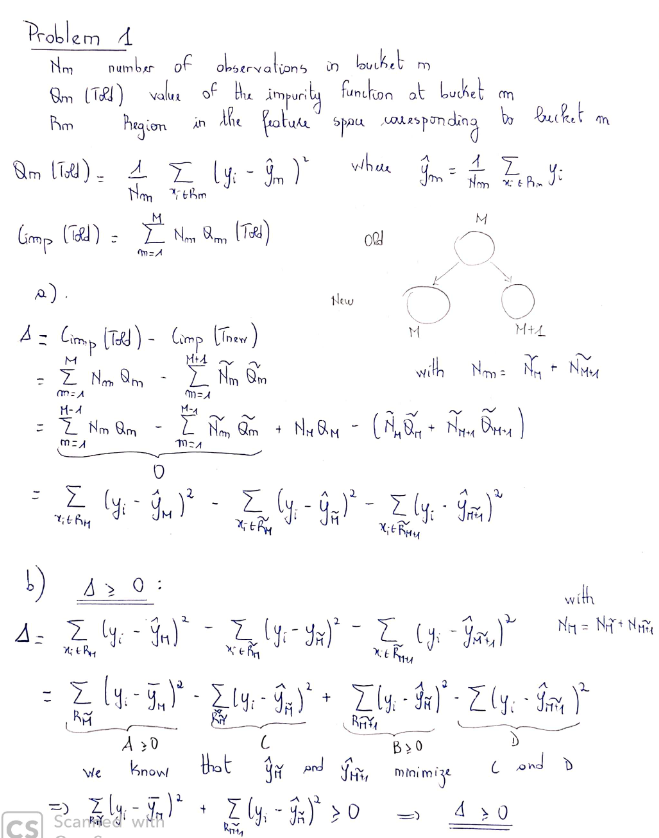
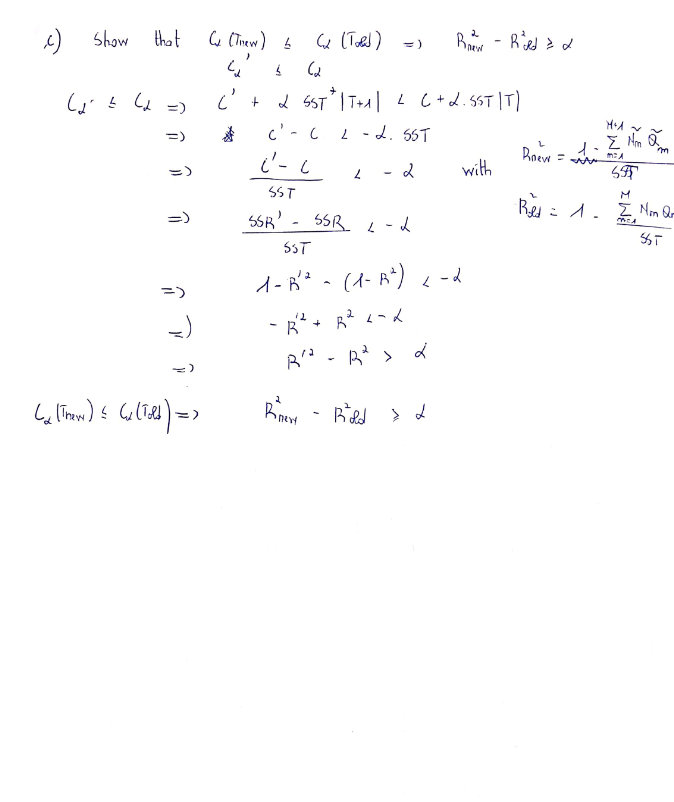
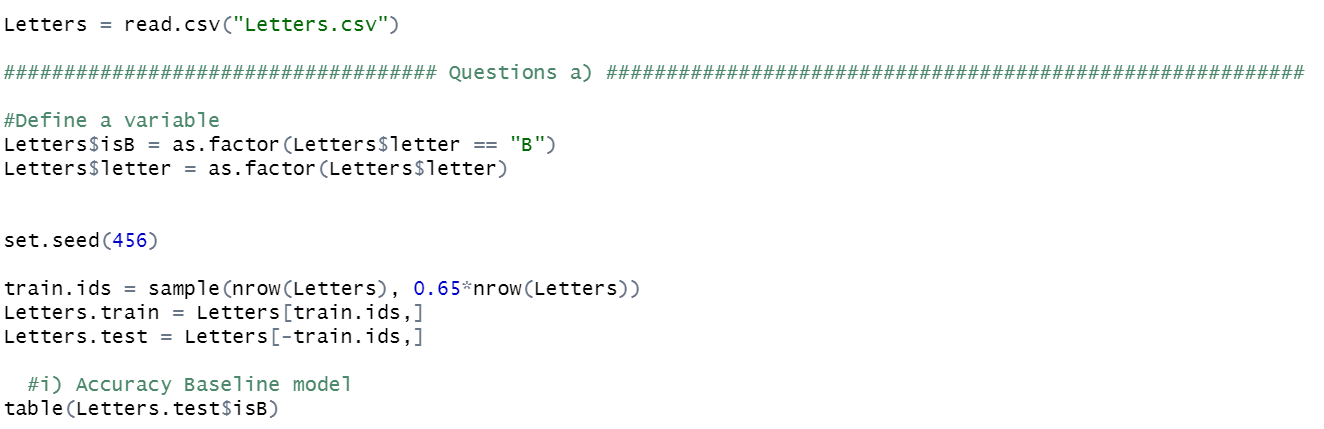
Homework 3



