Heart Disease Prediction with Linear methods

Tianran Zhang (tiz4001)

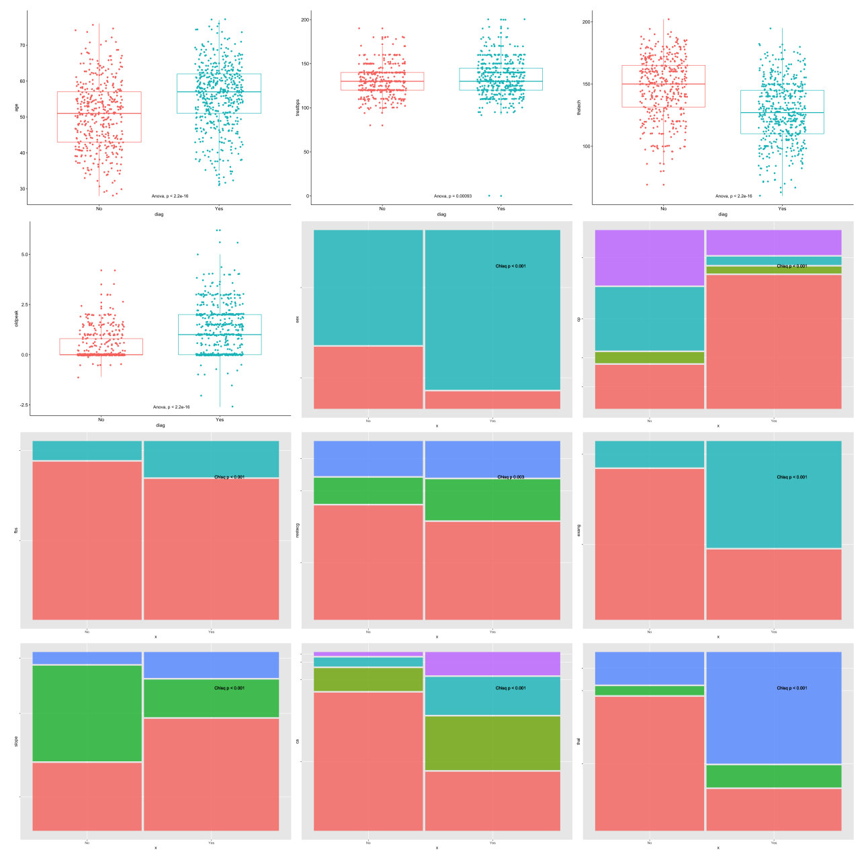
Weill Cornell Medicine

ABSTRACT

In the United States, cardiovascular diseases are the number 1 cause of death in adults. The final model has an overall 85% accuracy in detecting the likelihood of heart disease, and an overall 64% accuracy in predicting the severity of heart disease.

OBJECTIVE

Conduct one or more test out of 12 tests with a relatively low cost and use these results to detect the presence as well as the severity of CVD with a high accuracy.



METHOD

Generalized Linear Regression (GLM) for Binary data & Ordinary data

- **Binary outcome: logit link function**

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = \log\frac{\pi}{1-\pi} = \beta_{k0} + \beta_{k1}x_1 + \ \dots + \beta_{kp}x_p$$

- **Ordinary outcome: (k - 1) logit link functions**

$$\log\left(\frac{P(Y=k)}{P(Y=0)}\right) = \log\frac{\pi_k}{\pi_1} = \beta_{k0} + \beta_{k1}x_1 + \ \dots + \beta_{kp}x_p$$

