Shraddha

Assignment-2

Chapter-4

Dated: Recitation Exercises

Example 4

p = 1 feature X.

X = evenly distributed on [0, 1]

⇒ Predict observation that are within 10% of the range of X. Closest to test observation.

⇒ for X = 0.6 range [0.55, 0.65]

Answer:-

Available observation we can use for prediction is

For, X. Range = [0.55, 0.65]

⇒ [(0.65 - 0.55) / (1 - 0)) * 100

⇒ 0.10 * 100

⇒ 10%.

So on average we can use 10% of observation for prediction.

(b) p = 2 Features x_1, x_2 uniformly distributed on [0, 1] × [0, 1]

⇒ Predict test observation response within 10% of x_1 and 10% of x_2

Dated:

⇒ For $x_1 = 0.6$ $x_2 = 0.39$
we use Range $[0.55, 0.65]$ For x_1
$[0.3, 0.4]$ for x_2

Answer:-

$[(0.65 - 0.55)/(1-0) * 100] +$
$[(0.4 - 0.3)/(1-0)) * 100]$

⇒ $[0.10 * 100] + [0.1 * 100]$
⇒ $10\% + 10\%$
⇒ 1%

So, Fraction of the availble observation can be used for prediction is 1%.

(C) ⇒ P = 100
⇒ uniformaly distributed on each features in range $[0, 1]$

⇒ predict test observation's response using 10% of each features.

Answer:-

⇒ (observation feature range)^P Percentage.

⇒ $((1/10)^{100}) + 100\%$

⇒ $((0.1)^{100}) + 100\%$.

≈ 0%

Dated:

(d) Argue drawback of KNN for large P

⇒ From above (a - c) result we can conclude that
As value of P increase the fraction of availble observation to make prediction decrease.
⇒ means, when P = 100 percentage of observation for predict is almost 0 that shows KNN perform poor when number of observation are large.

(e) ⇒ P = 1, 2, 100
⇒ predict test observation by creating a p-dimensional hypercube centered around test observation.
⇒ 10 % of training observation.

Answer:-
length of each side of the hypercube

For P = 1
hypercube is a line
length = $(0.1)^{(1)} = 0.1$

P = 2
hypercube is square
length = $(0.1)^{(1/2)} = 0.316$

Dated:

For $p = 100$

hypercube is a dimentional cube.

length $= (0.1)^{\wedge}(1/100)$

$= 0.977$

$\Rightarrow$ When we predict test observation that contain 10% of training observation means we have large number of features so it will be better to use all the features.

(6) Chapter-4 Exercise-6

X_1 = hours X_2 = undergrad GPA

y = A grade.

$\Rightarrow$ logistic Regression

$\Rightarrow$ Estimated co-efficient $B_0 = -6$

$B_1 = 0.05$

$B_2 = 1$.

(9) Probability For Students who Studies

For $X_1 = 40$ hour

$X_2 = 3.5$ GPA

gets an A ?

Answer :-

Dated:

$$P(X) = e^{B_0 + B_1X_1 + B_2X_2} / 1 + e^{B_0 + B_1X_1 + B_2X_2}$$

$$= \frac{e^{-6 + (0.05)(40) + (1)(3.5)}}{1 + e^{-6 + (0.05 \times 40) + (1 \times 3.5)}}$$

$$\Rightarrow \frac{e^{-6 + 2 + 3.5}}{1 + e^{-6 + 3.5}} = \frac{e^{-0.5}}{1 + e^{-0.5}}$$

$$= \frac{0.6065}{1.6065}$$

$$= 0.3775$$

$$\boxed{P(X) = 37.75\%}$$

(b) How many hours student (x) need to study to have 50% of chance of getting A ?

$P(X) = 50\%$ $\quad X_2 = 3.5$

$X_1 = ?$

$$P(X) = e^{B_0 + B_1X_1 + B_2X_2} / 1 + e^{B_0 + B_1X_1 + B_2X_2}$$

$$0.5 = e^{-6 + 0.05X_1 + 3.5} / 1 + e^{-6 + 0.05X_1 + 3.5}$$

Dated:

$$= 0.5 = 2.718^{0.005}$$

$$= 0.5 = \frac{e^{0.005 x_1 - 2.5}}{1 + e^{0.005 x_1 - 2.5}}$$

$x_1 = 50$

So, Student need to study for 50 hours to have 50% of chance of getting A

Dated:

Chapter 4 Exercise-7

mean value for company issued stock dividend was $x = 10$

those who didn't was $x = 0$

$\Rightarrow$ variance was $\sigma^2 = 36$

$\Rightarrow$ 80% company issued dividend.

$\Rightarrow$ X follows normal distribution.

Answer Bayes' theorem

$$P(Y=k/X=x) = \frac{[\Pi k \; f_k(x)]}{\sum_{l=1}^{k} \Pi l \; f_l(x)}$$

$$= f_k(x) = \left[1/\sqrt{2\pi\sigma_k^2}\right] * e^{-(2\cdot 1/2\sigma^2)(x-\mu)^2}$$

$$= f_{yes}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

$$= \frac{1}{\sqrt{2\pi(36)}} \; e^{\frac{-1}{2(36)}(x-10)^2} \quad \because (x_{yes} = 10)$$

$$\boxed{f_{yes}(x) = \frac{e^{-1/2}}{\sqrt{72\pi}}} \quad \because x = 4$$

Dated: $f_{no}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{(-x-\mu)^2}{2\sigma^2}}$

$\boxed{f_{no(x)} = \frac{1}{\sqrt{2\pi}} \cdot e^{-2/9}}$ $\therefore x = 4$ $\therefore \mu = 0 \Rightarrow x_{no} = 0$

$P(Y=yes / x=4) = \frac{(\pi_{yes} f_{yes}(x))}{(\pi_{no} F_{no(4)} + \pi_{yes} f_{yes(x)})}$

$= \frac{0.8 \times e^{-12} / \sqrt{2\pi}}{0.2 \times e^{-2/8} / \sqrt{2\pi} + 0.8 \times e^{-12} / \sqrt{2\pi}}$

$\boxed{P(Y=yes / x=4) = 0.75}$ $\boxed{\text{for } x=4 \text{ probability is } 75\%.}$

Chapter -4 Exercise-9

a) On average what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

Answer

$$odds = \frac{P(x)}{[1 - P(x)]}$$

$$ODDS = 0.37$$

Dated:

$$0.37 = \frac{P(x)}{1 - P(x)}$$

$$\Rightarrow 0.37 - 0.37\,P(x) = P(x)$$

$$\Rightarrow P(x) + 0.37\,P(x) = 0.37$$

$$\Rightarrow P(x)(1.37) = 0.37$$

$$\Rightarrow P(x) = \frac{0.37}{1.37}$$

$$\boxed{P(x) = 0.2700} \Rightarrow 27\%$$

(b) individual has a 16% chance of defaulting on her credit card payment what are the odds that she will default?

$$P(x) = 16\% = 0.16$$

$$\text{Odds} = \frac{P(x)}{1 - P(x)}$$

$$= \frac{0.16}{1 - 0.16}$$

$$= \frac{0.16}{0.84}$$

$$\boxed{\text{Odds} = 0.19}$$

Dated: Chapter-5

Exercise-2

bootstrap Sample from a set of n observations.

a) What is the probability that the first bootstrap observation is not jth observation from the original sample?

Answer: number of observation = n, Every observation in original sample is independent.

$P \Rightarrow$ Probability that first bootstrap observation is jth observation from original sample.

$$= \frac{1}{n}$$

$P' \Rightarrow$ First bootstrap observation is not jth observation

$$= 1 - P$$
$$= 1 - \frac{1}{n}$$

Answer $= P' \Rightarrow \frac{n-1}{n}$

(b) Probability that Second bootstrap observation is not the jth observation from the original Sample?

Dated:

Answer

⇒ total number of observation = n

⇒ observation is part of bootstrap

⇒ In original sample, every observation has equal probability of appearing in each observation.

⇒ So Probability that second bootstrap observation is jth observation from original sample is $= \frac{1}{n}$

$\bar{p}$ ⇒ not jth observation from original sample $= 1 - \frac{1}{n}$

$= \frac{n-1}{n}$

(c) jth Observation is not in bootstrap sample is $\left(1 - \frac{1}{n}\right)^n$.

Answer :-

⇒ We obtain bootstrap sample from n observation

⇒ p_1 - jth observation not in First bootstrap sample $= 1 - \frac{1}{n}$

Dated:

P_2 = Probability that observation not in second bootstrap sample $= 1 - \frac{1}{n}$

P_n = Probability jth observation not in n^{th} bootstrap sample $= 1 - \frac{1}{n}$

So

Answer :-

P (observation is not in bootstrap sample) =

$\Rightarrow P_1 * P_2 * \ldots \cdot P_n$

$= \left(1 - \frac{1}{n}\right) * \left(1 - \frac{1}{n}\right) * \ldots \left(1 - \frac{1}{n}\right)$ ...n time

$\Rightarrow \left(1 - \frac{1}{n}\right)^n$

(d) $n = 5$

P (jth observation is in bootstrap sample

$\Rightarrow \left(1 - \frac{1}{n}\right)^n$

$\Rightarrow \left(1 - \frac{1}{5}\right)^5$

$= \left(\frac{5-1}{5}\right)^5 \Rightarrow \left(\frac{4}{5}\right)^5 \Rightarrow 0.672$

Dated:

(e) $n = 100$

P(jth observation is in bootstrap sample)

$\Rightarrow \left(1 - \frac{1}{n}\right)^n$

$\Rightarrow \left(1 - \frac{1}{100}\right)^{100}$

$\Rightarrow 0.634$

(f) $n = 10,000$

P(jth observation is in bootstrap sample)

$\Rightarrow \left(1 - \frac{1}{n}\right)^n$

$\Rightarrow \left(1 - \frac{1}{10,000}\right)^{10,000}$

$\Rightarrow 0.63214$

(g) Creat Plot that display.

For each integer value of n from 1 to 100,000 the probability that the jth observation is in the bootstrap sample.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

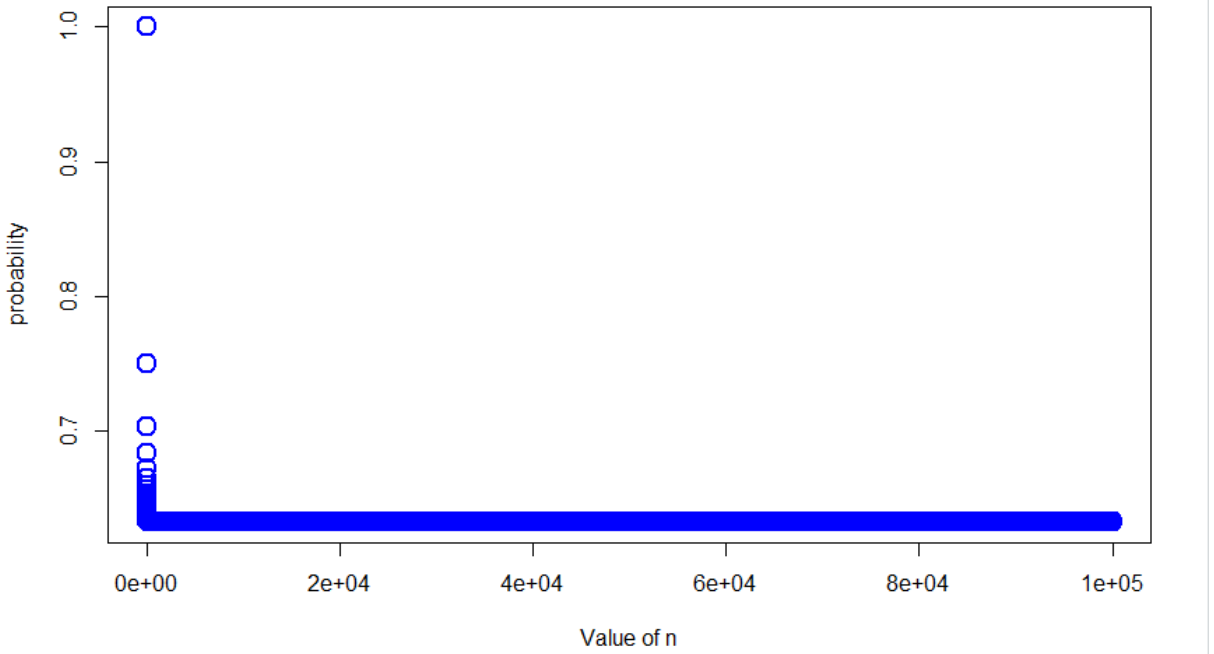
Go to file/function | Addins

Source

Console | Terminal | Background Jobs

R 4.2.1 · ~/

```
> plot(1:100000,1-(1-1/1:100000)^(1:100000),xlab="value of n",ylab="probability",col="blue",cex=2,lwd=2)
>
```



n = seq(1, 100000)

Dated:

(b) Numerically the probability that a bootstrap sample of size n=100 contain the jth observation.

j = 4. We repeatedly create bootstrap sample, and each time we record whether or not the forth observation is contained in bootstrap sample.