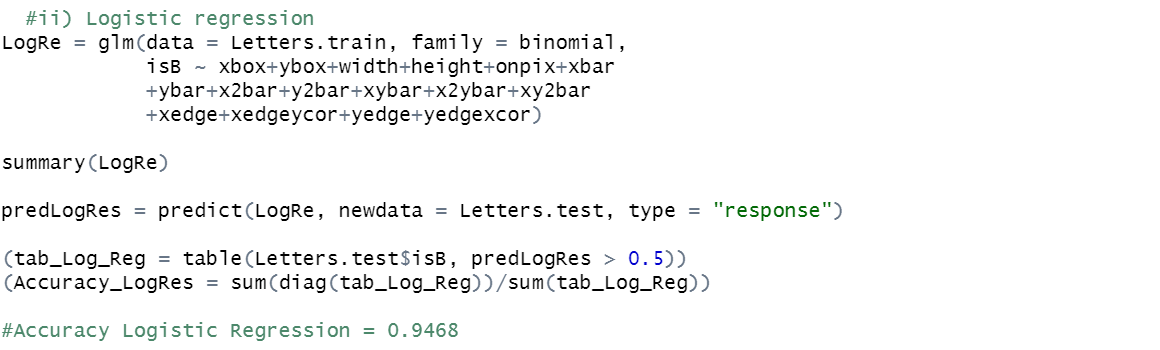


Problem 2: Framingham Heart Study



**i) Baseline method**

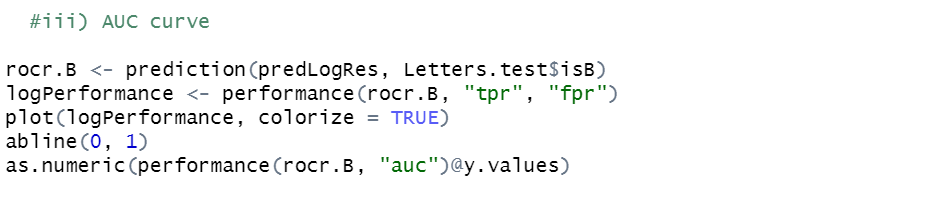
It always predict the most frequent outcome, “**not B”.**

