Homework 2: Smoothers, Generalized Additive Models, and Storytelling

Harvard CS 109B, Spring 2017

Nikhila Ravi

Feb 2017

Problem 1: Heart Disease Diagnosis

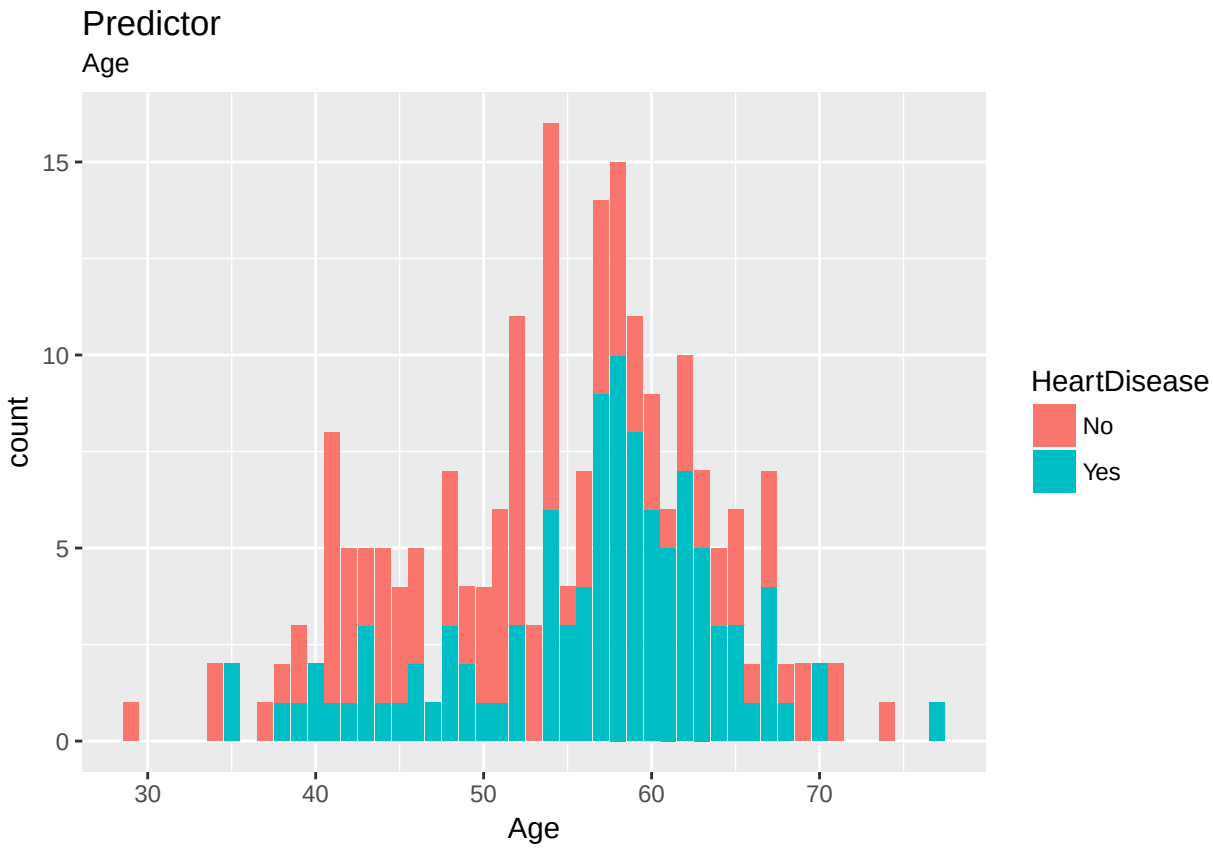
1. Inspect the data and the predictors

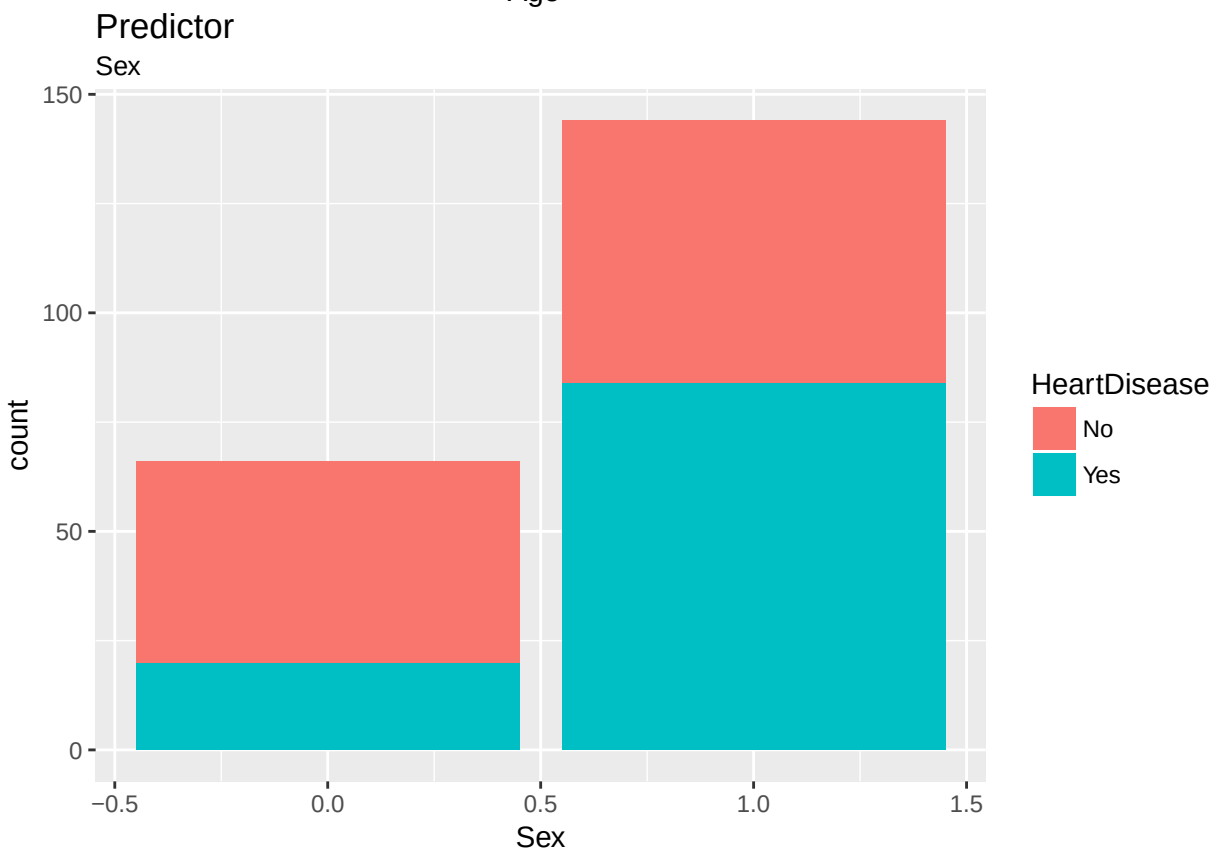
```
heart_train <- read.csv("./CS109b-hw2_q1_datasets/dataset_1_train.txt")
heart_test <- read.csv("./CS109b-hw2_q1_datasets/dataset_1_test.txt")
```

