Scientia Iranica

Supplementary Data

A New Approach to Estimating Destinations in Open Automated Fare Collection Systems based on errors-against-errors strategy

Mostafa Shafaati[1], Mahmoud Saffarzadeh*[2]

[1] PHD candidate, Department of Civil and Environment Engineering, Tarbiat Modares University, Tehran, Iran, email: (mostafa.shafaati@modares.ac.ir), Phone: +989127326455

[2] Professor, Faculty of Civil and environment Engineering, Tarbiat Modares University, Tehran, Iran, email: (saffar_m@modares.ac.ir), Phone: +989121125096

This file includes some tables and figures that were less important or complementary to the main content. All tables and figures in this file are referred to in the main file. There are 4 series of tables along with 3 series of figures. Tables T1 represent the validation results of each model for every single studied stop. We remind the readers of the fact that in order to summarize the paper, we started modeling from stop 7 and ended it in stop 20. We have included the number of stops in the figures of F1- F3 as well. Tables T2, and T3 demonstrate the validation results of the third approach: the aggregated approach (section 4.2.3 in the main file). As we have explained in the main file, both tables satisfy equation 13. Finally, table T4 shows the definitions of the concepts used or introduced in our paper.

Tables

Tables T1. Validation results of training each model

Stop 7

Classes	Precision	Recall	f1-score	Support
0	0.956	0.957	0.956	55182
1	0.318	0.311	0.314	3564
Macro avg.	0.637	0.634	0.635	58746
Weighted avg.	0.917	0.918	0.917	58746
Accuracy: 0.918				58746

Stop 8

Classes	Precision	Recall	f1-score	Support
0	0.942	0.944	0.943	54001
1	0.346	0.337	0.342	4745
Macro avg.	0.644	0.641	0.642	58746
Weighted avg.	0.894	0.895	0.894	58746
Accuracy: 0.895				58746

Stop 9

Classes	Precision	Recall	f1-score	Support
0	0.934	0.935	0.934	53217
1	0.367	0.362	0.364	5529
Macro avg.	0.650	0.649	0.649	58746
Weighted avg.	0.880	0.881	0.881	58746
Accuracy: 0.881				58746

Stop 10

Classes	Precision	Recall	f1-score	Support
0	0.929	0.930	0.929	52488
1	0.404	0.400	0.402	6258
Macro avg.	0.666	0.665	0.665	58746
Weighted avg.	0.873	0.873	0.873	58746
Accuracy: 0.871				58746

Stop 11

Classes	Precision	Recall	f1-score	Support
0	0.912	0.912	0.912	51019
1	0.419	0.418	0.419	7727
Macro avg.	0.666	0.665	0.665	58746
Weighted avg.	0.847	0.847	0.847	58746
Accuracy: 0.871				58746

Stop 12

Classes	Precision	Recall	f1-score	Support
0	0.894	0.896	0.895	49107
1	0.464	0.458	0.461	9639
Macro avg.	0.679	0.677	0.678	58746
Weighted avg.	0.823	0.824	0.824	58746
Accuracy: 0.824				58746

Stop 13

Classes	Precision	Recall	f1-score	Support
0	0.867	0.866	0.866	46622
1	0.487	0.491	0.489	12124
Macro avg.	0.677	0.678	0.678	58746
Weighted avg.	0.789	0.788	0.789	58746
Accuracy: 0.788				58746

Stop 14

Classes	Precision	Recall	f1-score	Support
0	0.851	0.853	0.852	45065
1	0.511	0.507	0.509	13681
Macro avg.	0.681	0.680	0.681	58746
Weighted avg.	0.772	0.772	0.772	58746
Accuracy: 0.772				58746

Stop 15

Classes	Precision	Recall	f1-score	Support
0	0.830	0.833	0.832	43384
1	0.524	0.519	0.521	15362
Macro avg.	0.677	0.676	0.677	58746
Weighted avg.	0.750	0.751	0.751	58746
Accuracy: 0.751				58746

Stop16

Classes	Precision	Recall	f1-score	Support
0	0.795	0.797	0.796	39745
1	0.573	0.570	0.571	19001
Macro avg.	0.684	0.683	0.683	58746
Weighted avg.	0.723	0.723	0.723	58746
Accuracy: 0.723				58746

Stop 17

Classes	Precision	Recall	f1-score	Support
0	0.761	0.761	0.761	35585
1	0.633	0.632	0.632	23161
Macro avg.	0.697	0.697	0.697	58746
Weighted avg.	0.710	0.710	0.710	58746
Accuracy: 0.710				58746

Stop 18

Classes	Precision	Recall	f1-score	Support
0	0.725	0.729	0.727	28207
1	0.749	0.745	0.747	30539
Macro avg.	0.737	0.737	0.737	58746
Weighted avg.	0.738	0.737	0.737	58746
Accuracy: 0.737				58746