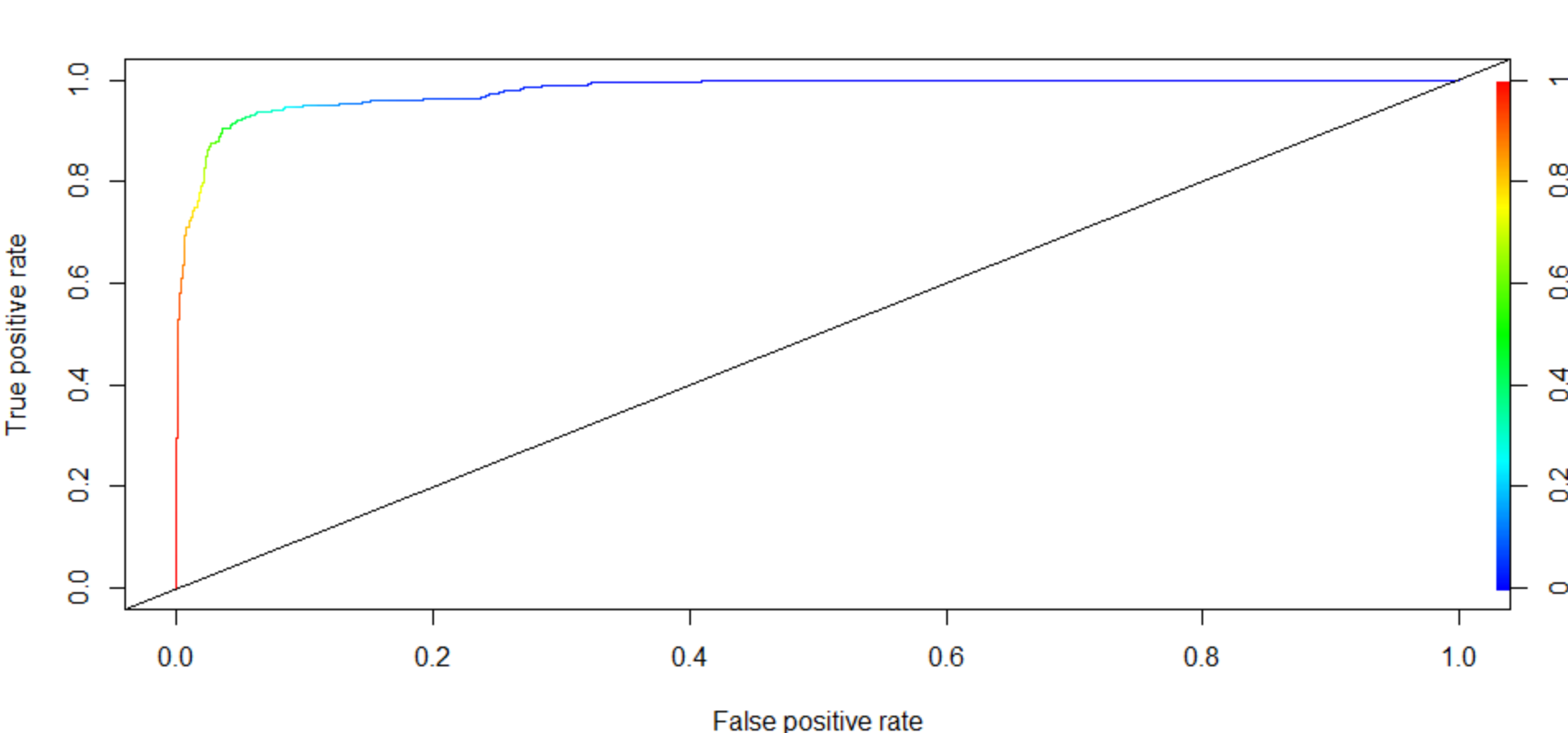
**ii) Logistic regression to predict whether or not the letter is B**

P(Y=1/X=x)=1/(1+exp(-(xbox+ybox+width+height+onpix+xbar+ybar+x2bar+y2bar+xybar+x2ybar+xy2bar+xedge+xedgeycor+yedge+yedgexcor)))

|  |  |  |
| --- | --- | --- |
|  | **No B P(Y=1)<0.5** | **B P(Y=1)>0.5** |
| **No B** | **760** | **28** |
| **B** | **30** | **273** |