⇒ Comment on the results

⇒ Store = rep(NA, 10000)

⇒ for(i in 1:10000){

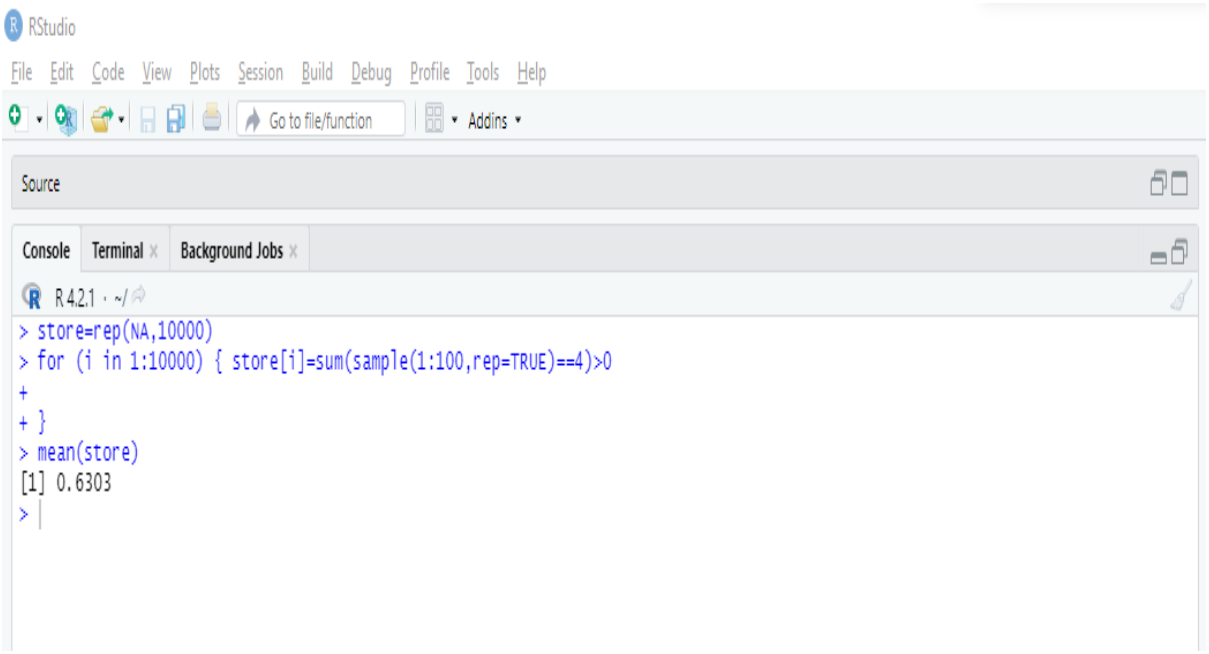
Store[i] = sum(sample(1:100, rep=TRUE
==4)>0

⇒ mean(store) = 0.6303

Answer We got mean value 0.6303 the observed probability of

$1-(1-1/100)^{100} = 63.4\%$ is

Compareable to our prediction.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Source
Console
Terminal
Background Jobs
R 4.2.1 · ~/
> store=rep(NA,10000)
> for (i in 1:10000) { store[i]=sum(sample(1:100,rep=TRUE)==4)>0
+
+ }
> mean(store)
[1] 0.6303
>

Dated:

Exercise - 3

a) Explain how k-fold cross validation is implemented.

Answer:

⇒ k-Fold Cross Validation, divides the n observation into k groups/folds of approximately equal size randomly.

⇒ First Fold is treated as validation set

⇒ Remining k-1 folds are used for fitting model. and mse1 is calculated

⇒ Each group considered as validation set once and a training set for (k-1) times

⇒ This Process is repeated k times

⇒ We Compute Test error By

$$CV(k) = \left(\frac{1}{k}\right) \sum_{1} mse_i$$

For $i = 1$ to k.

(b) What are the advantages and disadvantages of k-fold cross validation relative to:

Dated:

1) Validation set apporch ?

Advantages :-

This apporch is easy to understand as well easy to implement.

Disadvantages :-

⇒ Estimation of test error rate is highly vary depending on which observation are included in training set and which are included in validation set.

⇒ We use observation included in training & fit model

⇒ Training set is used for fitting model, so if few observation for training, the model will perform worse and overestimate test error.

2) LOOCV ?

Advantages :-

⇒ This model is subset of k-fold cross-validation with $k = n$.

⇒ So in this model, each observation considered once in validation set and $(n-1)$ times in training set.

⇒ This apporch gives approximately unbiased test error estimation.

Dated:

Disadvantages:-

⇒ This model requires fitting of statistical learning method n times.

⇒ K-fold cross validation required K times.

⇒ So LOOCV can be expensive.

⇒ LOOCV has greater variance than K-fold cross validation model.

Assignment-2 Shraddha

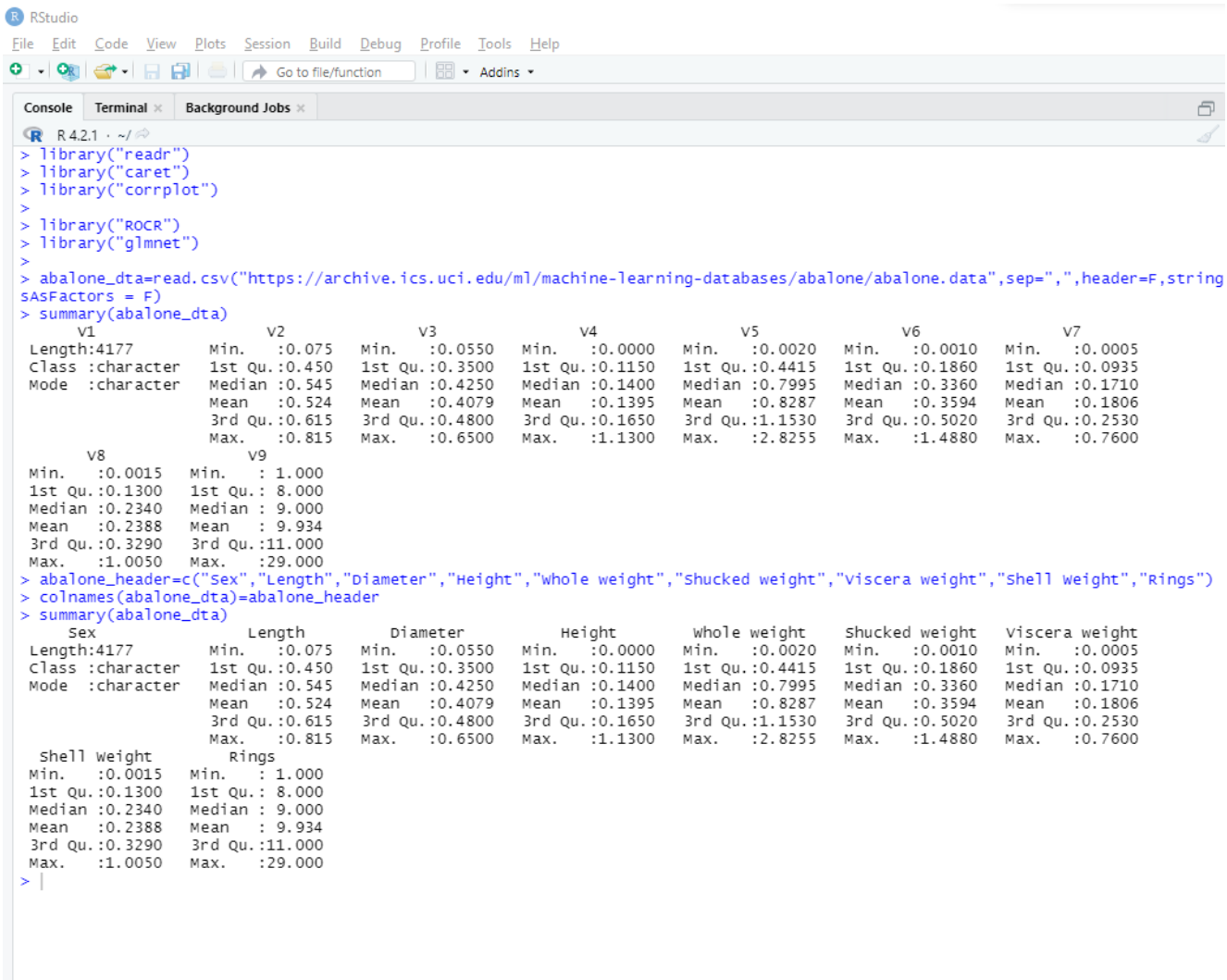
Practicum Problems

Problem 1:

Install packages caret, corrplot, glmnet , readr , tidyverse , ROCR using command install. Packages("packages")

Answer:

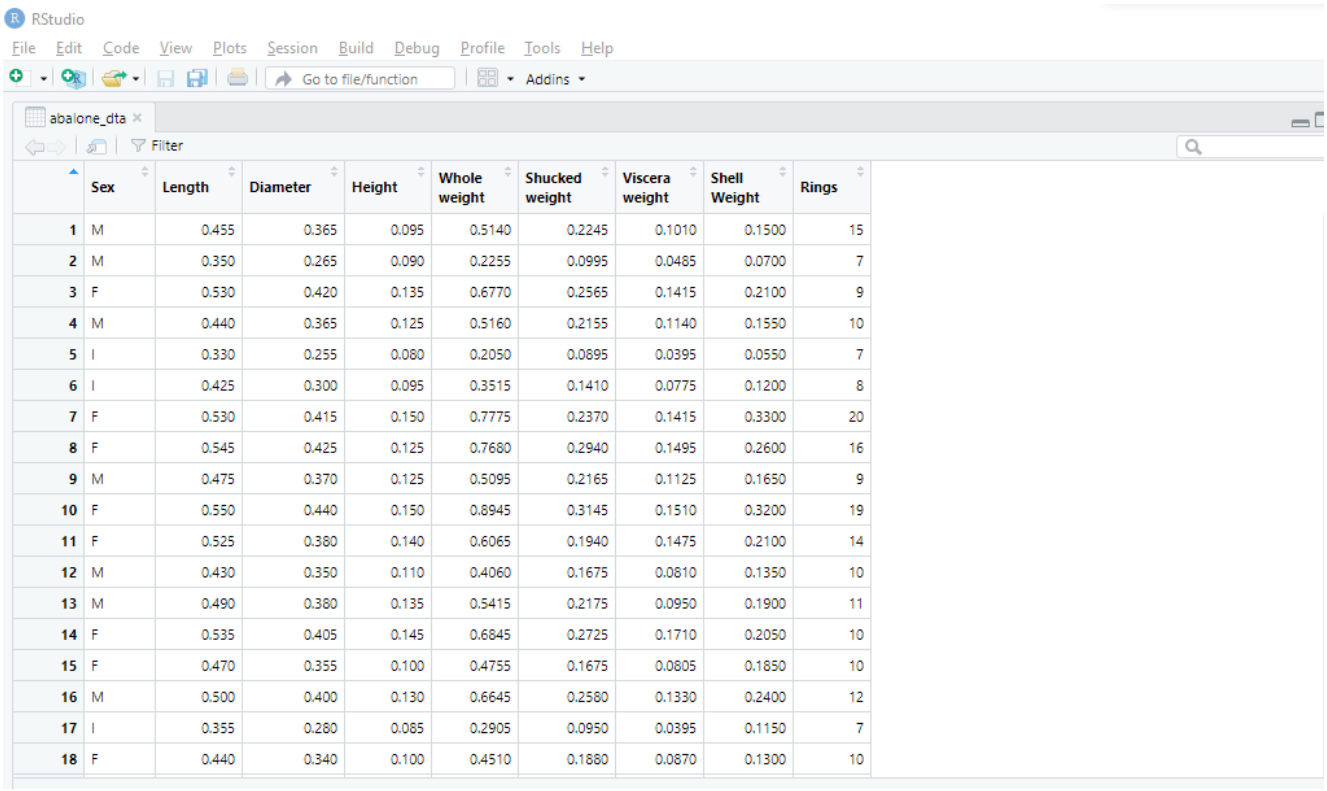
Load abalone Sample Dataset



Remove all observations in the Infant category

view(abalone data)