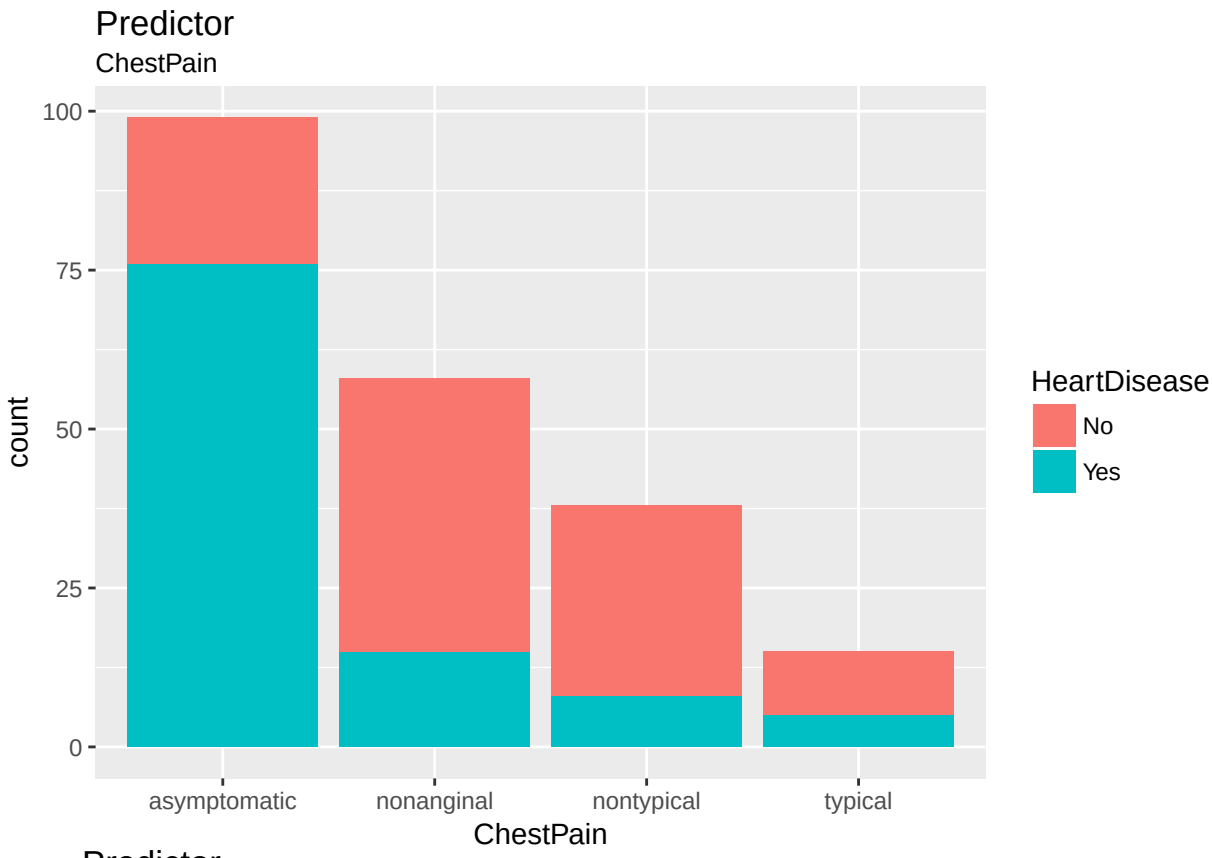
By visual inspection, do you find that the predictors are good indicators of heart disease in a patient?