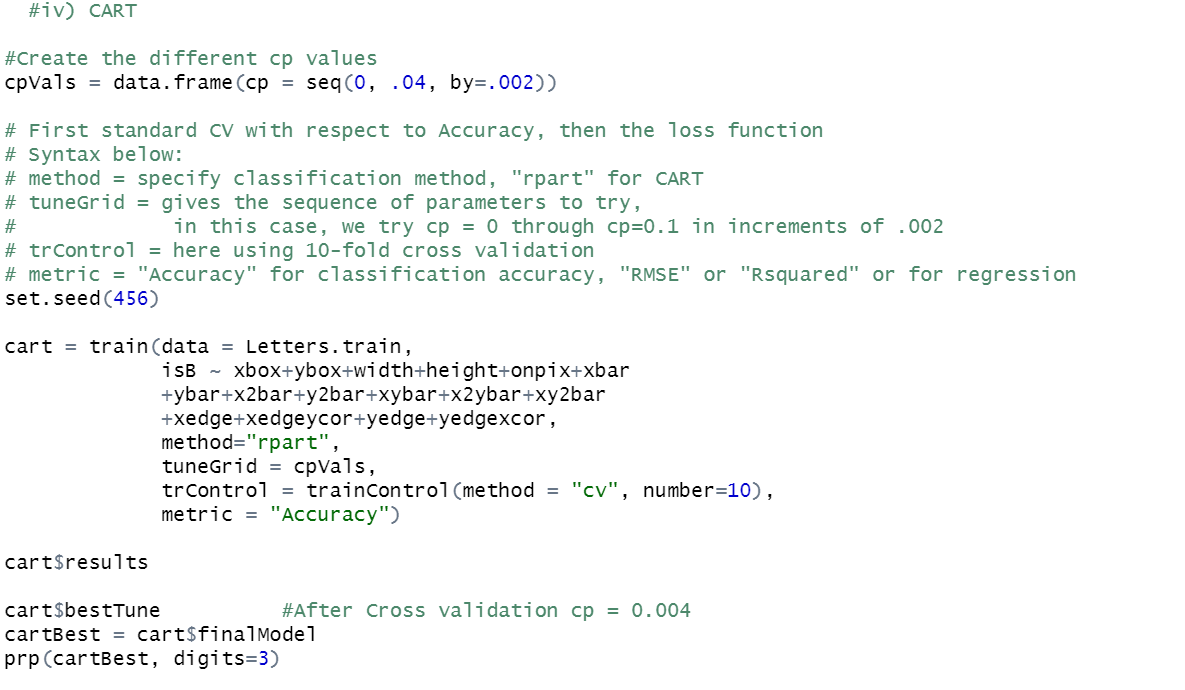
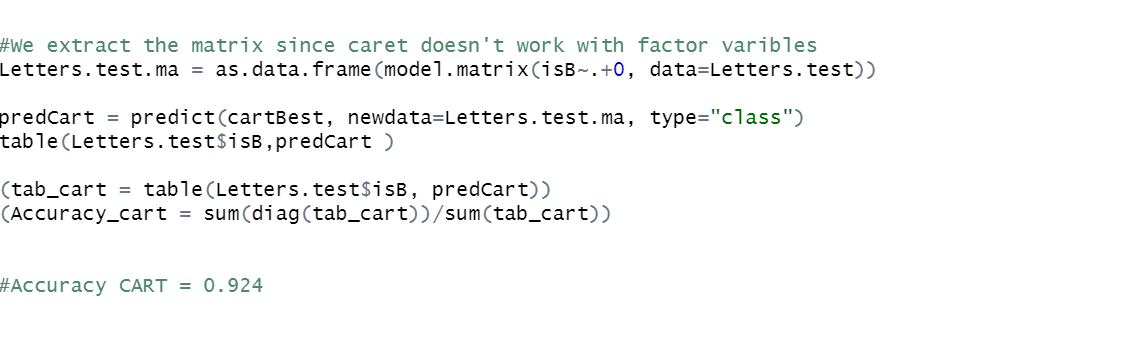
**iii) AUC of the logistic regression**





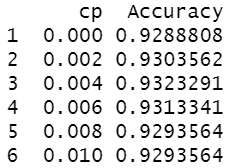
AUC = 0.9796

The regression model is a good fit to our data.

**iv) Construct a CART tp predict whether or not a letter is a B**

In this study we decide to focus on accuracy as our metric. We will a 10-fold Cross validation to choose the best cp values.

In this case, we try cp = 0 through cp=0.1 in increments of 0.002. Then we select the cp that yield to the best accuracy on the trainning set.

We find cp = 0.004

|  |  |  |
| --- | --- | --- |
|  | **No B P(Y=1)<0.5** | **B P(Y=1)>0.5** |
| **No B** | **768** | **20** |
| **B** | **62** | **241** |

**v) Random Forest**



|  |  |  |
| --- | --- | --- |
|  | **Good risk P(Y=1)<0.16** | **Bad risk P(Y=1)>0.16** |
| **No CHD** | **781** | **7** |
| **CHD** | **24** | **279** |

**vii) Compare the models**

In this case accuracy is much more important than interpretability since there is no clear meaning to pixels

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Baseline model** | **Regression model** | **CART** | **Random Forest** |
| **Accuracy** |  |  |  |  |

**Baseline model < CART < Regression model < Random Forest**

**Random Forest** has the best performance on the test set but is the less interpretable. In this case we are going to choose accuracy over interpretability because there no much information we could take from the features; they are not explicit.