Alternative methods

- LDA, SVM & KNN

Evaluate factor contributions in diagnose

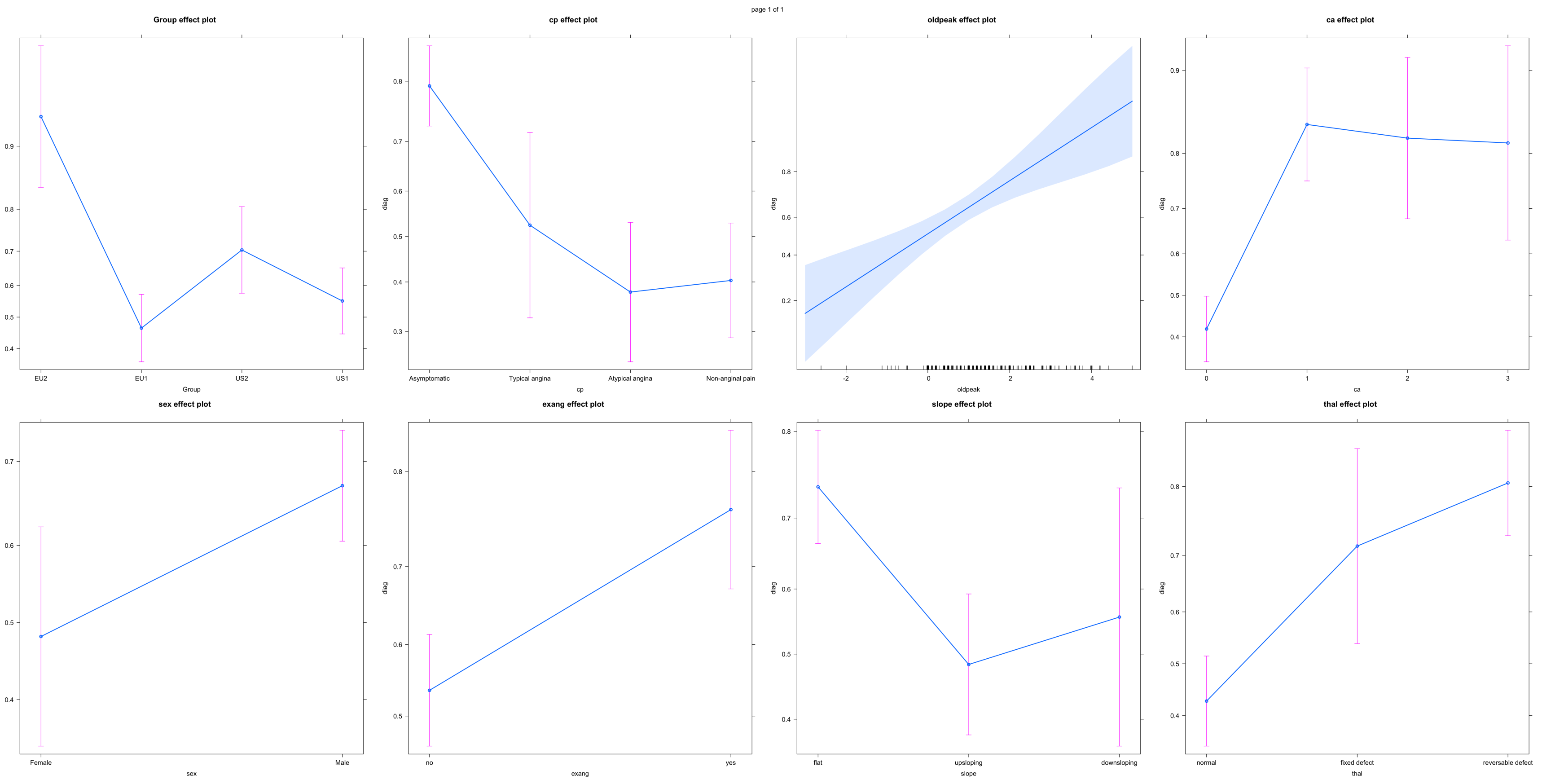
- PCA

PCA - Biplot
Dim1 (25.6%)