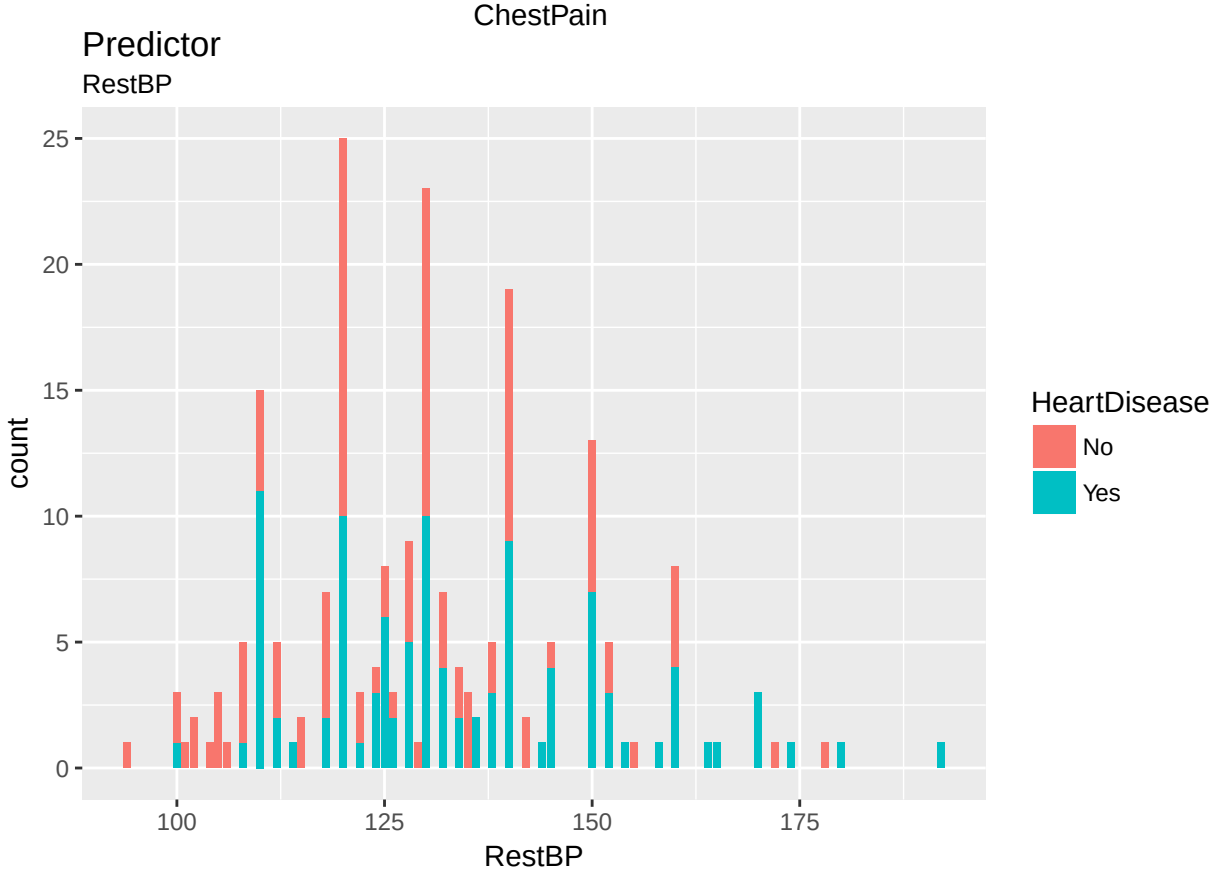
The predictors are all related to different aspects of a patients health. Let's plot each predictor separately with the outcome.

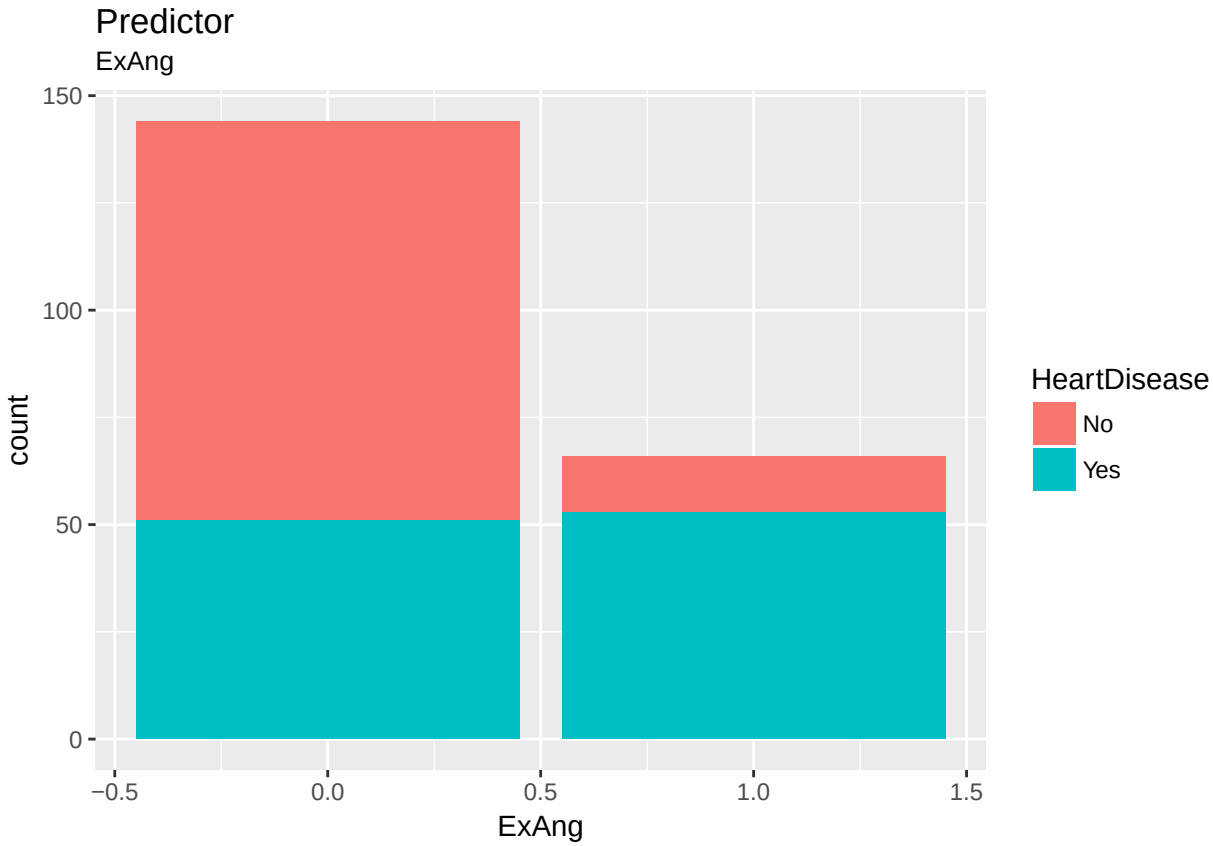
```
library(ggplot2)
for (pred in names(heart_train)) {
    if (pred != "HeartDisease") {
        print(ggplot(heart_train, aes_string(x = pred, fill = "HeartDisease")) +
            geom_bar() + ggtitle("Predictor ", pred))
    }
}
```