Assignment-2 Shraddha



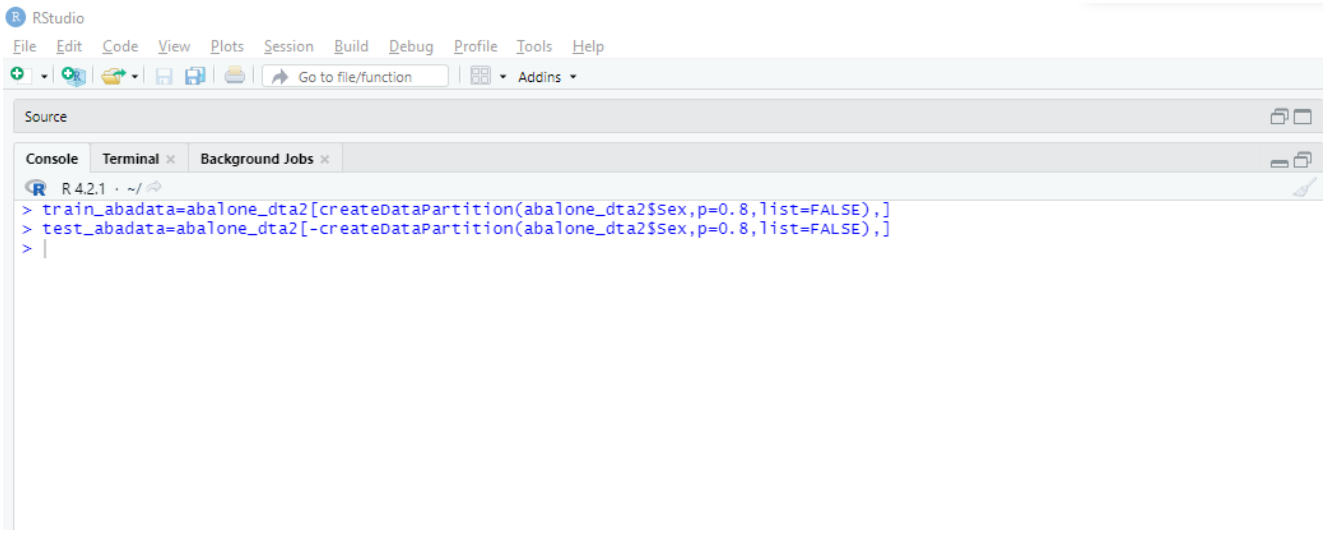
	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell Weight	Rings
1	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500	15
2	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700	7
3	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100	9
4	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550	10
5	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550	7
6	I	0.425	0.300	0.095	0.3515	0.1410	0.0775	0.1200	8
7	F	0.530	0.415	0.150	0.7775	0.2370	0.1415	0.3300	20
8	F	0.545	0.425	0.125	0.7680	0.2940	0.1495	0.2600	16
9	M	0.475	0.370	0.125	0.5095	0.2165	0.1125	0.1650	9
10	F	0.550	0.440	0.150	0.8945	0.3145	0.1510	0.3200	19
11	F	0.525	0.380	0.140	0.6065	0.1940	0.1475	0.2100	14
12	M	0.430	0.350	0.110	0.4060	0.1675	0.0810	0.1350	10
13	M	0.490	0.380	0.135	0.5415	0.2175	0.0950	0.1900	11
14	F	0.535	0.405	0.145	0.6845	0.2725	0.1710	0.2050	10
15	F	0.470	0.355	0.100	0.4755	0.1675	0.0805	0.1850	10
16	M	0.500	0.400	0.130	0.6645	0.2580	0.1330	0.2400	12
17	I	0.355	0.280	0.085	0.2905	0.0950	0.0395	0.1150	7
18	F	0.440	0.340	0.100	0.4510	0.1880	0.0870	0.1300	10

Keeping Male/Female Classes


```
> unique(abalone_dta$Sex)
[1] "M" "F" "I"
> abalone_dta2=abalone_dta[abalone_dta$Sex!="I",]
> unique(abalone_dta$Sex)
[1] "M" "F" "I"
> unique(abalone_dta2$Sex)
[1] "M" "F"
> abalone_dta2$Sex=factor(abalone_dta2$Sex)
> str(abalone_dta2)
'data.frame':   2835 obs. of  9 variables:
 $ Sex           : Factor w/ 2 levels "F","M": 2 2 1 2 1 1 2 1 1 2 ...
 $ Length        : num  0.455 0.35 0.53 0.44 0.53 0.545 0.475 0.55 0.525 0.43 ...
 $ Diameter      : num  0.365 0.265 0.42 0.365 0.415 0.425 0.37 0.44 0.38 0.35 ...
 $ Height        : num  0.095 0.09 0.135 0.125 0.15 0.125 0.125 0.15 0.14 0.11 ...
 $ Whole weight  : num  0.514 0.226 0.677 0.516 0.777 ...
 $ Shucked weight: num  0.2245 0.0995 0.2565 0.2155 0.237 ...
 $ Viscera weight: num  0.101 0.0485 0.1415 0.114 0.1415 ...
 $ Shell Weight  : num  0.15 0.07 0.21 0.155 0.33 0.26 0.165 0.32 0.21 0.135 ...
 $ Rings         : int  15 7 9 10 20 16 9 19 14 10 ...
>
```

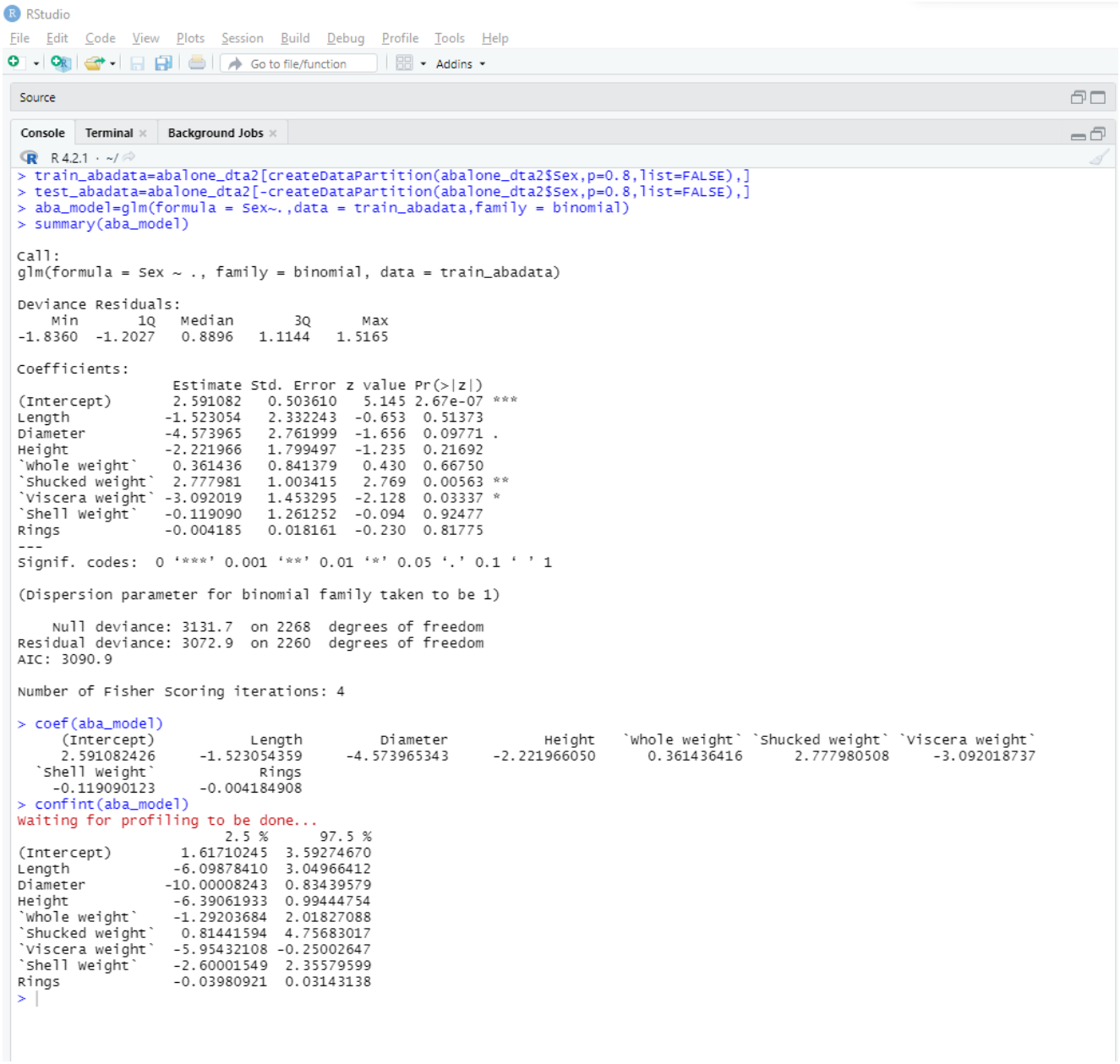
Assignment-2 Shraddha

Using the caret package, use createDataPartition to perform an 80/20 test-train split (80% training and 20% testing)



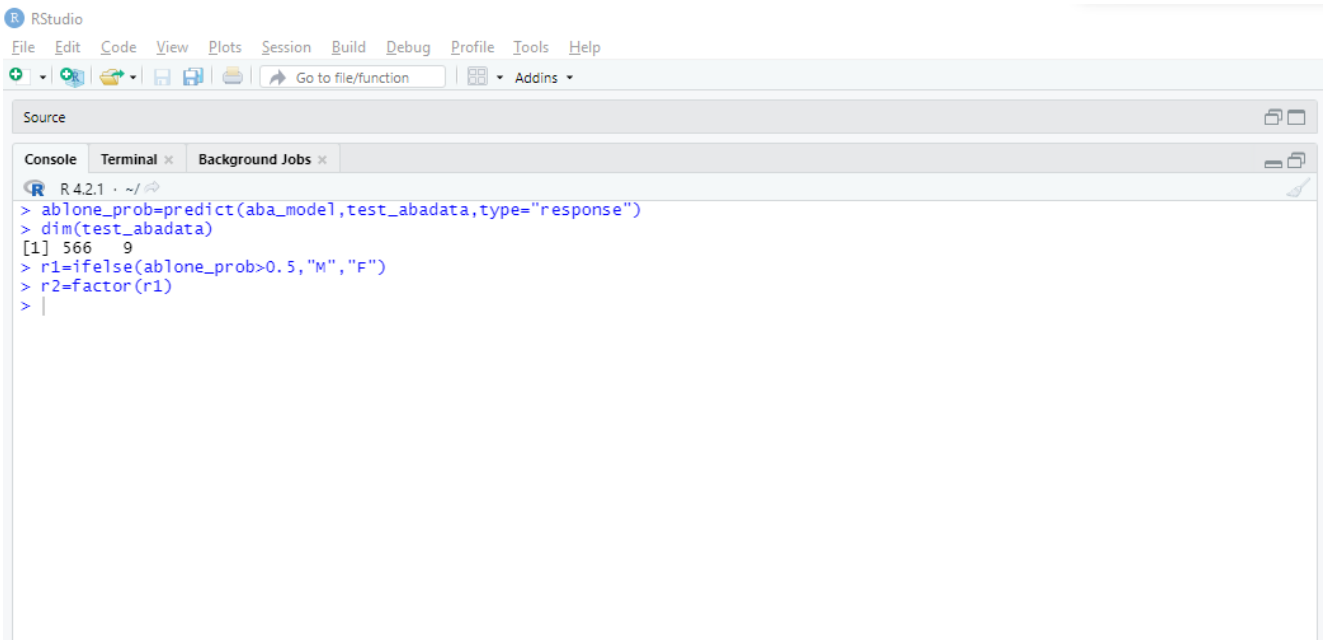
Fit a logistic regression using all feature variables via glm, and observe which predictors are relevant

Assignment-2 Shraddha

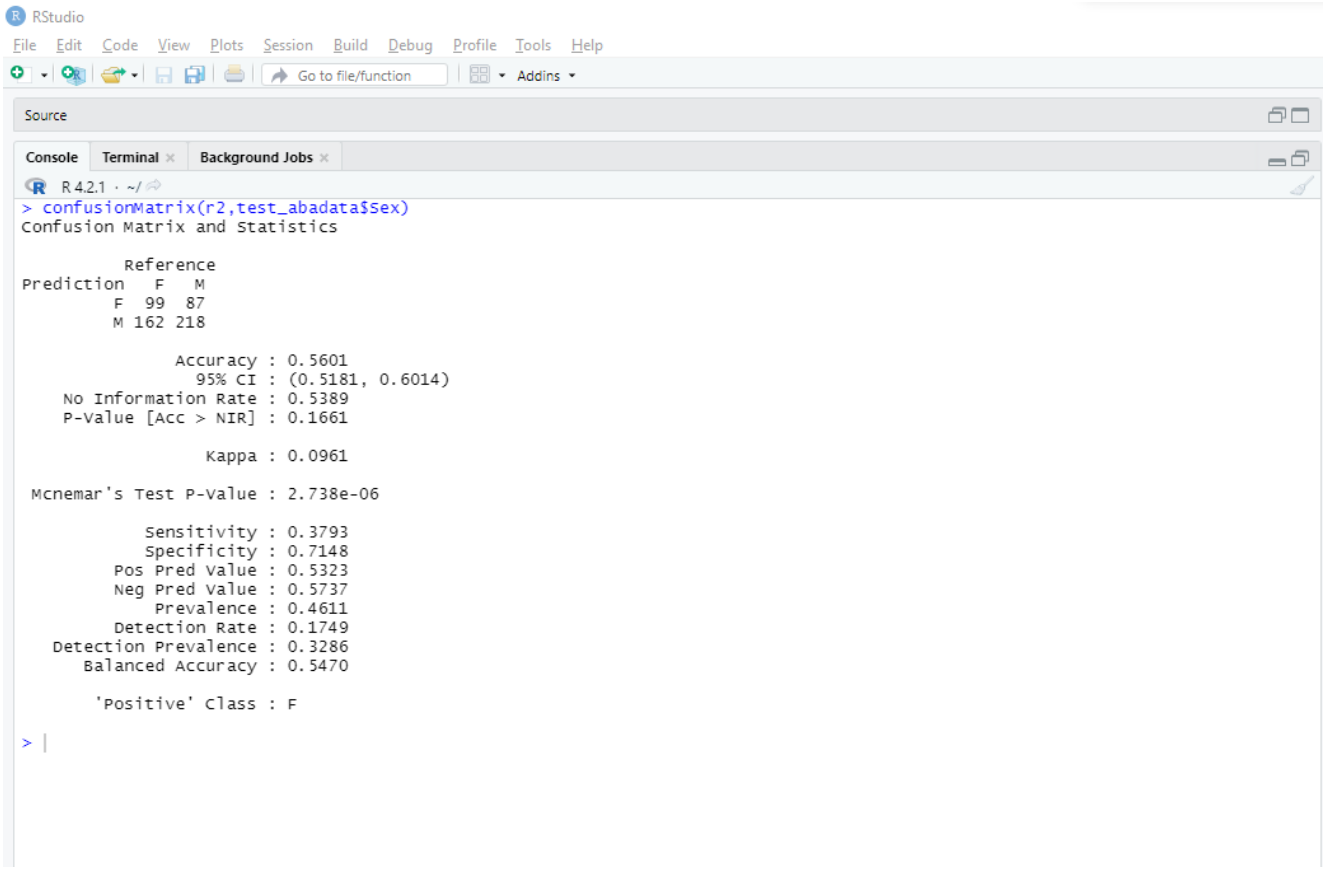


The confidence interval result in above image shows that, only “Shucked Weight” predictor has value 0 in confidence interval range which is align with the basic assumption of Null Hypothesis which represent there is no relationship between X and Y. So Null Hypothesis holds true for all the predictors except “Shucked Weight” means this predictor is relevant. Other predictors do not hold relationship with response variable.

