Statistical Learning - Assignment 2

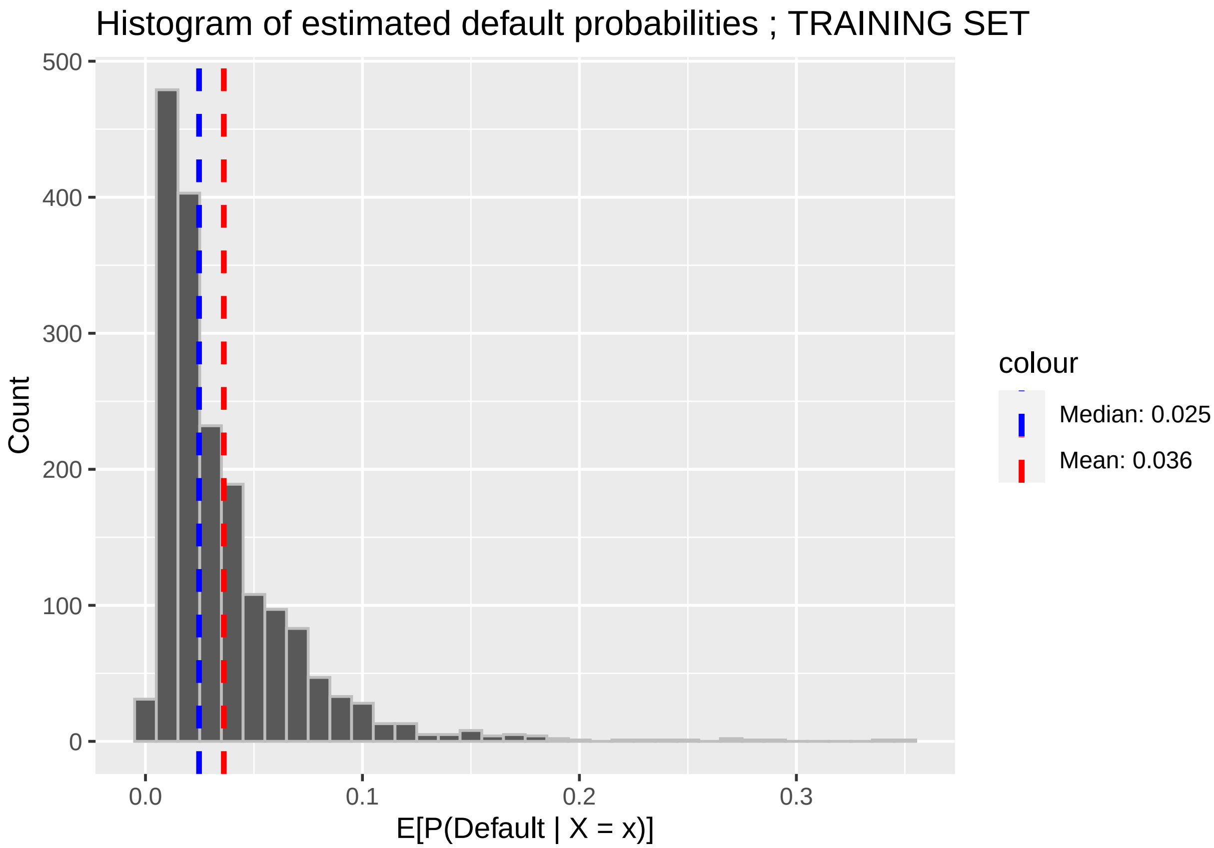
Francis Frégeau

Winter 2021

Q1

a.

```
## [1] Deviance Residuals:
##          0%         25%         50%         75%        100%
## -0.9278069 -0.3002828 -0.2161747 -0.1598389  3.0744514
## [1]
## [1] Coefficients:
##              Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -7.000170  0.6416842 -10.909059 0.000000e+00
## RevenueVol   2.895128  0.7304548   3.963459 7.387159e-05
## DebtRatio    4.018284  0.8306645   4.837433 1.315265e-06
## DiffRate    22.276542  8.5870620   2.594198 9.481177e-03
## [1]
## [1] Deviances:
##          Deviance Degrees of Freedom Pr(> ChiSq)      AIC     AICC      BIC
## Null     559.3741               1799           0 561.3741 561.3719 519.8519
## Residual 512.3564               1796           0 520.3564 520.3341 542.3385
## [1]
## [1] Saturated model deviance: 0
## [1]
## [1] Number of Newton-Raphson iterations: 10
```