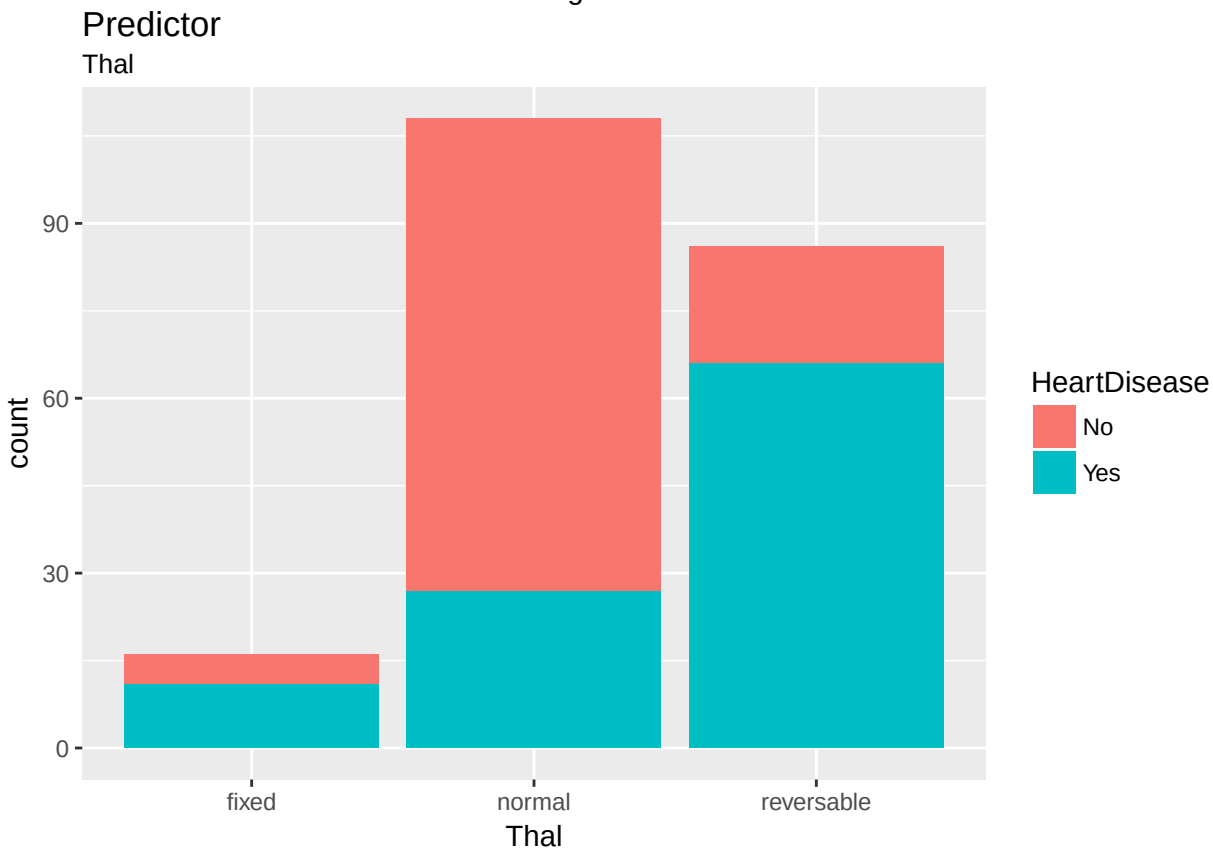

Predictor
Age
count
15
10
5
0
30
40
50
60
70
Age
HeartDisease
No
Yes


Predictor
Sex
count
150
100
50
0
−0.5
0.0
0.5
1.0
1.5
Sex
HeartDisease
No
Yes


Predictor
ChestPain
100
75
50
25
0
count
asymptomatic
nonanginal
nontypical
typical
ChestPain
HeartDisease
No
Yes


Predictor
RestBP
25
20
15
10
5
0
count
100
125
150
175
RestBP
HeartDisease
No
Yes





Examining the plots of the occurence of heart disease based on each predictor, there are a few predictors which

are good indicators of the incidence of heart disease. For example there appears to be greater proportion of heart disease patients among those aged between 55-65 (although there aren't equal numbers of people in each age group so this may be misleading). The occurence of asymptomatic chest pain is a clear indicator of heart disease with 75% of patients reporting asymptomatic chest pain also having heart disease. The presence of ExAng is also a clear indicator of heart disease with approximatley 80% of patients who report ExAng (exercise induced angina) also having heart disease. Finally the presence of reversable Thal (thallium scan) is also a clear indicator of heart disease with approx 86% of patients with reversable Thal also reporting Heart Disease.

