The Selection of a Model for Airlines Customer Satisfaction

Joanna Krężel
Anna Matysek
Piotr Mikler
Adam Szczerba

26.01.2022

Agenda

1. Data
2. Data preprocessing
3. Baseline model - Adaline
4. Random forest
5. Logistic regression
6. Neural network
7. Conclusion

Data

Feature	Description	Variable type
Satisfaction	Overall satisfaction	factor, 2 levels
Gender	Gender of the passenger	factor, 2 levels
Customer type	Loyalty of the passenger	factor, 2 levels
Age	The age of a passenger	continuous
Type of travel	Flight purpose	factor, 2 levels
Class	Travel class in the plane	factor, 3 levels
Flight distance	Distance of the journey	continuous
Seat comfort	Survey note for seat comfort	factor, 5 levels
Departure/arrival time convenience	Survey note for departure/arrival time convenience	factor, 5 levels
Food and drink	Survey note for food and drinks	factor, 5 levels
Gate location	Survey note for gate location	factor, 5 levels
Inflight WiFi service	Survey note for the inflight wifi	factor, 5 levels
Inflight entertainment	Survey note for inflight entertainment	factor, 5 levels
Online support	Survey note for online support	factor, 5 levels
Ease of online booking	Survey note for online booking	factor, 5 levels
On-board services	Survey note for on-board service	factor, 5 levels
Leg room	Survey note for leg room/space	factor, 5 levels
Baggage handling	Survey note for baggage handling	factor, 5 levels
Checkin service	Survey note for check-in service	factor, 5 levels
Cleanliness	Survey note for cleanliness	factor, 5 levels
Online boarding	Survey note for online boarding	factor, 5 levels
Departure delay	Delay upon departure	continuous
Arrival delay	Delay upon arrival	continuous

Figure 1: The list of variables

Data

```
      satisfaction      Gender             Customer Type        Age
satisfied   :71087   Female:65899   Loyal Customer   :106100   Min.    : 7.00
dissatisfied:58793   Male  :63981   disloyal Customer: 23780   1st Qu.:27.00
                                                               Median :40.00
                                                               Mean   :39.43
                                                               3rd Qu.:51.00
                                                               Max.   :85.00

       Type of Travel       Class       Flight Distance Seat comfort
Personal Travel:40187   Eco     :58309   Min.    :  50   0: 4797
Business travel:89693   Business:62160   1st Qu.:1359    1:20949
                        Eco Plus: 9411   Median :1925    4:28398
                                         Mean   :1981    5:17827
                                         3rd Qu.:2544    2:28726
                                         Max.   :6951    3:29183

Departure/Arrival time convenient Food and drink Gate location Inflight wifi service
0: 6664                           0: 5945        2:24518       2:27045
1:20828                           1:21076        3:33546       0:  132
2:22794                           2:27146        4:30088       3:27602
3:23184                           3:28150        1:22565       4:31560
4:29593                           4:27216        5:19161       5:28830
5:26817                           5:20347        0:    2       1:14711

Inflight entertainment Online support Ease of Online booking On-board service
4:41879                2:17260        3:22418                3:27037
2:19183                3:21609        2:19951                4:40675
0: 2978                4:41510        1:13436                1:13265
3:24200                5:35563        5:34137                2:17174
5:29831                1:13937        4:39920                5:31724
1:11809                0:    1        0:   18                0:    5

Leg room service Baggage handling Checkin service Cleanliness Online boarding
0:  444          3:24485          5:27005         3:23984     2:18573
4:39698          4:48240          2:15486         4:48795     3:30780
3:22467          1: 7975          4:36481         1: 7768     5:29973
2:21745          2:13432          3:35538         2:13412     4:35181
5:34385          5:35748          1:15369         5:35916     1:15359
1:11141                           0:    1         0:    5     0:   14

Departure Delay in Minutes Arrival Delay in Minutes
Min.    :   0.00           Min.    :   0.00
1st Qu.:   0.00           1st Qu.:   0.00
Median :   0.00           Median :   0.00
Mean   :  14.71           Mean   :  15.09
3rd Qu.:  12.00           3rd Qu.:  13.00
Max.   :1592.00           Max.   :1584.00
                          NA's   :393
```

Figure 2: Summary of the data set

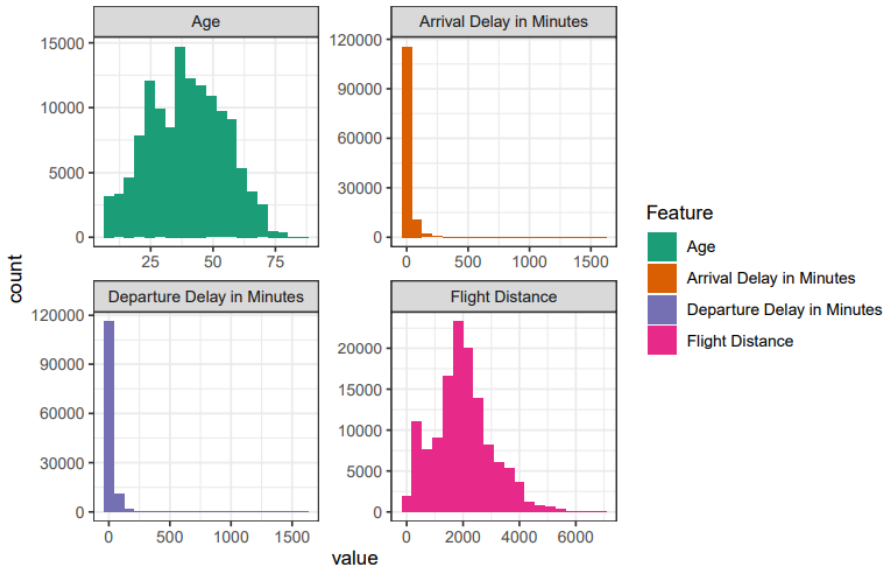


Figure 3: Histograms for selected features

Data preprocessing

Feature	NA_count	pct_of_data
ScheduleNote	6664	5.13%
FoodNote	5945	4.58%
SeatNote	4797	3.69%
EntertainmentNote	2978	2.29%
LegRoomNote	444	0.34%
ArrivalDelay	393	0.3%
WifiNote	132	0.1%
eBookingNote	18	0.01%
eBoardingNote	14	0.01%
CleanNote	5	0%
ServiceNote	5	0%
GateNote	2	0%
CheckInNote	1	0%
eSupportNote	1	0%