Dim2 (11.1%)
cos2

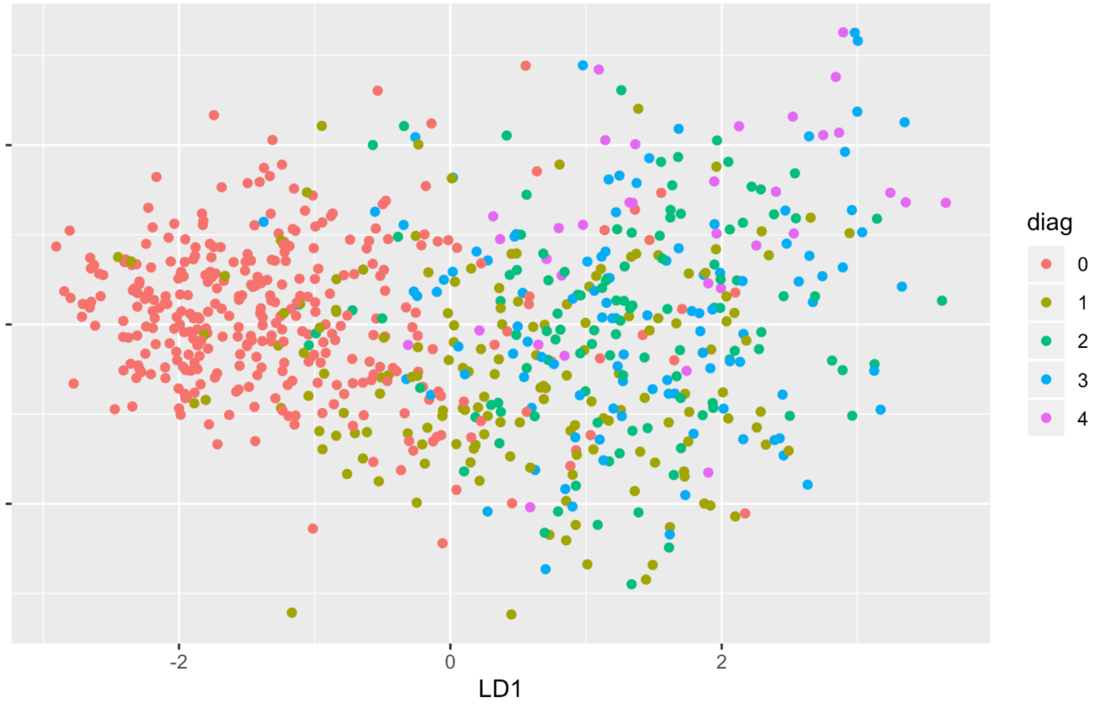
RESULT

The data is fitted with two models

Step 1. Detect the presence of the disease.

