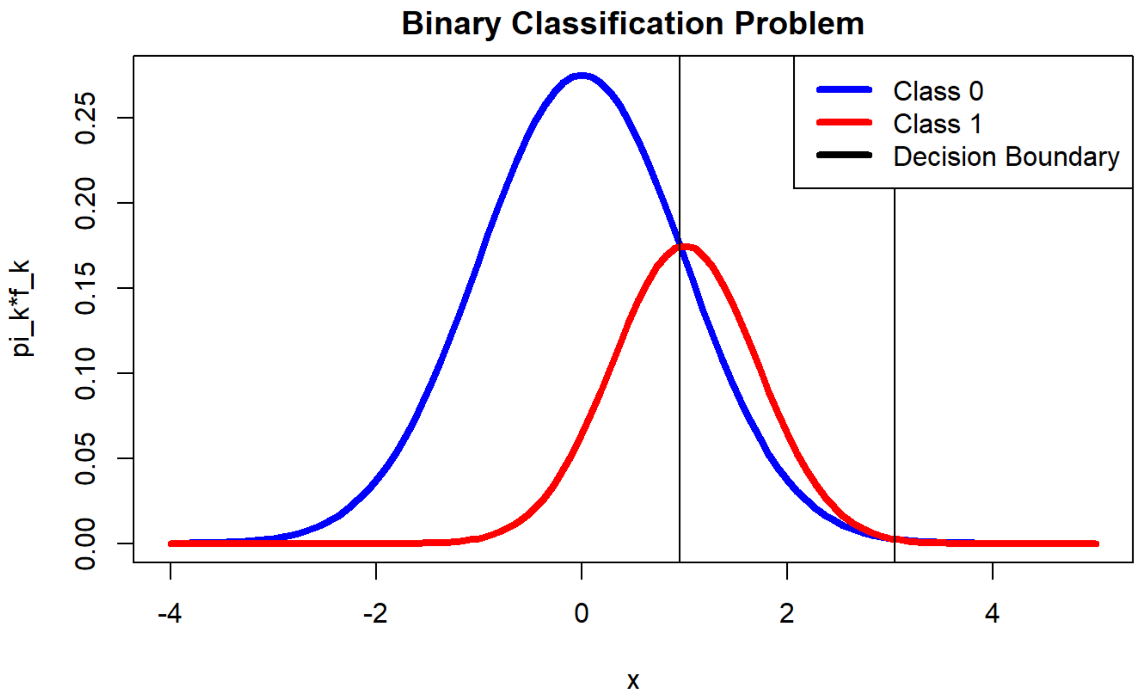
STAT462 – Assignment 1

NDU31

NICOLE DUNN

Question 1

(a) Plot Pi$_0$f$_0$(x) and Pi$_1$f$_1$(x) in the same figure.
(b) Find the Bayes decision boundary (Hint: Pi$_0$f$_0$(x) = Pi$_1$f$_1$(x) on the boundary).
(c) Using Bayes classifier, classify the observation X = 3. Justify your prediction.
(d) What is the probability that an observation with X = 2 is in class 1?



a & b)

Bayes Decision Boundary 1:

x = 0.9545803

Bayes Decision Boundary 2:

x = 3.045439

c)

Probability of Class 0 at x = 3:

0.4883885

Probability of Class 1 at x = 3:

0.5116115

With these calculations provided from R, I can classify x = 3 in Class 1. This is because the probability of x = 3 at Class 1 is higher than that of the probability of x = 3 at Class 0. Visually we can also see in the graph and with the numerical values for the Bayes Decision boundary that x = 3 falls within the two boundaries to classify Class 1.

d)

Probability of Class 1 at x = 2:

0.6333126

Continuing with the same equation modified for x =2, we find that x = 2 should be classified as Class 1 as it has a majority probability.