Figure 4: NA breakdown per feature

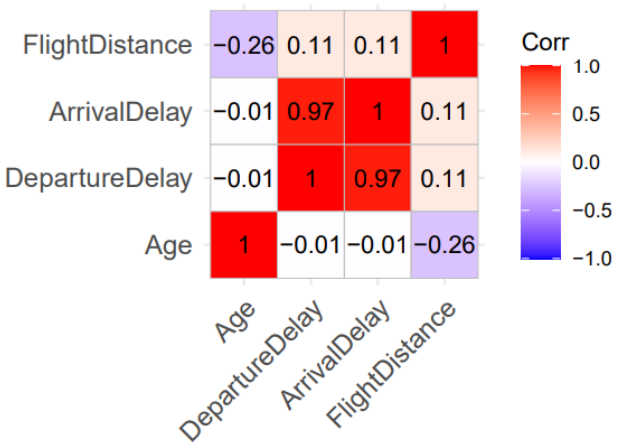


Figure 5: Correlation between continuous scale variables

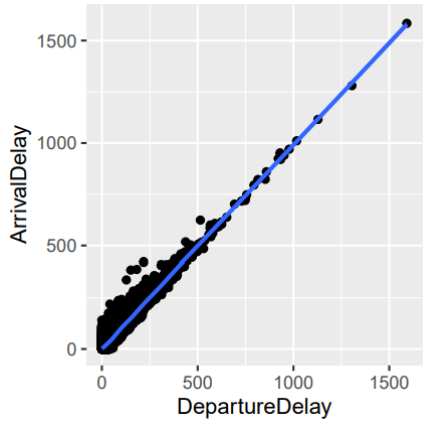


Figure 6: Relationship between departure delay and arrival delay

Data binning

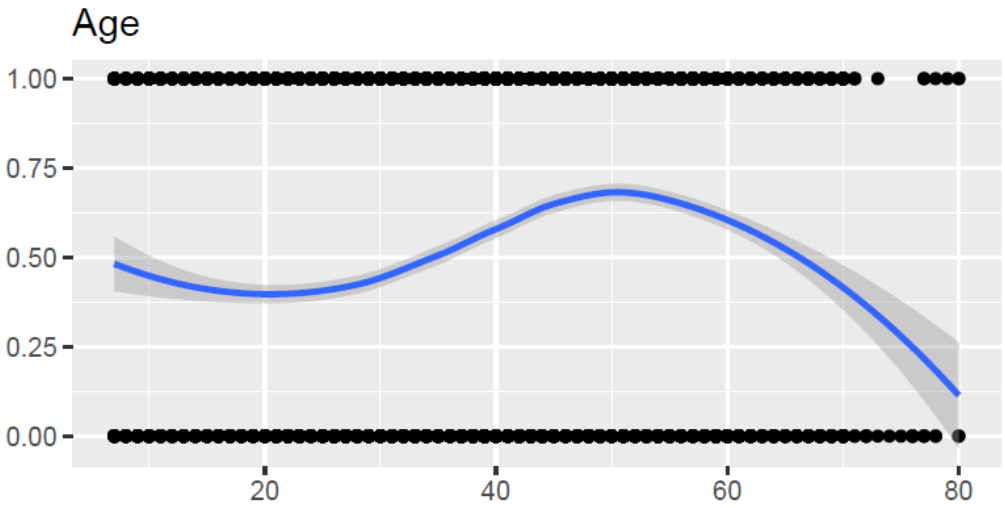


Figure 7: Loess estimator for IsSatisfied as function of Age

WOE Table for Age

	Final.Bin	Total.Count	Total.Distr.	0.Rate	WOE	IV
1	<= 28	30896	25.9%	58.6%	-51.4	0.068
2	<= 40	29312	24.6%	49.0%	-12.6	0.004
3	<= 59	47672	40.0%	33.9%	50.1	0.096
4	<= Inf	11375	9.5%	53.2%	-29.3	0.008
6	Total	119255	100.0%	45.9%	NA	0.177

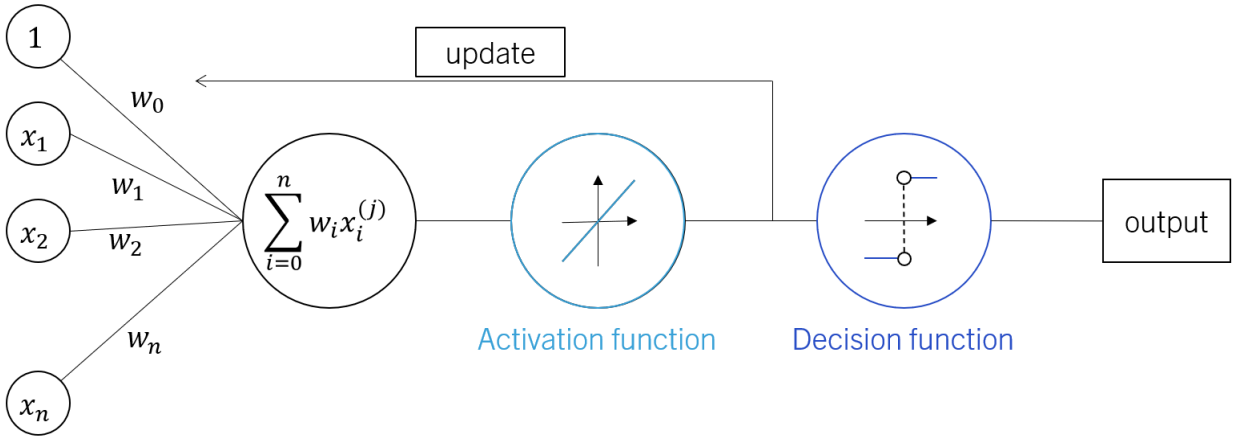
Figure 8: WOE Table for Age

Table 7: Information Value for all variables.

varName	IV
IsPersonalTravel	0.0532949
FlightDistance	0.1433763
Age	0.1769584
IsFemale	0.1893377
FoodNote	0.2330999
WifiNote	0.2920904
CheckInNote	0.3607163
Class	0.4001969
IsLoyal	0.4537176
CleanNote	0.4555277
BaggageNote	0.4718849
LegRoomNote	0.5714645
eBoardingNote	0.5948280
ServiceNote	0.6151573
eSupportNote	0.9223110
eBookingNote	1.1114520
SeatNote	1.4504018
EntertainmentNote	2.2947658

Figure 9: Information Value for all variables

Baseline model



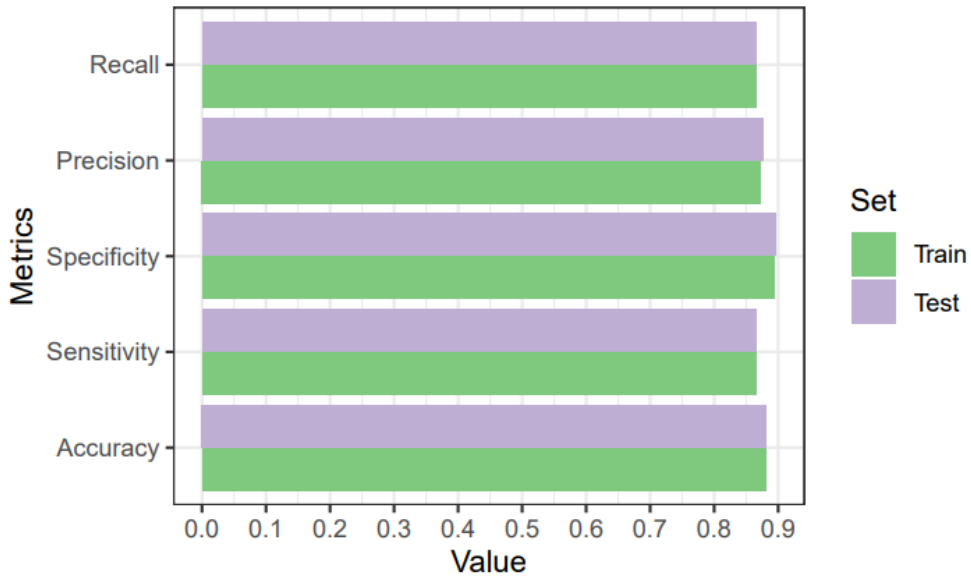


Figure 10: Performance metrics for the baseline Adaline model

Random forest - model fitting

```
frm<-IsSatisfied~Class+IsFemale+IsLoyal+Age+FlightDistance+
      EntertainmentNote+SeatNote+eBookingNote+eSupportNote+
      ServiceNote+eBoardingNote+LegRoomNote+BaggageNote+
      CleanNote+CheckInNote+WifiNote+FoodNote
```

```
##
## Call:
##  randomForest(formula = frm, data = Airlines_binned_train, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 6.18%
## Confusion matrix:
##       0     1 class.error
## 0 40963  2777  0.06348880
## 1  3123 48541  0.06044828
```

