P8106 Midterm Project

Shihui Zhu

Contents

I. Introduction

Motivation

Hepatitis C is a liver infectious disease caused by the hepatitis C virus (HCV) which is spread through inter-personal blood contact. The traditional approach of its diagnostic pathways are based on expert rules ("if... then... else") (Hoffmann), and the structure of which can be viewed as a decision tree. However, the diagnosis of HCV can be generally viewed as a classification problem and can be approached by multiple machine learning algorithms. Therefore, applying machine learning algorithms may help scientist to find potential new and automated diagnostic pathways. Since HCV is an infectious disease, an earlier diagnosis and treatment can help reduce its spread and induce better treatment outcome. In this project, multiple machine learning algorithms including regressions, discrimination, and tree methods are applied to the HCV dataset in order to determine which is the most efficient and accurate model for HCV diagnosis.

Research Question:

Which model is the best in diagnosing HCV disease (in an early stage) based on laboratory results?

Data Description

The dataset is an online dataset obtained from the *UCI Machine Learning Repository*, donated by Ralf Lichtinghagen, etc. al. It recorded the laboratory information from 615 blood donors and patients with HCV. The morphological pictures of the HCV patients ranged from chronic hepatitis C infection without fibrosis to end stage liver cirrhosis with a need for liver transplantation (LTX) (Hoffmann). There are total of 13 attributes in this dataset, including 12 continous and binary predictors, and one nominal outcome. The details of the variables are listed below:

- **Index**: Patient ID/No.
- **Category**: The categorical response variable, diagnosis ('0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis')
- **Age**: numerical, in years
- **Sex**: categorical(binary), sex (F = female, M = male)
- The other 10 numerical variables are biochemicals used for liver disease tests, albumin, bilirubin, choline esterase etc. al., abbreviated as **ALB**, **ALP**, **ALT**, **AST**, **BIL**, **CHE**, **CHOL**, **CREA**, **GGT**, and **PROT**.

Data Cleaning

The data is already prepared in .csv format. However, the dataset has 31 missing values and are considered to be missing-at-random (MAR). Due to the concern of the data size, we applied the bagging imputation to the original dataset to accommodate those missing values. Since our purpose is to predict early HCV diagnosis result, we recoded the nominal response variable (**Category**) in to a response variable of HCV patient(**Patient**) and non-HCV blood donors (**Donor**). The *tidyverse* package is used for data cleaning this step.

For training and testing purpose, the original data was randomly divded into two subsets: training set (75%) and the testing set (25%). The exact same training and testing set was used for the training of all models to ensure the reproducibility of the process.

II. Exploratory analysis/visualization

We applied both visualization and numerical analysis to get an overview of the preprocessed data.

Table 1: **Table 1. Summary of Dataset**

Variable	**Overall**, N = 615	**Donor**, N = 540	**Patient**, N = 75	**p-value**
Age	47 (39, 54)	47 (39, 54)	49 (40, 58)	0.2
Sex				0.076
F	238 (39%)	216 (40%)	22 (29%)	
M	377 (61%)	324 (60%)	53 (71%)	
ALB	42.0 (38.8, 45.2)	42.1 (39.1, 45.3)	40.0 (34.5, 44.0)	<0.001
ALP	66 (53, 79)	67 (55, 80)	53 (36, 69)	<0.001
ALT	23 (16, 33)	23 (17, 33)	15 (6, 41)	0.005
AST	26 (22, 33)	25 (21, 30)	69 (43, 112)	<0.001
BIL	7 (5, 11)	7 (5, 10)	14 (10, 27)	<0.001
CHE	8.26 (6.94, 9.59)	8.32 (7.08, 9.62)	7.10 (4.00, 9.34)	<0.001
CHOL	5.30 (4.62, 6.06)	5.40 (4.69, 6.17)	4.45 (3.77, 5.22)	<0.001
CREA	77 (67, 88)	78 (68, 89)	71 (61, 81)	0.003
GGT	23 (16, 40)	22 (15, 32)	66 (40, 120)	<0.001
PROT	72.2 (69.3, 75.4)	72.1 (69.3, 75.1)	72.8 (69.8, 78.2)	0.064

The mean and 95% CI intervals were listed for each continuous predictors, and the percentage distribution was listed for categorical variable. There are much more non-HCV donor than patient in this dataset (540 v.s. 75), so our dataset contains more negative cases than positive cases. Among patients, males are more than two times of females.