Question 2

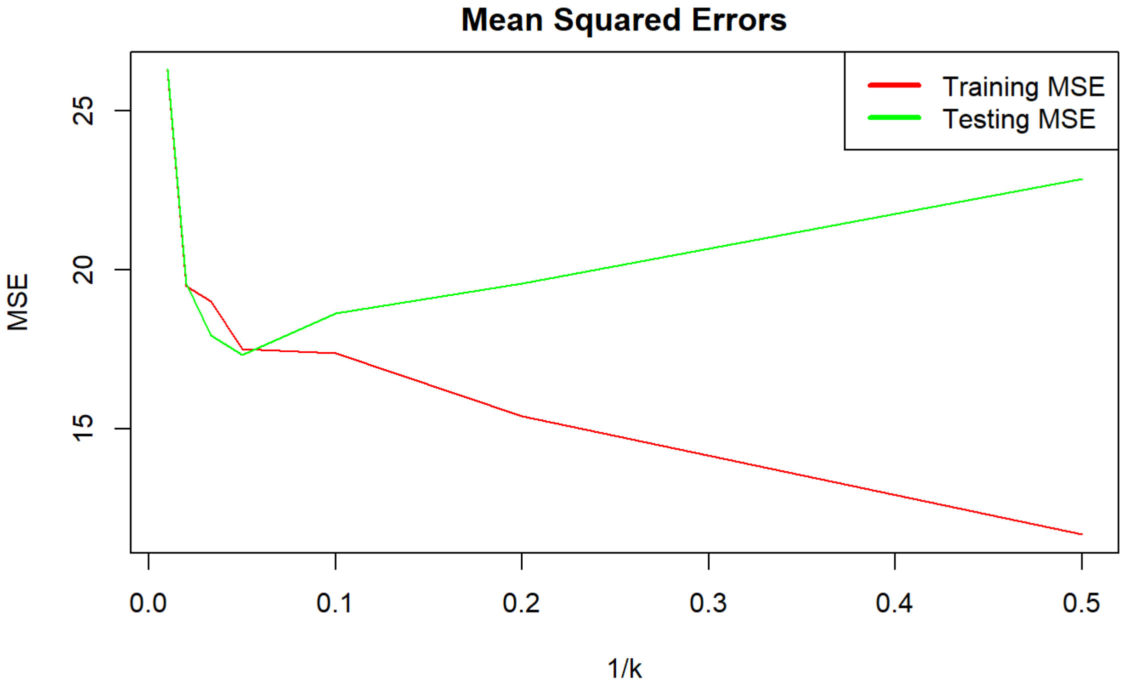
(a) Perform kNN regression with k = 2; 5; 10; 20; 30; 50 and 100, (learning from the training data) and compute the training and testing MSE for each value of k.
(b) Which value of k performed best? Explain.
(c) Plot the training data, testing data and the best kNN model in the same figure.
(The points() function is useful to plot the kNN model because it is discontinuous.)
(d) Describe the bias-variance trade-off for kNN regression.

a)