Figure 11: The Random Forest model

Random forest - model fitting

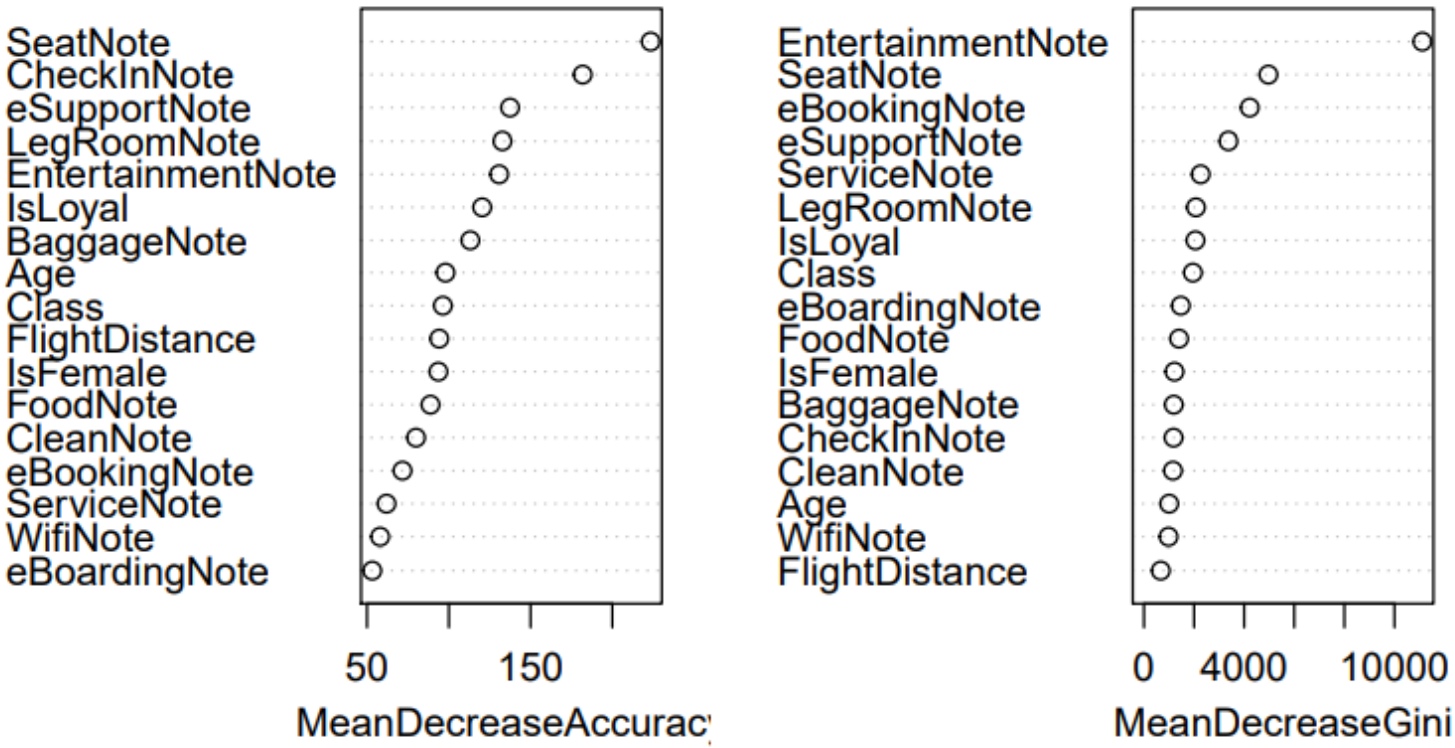


Figure 12: Importance of variables measured by the mean decrease of accuracy and Gini score