Stop 19

Classes	Precision	Recall	f1-score	Support
0	0.722	0.723	0.722	25507
1	0.787	0.787	0.787	33239
Macro avg.	0.755	0.755	0.755	58746
Weighted avg.	0.759	0.759	0.759	58746
Accuracy: 0.759				58746

Stop 20

Classes	Precision	Recall	f1-score	Support
0	0.715	0.717	0.716	20262
1	0.851	0.849	0.850	38484
Macro avg.	0.783	0.783	0.783	58746
Weighted avg.	0.804	0.804	0.804	58746
Accuracy: 0.804				58746

Table T2. The confusion matrix to evaluate the model's performance regarding the whole studied stops (The first validation set)

actual / predicted	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
7	7	57	32	**136**	**37**	**40**	**50**	**23**	**29**	**36**	**70**	**60**	**25**	**19**	**44**	**25**	**92**	**15**	**72**	**11**	**23**	**25**	**11**	**9**	**28**	**104**
8	3	13	4	39	**23**	**19**	**35**	**28**	**7**	**22**	**25**	**24**	**11**	**15**	**22**	**16**	**47**	**12**	**26**	**3**	**10**	**10**	**1**	**3**	**8**	**27**
9	1	1	2	6	7	**21**	**31**	**19**	**11**	**11**	**10**	**19**	**12**	**10**	**10**	**21**	**24**	**8**	**15**	**1**	**8**	**4**	**2**	**2**	**5**	**13**
10	0	0	2	3	9	17	**26**	**18**	**10**	**11**	**21**	**15**	**9**	**6**	**9**	**17**	**28**	**9**	**13**	**1**	**6**	**2**	**0**	**1**	**4**	**11**
11	1	6	0	10	23	27	41	**22**	**16**	**45**	**39**	**38**	**20**	**17**	**32**	**43**	**56**	**15**	**33**	**5**	**10**	**4**	**3**	**4**	**6**	**23**
12	2	8	8	38	12	18	40	23	**23**	**21**	**47**	**33**	**19**	**24**	**51**	**61**	**81**	**25**	**52**	**11**	**16**	**10**	**3**	**3**	**14**	**39**
13	0	3	3	16	9	17	32	13	18	**57**	**53**	**82**	**27**	**18**	**77**	**70**	**90**	**21**	**60**	**5**	**12**	**15**	**3**	**3**	**15**	**26**
14	2	3	3	30	9	12	13	8	9	31	**32**	**50**	**21**	**15**	**55**	**47**	**84**	**14**	**45**	**6**	**16**	**10**	**7**	**5**	**7**	**19**
15	0	0	2	25	12	16	21	9	18	39	41	**33**	**26**	**24**	**75**	**65**	**106**	**26**	**43**	**5**	**11**	**10**	**5**	**1**	**6**	**17**
16	0	2	4	31	9	8	30	19	18	41	63	75	**48**	**42**	**102**	**94**	**135**	**45**	**81**	**8**	**33**	**13**	**9**	**6**	**11**	**49**
17	0	0	2	32	7	33	47	26	24	85	72	78	60	**56**	**146**	**159**	**221**	**82**	**113**	**17**	**36**	**16**	**13**	**12**	**18**	**40**
18	3	1	3	29	6	9	32	37	36	46	92	151	118	104	**283**	**291**	**435**	**139**	**229**	**28**	**98**	**54**	**20**	**19**	**44**	**121**
19	6	32	18	60	21	16	30	12	15	30	57	48	47	45	90	**116**	**240**	**101**	**148**	**17**	**50**	**25**	**15**	**16**	**29**	**67**
20	0	4	4	15	4	5	4	2	7	4	24	86	92	110	145	193	**338**	**131**	**201**	**25**	**69**	**36**	**11**	**27**	**68**	**137**

Table T3. The confusion matrix to evaluate the model's performance regarding the whole studied stops (The second validation set)