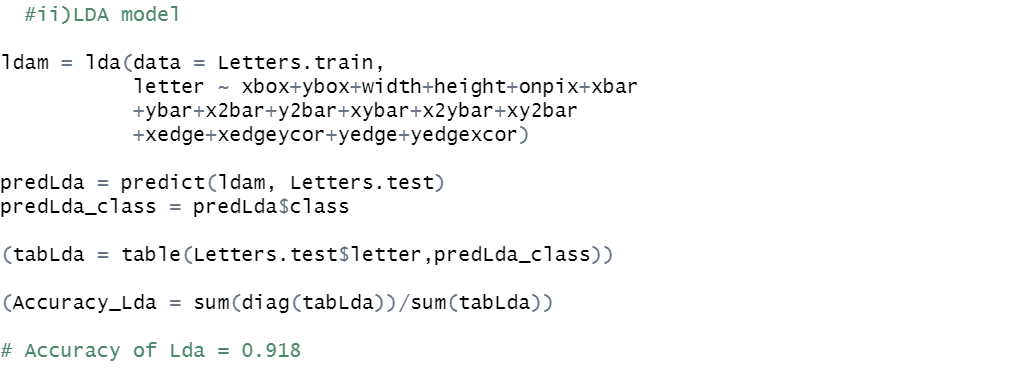
1. **Original problem of interest**

We will try to find the best model to predict “Letter”

**i) Baseline model**

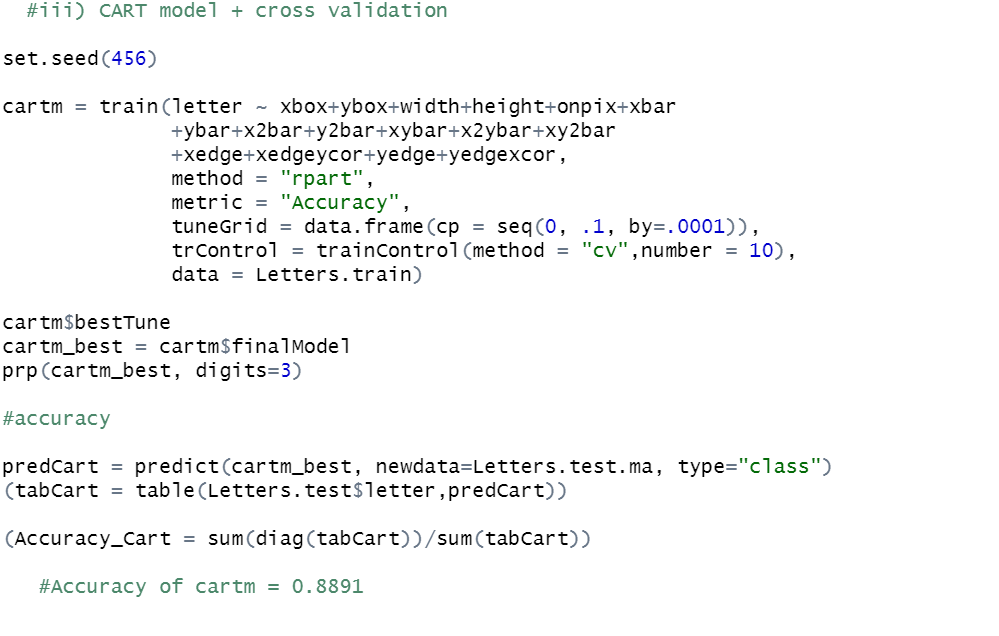
|  |  |  |  |
| --- | --- | --- | --- |
| **A** | **B** | **P** | **R** |
| **546** | **463** | **520** | **496** |

Considering a base line model that would always predict the most frequent class A, we would have this accuracy on the test set

**ii) LDA**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **P** | **R** |
| **A** | **227** | **5** | **1** | **10** |
| **B** | **0** | **272** | **0** | **31** |
| **P** | **0** | **6** | **273** | **4** |
| **R** | **0** | **31** | **1** | **230** |