Random forest - performance

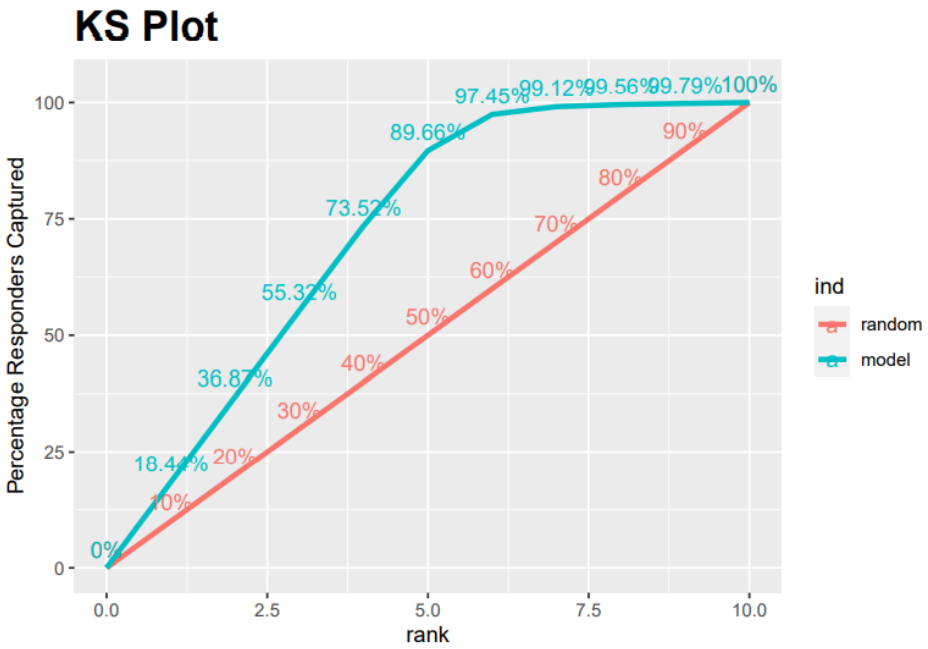


Figure 13: The KS Plot for the random forest

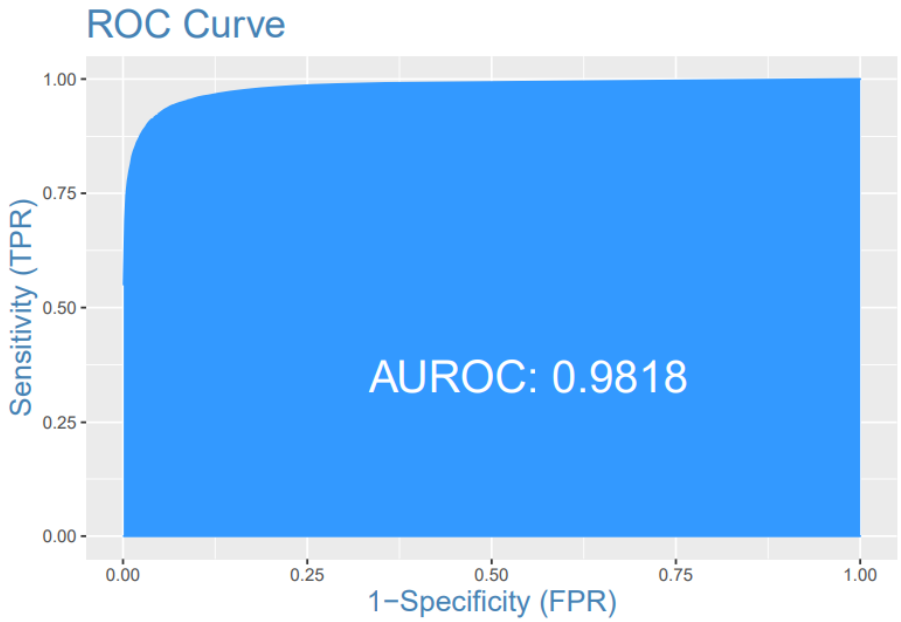


Figure 14: The ROC for the random forest

Random forest - validation

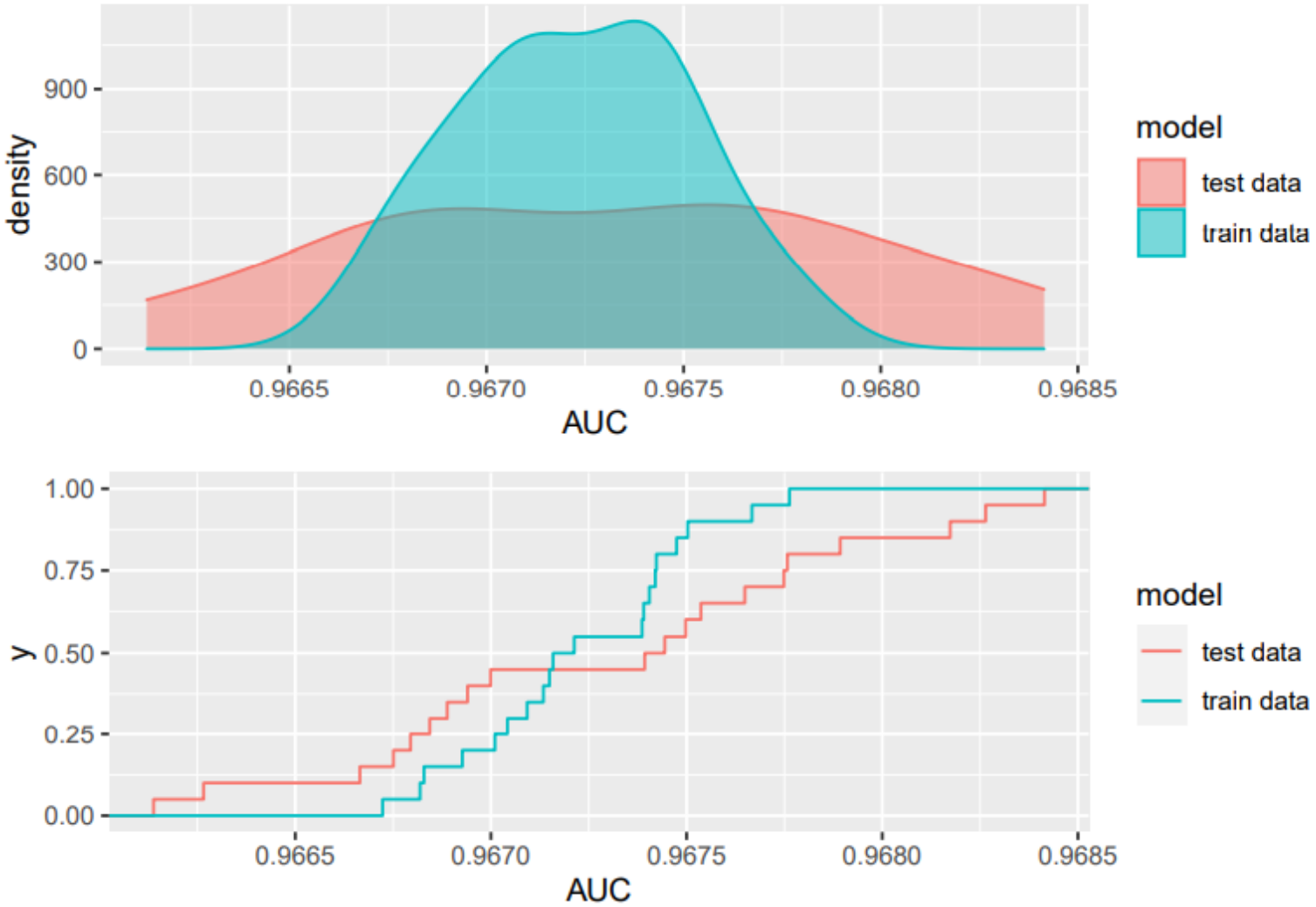


Figure 15: The results of the cross validation for the random forest

Logistic regression 1 - model fitting

```
frm<-IsSatisfied~Class+IsFemale+IsLoyal+Age+FlightDistance+
      EntertainmentNote+SeatNote+eBookingNote+eSupportNote+
      ServiceNote+eBoardingNote+LegRoomNote+BaggageNote+
      CleanNote+CheckInNote+WifiNote+FoodNote
```

```
##
## Call:
## glm(formula = frm, family = "binomial", data = Airlines_binned_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1699  -0.3623   0.0313   0.3099   3.6301
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -7.21962    0.08599 -83.961  < 2e-16 ***
## ClassBusiness       1.34302    0.02605  51.550  < 2e-16 ***
## ClassEcoPlus        0.03494    0.04385   0.797 0.425568
## IsFemale1           0.97658    0.02355  41.466  < 2e-16 ***
## IsLoyal1            1.91910    0.03742  51.283  < 2e-16 ***
## Age30s             -0.02214    0.03196  -0.693 0.488560
## Age40s50s           0.21030    0.03105   6.774 1.25e-11 ***
## Age60plus          -0.37541    0.04313  -8.704  < 2e-16 ***
## FlightDistanceM    -0.19019    0.02919  -6.516 7.21e-11 ***
## FlightDistanceH    -0.13764    0.03939  -3.494 0.000475 ***
## EntertainmentNoteM  1.82628    0.02847  64.152  < 2e-16 ***
## EntertainmentNoteH  2.99102    0.04336  68.979  < 2e-16 ***
