RESULTS BY MODELS :

**WITH ALL THE PREDICTORS :**

Linear regression :

MSE = 0.1540146

Logistic regression :

MSE = 0. 1540146

Cross Validation =

- for K = 10 : 0.1341752

- for leave-one-out (more precise) : 0.1319878

Polynomial regression + splines : *later in the tests*

Tree-based method :

«trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches seen »

BUT we’ll use it to have a general impression and view of the situation.

Classification tree :

Prunning tree :  
«  A smaller tree with fewer splits might lead to lower variance and better interpretation at the cost of a little bias.»

Size = 9

-> not do again after best subset selection : already « blurry » enough

**AFTER BEST SUBSET SELECTION :**

**Subset of predictor selected :**

BIC : 7 variables

*image*

Cp : 9 variables

*image*

Adjusted R square : 10 variables

*image*

*Therefore, the determination of which model of a given size is best must be made using only the training observations .*

MSE : We find thhe best model has 8 variables.

image

CROSS VALIDATION :

We find the best model has 11 variables, with cv = 0.1294388

However we can notice that the model with 7 variables has a cv = 0.1305488, which is very close to the previous one : difference = 0.00111.

Indeed, having 7 variables is more comfortables than having 11.

*image*

Non-linear regression :

Polynomial regression + splines :