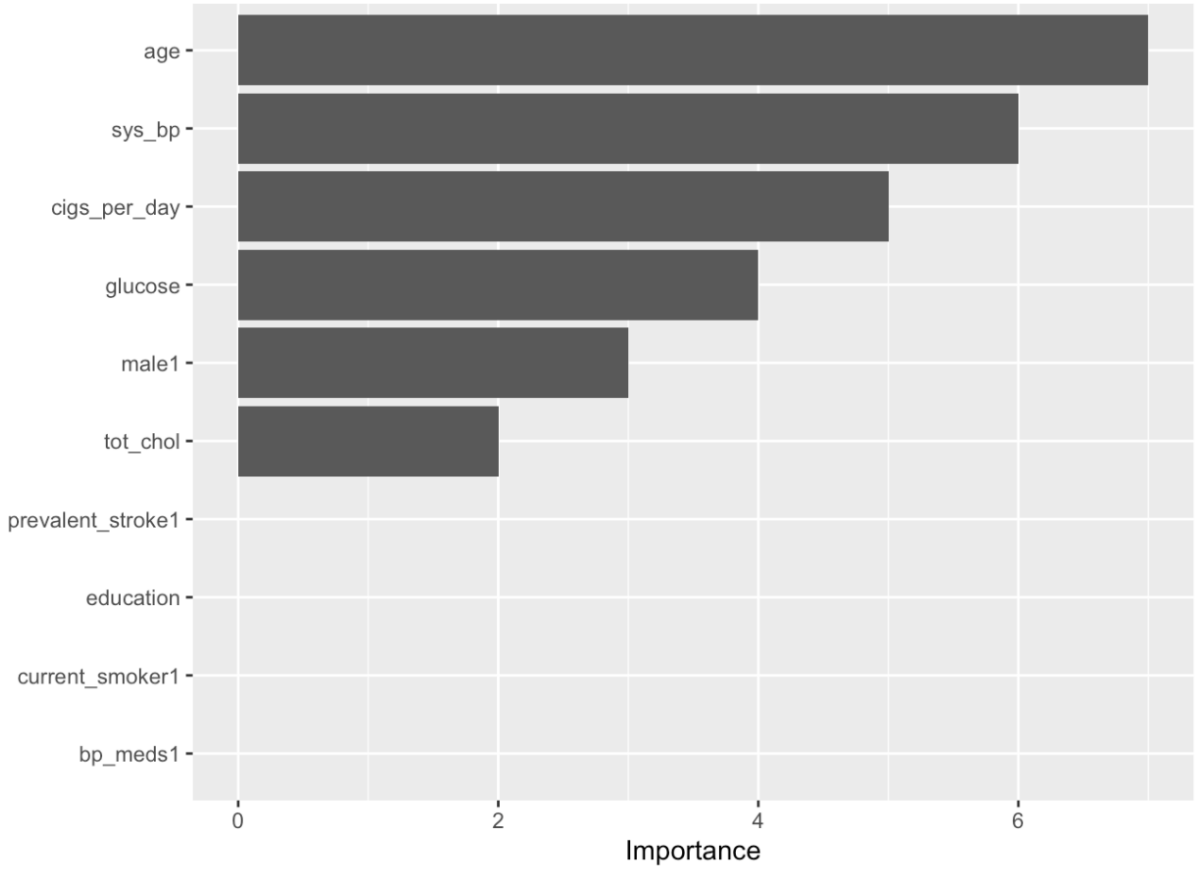


The box plot of Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve of the fitted models with training data shows that the MARS, GLM and GLMN model have relatively better performance among the seven models.

Test data performance:



The MARS model also gives the highest AUC of 0.729 among all the seven models in predicting with the test data. Thus, we decide that MARS is the best model in predicting the 10 year risk of coronary heart disease with the data.



Our best MARS model has degree=1 and nprune=7 based on corss-validation statistics. Among all of the predictors, age, systolic blood pressure, the number of cigarettes smoked per day are the three most important variables in predicting the 10 year risk of coronary heart disease. The details of the model are shown below. The GCV is 0.1158595, RSS 340.074, GRSQ 0.09708503 and QSq 0.1055206. Since MARS is a non-parametric method, there is no assumption to be considered. Limitations of MARS models include no missing data allowed, potential arbitrariness in variables selection as we have highly correlated variables and the resulting fitted function might not be very smooth and thus affect interpretability.

```
## Call: earth(x=matrix[2991,15], y=factor.object, keepxy=TRUE,
##             glm=list(family=function.object, maxit=100), degree=1, nprune=8)
##
## GLM coefficients
##                             yes
## (Intercept)        -0.64971124
## male1               0.46765484
## h(64-age)          -0.06911312
## h(age-64)           0.11503128
## h(40-cigs_per_day) -0.02528390
## h(tot_chol-382)     0.03933832
## h(sys_bp-130)       0.02191011
## h(glucose-89)       0.00871567
##
## GLM (family binomial, link logit):
##  nulldev   df       dev   df   devratio     AIC iters converged
##  2522.91 2990   2229.83 2983      0.116    2246     5         1
##
## Earth selected 8 of 24 terms, and 6 of 15 predictors (nprune=8)
## Termination condition: RSq changed by less than 0.001 at 24 terms
## Importance: age, sys_bp, cigs_per_day, glucose, male1, tot_chol, ...
## Number of terms at each degree of interaction: 1 7 (additive model)
## Earth GCV 0.1148495    RSS 340.078    GRSq 0.09708503    RSq 0.1055206
```

Conclusion

The MRAS model fits our data best and gives the most accurate prediction. It points out that age, systolic blood pressure, the number of cigarettes smoked per day, glucose level, sex and total cholesterol level are important risk factors of heart disease. The results confirms with common sense and literature records. People should be more attentive of their physiological index such as blood pressure and living habits such as smoking as they grow older in prevent of heart diseases.