## SeatNoteM           0.85488    0.03330  25.670  < 2e-16 ***
## SeatNoteH           5.13222    0.11098  46.246  < 2e-16 ***
## eBookingNoteH       1.64792    0.06294  26.180  < 2e-16 ***
## eBookingNoteM       0.85178    0.06052  14.073  < 2e-16 ***
## eSupportNoteM       0.22928    0.03314   6.918 4.59e-12 ***
## eSupportNoteH       0.47526    0.03731  12.739  < 2e-16 ***
## ServiceNoteH        0.72232    0.04724  15.290  < 2e-16 ***
## ServiceNoteM        0.22391    0.04594   4.874 1.09e-06 ***
## eBoardingNoteH      0.19938    0.03785   5.267 1.38e-07 ***
## eBoardingNoteM      0.32782    0.03447   9.509  < 2e-16 ***
## LegRoomNoteM        0.03922    0.04776   0.821 0.411536
## LegRoomNoteH        0.75533    0.04772  15.827  < 2e-16 ***
## BaggageNoteM        0.10721    0.03359   3.192 0.001412 **
## BaggageNoteH        0.46100    0.03863  11.932  < 2e-16 ***
## CleanNoteM          0.06595    0.03497   1.886 0.059276 .
## CleanNoteH          0.48344    0.03958  12.214  < 2e-16 ***
## CheckInNoteM        0.33644    0.02874  11.706  < 2e-16 ***
## CheckInNoteH        0.90613    0.03728  24.307  < 2e-16 ***
## WifiNoteH          -0.13901    0.05062  -2.746 0.006025 **
## WifiNoteM           0.12726    0.04902   2.596 0.009435 **
## FoodNoteM          -0.48163    0.03440 -14.000  < 2e-16 ***
## FoodNoteH          -0.51322    0.04304 -11.924  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 131599  on 95403  degrees of freedom
## Residual deviance:  51904  on 95370  degrees of freedom
## AIC: 51972
##
## Number of Fisher Scoring iterations: 7
```