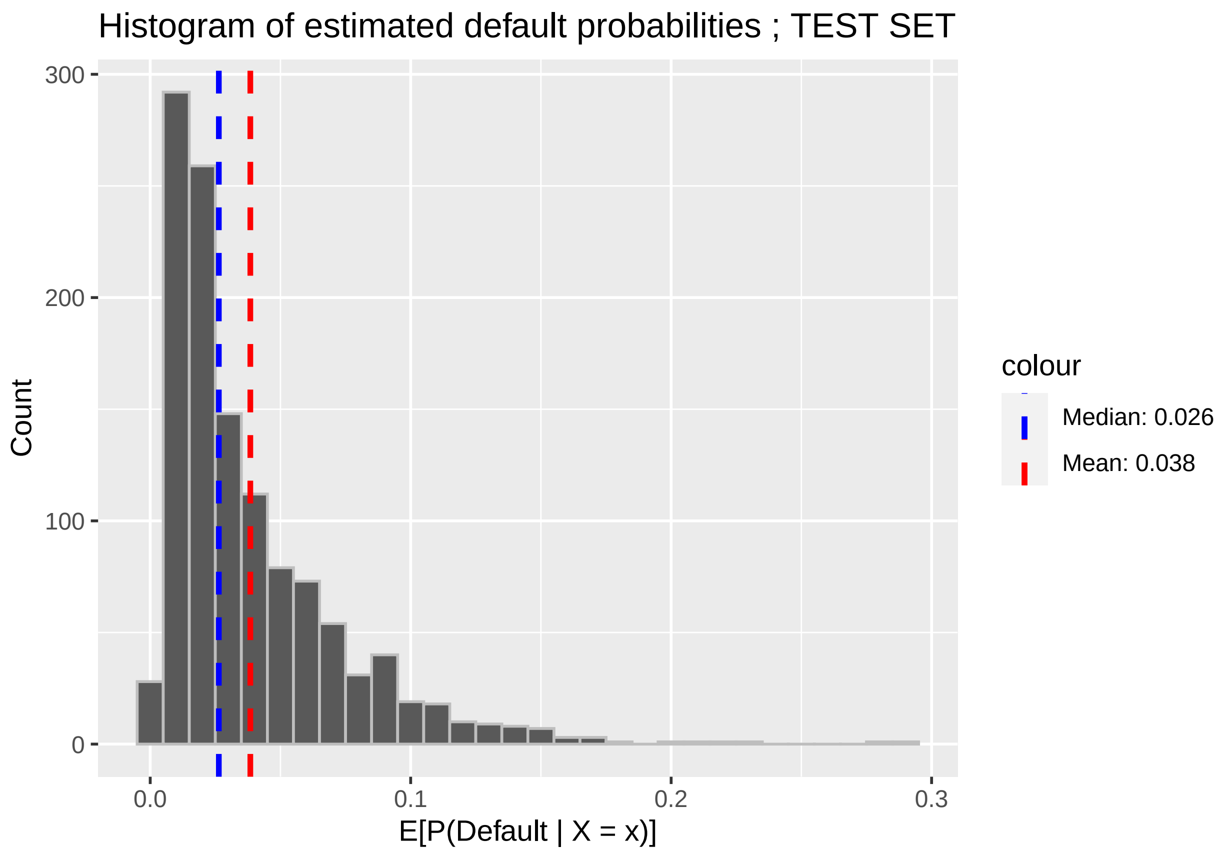

Histogram of estimated default probabilities ; TRAINING SET
Count
500
400
300
200
100
0
0.0
0.1
0.2
0.3
E[P(Default | X = x)]
colour
Median: 0.025
Mean: 0.036

b.

```
##
## Call:
## glm(formula = InDefault ~ ., family = binomial(), data = as.data.frame(TrainData))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9278  -0.3003  -0.2162  -0.1598   3.0745
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.0002     0.6417 -10.909  < 2e-16 ***
## RevenueVol    2.8951     0.7305   3.963 7.39e-05 ***
## DebtRatio     4.0183     0.8307   4.837 1.32e-06 ***
## DiffRate     22.2766     8.5871   2.594  0.00948 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 559.37  on 1799  degrees of freedom
## Residual deviance: 512.36  on 1796  degrees of freedom
## AIC: 520.36
##
## Number of Fisher Scoring iterations: 7
```

The results are virtually the same.

c.


Histogram of estimated default probabilities ; TEST SET
300
200
100
0
Count
0.0
0.1
0.2
0.3
E[P(Default | X = x)]
colour
Median: 0.026
Mean: 0.038

d.