Assignment-2 Shraddha



Create Confusion Matrix



Assignment-2 Shraddha

Plotting ROC Curve

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Source

Console | Terminal | Background Jobs

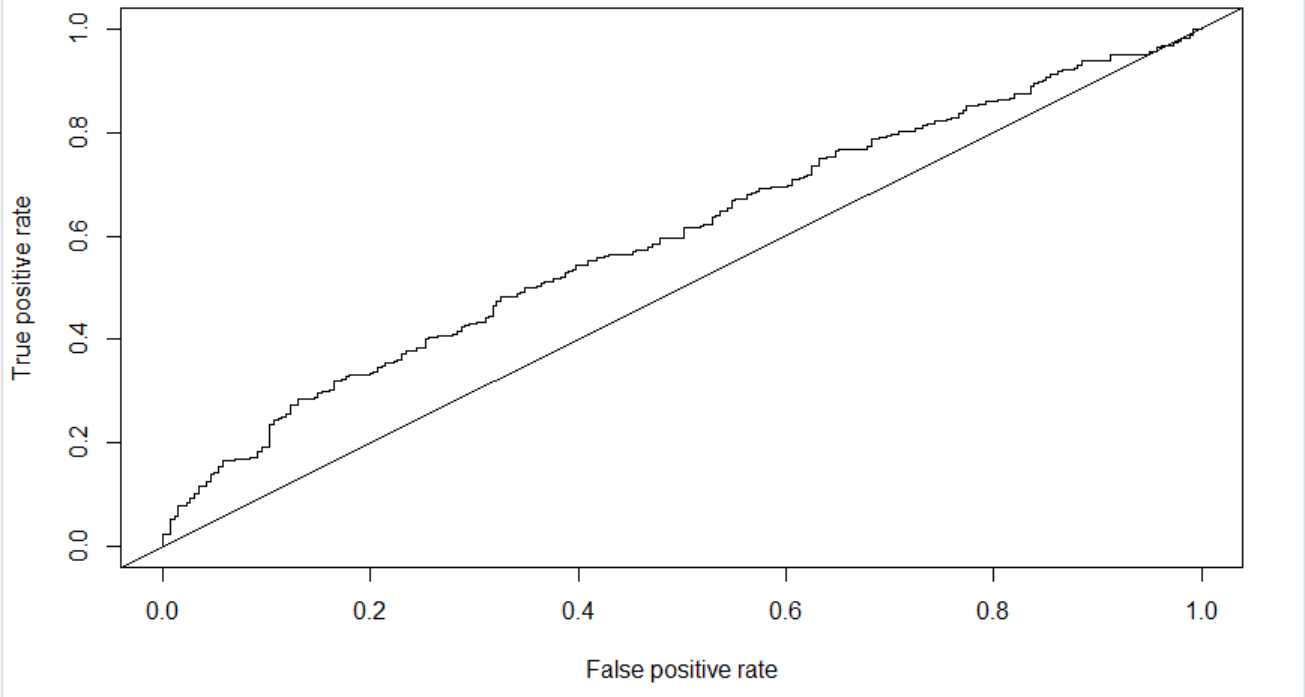
R 4.2.1 · ~/

```
> Roc_prediction=prediction(ablone_prob,test_abadata$Sex)
> Roc_Performance=performance(Roc_prediction,measure = "tpr",x.measure = "fpr")
> plot(Roc_Performance)
> abline(0,1)
> auc_performance=performance(Roc_prediction,measure = "auc")
> auc_performance@y.values
[[1]]
[1] 0.595955

>
```

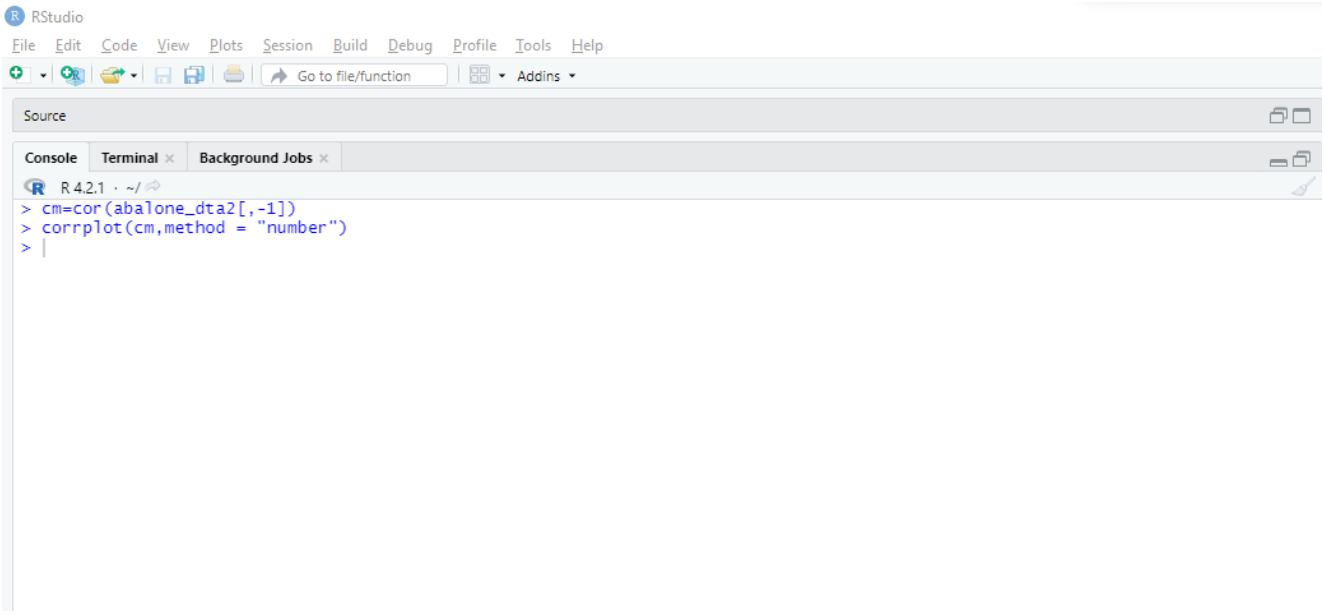
Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export | Publish



Assignment-2 Shraddha

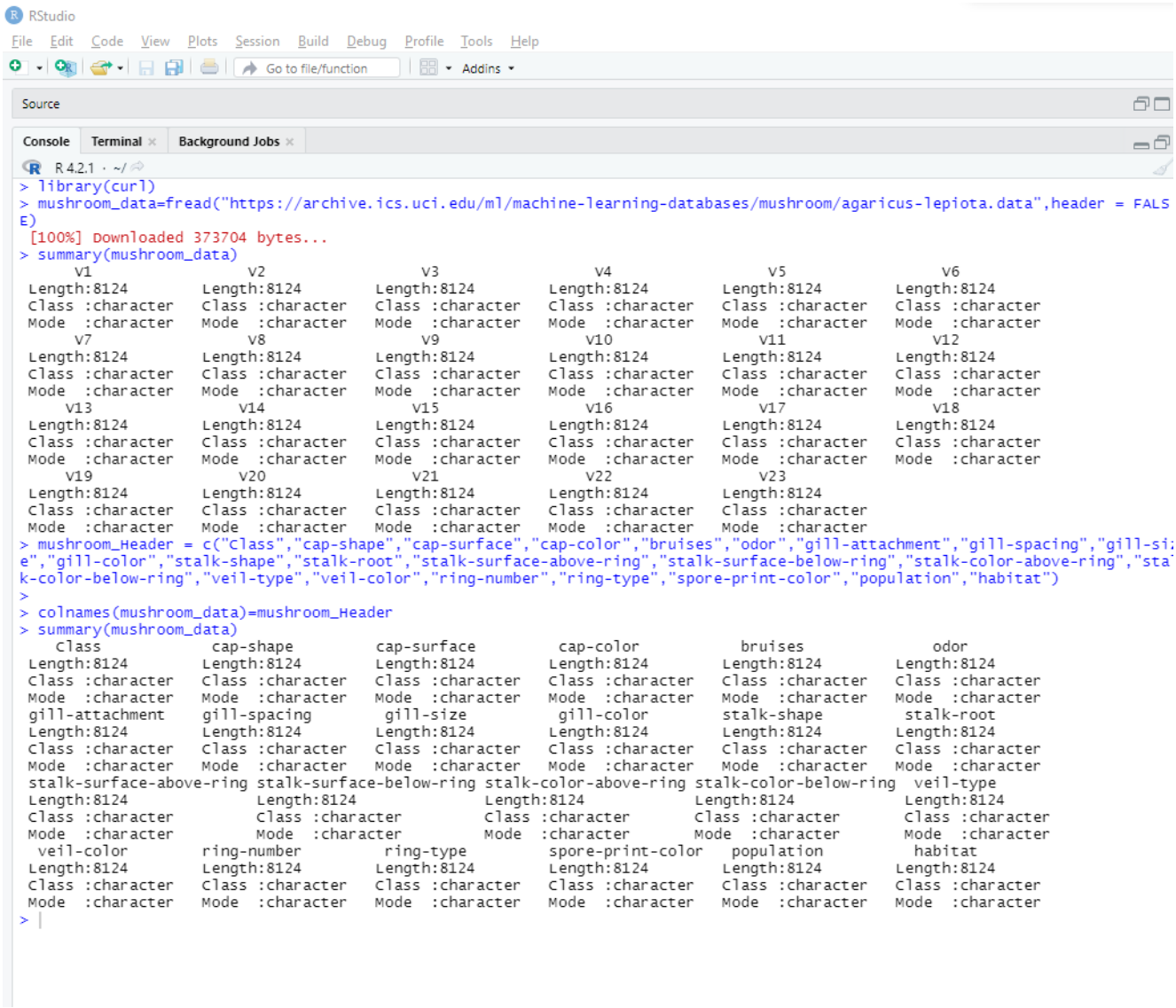
Plotting Corelation between predictors



From the above result we can see that there is a positive linear relation for all factors except the rings predictor which has a weak uphill relationship, while others have a high uphill connection.

Problem-2

Load the mushroom sample dataset from the UCI Machine Learning Repository using R Dataframe.



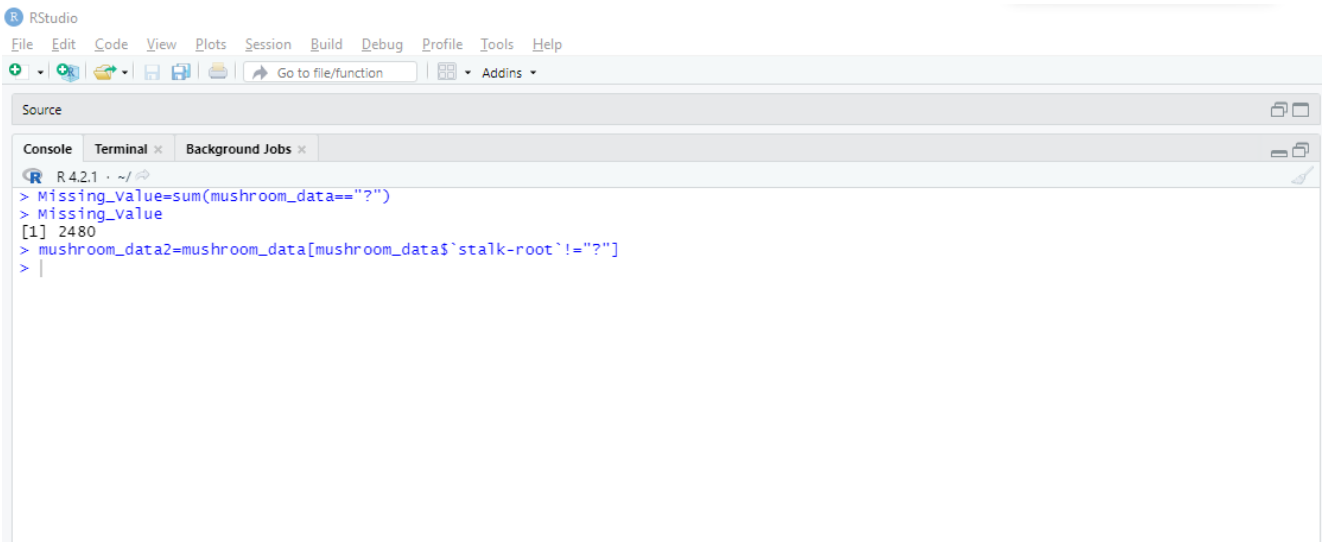
Assignment-2 Shraddha

Class Distribution and convert class attribute to a factor

```
> table(mushroom_data$Class)

   e    p
4208 3916
> mushroom_data$Class=factor(mushroom_data$Class)
> dim(mushroom_data)
[1] 8124   23
> str(mushroom_data)
Classes 'data.table' and 'data.frame':  8124 obs. of  23 variables:
 $ Class                   : Factor w/ 2 levels "e","p": 2 1 1 2 1 1 1 1 2 1 ...
 $ cap-shape               : chr  "x" "x" "b" "x" ...
 $ cap-surface             : chr  "s" "s" "s" "y" ...
 $ cap-color               : chr  "n" "y" "w" "w" ...
 $ bruises                 : chr  "t" "t" "t" "t" ...
 $ odor                    : chr  "p" "a" "l" "p" ...
 $ gill-attachment         : chr  "f" "f" "f" "f" ...
 $ gill-spacing            : chr  "c" "c" "c" "c" ...
 $ gill-size               : chr  "n" "b" "b" "n" ...
 $ gill-color              : chr  "k" "k" "n" "n" ...
 $ stalk-shape             : chr  "e" "e" "e" "e" ...
 $ stalk-root              : chr  "e" "c" "c" "e" ...
 $ stalk-surface-above-ring: chr  "s" "s" "s" "s" ...
 $ stalk-surface-below-ring: chr  "s" "s" "s" "s" ...
 $ stalk-color-above-ring  : chr  "w" "w" "w" "w" ...
 $ stalk-color-below-ring  : chr  "w" "w" "w" "w" ...
 $ veil-type               : chr  "p" "p" "p" "p" ...
 $ veil-color              : chr  "w" "w" "w" "w" ...
 $ ring-number             : chr  "o" "o" "o" "o" ...
 $ ring-type               : chr  "p" "p" "p" "p" ...
 $ spore-print-color       : chr  "k" "n" "n" "k" ...
 $ population              : chr  "s" "n" "n" "s" ...
 $ habitat                 : chr  "u" "g" "m" "u" ...
 - attr(*, ".internal.selfref")=<externalptr>
> |
```