Figure 16: The first logistic regression model

Logistic regression 1 - performance

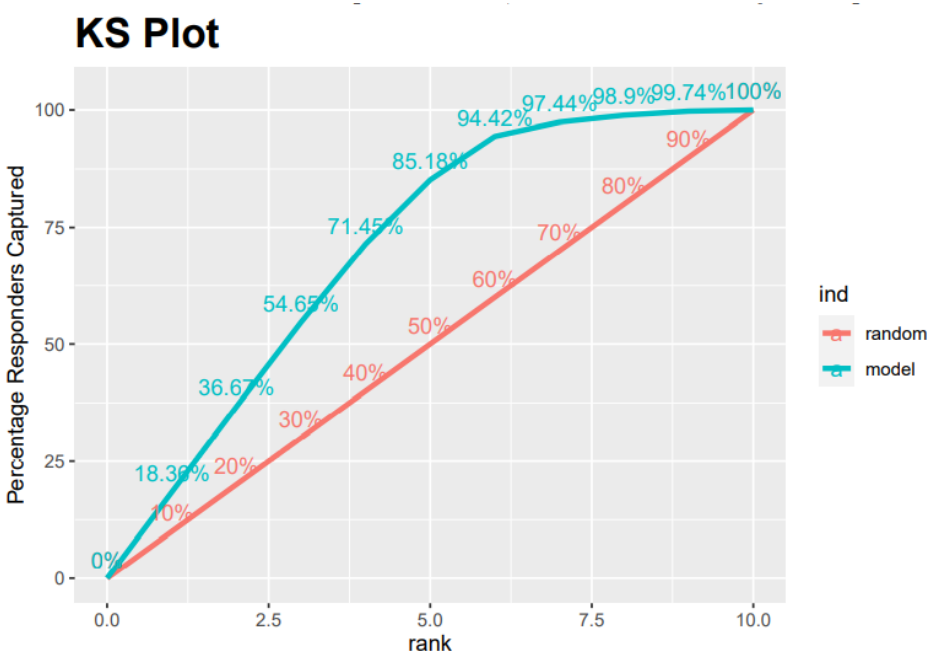


Figure 17: The KS plot for the first logistic regression model

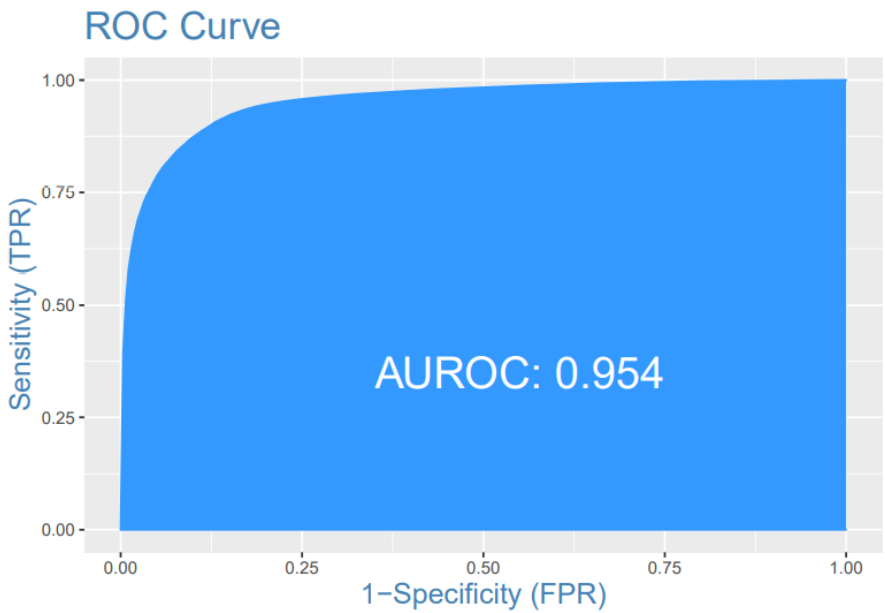


Figure 18: The ROC for the first logistic regression model

Logistic regression 2 - model fitting

```
frm2<-IsSatisfied~Class+IsFemale+IsLoyal+Age+FlightDistance+
      EntertainmentNote+SeatNote+eBookingNote+eSupportNote+
      ServiceNote+eBoardingNote+LegRoomNote+BaggageNote+
      CleanNote+CheckInNote+FoodNote+IsFemale*IsLoyal+IsFemale*Age+
      IsFemale*Class+Class*FlightDistance
```

```
##
## Call:
## glm(formula = frm2, family = "binomial", data = Airlines_binned_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1219  -0.3482   0.0296   0.3028   3.5915
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -7.23908    0.09822 -73.699  < 2e-16 ***
## ClassBusiness               1.74811    0.06208  28.159  < 2e-16 ***
## ClassEcoPlus                0.42848    0.11666   3.673 0.000240 ***
## IsFemale1                   1.15594    0.07710  14.992  < 2e-16 ***
## IsLoyal1                    1.40869    0.05355  26.307  < 2e-16 ***
## Age30s                      0.21245    0.04727   4.494 6.99e-06 ***
## Age40s50s                   0.55289    0.04442  12.447  < 2e-16 ***
## Age60plus                  -0.33849    0.06615  -5.117 3.10e-07 ***
## FlightDistanceM            -0.22820    0.04753  -4.801 1.58e-06 ***
## FlightDistanceH            -0.44938    0.08464  -5.310 1.10e-07 ***
## EntertainmentNoteM          1.75566    0.02894  60.658  < 2e-16 ***
## EntertainmentNoteH          2.84328    0.04440  64.040  < 2e-16 ***
## SeatNoteM                   0.84514    0.03383  24.985  < 2e-16 ***
## SeatNoteH                   5.15478    0.11218  45.949  < 2e-16 ***
## eBookingNoteH               1.55519    0.05784  26.888  < 2e-16 ***
## eBookingNoteM               0.92479    0.05355  17.270  < 2e-16 ***
## eSupportNoteM               0.16166    0.03321   4.868 1.13e-06 ***
## eSupportNoteH               0.42142    0.03776  11.161  < 2e-16 ***
## ServiceNoteH                0.69248    0.04817  14.377  < 2e-16 ***
## ServiceNoteM                0.17398    0.04644   3.747 0.000179 ***
## eBoardingNoteH              0.22947    0.03721   6.168 6.94e-10 ***
## BaggageNoteH                0.50288    0.03998  12.577  < 2
## CleanNoteM                  0.10369    0.03610   2.872 0.00
## CleanNoteH                  0.55564    0.04111  13.516  < 2
## CheckInNoteM                0.34630    0.02908  11.909  < 2
## CheckInNoteH                0.94897    0.03817  24.862  < 2
## FoodNoteM                  -0.47547    0.03481 -13.660  < 2
## FoodNoteH                  -0.51726    0.04323 -11.966  < 2
## IsFemale1:IsLoyal1          1.08540    0.07105  15.276  < 2
## IsFemale1:Age30s           -0.39527    0.06544  -6.041 1.54
## IsFemale1:Age40s50s        -0.57123    0.06321  -9.038  < 2
## IsFemale1:Age60plus        -0.06771    0.09023  -0.750 0.45
## ClassBusiness:IsFemale1    -1.53533    0.05220 -29.412  < 2
## ClassEcoPlus:IsFemale1     -0.48121    0.09760  -4.930 8.20
## ClassBusiness:FlightDistanceM  0.49425  0.06107   8.093 5.82
## ClassEcoPlus:FlightDistanceM  -0.19410  0.10774  -1.802 0.07
## ClassBusiness:FlightDistanceH  0.73526  0.09554   7.696 1.41
## ClassEcoPlus:FlightDistanceH  -0.12560  0.20768  -0.605 0.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 131599  on 95403  degrees of freedom
## Residual deviance:  50220  on 95362  degrees of freedom
## AIC: 50304
##
## Number of Fisher Scoring iterations: 7
```