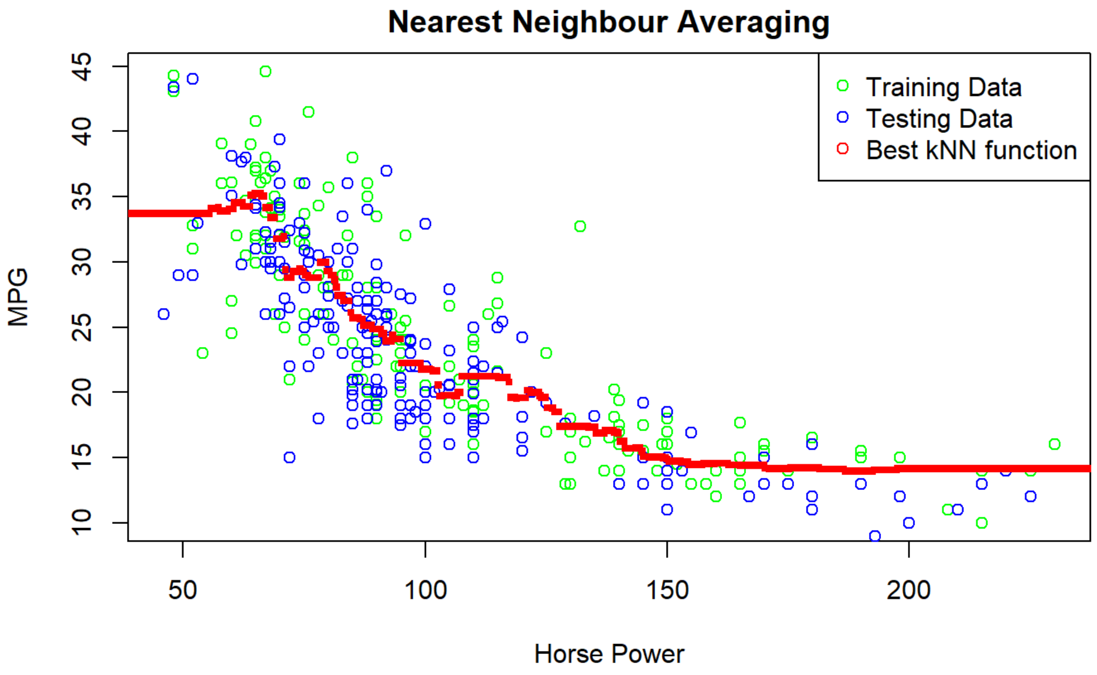
k	MSE_test	MSE_train
2	22.86349	11.67317
3	19.56322	15.39669
10	18.62914	17.38083
20	17.31858	17.49457
30	17.93018	18.99924
50	19.57374	19.47530
100	26.31542	26.25969

b)

The value of the given k's that performed the best would be k = 20. As we can see from the above table, it has the combined lowest MSE's compared to all the rest of the values of k. The testing mean squared error gives a much better indication of how well the model is performing though, and we can see from these values that k = 20 has the lowest MSE within the test column.



c)



d)

As the flexibility of $\hat{f}$ increases, its bias decreases and its variance increases – this is a bias variance trade off. This is seen both within the above graph and table. As the k value increases, we can see MSE_test start to decrease before starting to increase again after k = 20. We can tell that this would mean the model is underfitting when k < 20 and over fitting when k > 20. Underfitting means the model has high bias and low variance and overfitting means it has low bias and high variance.

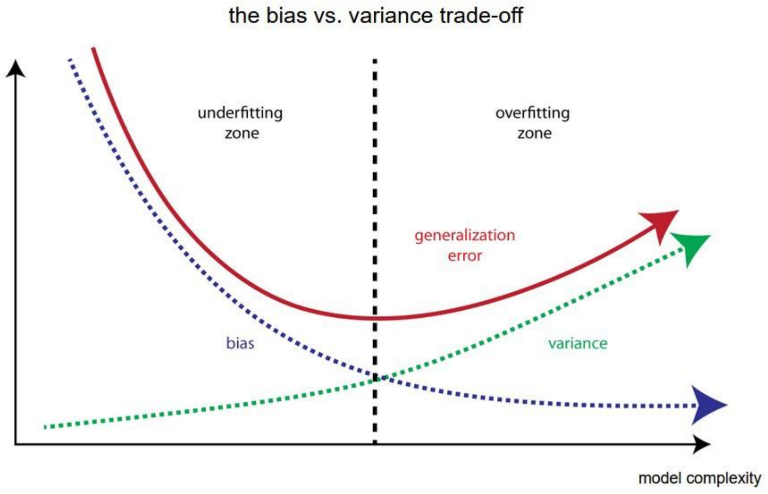


Figure 1 https://towardsdatascience.com/bias-and-variance-but-what-are-they-really-ac539817e171

Question 3

(a) Estimate the probability of a student getting a GPA value >= 7 in STAT318 if they study for 5 hours per week and attend all 36 classes.
(b) If a student attends 18 classes, how many hours do they need to study per week to have a 50% chance of getting a GPA value >= 7 in STAT318?

a)

Probability of GPA >= 7:

0.8581489

b)

How many hours needed to have a 50% chance of achieving a GPA >= 7:

log(0.5/(1-0.5)) = -16 + 1.4x1 + 0.3*18

x = 7.57142857143

7.6 hours

Question 4

(a) Perform multiple logistic regression using the training data. Comment on the model obtained.