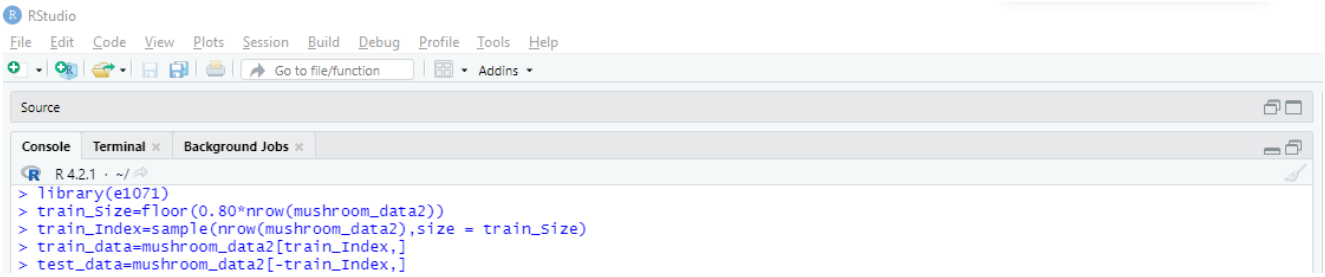
Number of Missing Values:



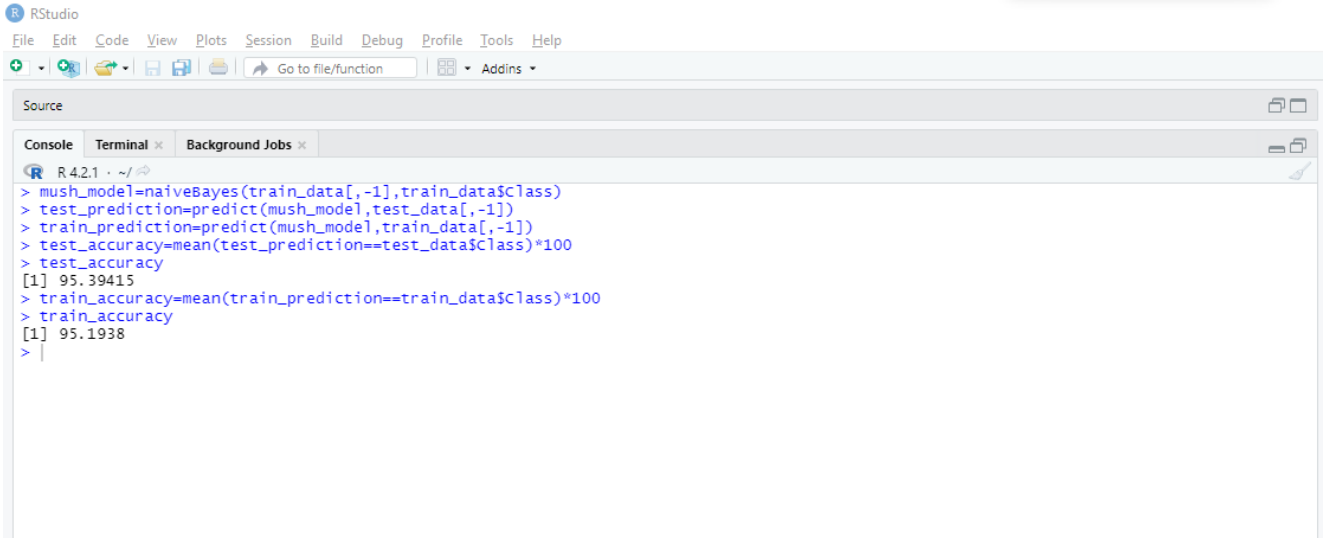
Removing the data with missing values do not impact on our dataset.

Create a Naive Bayes classifier using the e1071 package, using the sample function to split the data between 80% for training and 20% for testing.

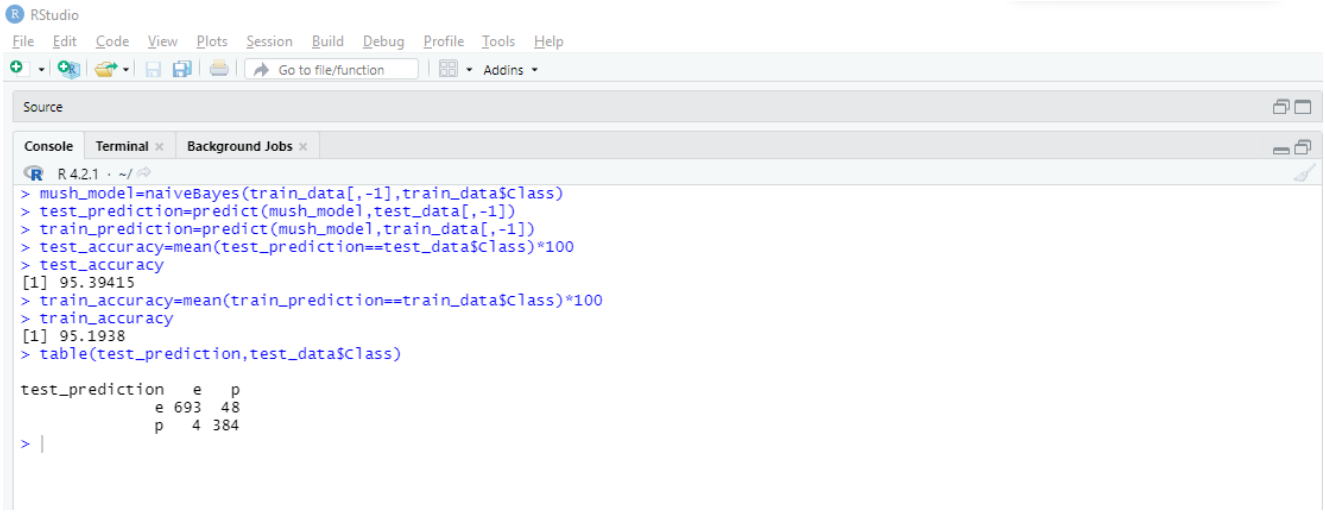
Assignment-2 Shraddha



Prediction of testing and training data



Use the table function to create a confusion matrix of predicted vs. actual classes - how many false positives did the model produce?

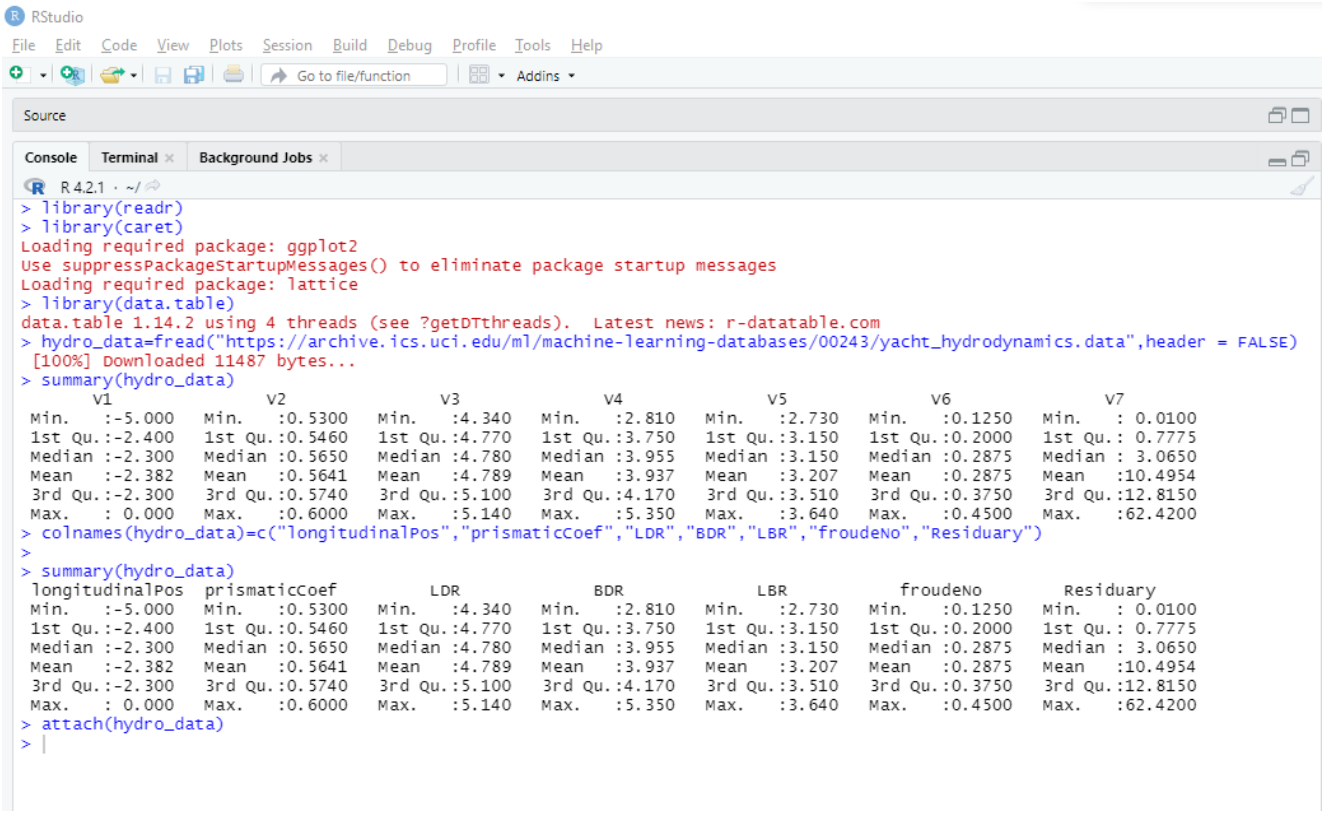


The result in above image shows TP = 693, FP = 48, FN = 4 and TN = 384

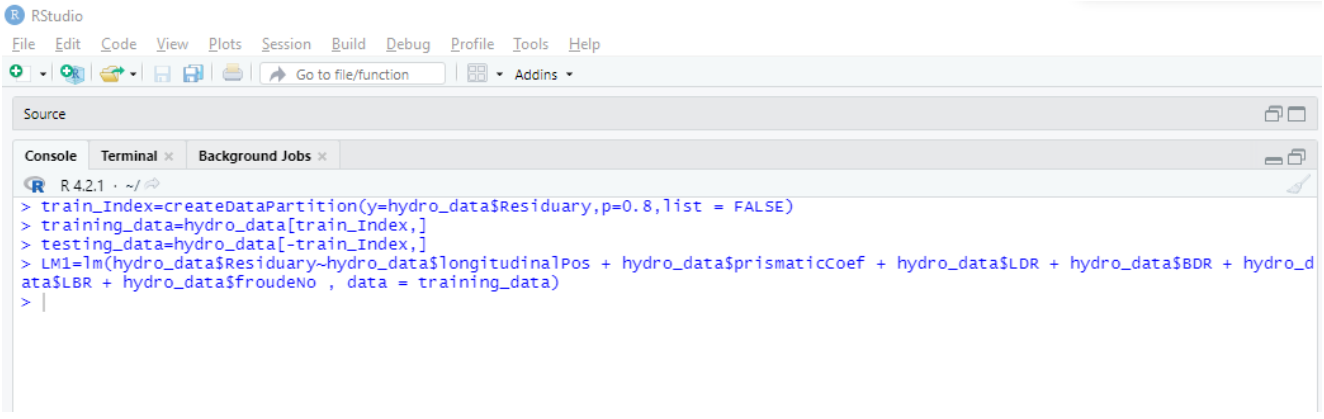
Assignment-2 Shraddha

Problem-3

Load the Yacht Hydrodynamics sample dataset from the UCI Machine Learning Repository into R using Dataframe.