Figure 19: The second logistic regression model

Logistic regression 2 - performance

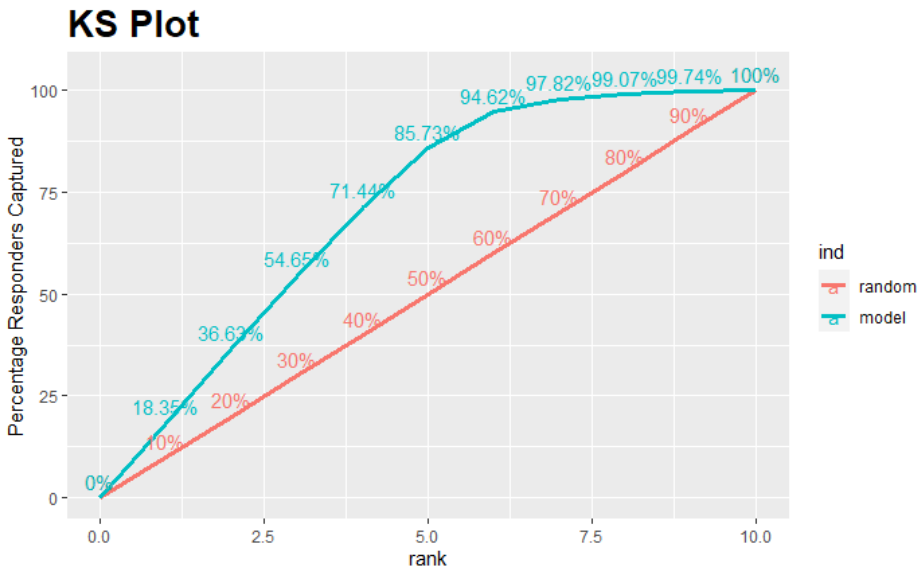


Figure 20: The KS plot for the second logistic regression model

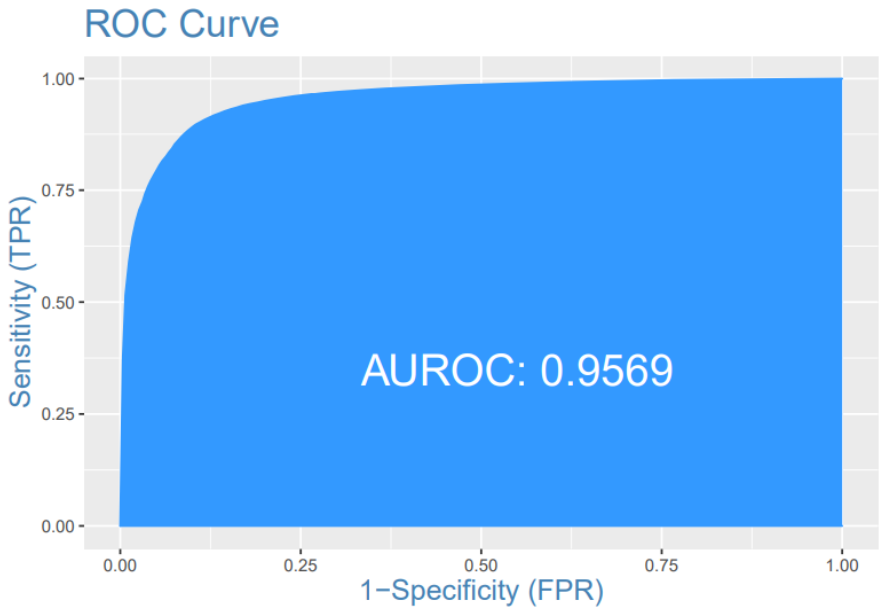


Figure 21: The ROC for the second logistic regression model

Logistic regression 2 - validation

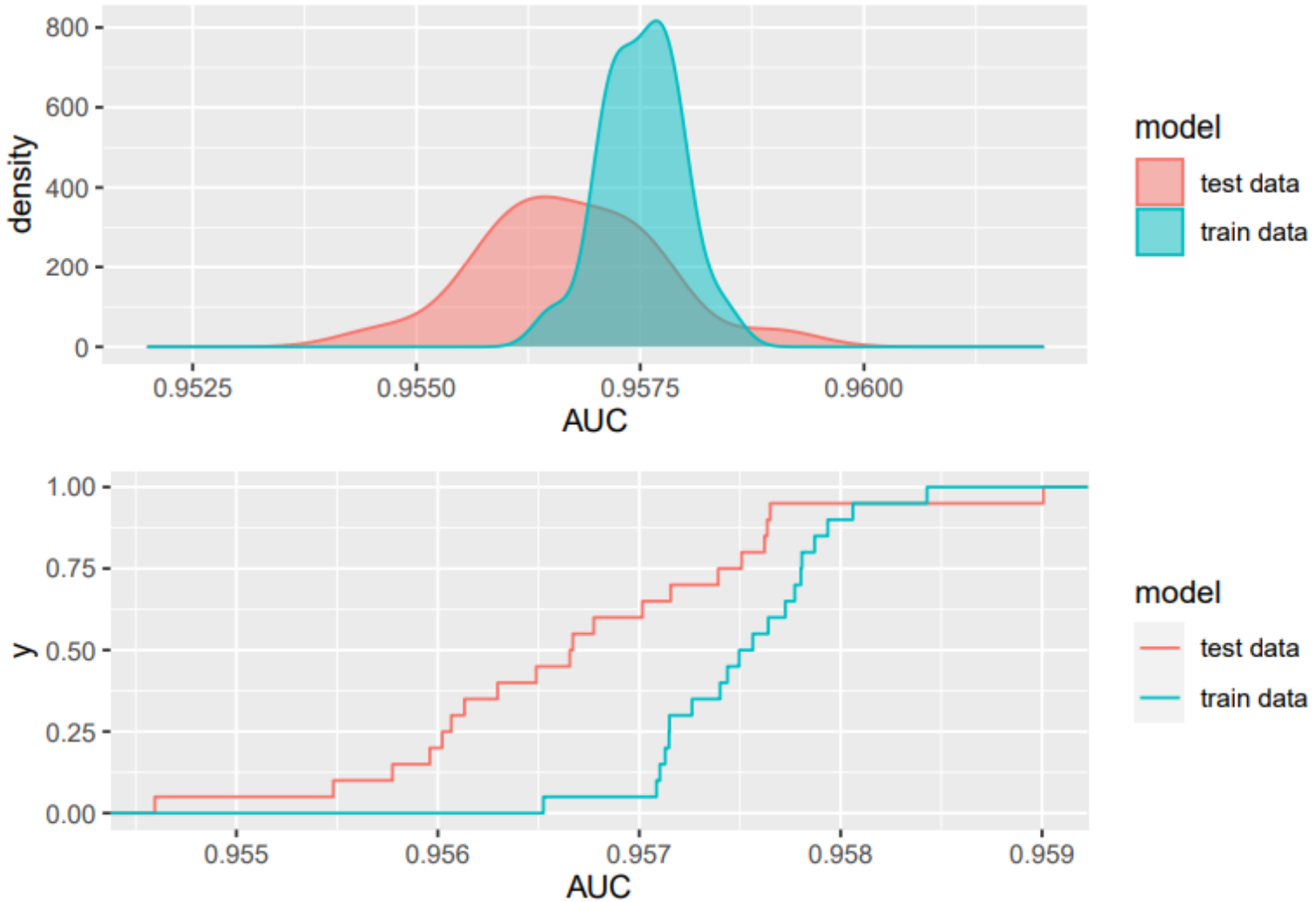


Figure 22: The results of the cross validation for the second logistic regression model

Neural network - model fitting & performance

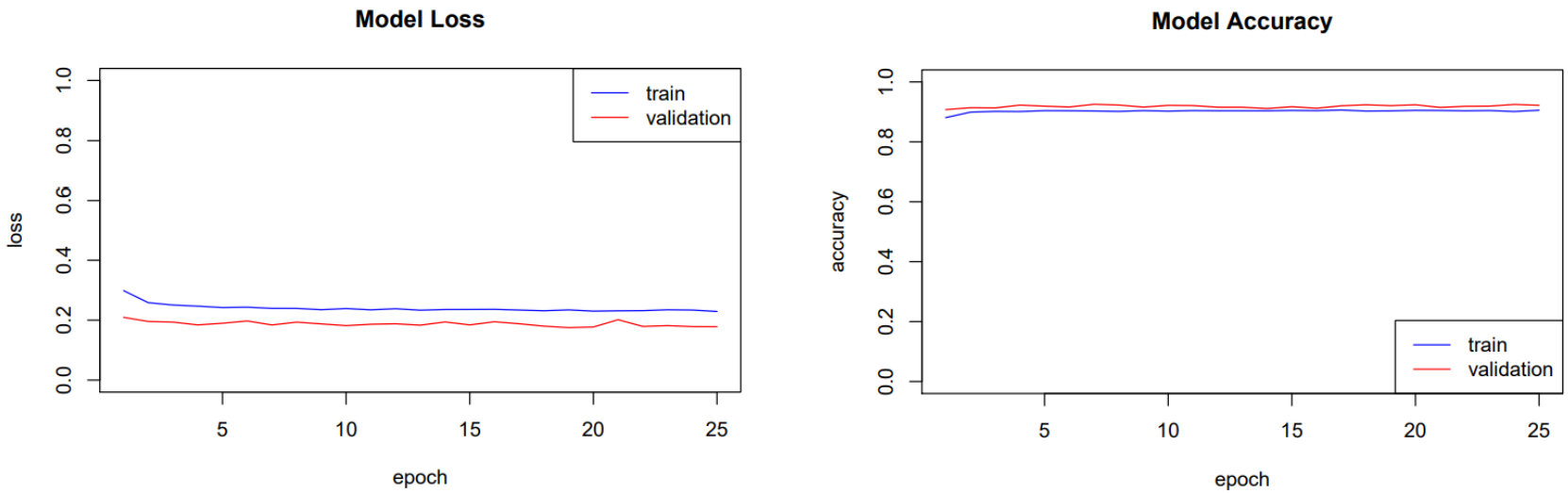


Figure 23: The entire accuracy of the model was achieved in the first epoch of neural network training