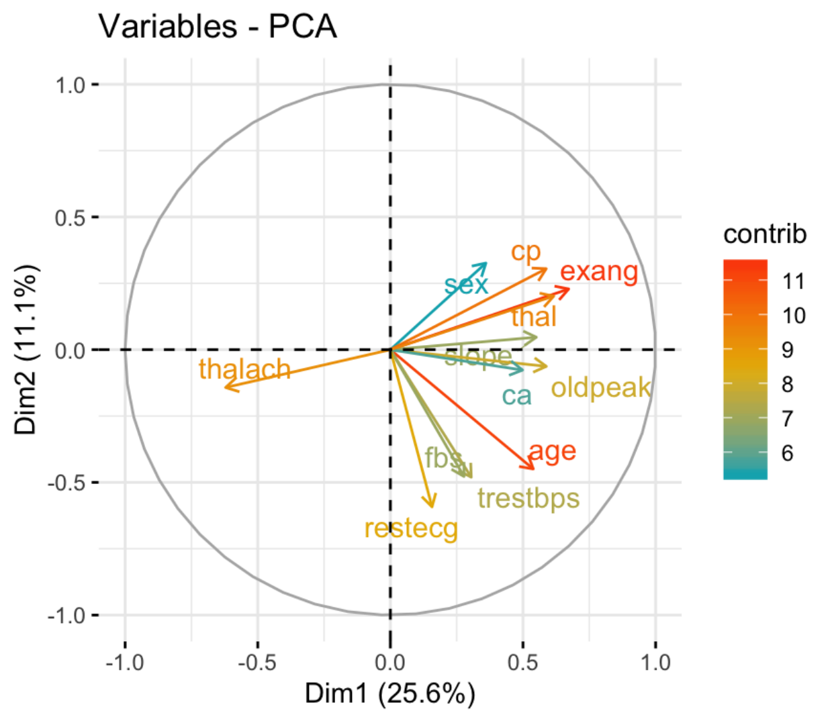


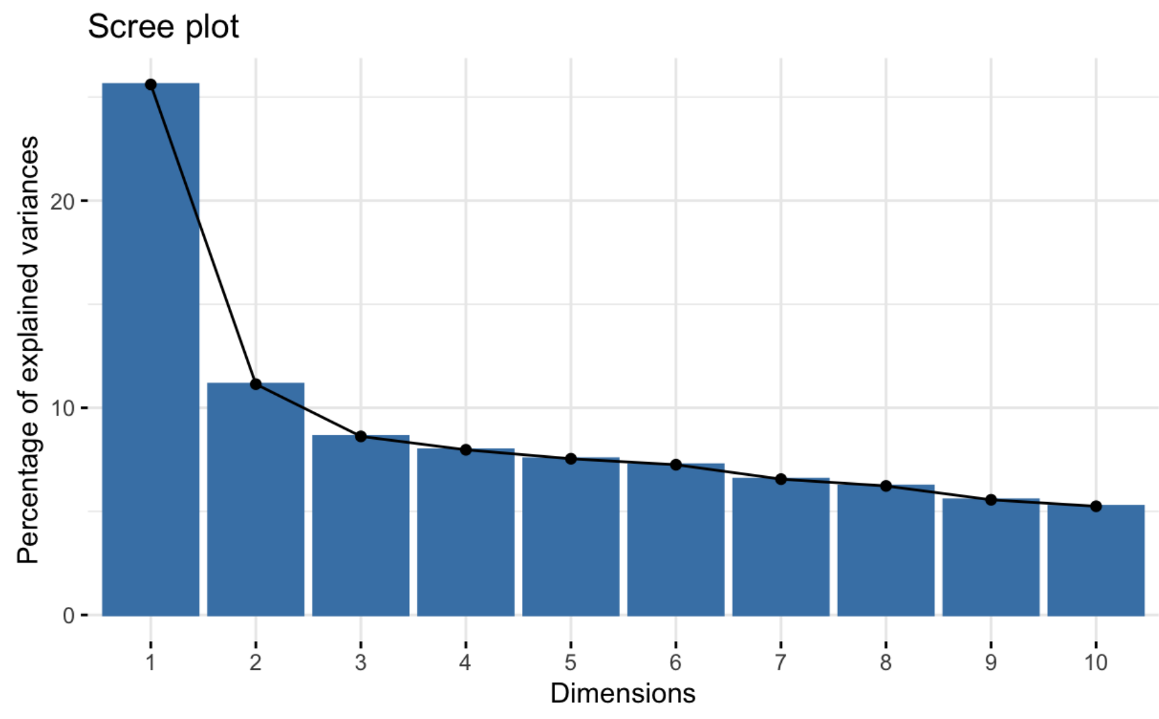
Step 2. Predict the severity of the heart disease.



Test Method	AUC	Accuracy Rate
Multinorm	0.7820	0.5870
Adjacent-category	0.7820	0.59870
LDA	0.8050	0.6410
SVM	0.7470	0.6030
KNN	0.7154	0.5272