Apply the generalized additive model (GAM) method to fit a binary classification model to the training set and report its classification accuracy on the test set.

```
# helper function for cross validation
library(boot)

# Function to compute k-fold cross-validation accuracy for a
# given classification model
cv_accuracy = function(model, data, k) {
    # Input: 'model' - a fitted classification model 'data' -
    # data frame with training data set used to fit the model 'k'
    # - number of folds for CV Output: 'cv_accuracy' -
    # cross-validation accuracy for the model

    acc <- 1 - cv.glm(data, model, K = k)$delta[1]
    return(acc)
}
```

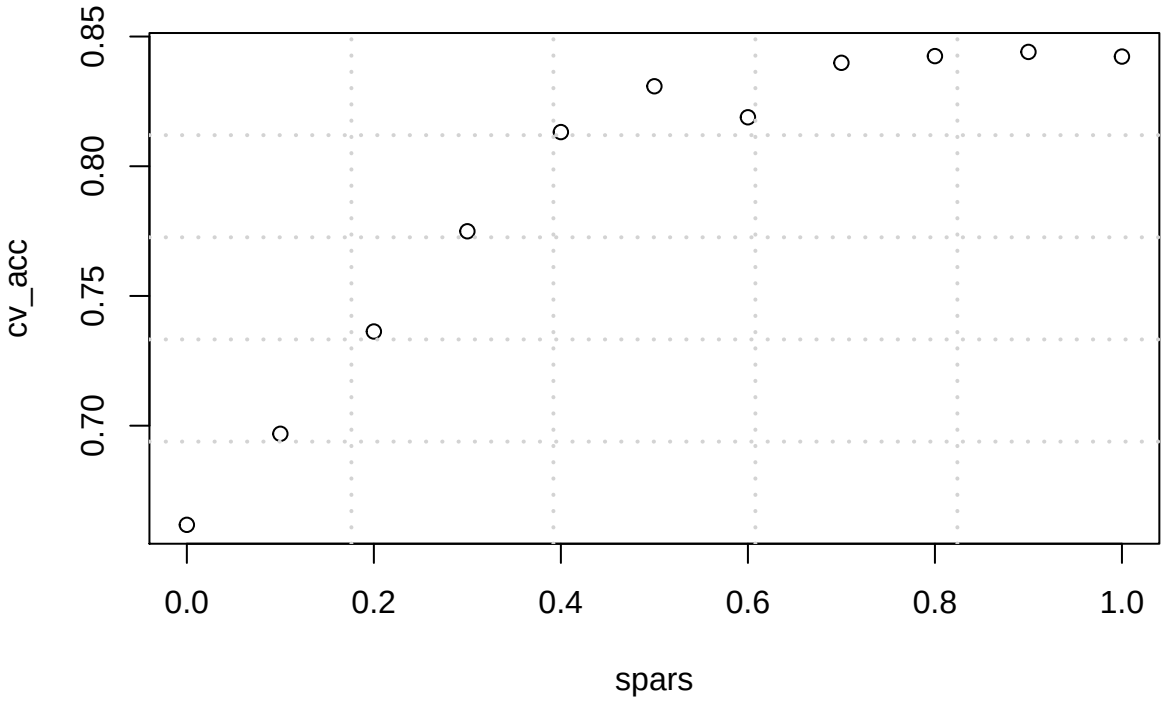
A smoothing spline can be fit to all the quantitative predictors. Unlike in sklearn, in R, the qualitative predictors can be left as they are - they will automatically be converted into dummy variables for each category.

For categorial variables, smoothing splines are non appropriate - nonparametric regression splines cannot handle the presence of categorical predictors without resorting to sample-splitting which can result in a substantial loss in efficiency (Ma, S., Racine, J., and Yang, L. (2014) Spline Regression in the Presence of Categorical Predictors. Journal of Applied Econometrics).

For the numerical predictors, the optimal value of the spar parameter can be determined using five fold cross validation. As the response is a binary outcome, the accuracy of prediction should be used to select the spar value instead of R^2.

```
library(gam)
# spans
spars <- seq(0, 1, by = 0.1)
num_param = length(spars)
cv_acc = rep(0, num_param)

# iterate through spans
for (i in 1:num_param) {
    heart.gam <- gam(HeartDisease ~ s(Age, spar = spars[i]) +
        s(RestBP, spar = spars[i]) + Sex + ChestPain + ExAng +
        Thal, data = heart_train, family = binomial(link = "logit"))
    cv_acc[i] <- cv_accuracy(heart.gam, heart_train, 5)
}
```

```
plot(spars, cv_acc)
grid(5, 5, lwd = 2)
```



From the plot of accuracy vs spar, it can be seen that the highest accuracy of approx 0.83 can be achieved with the optimal spar value.

```
best_spar = spars[which.max(cv_acc)]
cat("Span at which CV accuracy is maximised", best_spar)
```

```
## Span at which CV accuracy is maximised 0.9
```

This value of spar can be used to fit the GAM model.