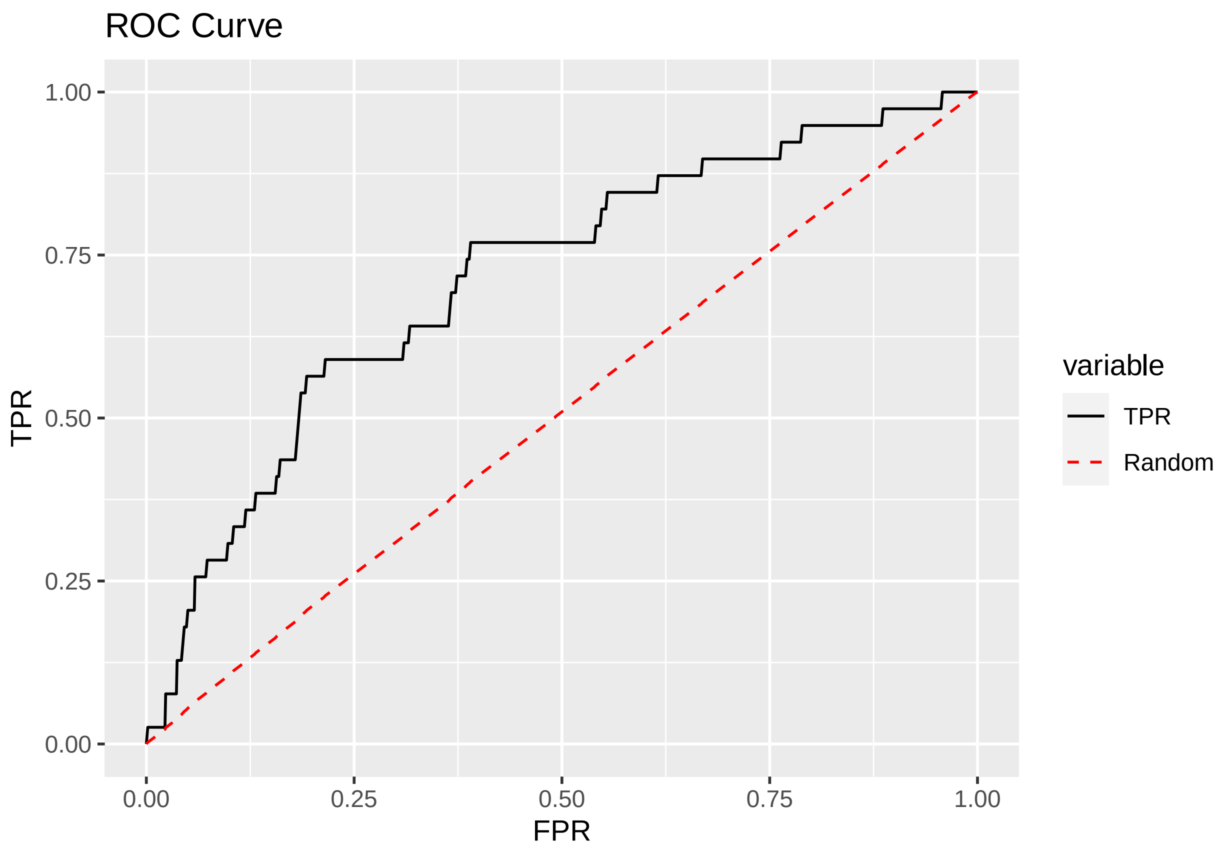
```
## $`1`
## $`1`$CM
##        1_Pred 0_Pred
## 1_True      0     65
## 0_True      0   1735
##
## $`1`$stats
##             Value
## Accuracy    0.964
## Sensitivity 0.000
## Specificity 1.000
## Precision     NaN
## NPV         0.964
## F1            NaN
## Kappa       0.000
##
##
## $`20`
## $`20`$CM
##        1_Pred 0_Pred
## 1_True     30     35
## 0_True    318   1417
##
## $`20`$stats
##             Value
## Accuracy    0.804
## Sensitivity 0.462
## Specificity 0.817
## Precision   0.086
## NPV         0.976
## F1          0.145
## Kappa       0.090
##
##
## $`50`
## $`50`$CM
##        1_Pred 0_Pred
## 1_True     60      5
## 0_True   1040    695
##
## $`50`$stats
##             Value
## Accuracy    0.419
## Sensitivity 0.923
## Specificity 0.401
## Precision   0.055
## NPV         0.993
## F1          0.103
## Kappa       0.037
```

e.