**iIi) CART model + Cross Validation**



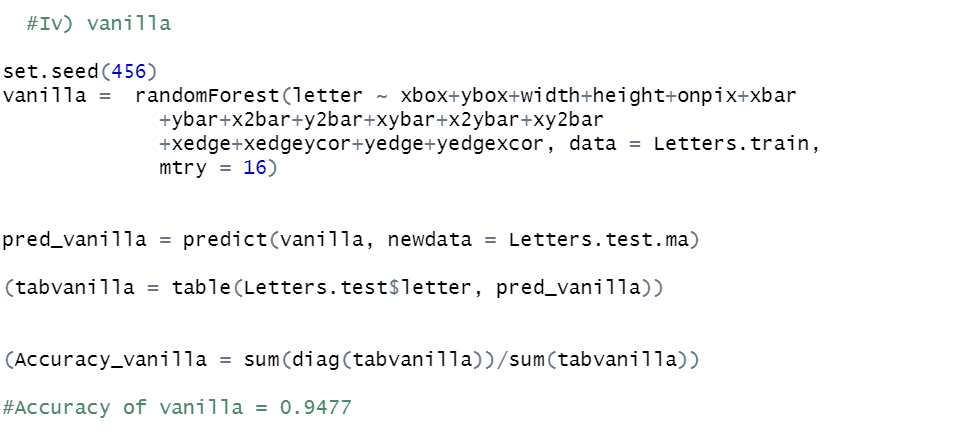
* Cross validation

First I decided to do a 10-fold cross validation to find the cp parameter with the best accuracy on the training set. I tried the same values of the first CART cp = 0 through cp=0.1 in increments of 0.002. We found cp = 0 with an accuracy of 0.886

Then, we changed the increment to 0.0001. We find a **cp = 0.0007** with an accuracy of 0.91 on the trainning set.

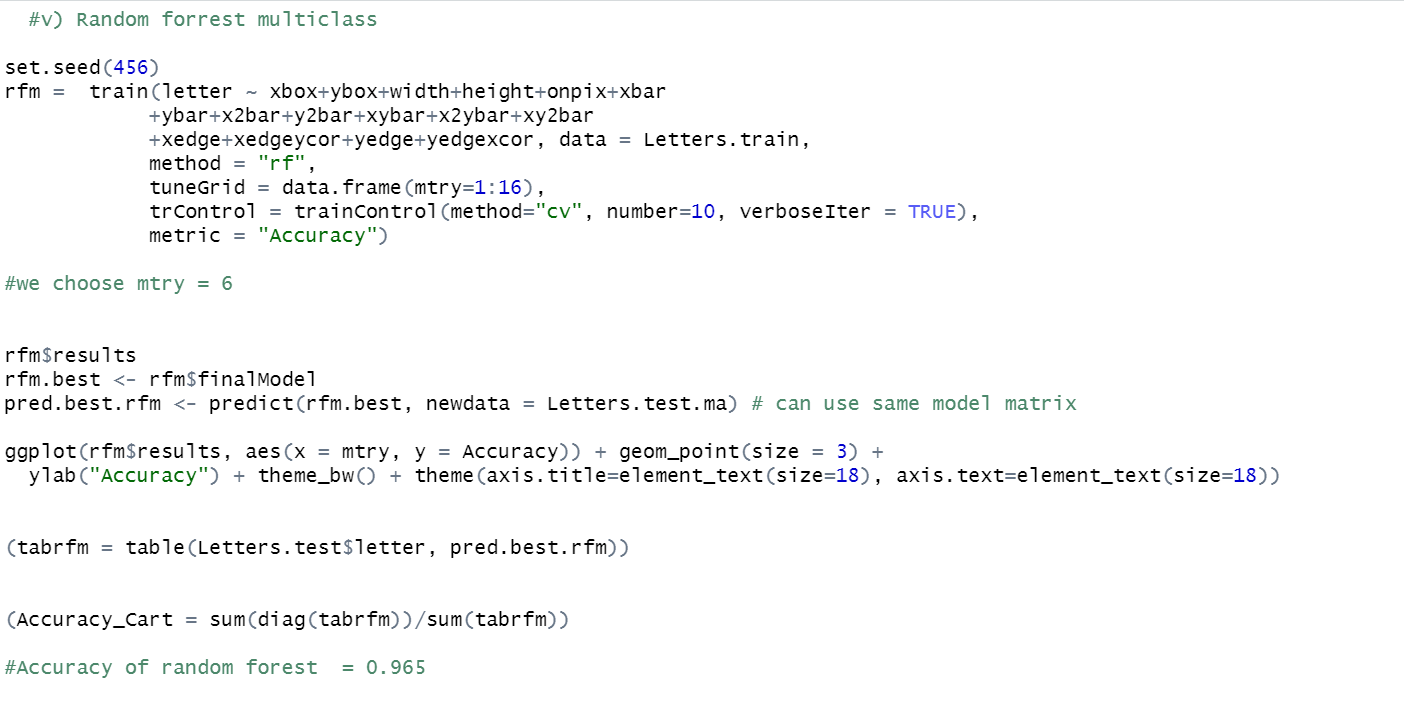
* Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **P** | **R** |
| **A** | **230** | **5** | **1** | **7** |
| **B** | **10** | **253** | **10** | **30** |
| **P** | **2** | **13** | **265** | **3** |
| **R** | **10** | **27** | **3** | **222** |