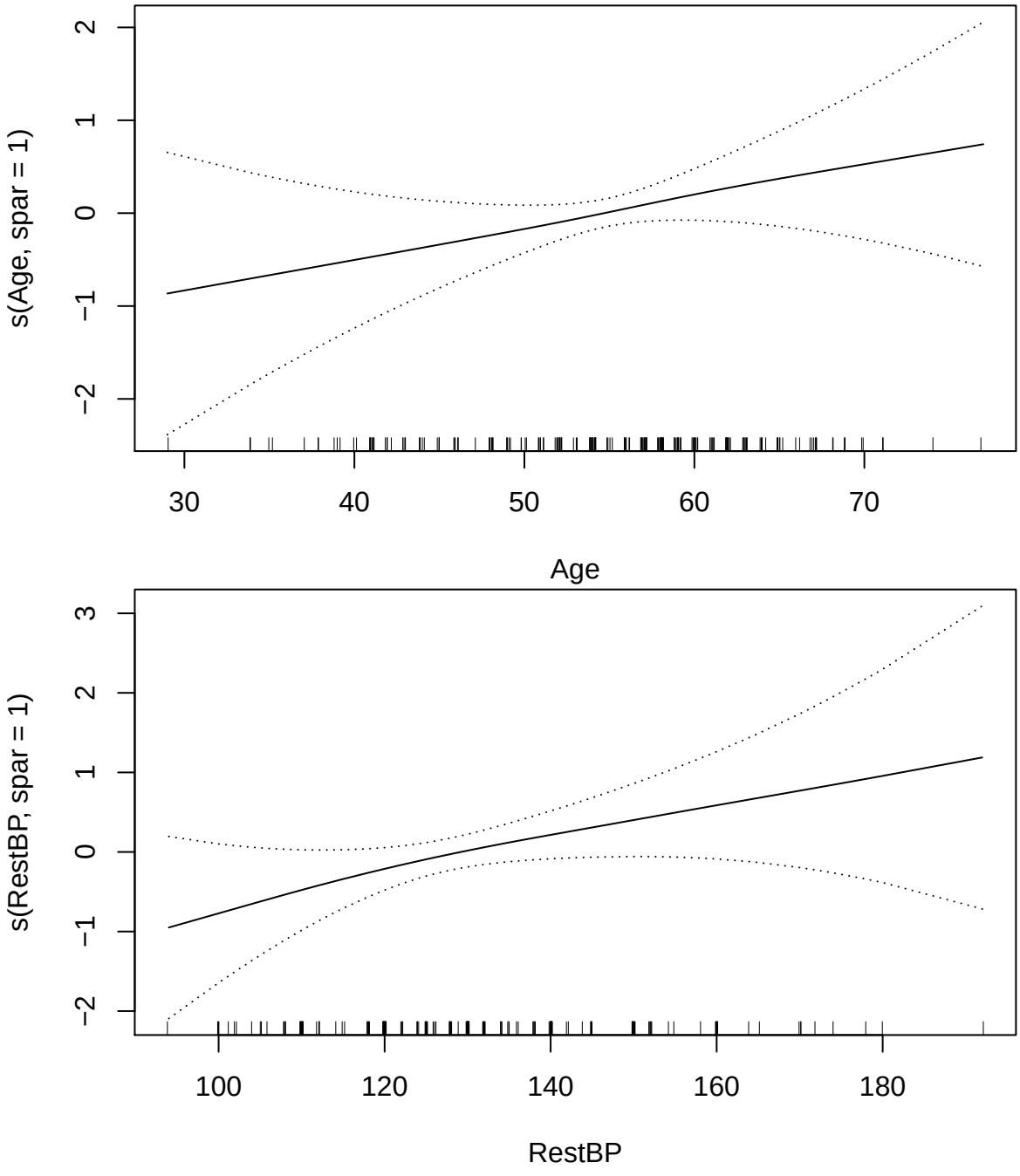
```
heart.gam <- gam(HeartDisease ~ s(Age, spar = 1) + s(RestBP,
    spar = 1) + Sex + ChestPain + ExAng + Thal, data = heart_train,
    family = binomial(link = "logit"))
accuracy <- function(model, test, column) {
    y <- predict(model, test)
    ybar <- c(y > 0)
    ytest <- c(test["HeartDisease"] == "Yes")
    compare <- c(ybar == ytest)
    return(sum(compare)/length(compare))
}

cat("accuracy on the test set", accuracy(heart.gam, heart_test,
    HeartDisease))
```