Statistical evidence is showed by the Wilcoxon rand sum test (for continous) and the Pearson's Chi-squared test (for categorical) since the following feature plots indicated that many of the continuous variables do not have a normal distribution. The p-value indicates that the distributions of continuous variables ALB, ALP, AST, BIL, CHE, CHOL, and GGT show greatest statistically important difference between the non-HCV blood donors and the HCV patients. This can be used as a reference for accessing the importance of predictors in our following analysis.

Then we looked at the distributions of each predictors, and accessed whether their distributions are differed by the response variable graphically. The following feature plots were generated for the all continuous vairables, **Age**, **ALB**, **ALP**, **ALT**, **AST**, **BIL**, **CHE**, **CHOL**, **CREA**, **GGT**, and **PROT**.

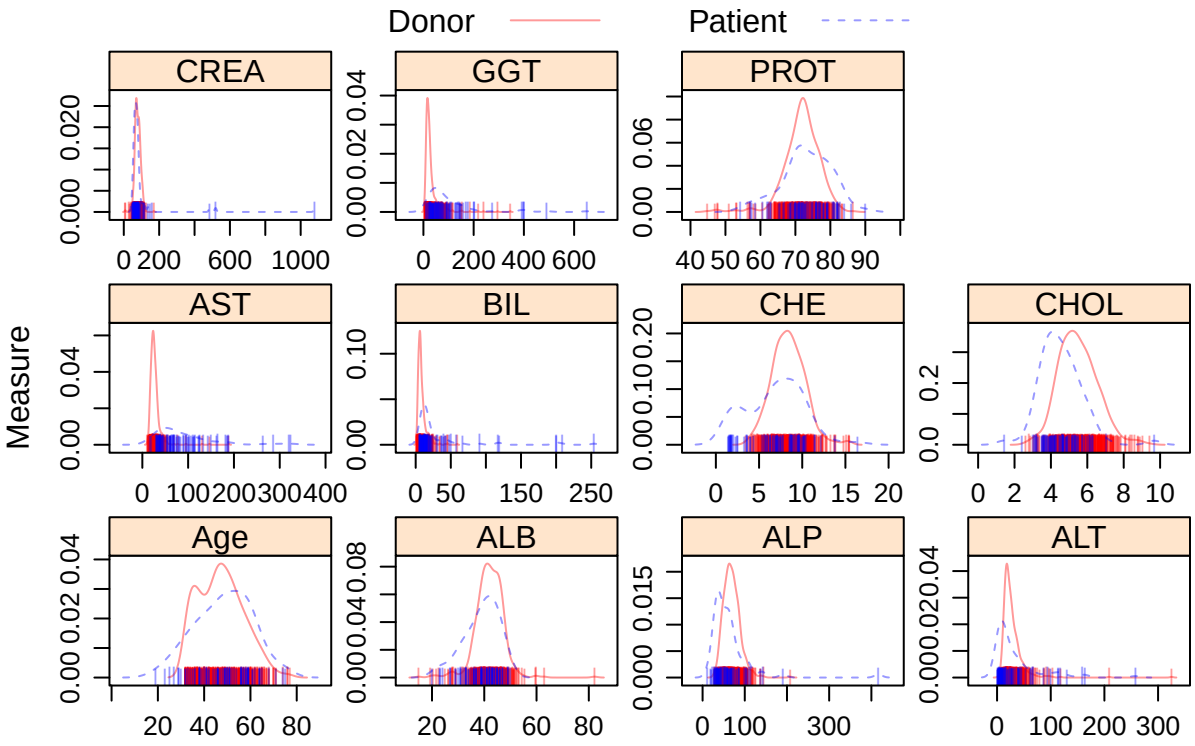


Figure 1. Density Plots

The figure echoed with our summary table above. As mentioned before, some of the numerical variables do not follow a normal distribution. For instance, the distribution of **Age** and **CHE** has two peaks. Also, the distributions of **CHOL** and **AST** showed the greatest difference between two categories, but other variables seem to have distributions that are similar between the two diagnosis groups. To summary, the dataset has shown some statistically important difference between the non-HCV blood donors and the HCV patients, but most of the differences is not visibly detectable among the predictors.