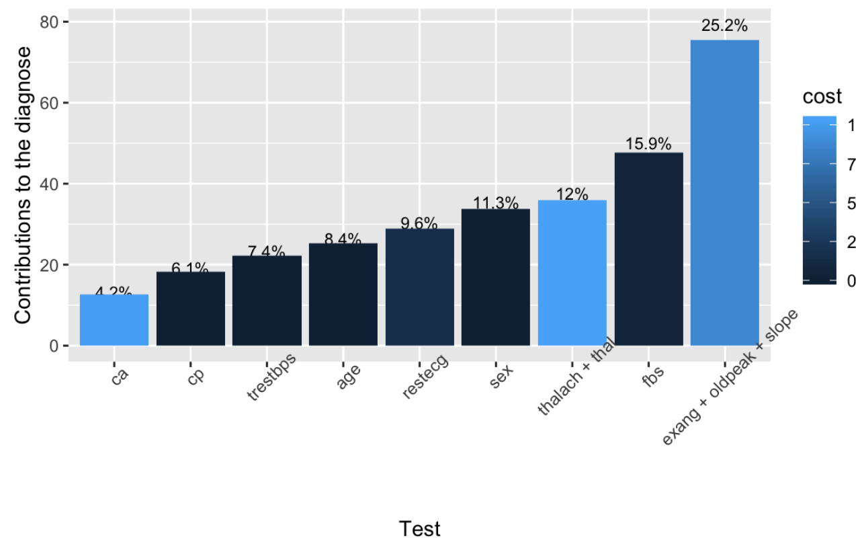
Step 3. Cost Consideration

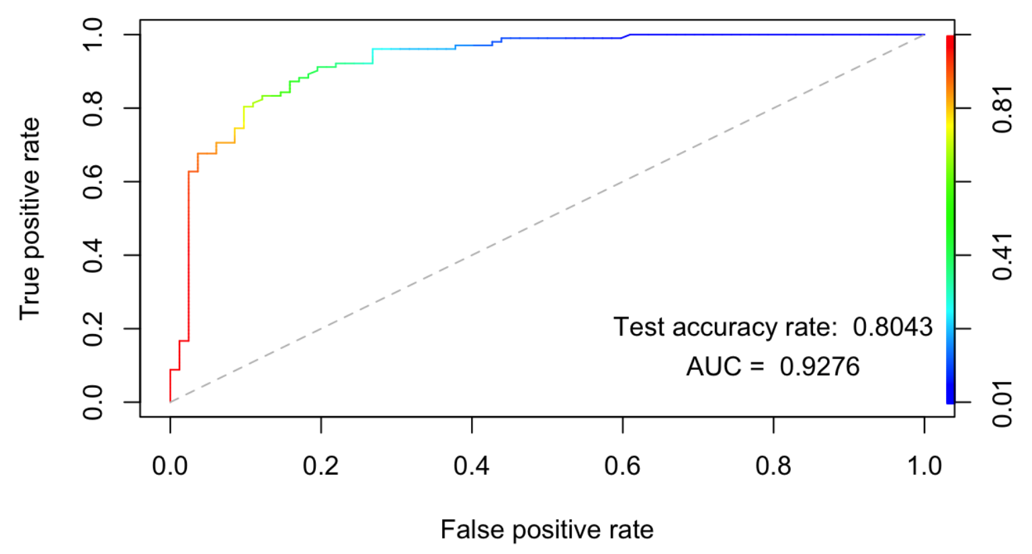




Selected predictors based on cost consideration:
exang + oldpeak + slope + cp + age + sex

Test	Result	Cost	Total Cost
cp	cp	$0	$87.3
exang or oldpeak or slope	exang oldpeak slope	$87.3	





Presence

Test Method	AUC	Accuracy Rate
Multinorm	0.6910	0.4800
LDA	0.7350	0.5200
SVM	0.7190	0.4870
KNN	0.6992	0.5018

Severity

CONCLUSION

Summary

	Without Cost Consideration		With Cost Consideration Total cost: $87.3	
Objective for CVD	Presence	Severity	Presence	Severity
Final Model	GLM	LDA	GLM	LDA
Significant Predictors	Group + sex + cp + exang + oldpeak + slope + ca + thal		exang + oldpeak + slope + cp + age + sex	
AC rate	84.78%	64.1%	80.43%	53.3%
AUC	92.76%	80.5%	85.40%	70.8%

Future Direction

- Adjust the model with more data
- Try non-linear methods

BIBLIOGRAPHY

1. Lecture Notes from Biostats II Chap 2.4.
2. LDA: Kassambara, A. (2018). *Machine Learning Essentials: Practical Guide in R*. sthda.
3. SVM: https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/
4. KNN: https://rpubs.com/maulikpatel/221668
5. PCA: Kassambara, A. (2017). *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra* (Vol. 2). STHDA.