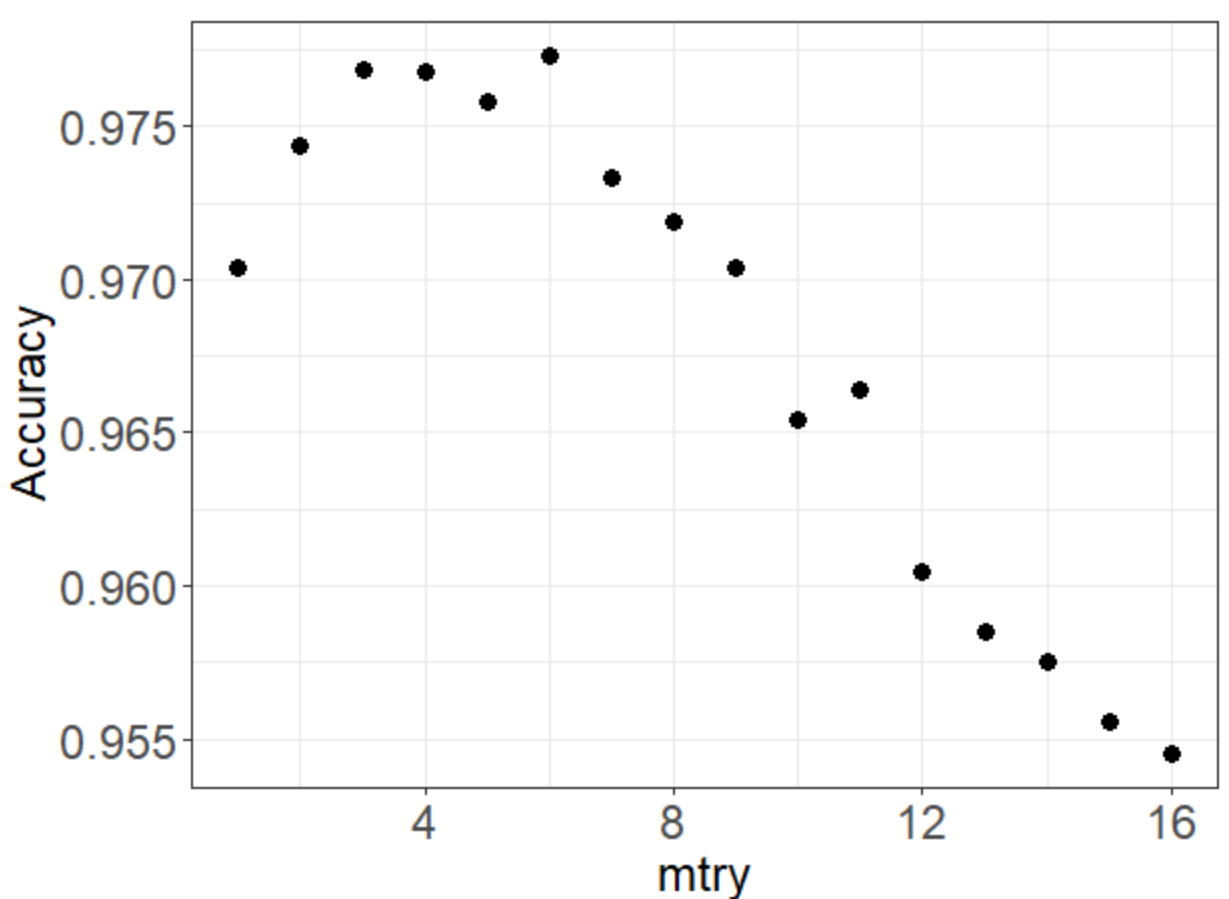
**iv) Vanilla**

Through the Random Forest model we will set **mtry = 16** to examine all the variables at each spit of fitted CART trees