III. Models

Predictors

There are only 12 predictors so we used them all in model training. They are:

- **Age**: numerical, in years
- **Sex**: categorical(binary), sex (F = female, M = male)
- 10 numerical variables from laboratory results, **ALB**, **ALP**, **ALT**, **AST**, **BIL**, **CHE**, **CHOL**, **CREA**, **GGT**, and **PROT**.

Techniques

Since the traditional approach in HCV diagnostic pathways is related to decision tree, two tree methods were used: the conditional inference trees (CTREE) and regression tree (RPART). The tree methods can be displayed graphically and more easily understood by physicians. Taking the diagonsis as a classification problem, other machine learning models were also trained for the purpose our project. We used the generalized additive model model (GAM), the generalized linear regression models, GLMNET (with penalization)

and GLM (without penalization), linear and quadratic discriminant analysis models (LDA, QDA), as well as naive bayes (NB). All the models were trained using the package *caret*. The regression models (GLM, LDA, etc. al.) can accept mixture of variables which is suitable for our case. NB is useful when predictor number is large. GAM model can include any quadratically penalized GLM and a variety of other models, which induces great flexibility. Linear regression model also assumed the independence of the predictors.

We used 10-folds cross-validation for all model training. We didn't apply repeated cross-validation due to constrain of the computation resource.

Tuning parameters

GLMNET Model The tuning parameters is tested within the train function of the `caret` package. We tested on different ranges of regularization parameter λ and α. We looked for the point where the best cross-validated ROC AUC is obtained. The result is alpha = 0.85, and lambda = 0.044.

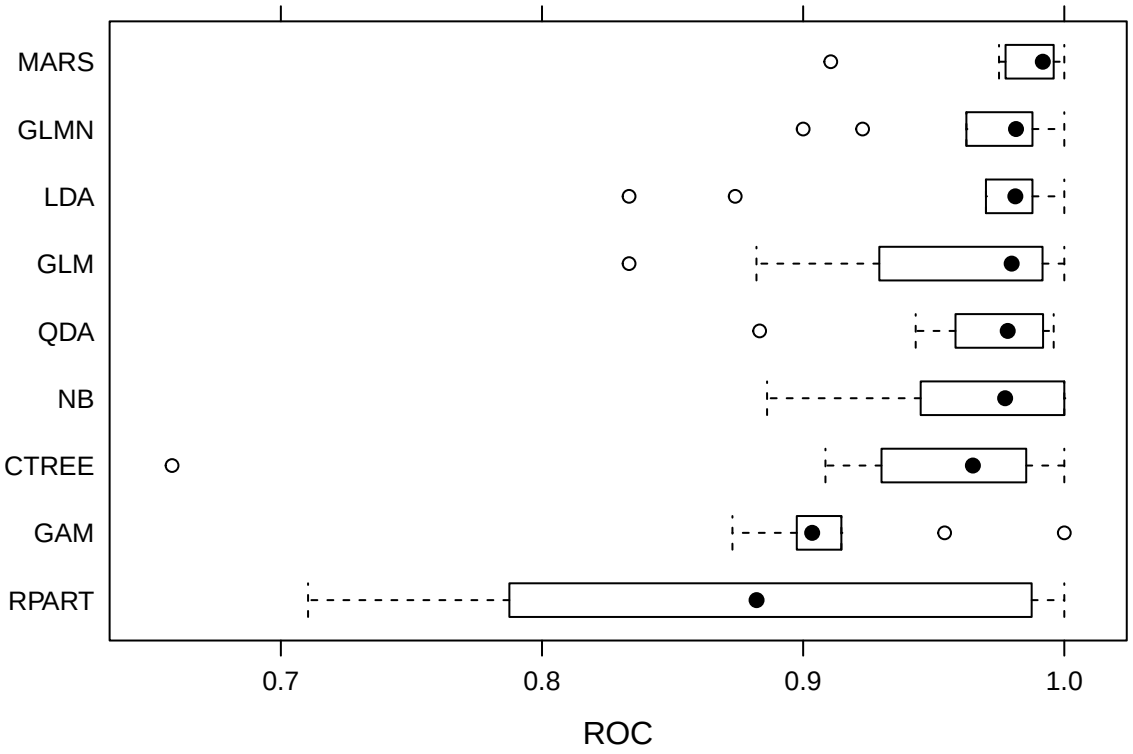
MARS Model MARS model can take a wide degree of features and number of terms. For simplicity, we only consider the performance of MARS in the first four degrees and all terms. The best tuned parameter is degree = 3 and nprune = 6.