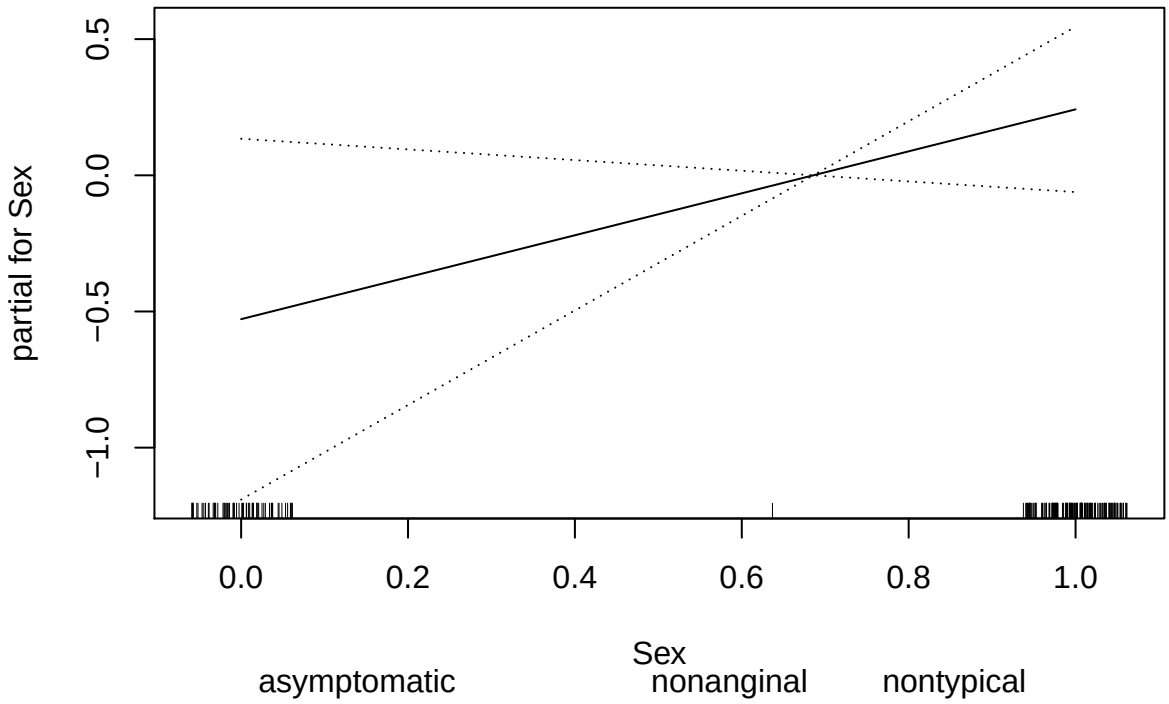
```
## accuracy on the test set 0.8131868
```

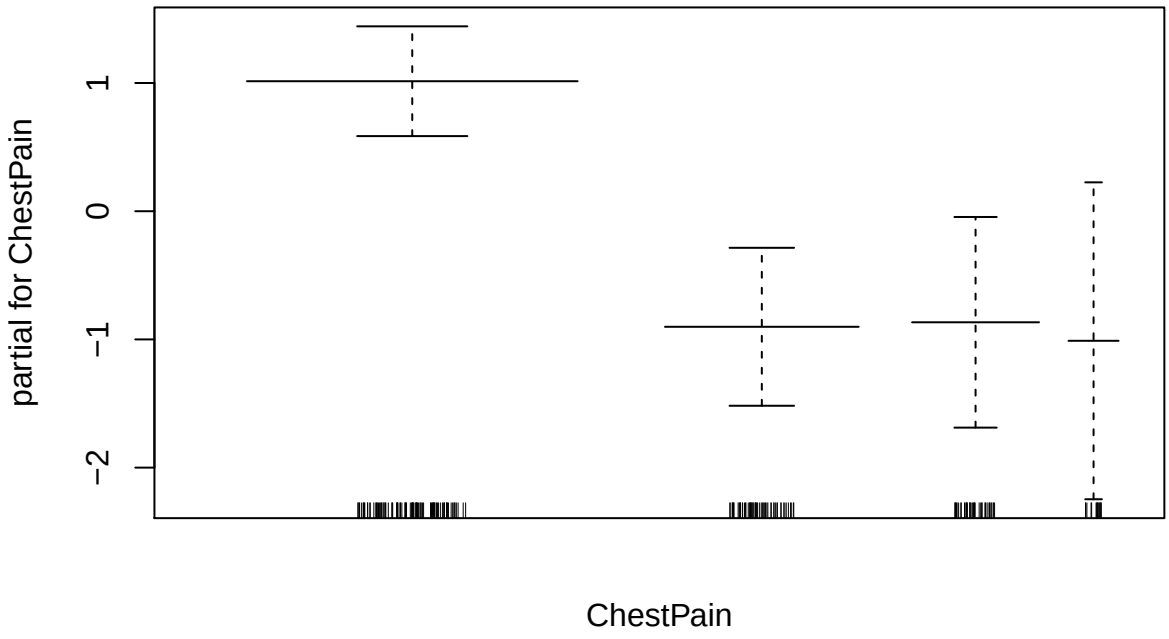
Do you find any benefit in modeling the numerical predictors using smoothing splines?