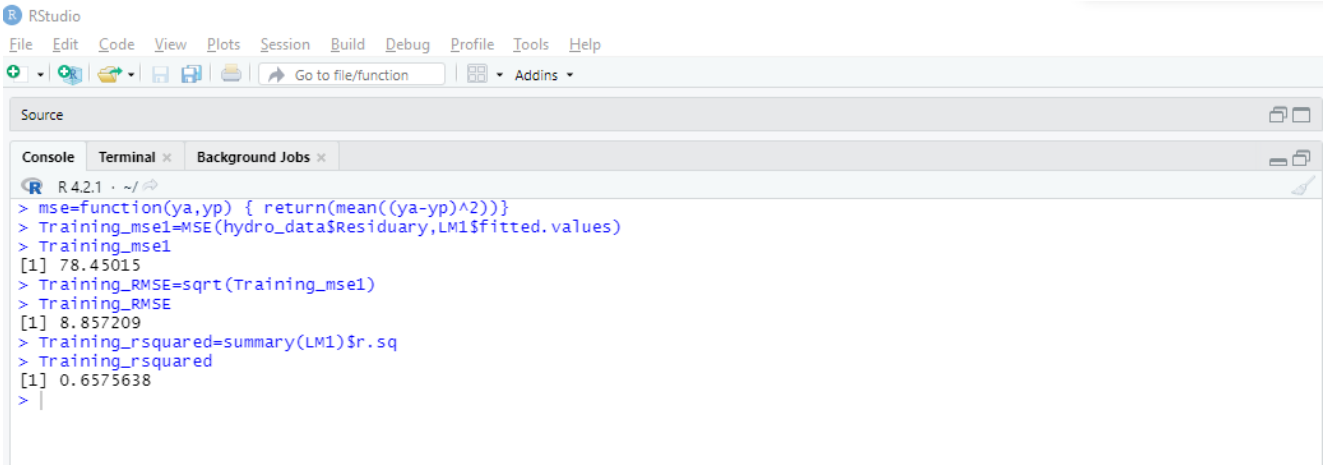


Use the caret package to perform a 80/20 test-train split (via the createDataPartition function), and obtain a training fit for a linear model

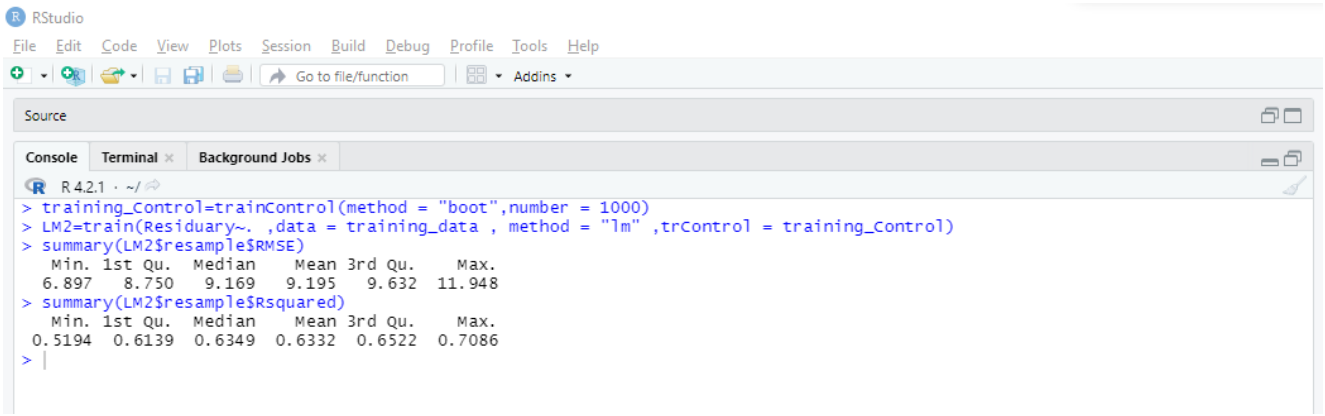


What is the training MSE/RMSE and R2 results?

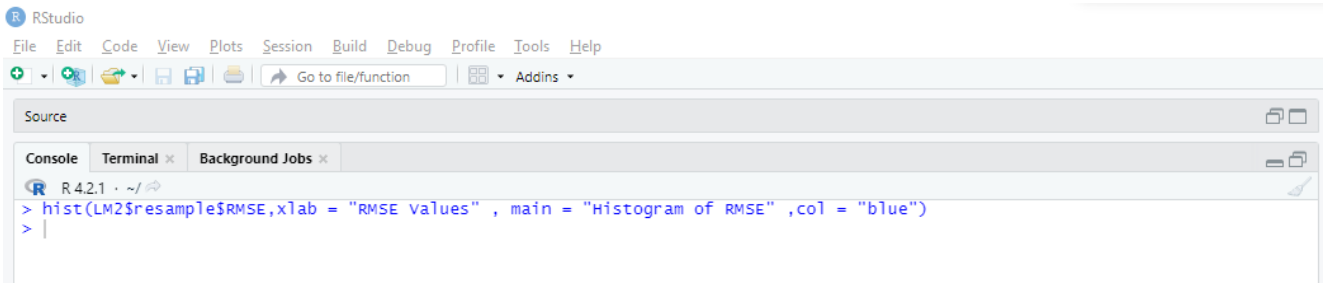
Assignment-2 Shraddha



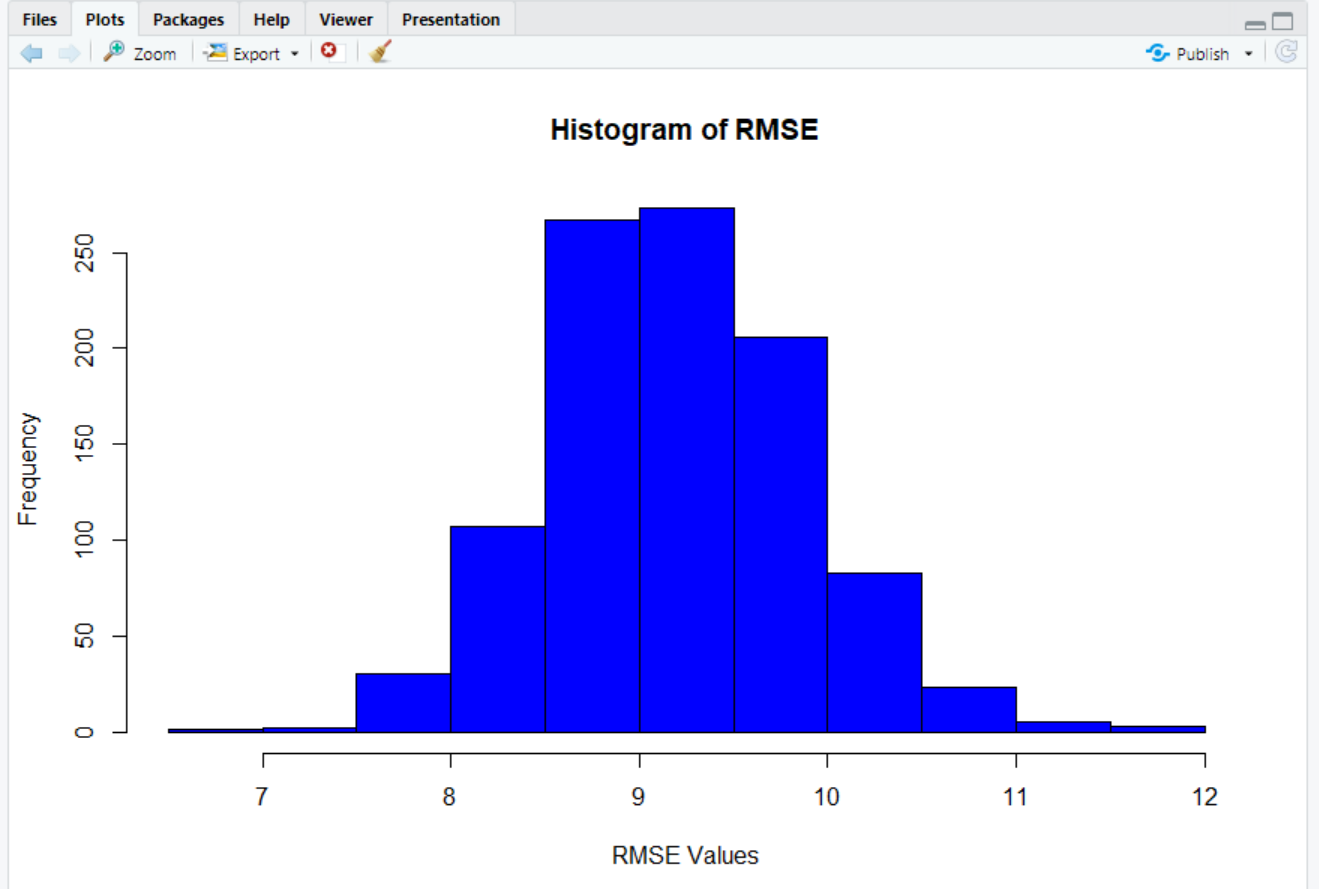
Next, use the caret package to perform a bootstrap from the full sample dataset with N=1000 samples for fitting a linear model (via the train Control method), resulting in a training MSE/RMSE and R2 for each resample.



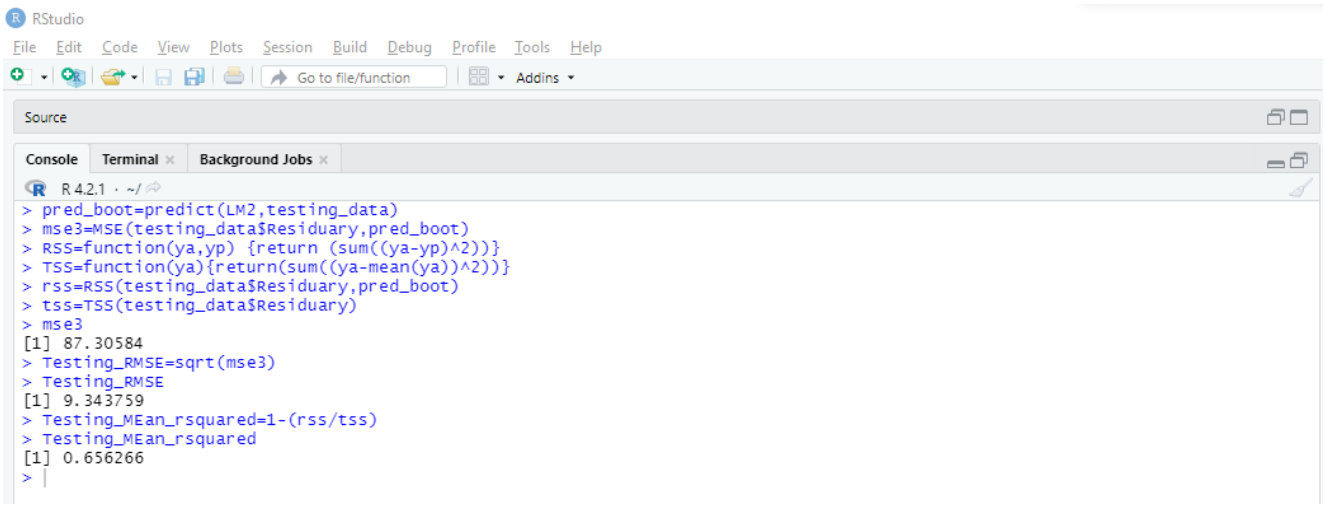
Plot a histogram of the RMSE values and provide a mean RMSE and R2 for the fit.



Assignment-2 **Shraddha**



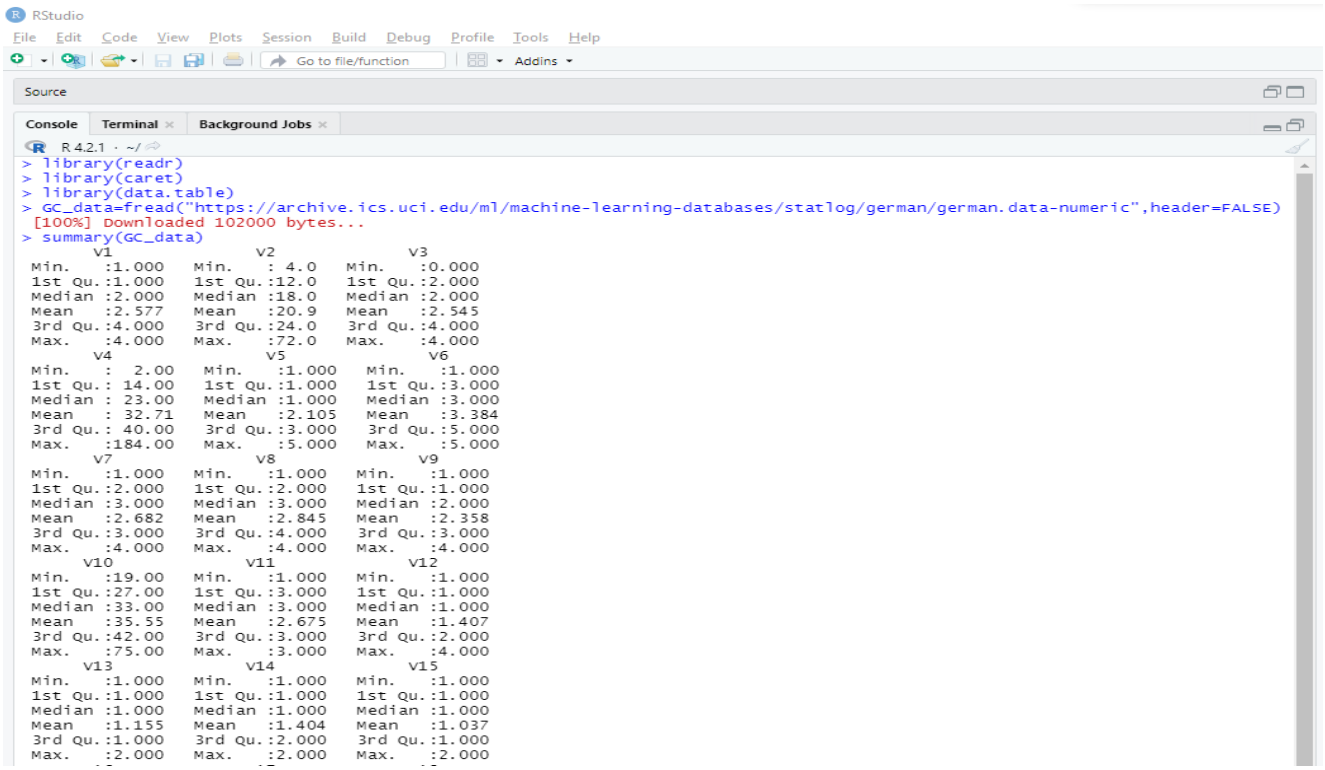
Assignment-2 Shraddha



The result of testing data set shows there is no difference between the actual and bootstrap models.