No. The dataset is too imbalanced, and as such, a very high level of accuracy can be achieved solely by setting $\forall x : \mathbb{E}\left[P\left(Y=1|X=x\right)\right]=0$. It would probably be best to use F1 since it is greatly affected by precision. The AUC could also be a good alternative since it will be penalized by both low sensitivity and specificity.

f.

```
## $stats
##          Value
## AUC  0.7135758
## Var  0.0891093
## Gini 0.4271517
##
## $plot
```



Q2

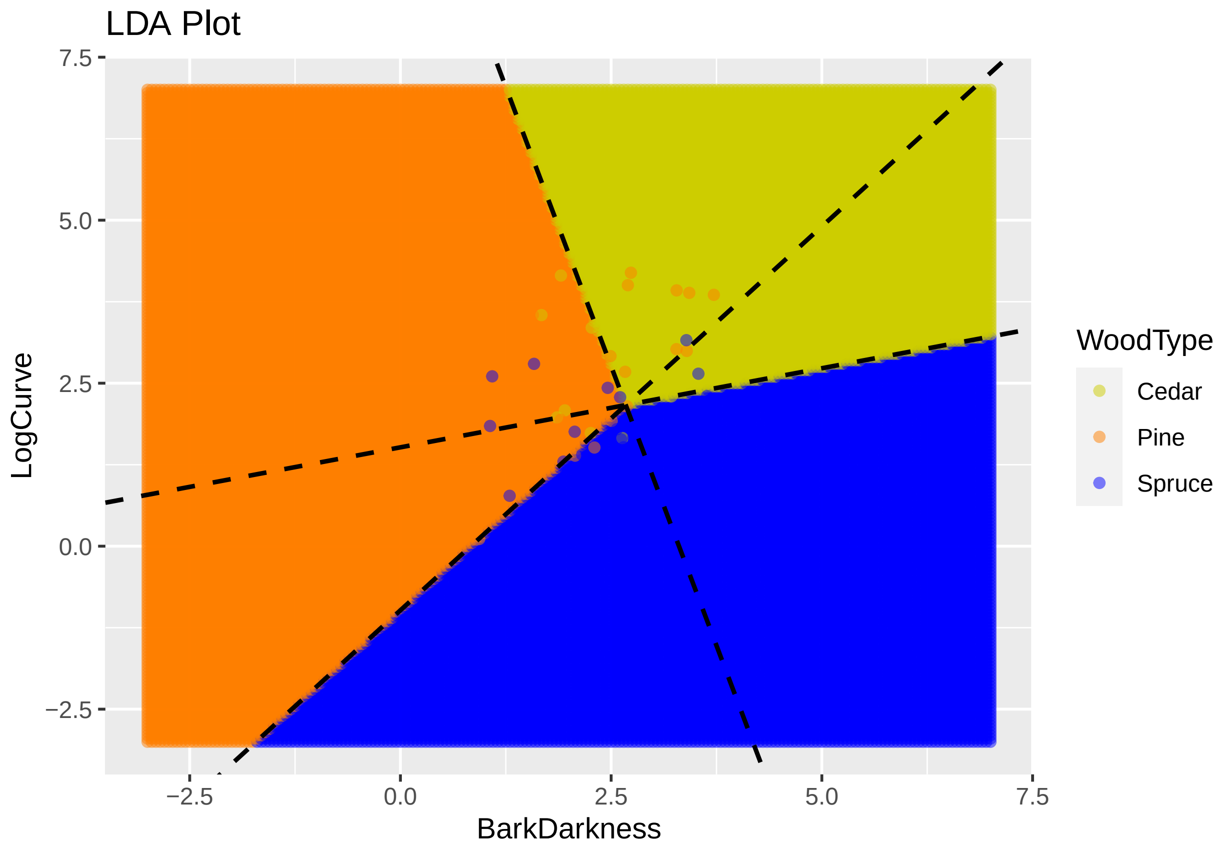
a.

```
## [1] Common covariance matrix:
## [1]
##              BarkDarkness  LogCurve
## BarkDarkness    0.9194027 0.2920257
## LogCurve        0.2920257 0.8180154
## [1]
## [1] Means accross classes:
## [1]
##    WoodType BarkDarkness  LogCurve  n     P
## 1:    Cedar     3.240607 3.2671749 43 0.215
## 2:     Pine     1.106614 2.1398566 89 0.445
## 3:   Spruce     3.030018 0.9863158 68 0.340
## [1]
## [1] In-sample accuracy: 0.845

## [1] Prediction for BarkDarkness = LogCurve = 3:

##        Cedar      Pine     Spruce Prediction
## 1: 0.7157304 0.2130577 0.07121188      Cedar
```

b.



c.

```
##              Coordinate
## BarkDarkness   2.673249
## LogCurve       2.164786
```