predicted \ actual	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
7	8	37	14	**114**	**30**	**24**	**31**	**17**	**24**	**19**	**58**	**44**	**18**	**9**	**38**	**29**	**63**	**14**	**41**	**5**	**13**	**17**	**3**	**7**	**26**	**64**
8	6	26	7	96	**30**	**39**	**39**	**23**	**24**	**33**	**43**	**43**	**19**	**22**	**51**	**26**	**74**	**13**	**52**	**7**	**17**	**12**	**2**	**2**	**11**	**54**
9	1	5	3	30	17	**32**	**25**	**24**	**24**	**17**	**29**	**19**	**18**	**9**	**25**	**23**	**30**	**8**	**26**	**2**	**5**	**5**	**3**	**2**	**6**	**17**
10	0	0	0	0	11	18	**29**	**17**	**12**	**21**	**11**	**18**	**7**	**6**	**2**	**9**	**17**	**3**	**5**	**2**	**3**	**1**	**1**	**0**	**3**	**9**
11	0	5	2	15	18	36	43	**31**	**15**	**50**	**37**	**34**	**18**	**17**	**33**	**36**	**62**	**16**	**49**	**2**	**10**	**4**	**2**	**3**	**12**	**19**
12	0	6	6	36	9	14	37	29	**15**	**26**	**31**	**46**	**15**	**23**	**43**	**45**	**64**	**14**	**34**	**4**	**14**	**5**	**2**	**0**	**13**	**32**
13	0	5	4	19	11	14	35	18	22	**49**	**67**	**77**	**33**	**31**	**84**	**71**	**111**	**33**	**67**	**8**	**20**	**13**	**5**	**2**	**22**	**40**
14	0	0	0	9	4	4	9	6	6	17	**16**	**33**	**17**	**7**	**32**	**23**	**46**	**14**	**14**	**6**	**6**	**6**	**1**	**2**	**3**	**17**
15	0	2	2	37	16	24	27	26	11	41	35	**35**	**31**	**43**	**62**	**83**	**127**	**59**	**67**	**11**	**13**	**9**	**4**	**6**	**9**	**30**
16	0	2	4	48	11	12	32	24	26	56	57	101	**54**	**45**	**140**	**146**	**206**	**59**	**108**	**11**	**44**	**19**	**11**	**7**	**12**	**61**
17	0	1	2	36	8	29	44	27	24	62	63	79	43	**55**	**106**	**128**	**171**	**58**	**82**	**17**	**30**	**24**	**9**	**1**	**11**	**52**
18	3	20	6	34	10	22	35	39	38	70	132	172	116	146	**289**	**329**	**525**	**157**	**246**	**28**	**108**	**29**	**23**	**36**	**46**	**137**
19	3	35	11	26	15	10	10	6	9	25	30	61	49	48	73	**108**	**179**	**69**	**91**	**17**	**51**	**22**	**15**	**8**	**26**	**57**
20	2	9	5	23	7	4	8	5	4	8	30	58	68	84	136	158	**352**	**114**	**198**	**18**	**73**	**47**	**17**	**24**	**50**	**122**

Table T4. Definitions of the concepts, terms, and variables presented in the paper

concept	Definition
P	Positive class (Those who alight at the current or previous stops)
N	Negative class (Those who don't alight at the current or previous stops)
TN	True Negative (The actual and predicted classes are both negative)
FP	False Positive (The actual class is negative but the predicted is positive)
FN	False Negative (The actual class is positive but the predicted is negative)
TP	True Positive (The actual and predicted classes are both positive)
A	$\lvert FN - FP \rvert$
Recall	$\frac{TP}{TP + FN}$
Specifity	$\frac{TN}{FP + TN}$

concept	Definition
G_mean	$\sqrt{Recall \times Specifity}$
Precision	$\frac{TP}{TP + FP}$
TPR	True Positive Rate = Recall
FPR	$False\ Positive\ Rate\ (FPR) = 1 - \frac{1}{Recall}$
F_{1_score}	$2\frac{precison \times recall}{precision + recall}$
Thresholding	Shifting decision threshold to solve imbalanced classification problems
ROC curve	A diagnostic plot that evaluate the probability predictions made by a model on a test dataset.
Precision-Recall curve	These plots are calculated and plotted for probability predictions by creating crisp class labels and calculating precision and recall for each threshold.
Diff	the estimated stop number – the actual stop number
N	The number of observations belonging to each Diff
ACEC	Accepted Classification Error Categories in the first validation approach. Each category is shown by a small English letter starting from category "a" where the Diff is 0, category "b" where the Diff is -1, 0, +1, and etc.
NMC	Number of Members in the Category in the first validation approach
PMC	Percentage of Members in the Category in the first validation approach
i	The absolute value of Diff
I_{1i}	$\begin{cases} N_i \quad if\ i = 0 \\ N_i - N_{-i}\ if\ i \neq 0 \end{cases}$ (The first index in the second validation approach)
I_{2i}	N_{-i}/N_i (The second index in the second validation approach)
O_{jk}	The number of observations when the predicted stop number is "j" and the actual stop number is "k" in the big confusion matrices of Tables T3, and T4
Compromised predictions	In the third validation approach, if the predicted stop number is equal or greater than the actual stop number, then that prediction is acceptable. We may consider to accept the predictions in which the predicted is smaller than the actual by 1 or 2 stops. These predictions are referred to as compromised predictions in this paper.
LR	The Last Relationship which the name given to the following inequality which should be satisfied in the third validation approach:

concept	Definition
	$\sum_{j=7}^{20}\sum_{k=2}^{j-3} O_{jk} \leq \sum_{j=7}^{20}\sum_{k=j-2}^{27} O_{jk}$