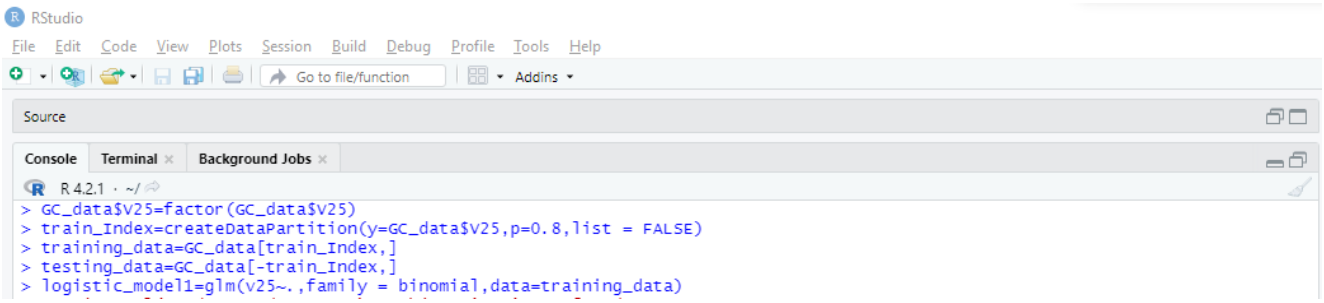
Problem 4

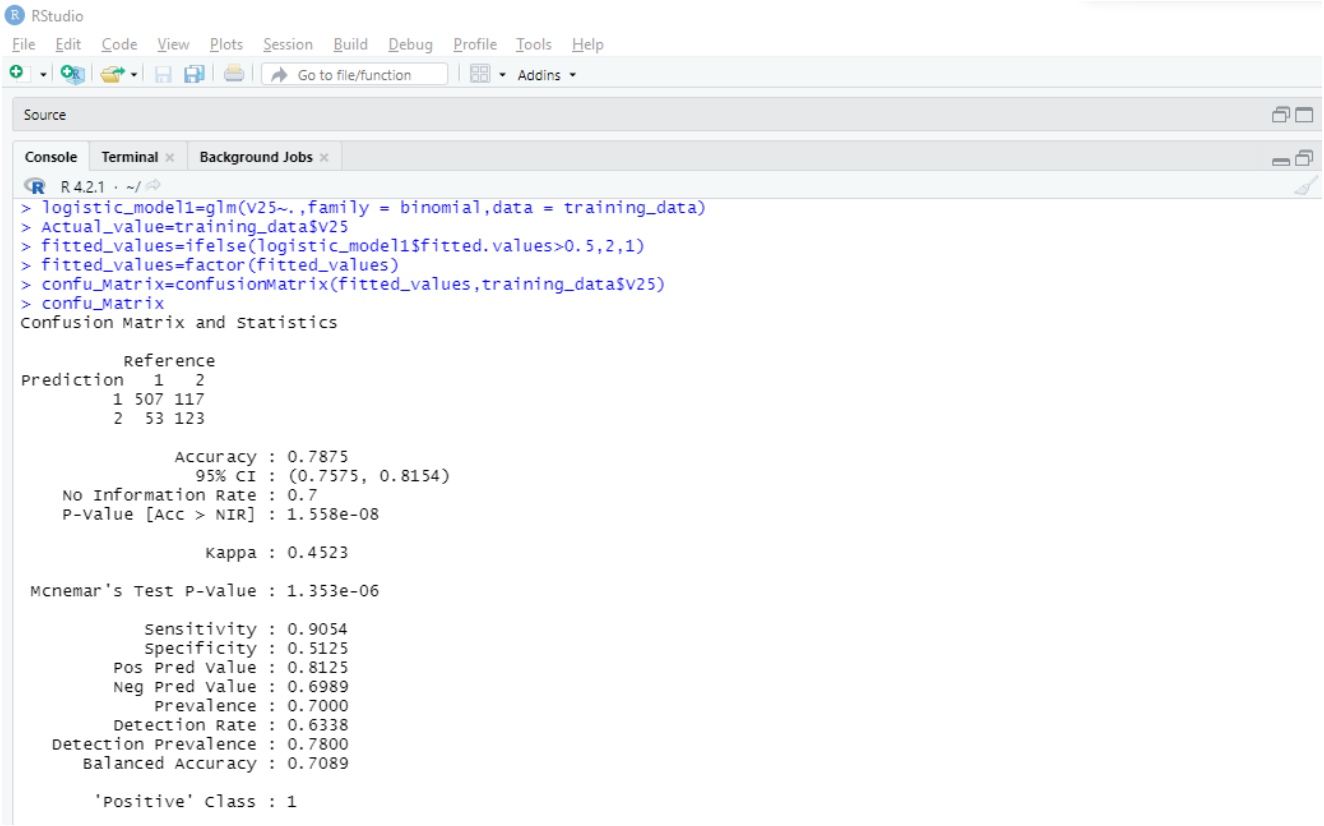
Load the German Credit Data sample dataset from the UCI Machine Learning Repository (German. Data-numeric) into R using a dataframe.



Assignment-2 Shraddha

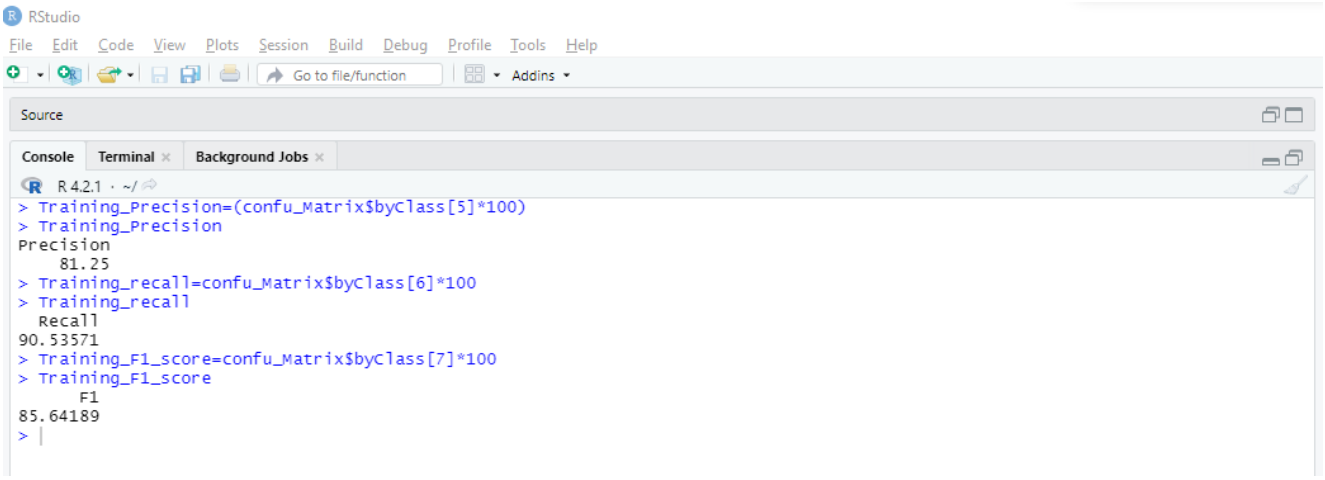
Perform a 80/20 test-train split (via the createDataPartition function), and obtain a training fit for a logistic model via the glm package.

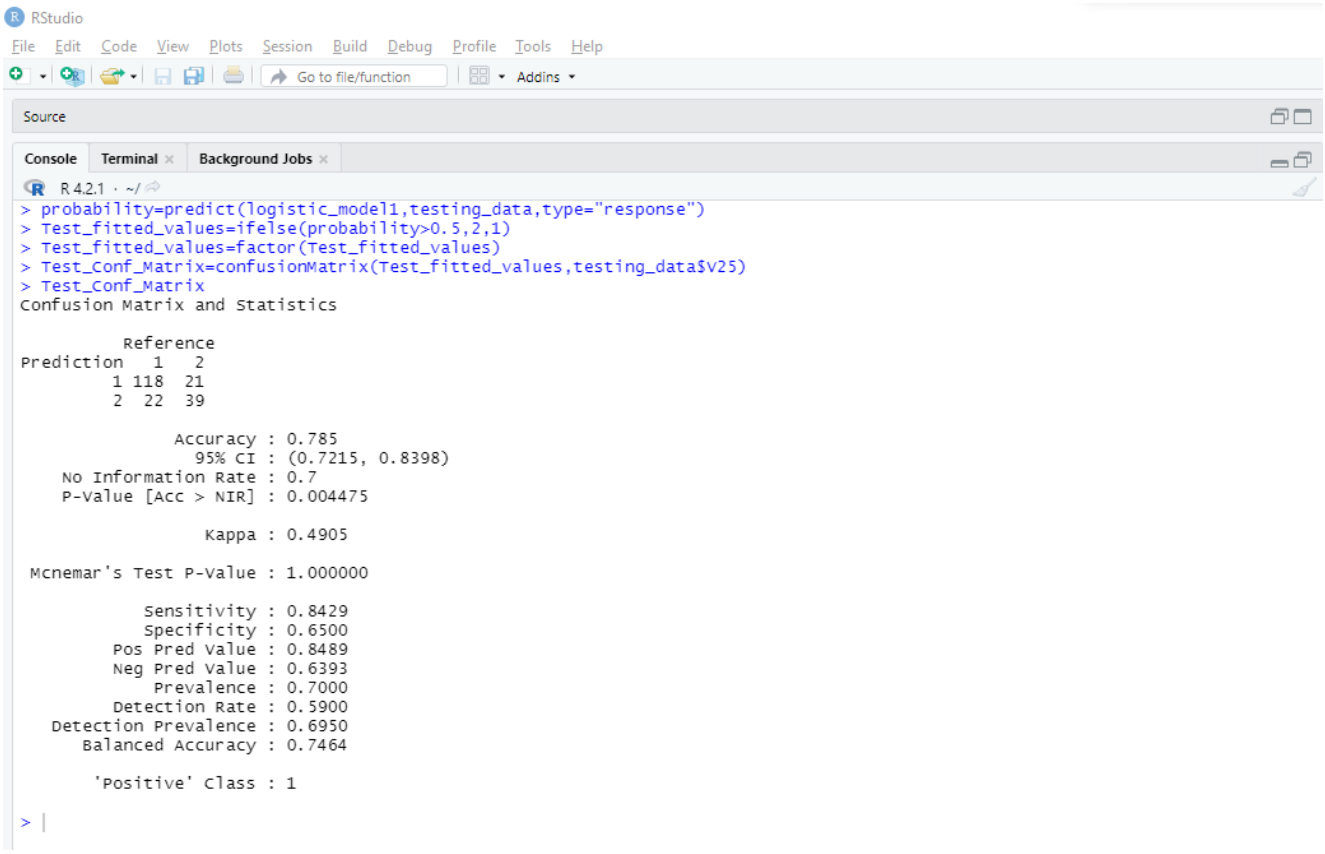




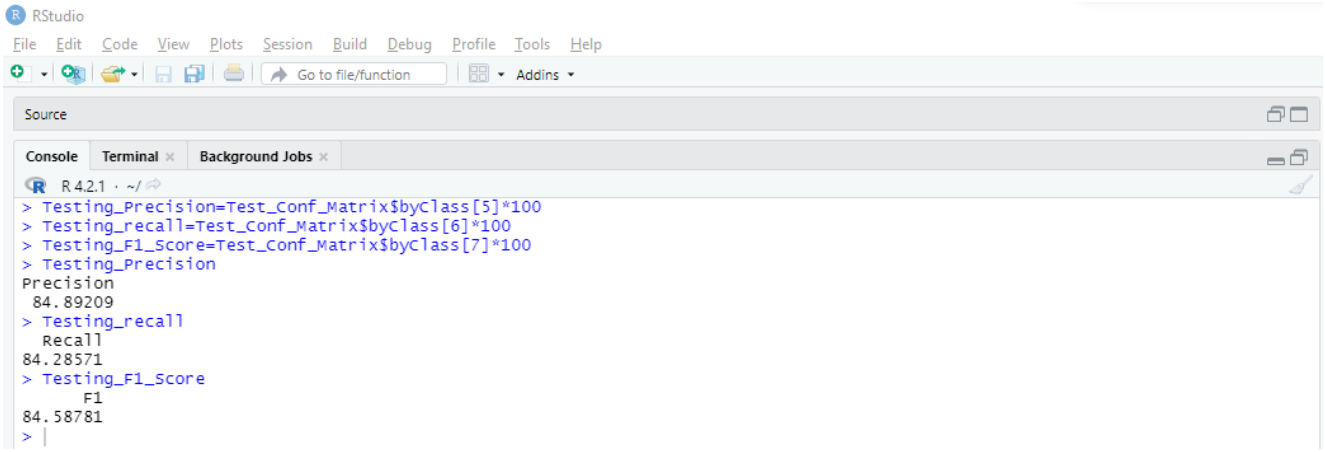
What is the training Precision/Recall and F1 results?