Neural network - model fitting & performance

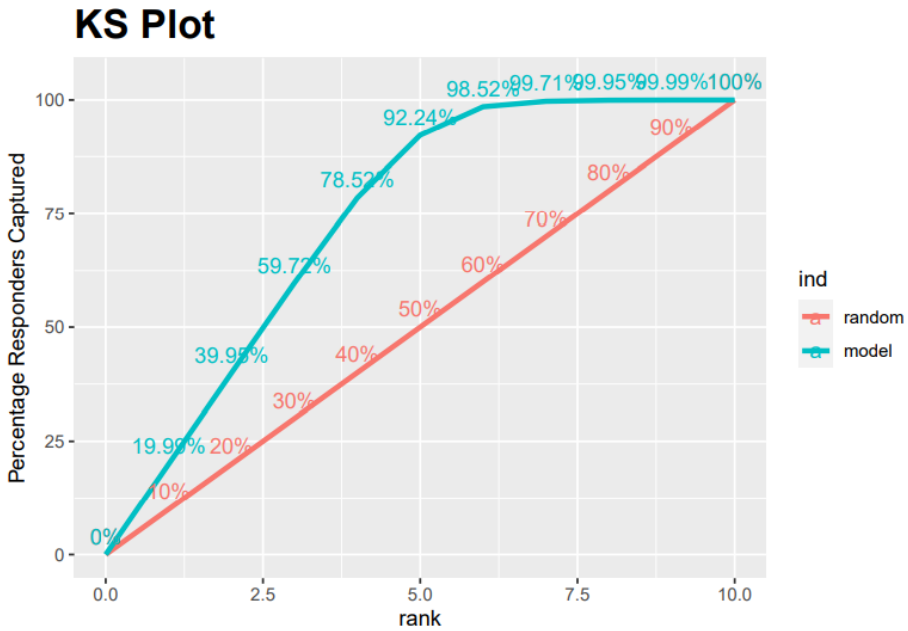


Figure 24: The KS plot for the neural network

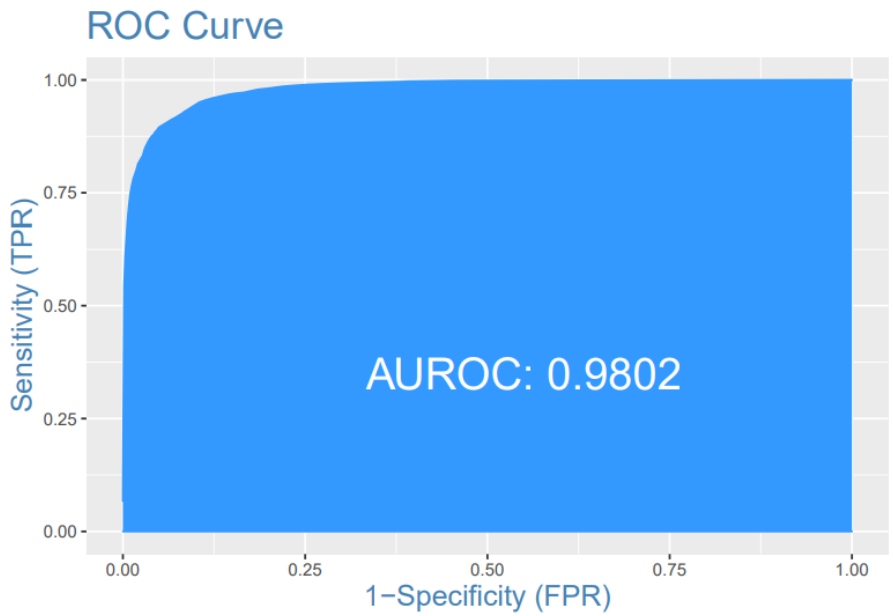


Figure 25: The ROC for the neural network

Neural network - validation

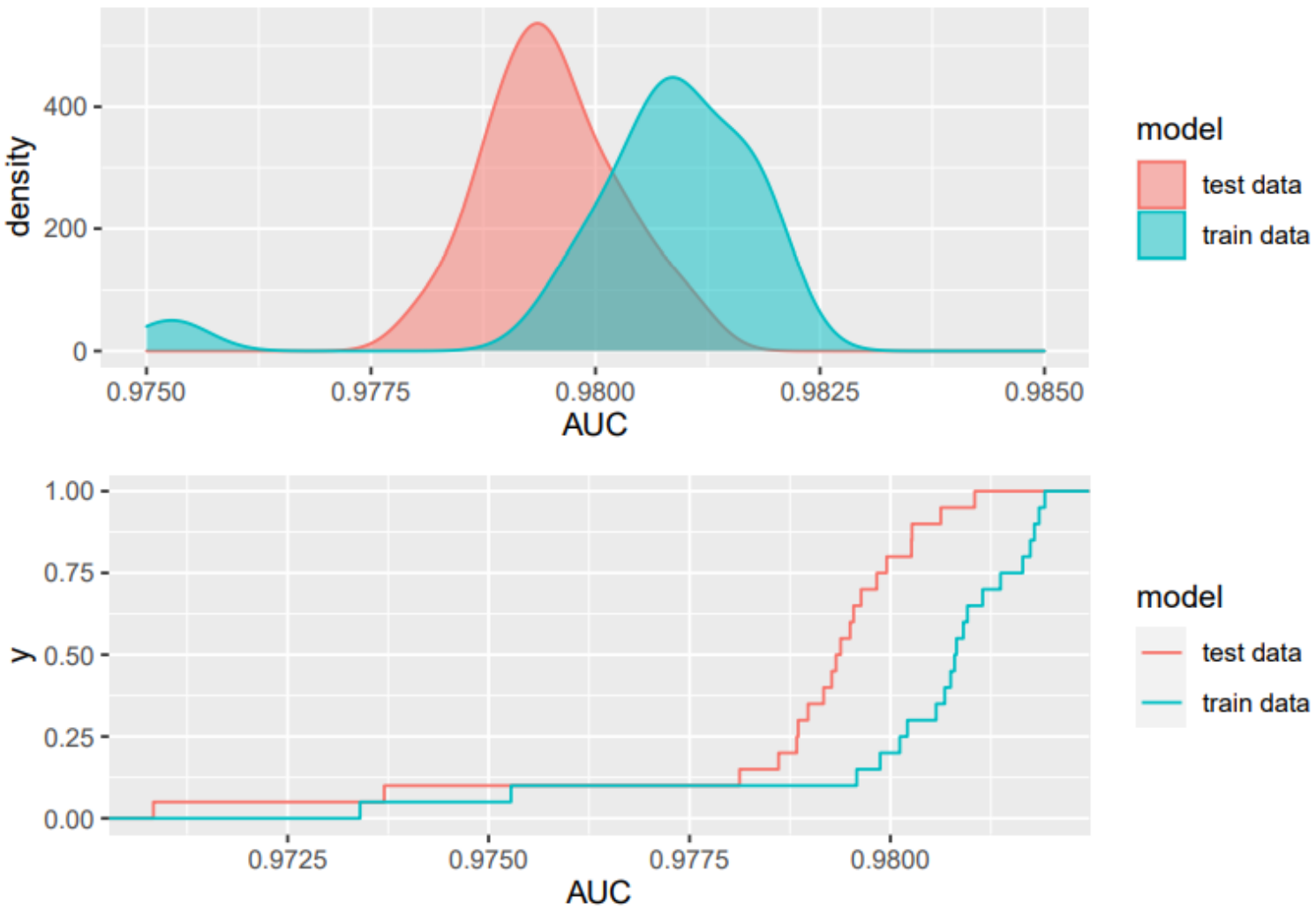


Figure 26: The results of cross validation for the neural network

Conclusion

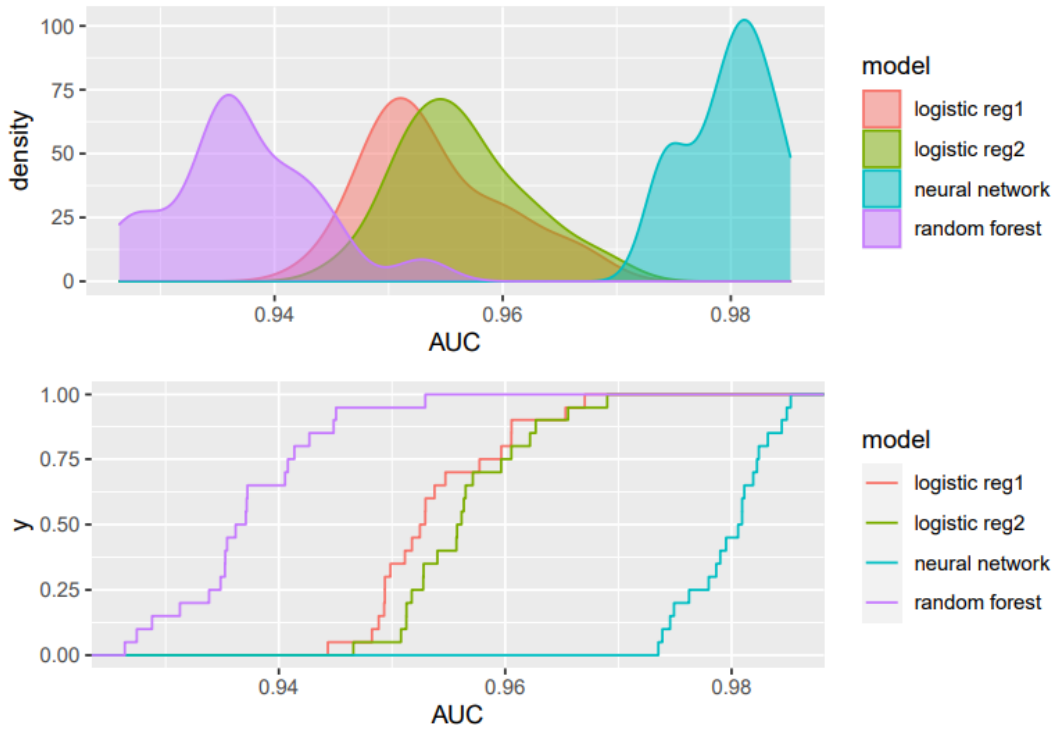


Figure 27: The kernel density for the observed areas under the curve (top) and the cumulative probability density functions (bottom) for the challenger models. All AUCs shown are for the test data only.

Model	Mean AUC on Test Data	Mean AUC on Training Data
logistic regression 1	0.9539931	0.9542947
logistic regression 2	0.9564216	0.9571064
random forest	0.9372222	0.9672647
neural network	0.9798081	0.9795287