(b) Suppose we classify observations using

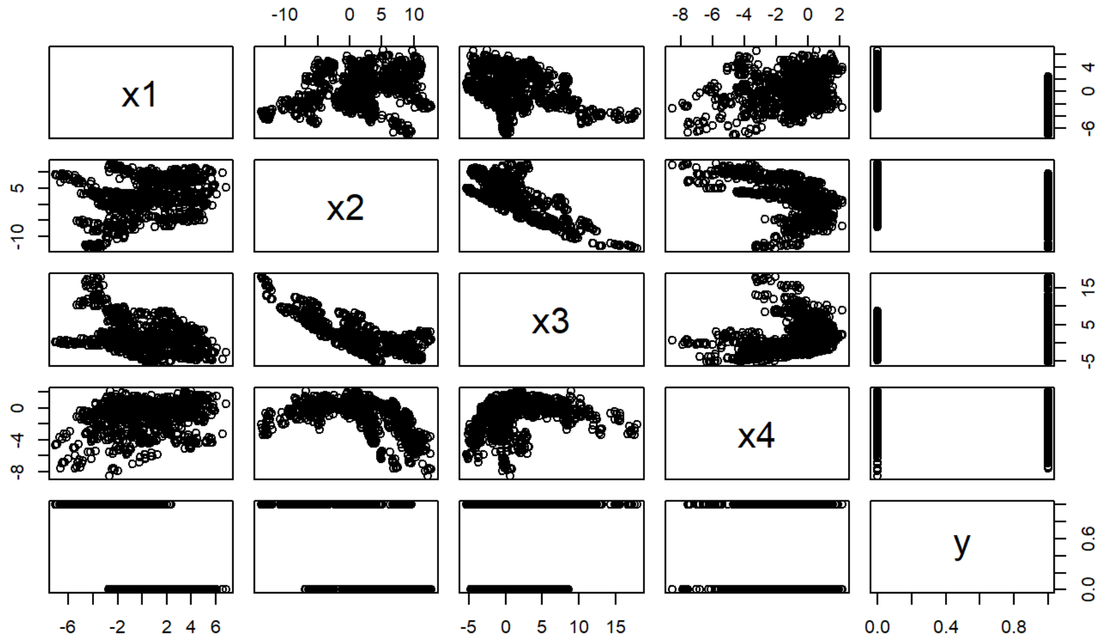
$$f(x) = \begin{cases} \textit{forged banknote} & \text{if } \Pr(Y = 1|X = x) > \theta \\ \textit{genuine banknote} & \text{otherwise.} \end{cases}$$