Assignment-2 Shraddha

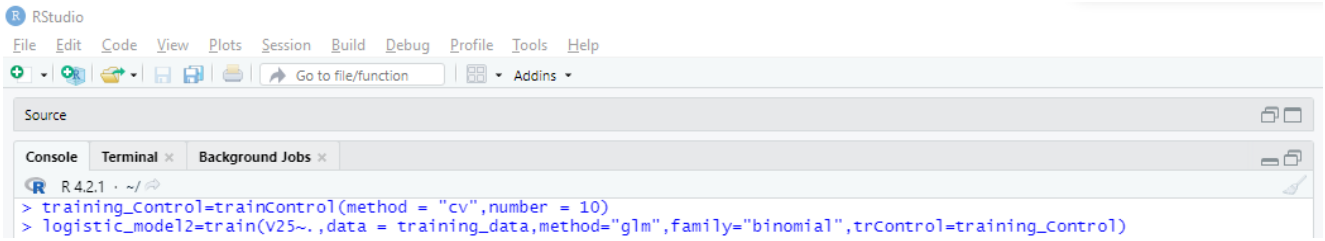




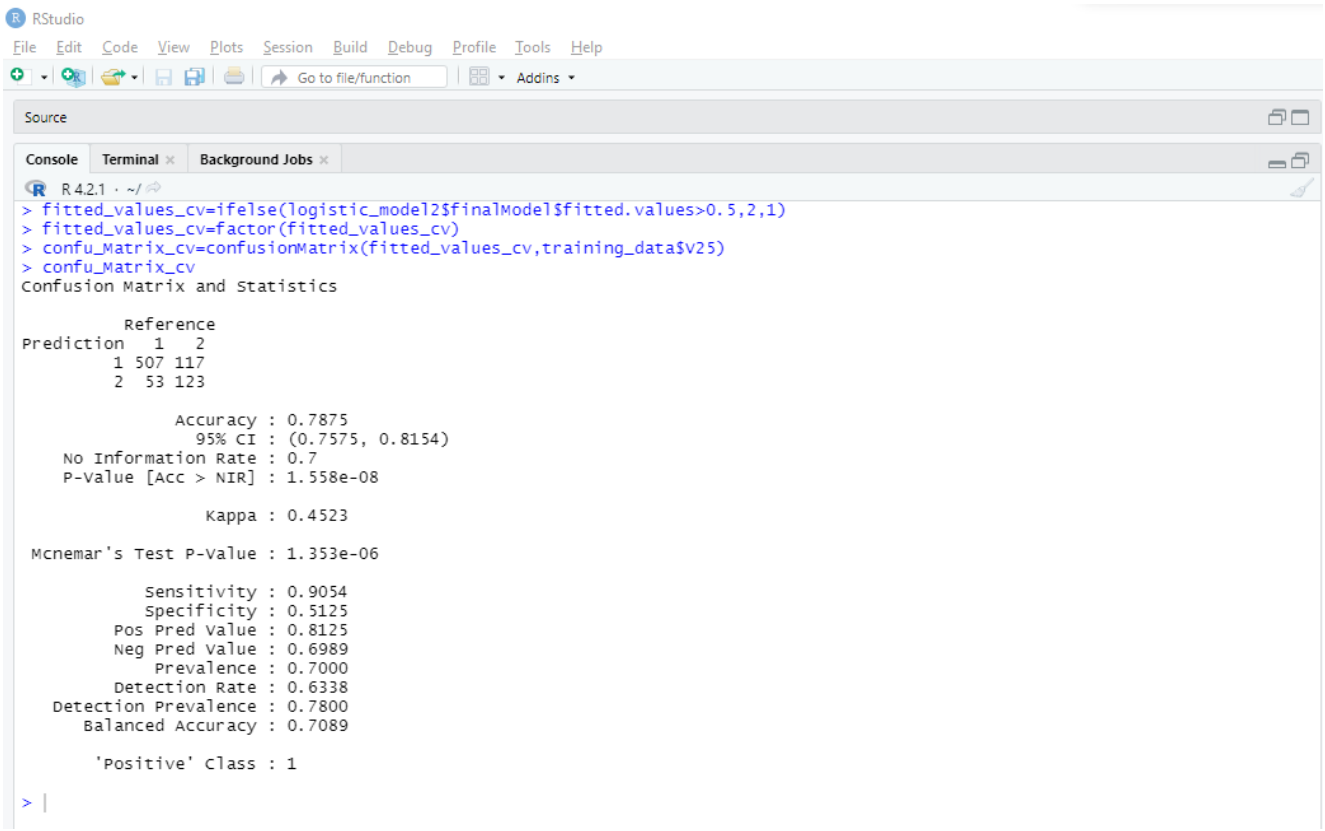
Assignment-2 Shraddha

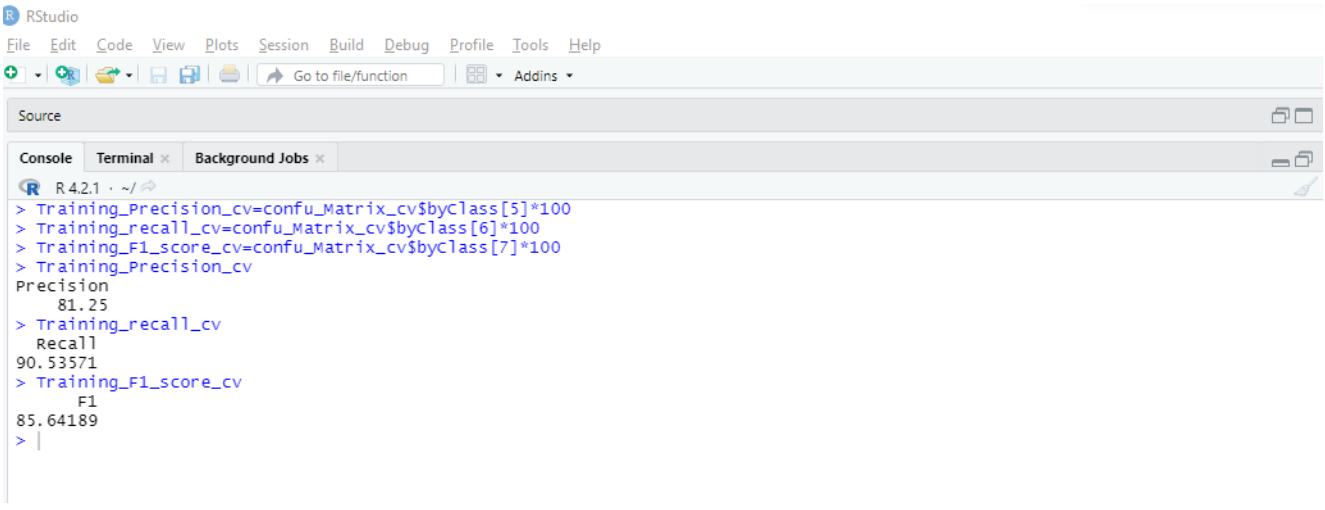


use the trainControl and train functions to perform a k=10 fold cross-validation fit of the same model, and obtain cross-validated training Precision/Recall and F1 values

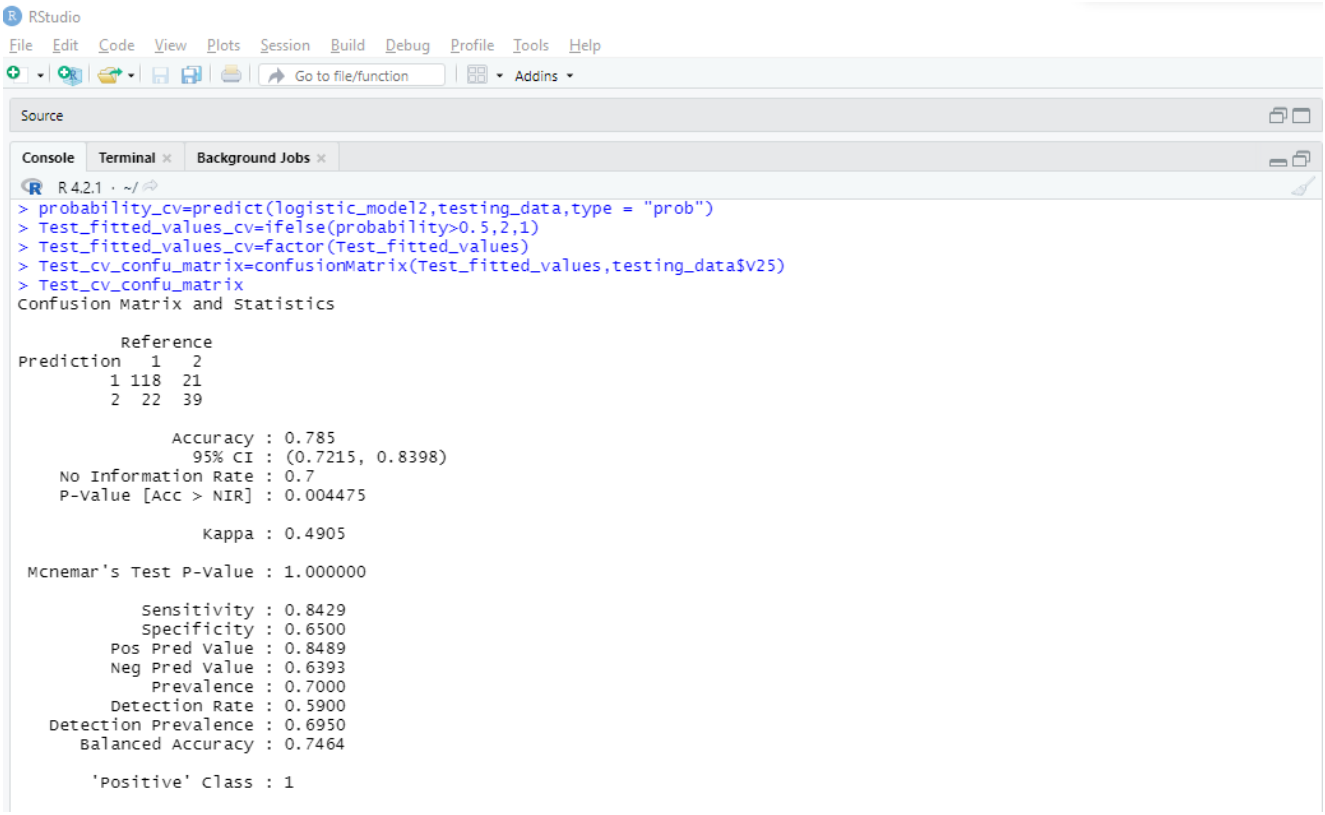


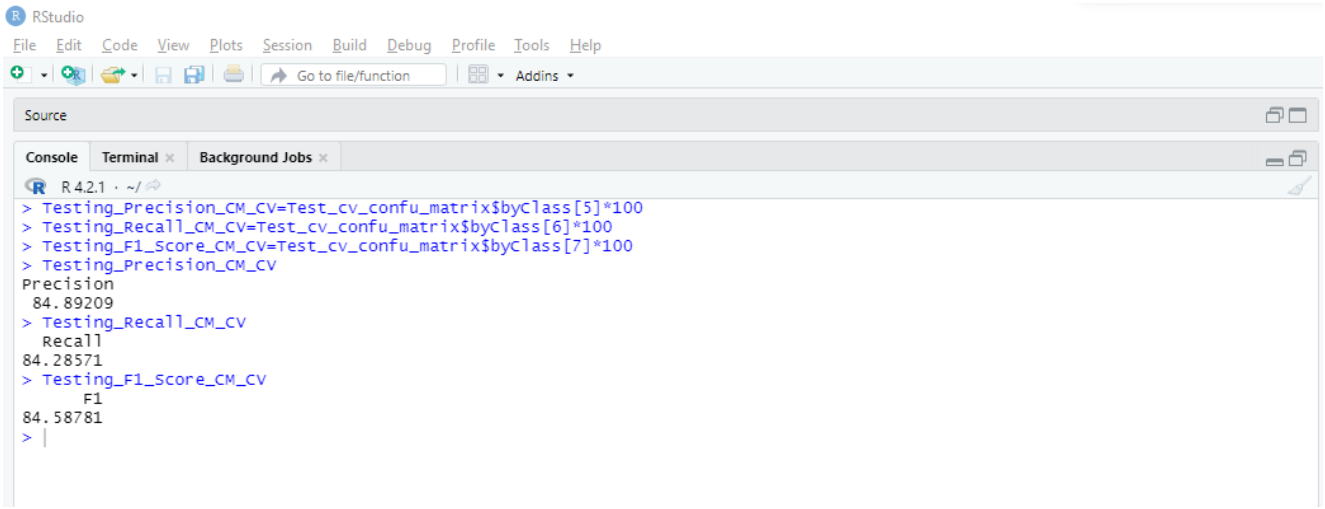
Assignment-2 Shraddha





Assignment-2 Shraddha





By looking at testing data we can say that there is no difference between original and bootstrap models.