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **P** | **R** |
| **A** | **236** | **4** | **1** | **2** |
| **B** | **2** | **280** | **2** | **19** |
| **P** | **0** | **2** | **278** | **3** |
| **R** | **1** | **19** | **2** | **240** |

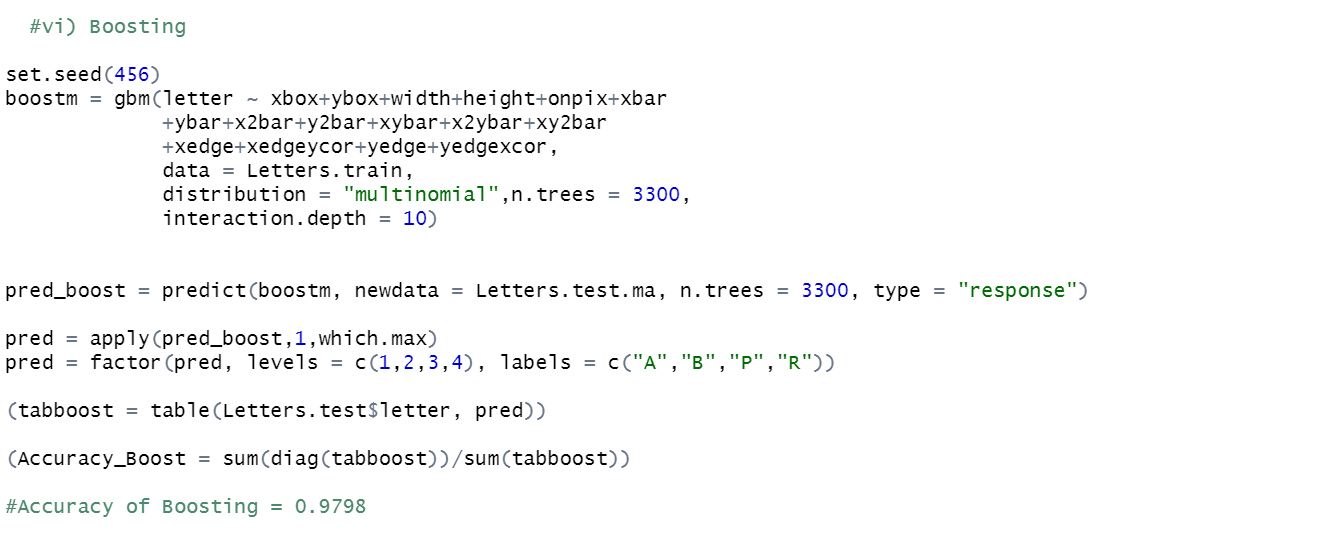
**v) Random Forest + Cross validation**

We will use a 10 fold cross validation to select mtry (number of variables examined at each split 1..16), number of trees = 500. We find mtry = 6 that yield to accuracy 0.9787 on the training set.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **P** | **R** |
| **A** | **238** | **2** | **1** | **2** |
| **B** | **1** | **286** | **0** | **16** |
| **P** | **0** | **2** | **279** | **2** |
| **R** | **0** | **10** | **2** | **250** |

**vi) Boosting**



We set the interaction depth to 10, run the method for 3300 iterations. With a 5-fold cross validation we have:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **P** | **R** |
| **A** | **240** | **2** | **0** | **1** |
| **B** | **0** | **295** | **0** | **8** |
| **P** | **0** | **1** | **277** | **5** |
| **R** | **0** | **5** | **0** | **257** |

**vii) Comparison**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **LDA** | **CART** | **Vanilla (bagging)** | **Random Forest** | **Boosting** |
| **Accuracy** |  |  |  |  | 0.980 |

CART < LDA < Vanilla < Random Forest < Boosting

Boosting is more accurate than Random Forest on the test set. Thus we recommend using the first one since for the same reasons as before we choose to give more importance in this study to accuracy than interpretability (Even if we find the most impactful features, it would be hard to give them a concrete explanation).