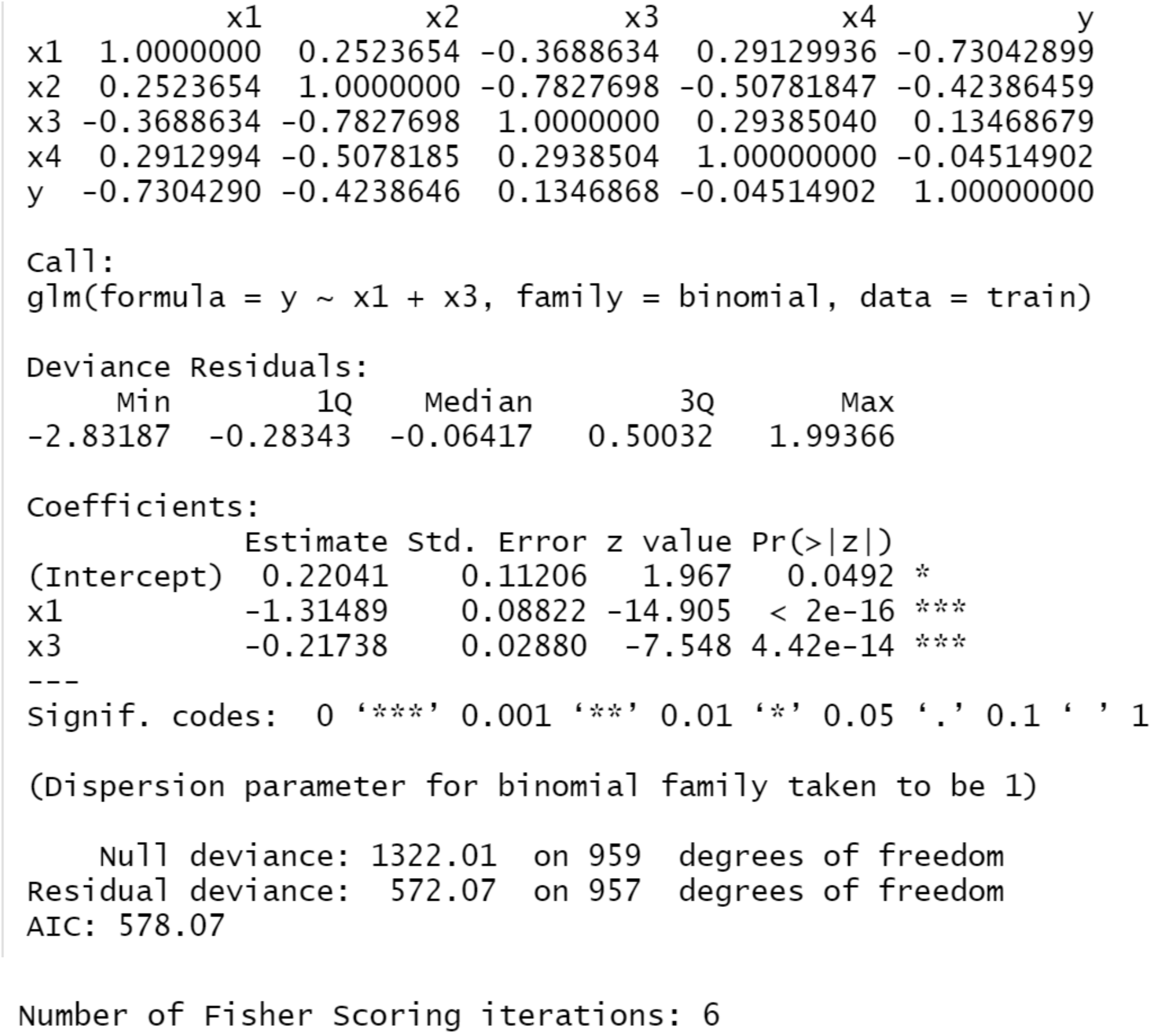
i. Plot the training data (using a different symbol for each class) and the decision boundary for θ = 0.5 on the same figure.

ii. Using θ = 0.5, compute the confusion matrix for the testing set and comment on your output.

iii. Compute confusion matrices for the testing set using θ = 0.3 and θ = 0.6. Comment on your output. Describe a situation when the θ = 0.3 model may be the preferred model.