NB and Tree models NB is trained by the Laplace correction parameter and the kernel density estimates. The CTREE is trained by minicriterion and RPART is trained by the complexity parameter (cp). The results are: Laplace correction(FL) = 1, adjust = 1.4 for NB model, minicriterion = 0.8199077 for CTREE model, and cp = 0.01947204 for the RPART model.

IV. Conclusions

Final Model

We address the performance of each models by the 10-folds cross-Validation ROC AUC scores based on the training set.

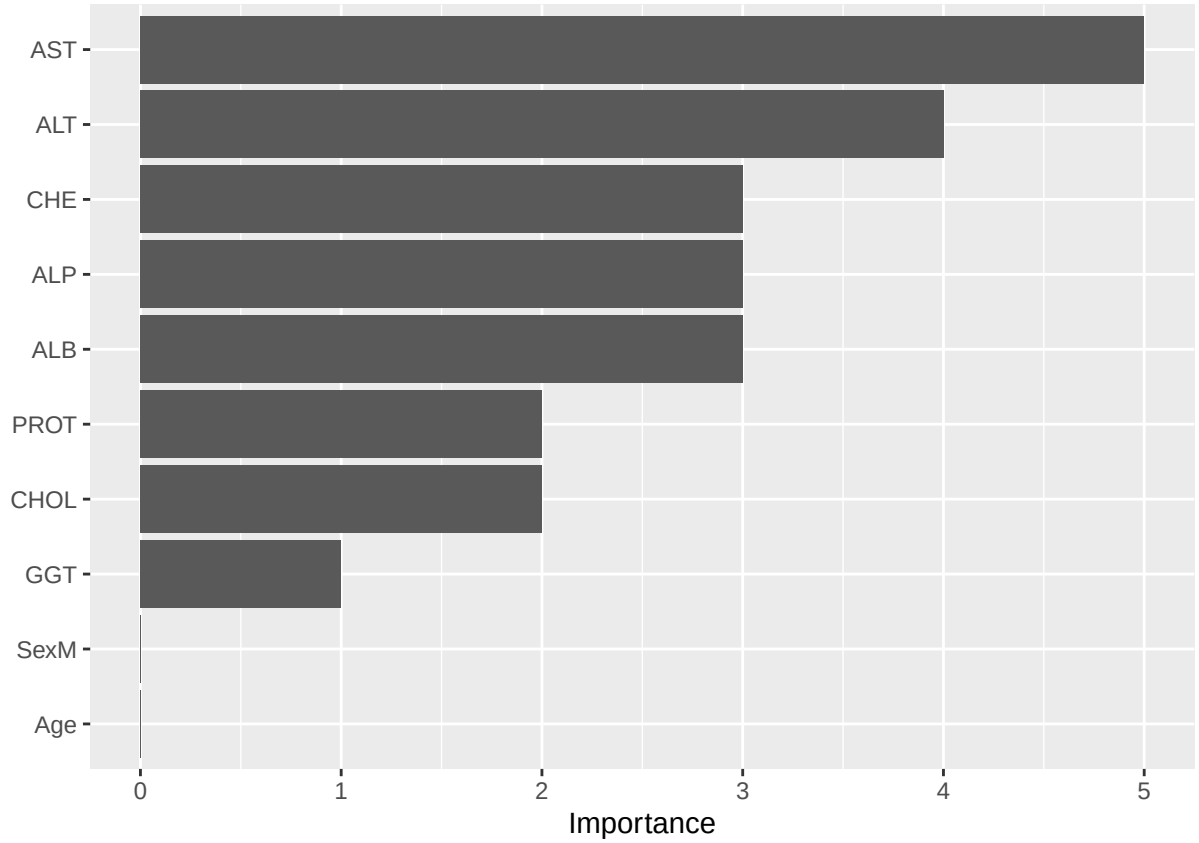


As shown by the figure above, all of the models perform generally well. The best model is the MARS model. It worthed notice that the tree methods were not outstanding comparing with the other models.