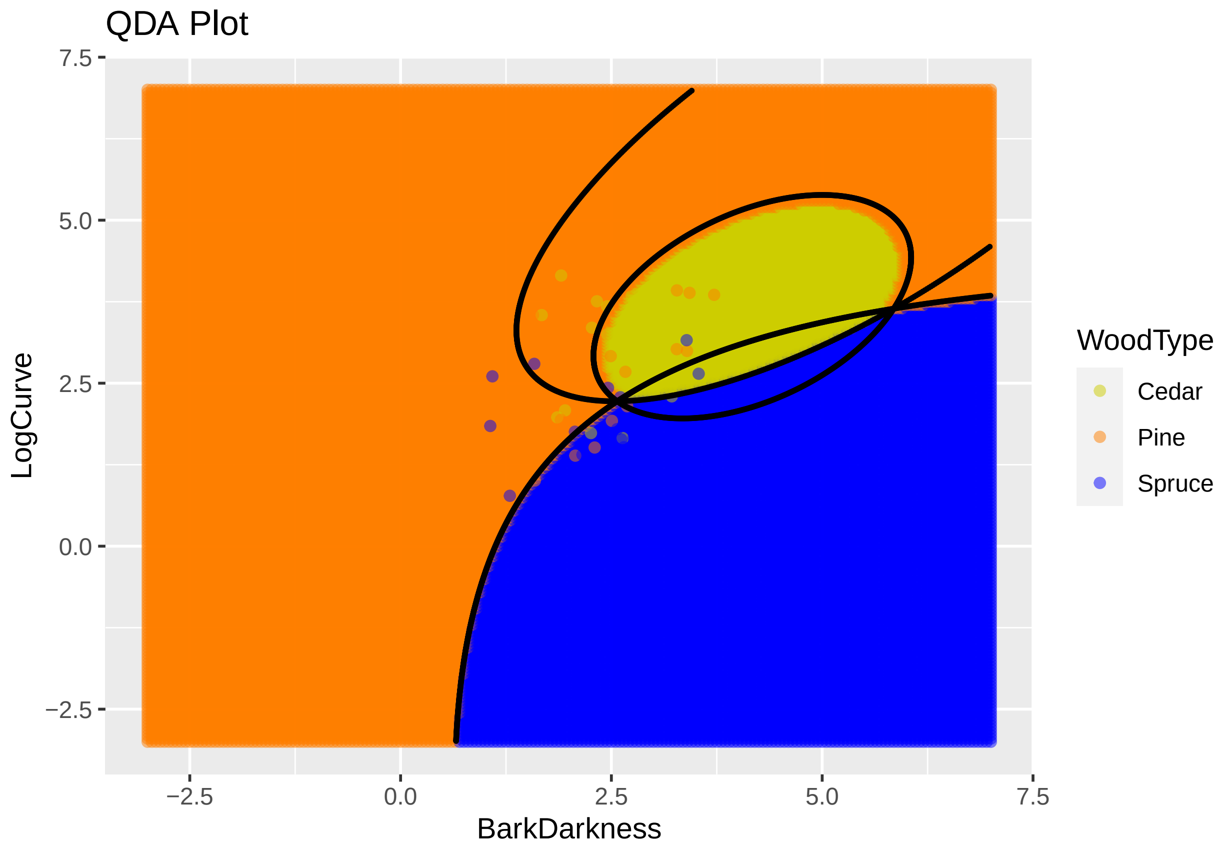
d.

```
## [1] Covariance matrices:
## [1]
## $Cedar
##              BarkDarkness   LogCurve
## BarkDarkness   0.14107773 0.06032774
## LogCurve       0.06032774 0.12105923
##
## $Pine
##              BarkDarkness  LogCurve
## BarkDarkness    0.4641453 0.2280475
## LogCurve        0.2280475 0.4311469
##
## $Spruce
##              BarkDarkness    LogCurve
## BarkDarkness  0.314179713 0.003650447
## LogCurve      0.003650447 0.265809323
##
## [1]
## [1] Means accross classes:
## [1]
##    WoodType BarkDarkness  LogCurve  n     P
## 1:    Cedar     3.240607 3.2671749 43 0.215
## 2:     Pine     1.106614 2.1398566 89 0.445
## 3:   Spruce     3.030018 0.9863158 68 0.340
## [1]
## [1] In-sample accuracy: 0.845

## [1] Prediction for BarkDarkness = LogCurve = 3:

##        Cedar       Pine       Spruce Prediction
## 1: 0.9815091 0.01806431 0.0004266103      Cedar
```

e.


QDA Plot
7.5
5.0
2.5
0.0
−2.5
LogCurve
−2.5
0.0
2.5
5.0
7.5
BarkDarkness
WoodType
Cedar
Pine
Spruce