a)





```
           x1         x2         x3          x4           y
x1  1.0000000  0.2523654 -0.3688634  0.29129936 -0.73042899
x2  0.2523654  1.0000000 -0.7827698 -0.50781847 -0.42386459
x3 -0.3688634 -0.7827698  1.0000000  0.29385040  0.13468679
x4  0.2912994 -0.5078185  0.2938504  1.00000000 -0.04514902
y  -0.7304290 -0.4238646  0.1346868 -0.04514902  1.00000000

Call:
glm(formula = y ~ x1 + x3, family = binomial, data = train)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.83187  -0.28343  -0.06417   0.50032   1.99366  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.22041    0.11206   1.967   0.0492 *  
x1          -1.31489    0.08822 -14.905  < 2e-16 ***
x3          -0.21738    0.02880  -7.548 4.42e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1322.01  on 959  degrees of freedom
Residual deviance:  572.07  on 957  degrees of freedom
AIC: 578.07

Number of Fisher Scoring iterations: 6
```

The estimated slope parameters are significant for both x1 and x3. Doing the below calculations with them shows that both have a higher chance of being genuine, but x1 has a much higher chance to be genuine than forged.

$$\frac{\exp(0.22041 - 1.31489(1))}{1 + \exp(0.22041 - 1.31489(1))} = 0.2507756043$$

$$\frac{\exp(0.22041 - 1.31489(0))}{1 + \exp(0.22041 - 1.31489(0))} = 0.5548805025$$

$$\frac{\exp(0.22041 - 0.21738(1))}{1 + \exp(0.22041 - 0.21738(1))} = 0.5007574994$$

$$\frac{\exp(0.22041 - 0.21738(0))}{1 + \exp(0.22041 - 0.21738(0))} = 0.5548805025$$

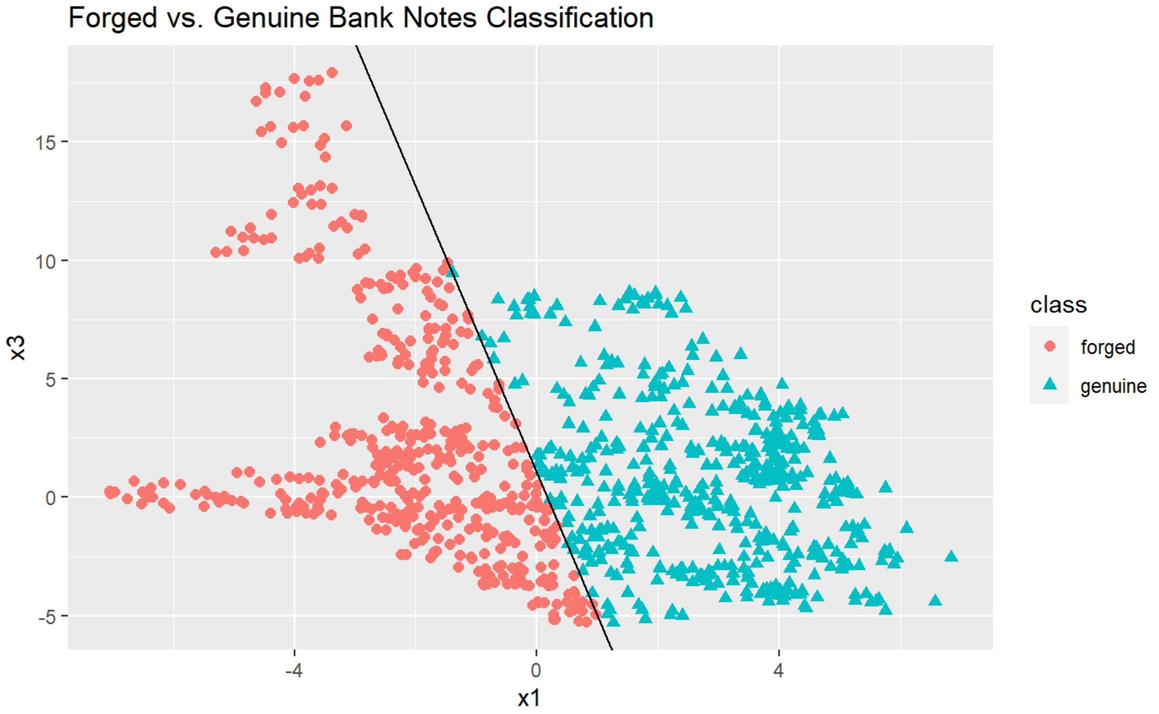
The standard error tells us how accurate the mean of our sample from that population is likely to be compared to the true population mean. When the standard error increases, it becomes more likely that any given mean is an inaccurate representation of the true population mean. Because of how small our standard errors are, we can see assume that our accuracy is relatively high.