The coeffients and variables used by the MARS model is provided below:

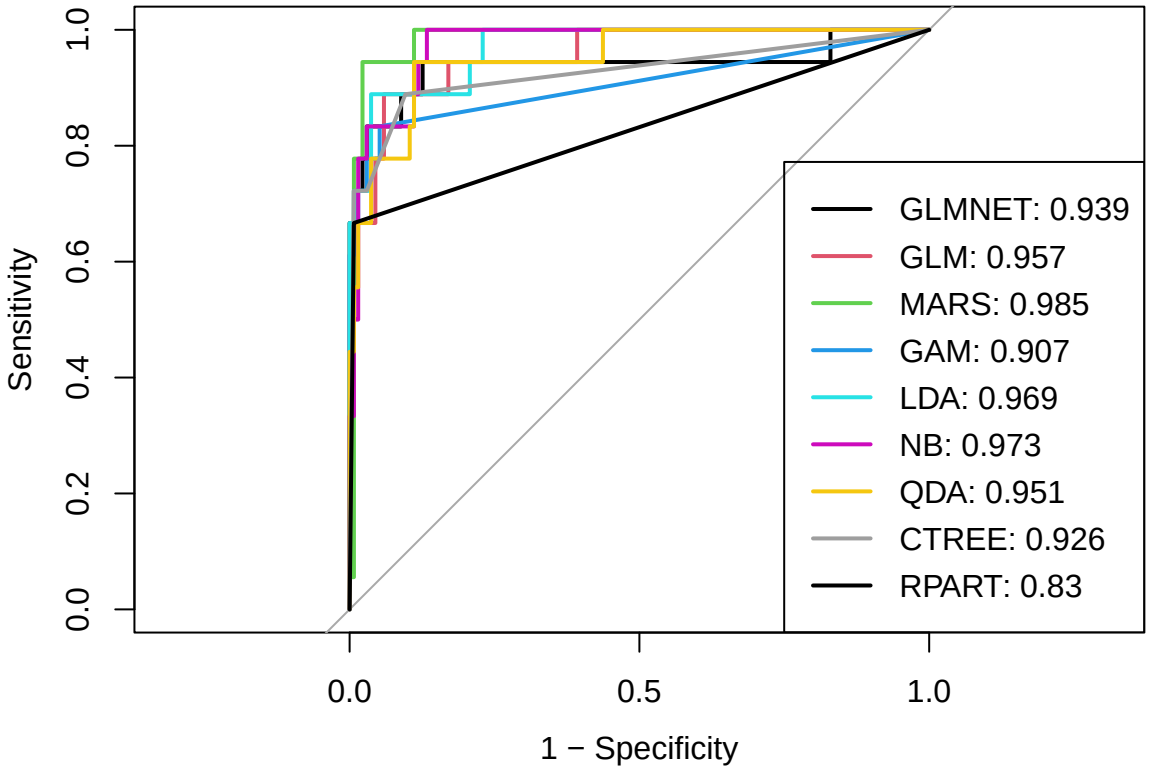
```
##                              (Intercept)                              h(77.2-AST)
##                               2.65321738                              -0.12336797
##                             h(18.2-ALT)                               h(138-GGT)
##                               0.42725242                              -0.02424103
## h(ALT-18.2) * h(4.88-CHOL) * h(PROT-72)    h(ALB-40.4) * h(45.8-ALP) * h(CHE-6)
##                               0.01752746                               0.03941077
```

The important variables are shown below:



From the figure above, we see that the variables **AST**, **ALT**, **CHE**, **ALP**, and **ALB** are the five most important features used in MARS. This also echoed with our hypothesis test at the beginning. The age and sex were not important and therefore not used at all in the model.

The following plots indicated the performance of models on testing set:



Not surprisingly, the MARS model has the best performance in prediction, and has a ROC AUC score as high as 0.98.

Therefore, we selected the MARS model to be our final model, and concluded it is the best in diagnosing HCV disease (in an early stage) based on laboratory results.

Limitation

- Our dataset is very unbalanced. The rare disease outcome made our model underestimate the potential effect of some predictors.
- Missing data: we assume data are missing-at-random but we never know whether this is the true scenario
- Model limitations: in terms of interpretability, MARS, GAM etc., al. are very limited comparing with tree-based method. They might appear to be confusing for physicians and hard to make sense biologically.

Citation

1. Georg Hoffmann, etc. al. "Using machine learning techniques to generate laboratory diagnostic pathways—a case study", *Journal of Laboratory and Precision Medicine.* https://jlpm.amegroups.com/article/view/4401/5424.