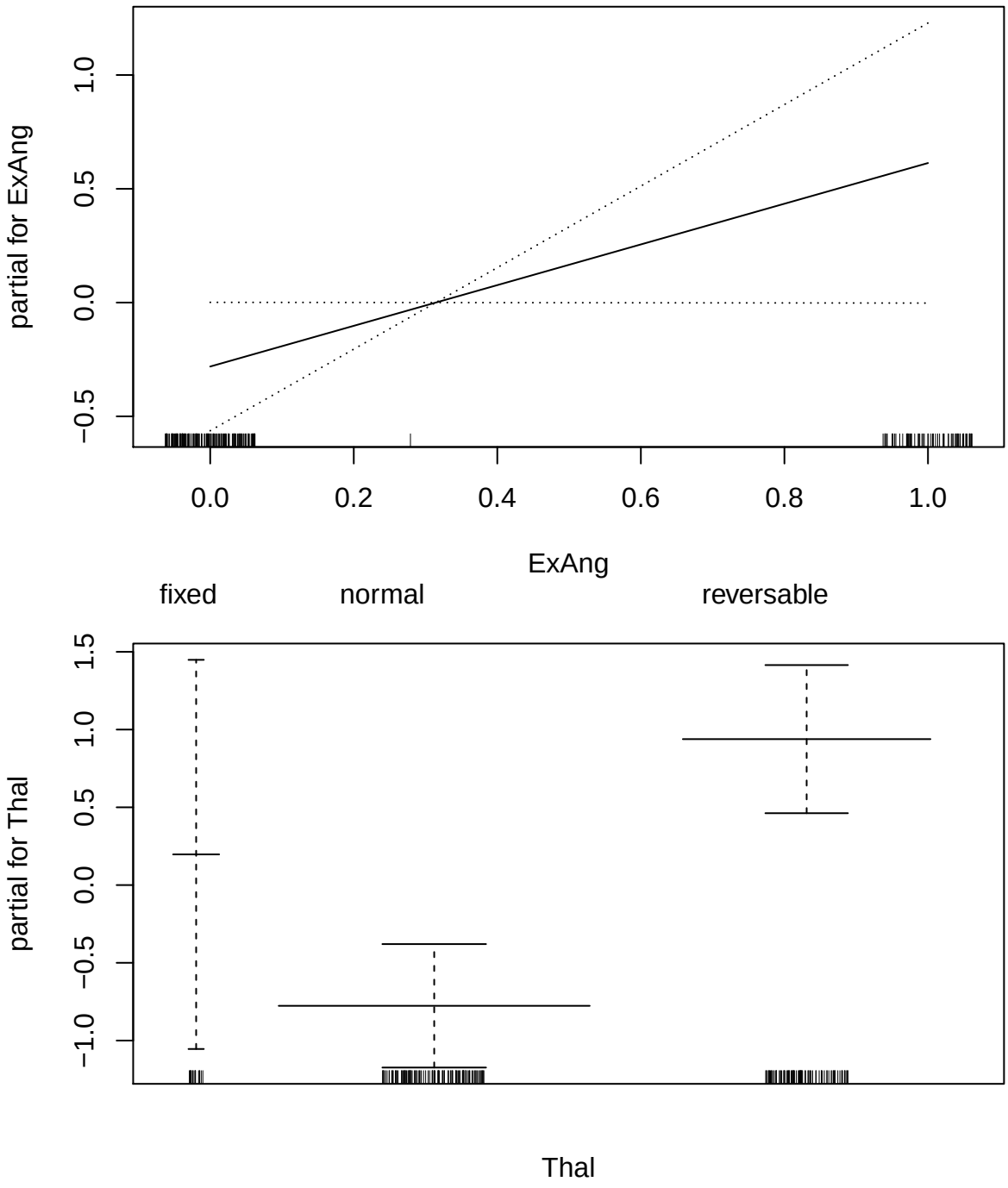
The smooth of each predictor in the model can be visualised.

```
plot.gam(heart.gam, se = TRUE)
```




partial for Sex
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0
Sex
asymptomatic
nonanginal
nontypical


partial for ChestPain
1
0
−1
−2
ChestPain



Smoothing splines for numerical predictors are useful because they allow non linear transformations of the predictor. However, for each of the two numerical predictors in this question, the result after transformation is a linear function of the predictor and hence there is no benefit in modelling the numerical predictors using smoothing splines.

```
heart.gam$coefficients
```

```
##         (Intercept)    s(Age, spar = 1) s(RestBP, spar = 1)
##         -4.30211973          0.03504879          0.02125104
##                 Sex ChestPainnonanginal ChestPainnontypical
##          0.77032145         -1.91529679         -1.88028711
```

```
##    ChestPaintypical               ExAng          Thalnormal
##         -2.02455513          0.89420797         -0.97366549
##      Thalreversable
##          0.74137489
```

Compare the fitted GAM with the following models: (i) a GAM with only the intercept term; (ii) a GAM with only categorical predictors; and (iii) a GAM with all predictors entered linearly.

```
heart.gam_intercept <- gam(HeartDisease ~ 1, family = binomial(link = "logit"),
    heart_train)
heart.gam_categorical <- gam(HeartDisease ~ Sex + ChestPain +
    ExAng + Thal, family = binomial(link = "logit"), heart_train)
heart.gam_linear <- gam(HeartDisease ~ Age + RestBP + Sex + ChestPain +
    ExAng + Thal, family = binomial(link = "logit"), heart_train)

anova(heart.gam_intercept, heart.gam, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: HeartDisease ~ 1
## Model 2: HeartDisease ~ s(Age, spar = 1) + s(RestBP, spar = 1) + Sex +
##     ChestPain + ExAng + Thal
##   Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
## 1    209.00     291.10
## 2    199.26     181.08 9.7418   110.02 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(heart.gam_categorical, heart.gam, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: HeartDisease ~ Sex + ChestPain + ExAng + Thal
## Model 2: HeartDisease ~ s(Age, spar = 1) + s(RestBP, spar = 1) + Sex +
##     ChestPain + ExAng + Thal
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    202.00     189.75
## 2    199.26     181.08 2.7418   8.6727  0.02729 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(heart.gam_linear, heart.gam, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: HeartDisease ~ Age + RestBP + Sex + ChestPain + ExAng + Thal
## Model 2: HeartDisease ~ s(Age, spar = 1) + s(RestBP, spar = 1) + Sex +
##     ChestPain + ExAng + Thal
##   Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
## 1    200.00     181.62
## 2    199.26     181.08 0.74182  0.53772   0.3557
```

Comparing the results of the likelihood ratio test the following conclusions can be drawn: The GAM model

fitted in the first part of the question (with smoothing splines for numerical predictors and all categorical predictors): - is better than a GAM model with only an intercept term at a significance level of 0 - is better than a GAM model with only categorical predictors at a significance level of 0.01 - is worse than a GAM model with all predictors entered linearly

This analysis shows that the linear combination of categorial and quanitiative predictors results in a model which is superior to models which only include a subset of these predictors, and the addition of smoothing splines provides no improvement in the modelling of numerical predictors.