The z-scores show that x1 is 14.9 standard deviations below the mean average and x3 is 7.5 below.

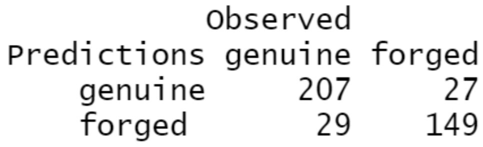
The p-value for both x1 and x3 are extremely small. Since the values are less than 0.1, we can say that both statistics are statistically significant predictor variables in the model.

To determine if a model is "useful" we can compute the Chi-Square statistic as X^2 = Null deviance – Residual deviance with p degrees of freedom.
$X^2 = 1322.01 - 572.07$
$X^2 = 749.94$
There are p = 2 predictor variables degrees of freedom. The P-Value is < .00001 with this Chi-Square score which shows that the model is very "useful."

b)
i.



ii.



```
              Observed
Predictions genuine forged
    genuine     207     27
    forged       29    149
```

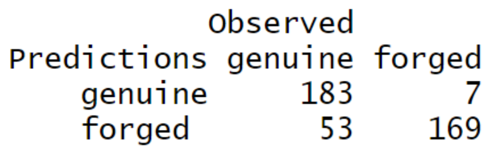
theta = 0.5

From this confusion matrix, we can see that we have 207 "true positive" genuine responses and 149 "true negative" forged responses. There are 27 "false positive" genuine responses and 29 "false negative" responses in contrast. From this we can calculate the recall, precision and accuracy of the results. Recall tells us how many notes we predicted correctly, precision tells us from how many notes we have predicted, how many are their true value and accuracy tells us how many we have predicted correctly. All these values should be as high as possible for an accurate model.

Recall.g = TP / TP + FN
= 207 / 207 + 29

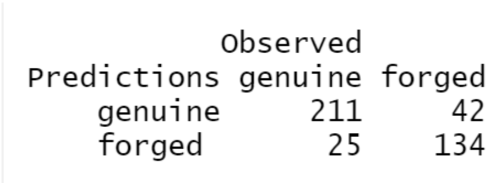
= 0.877
Precision.g = TP / TP + FP
= 207 / 207 +27
=0.885
Recall.f = TN / TN + FP
= 149 / 149 + 27
= 0.847
Precision.f = TN / TN + FN
= 149 / 149 +29
=0.837
Accuracy = TP + TN / All possibilities
= 207 + 149 / 207+149+27+29
=0.864
From all this we can see that our model has performed very well.

iii.



```
            Observed
Predictions genuine forged
    genuine     183      7
    forged       53    169
```

theta = 0.3



```
            Observed
Predictions genuine forged
    genuine     211     42
    forged       25    134
```

theta = 0.6

For both matrices I have calculated the recall, precision, accuracy and f-measure, these are as follows:

Theta	0.3	0.6
Recall.g	0.775	0.894
Precision.g	0.963	0.834
Recall.f	0.960	0.761
Precision.f	0.761	0.843
Accuracy	0.854	0.837
F-measure.g	0.859	0.863
F-measure.f	0.849	0.8

As we can see both models performed slightly worse with adjusted theta values with either recall or precision falling below the values observed with theta = 0.5 . Some values are much higher in comparison though. Overall, both theta versions still have relatively high accuracy. Adjusting the value of theta to 0.3 makes the model much stricter with its classification. A situation that may suit a stricter theta could be used in a medication setting where dose amount is closely monitored, or in terms of bank notes – how many values are being assessed in terms of forgery.