Confusion matrices for each model:

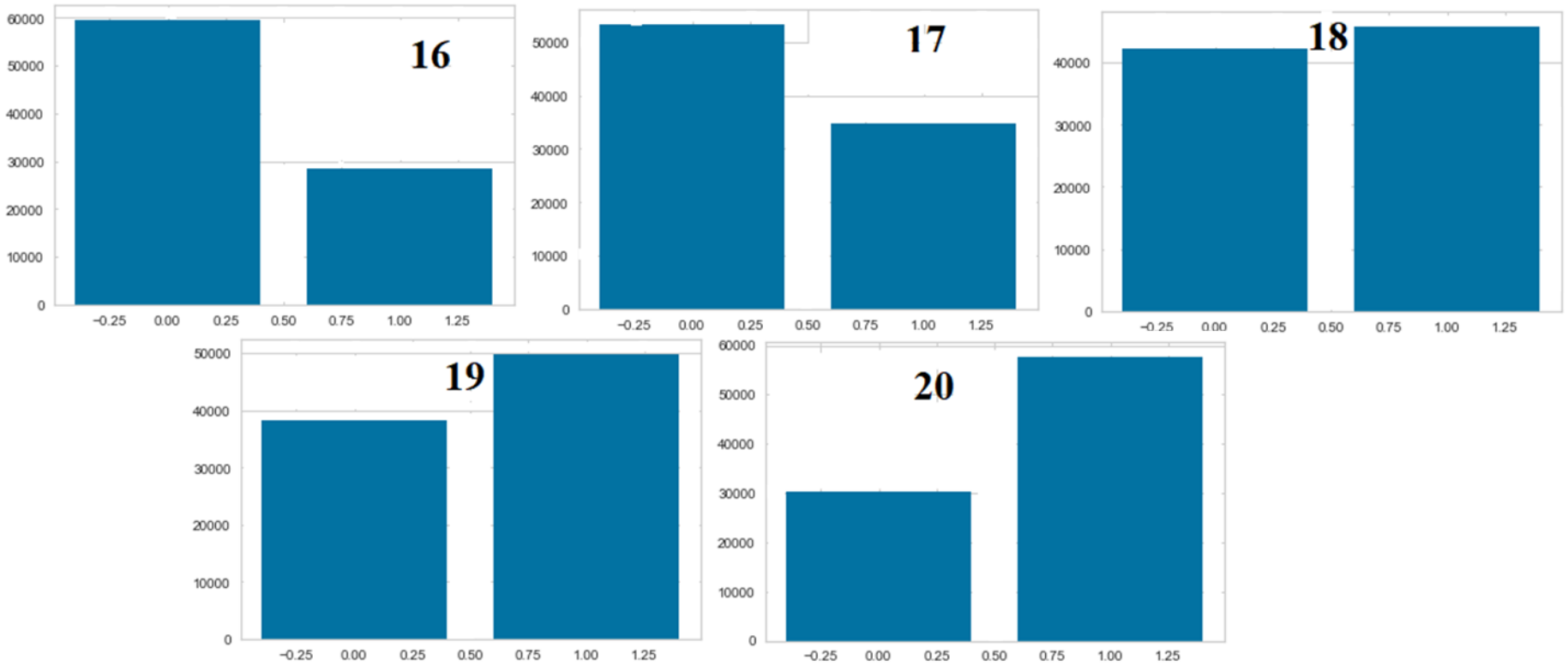
Stop 7: $\begin{bmatrix} 52805 & 2377 \\ 2456 & 1108 \end{bmatrix}$, Stop 8: $\begin{bmatrix} 50984 & 3017 \\ 3146 & 1599 \end{bmatrix}$, Stop 9: $\begin{bmatrix} 49766 & 3451 \\ 3528 & 2001 \end{bmatrix}$, Stop10: $\begin{bmatrix} 48789 & 3699 \\ 3755 & 2503 \end{bmatrix}$

Stop11: $\begin{bmatrix} 46545 & 4474 \\ 4498 & 3229 \end{bmatrix}$, Stop12: $\begin{bmatrix} 44011 & 5096 \\ 5225 & 4414 \end{bmatrix}$, Stop13: $\begin{bmatrix} 40359 & 6263 \\ 6174 & 5950 \end{bmatrix}$, Stop14: $\begin{bmatrix} 38431 & 6634 \\ 6741 & 6940 \end{bmatrix}$, Stop15: $\begin{bmatrix} 36144 & 7240 \\ 7392 & 7970 \end{bmatrix}$

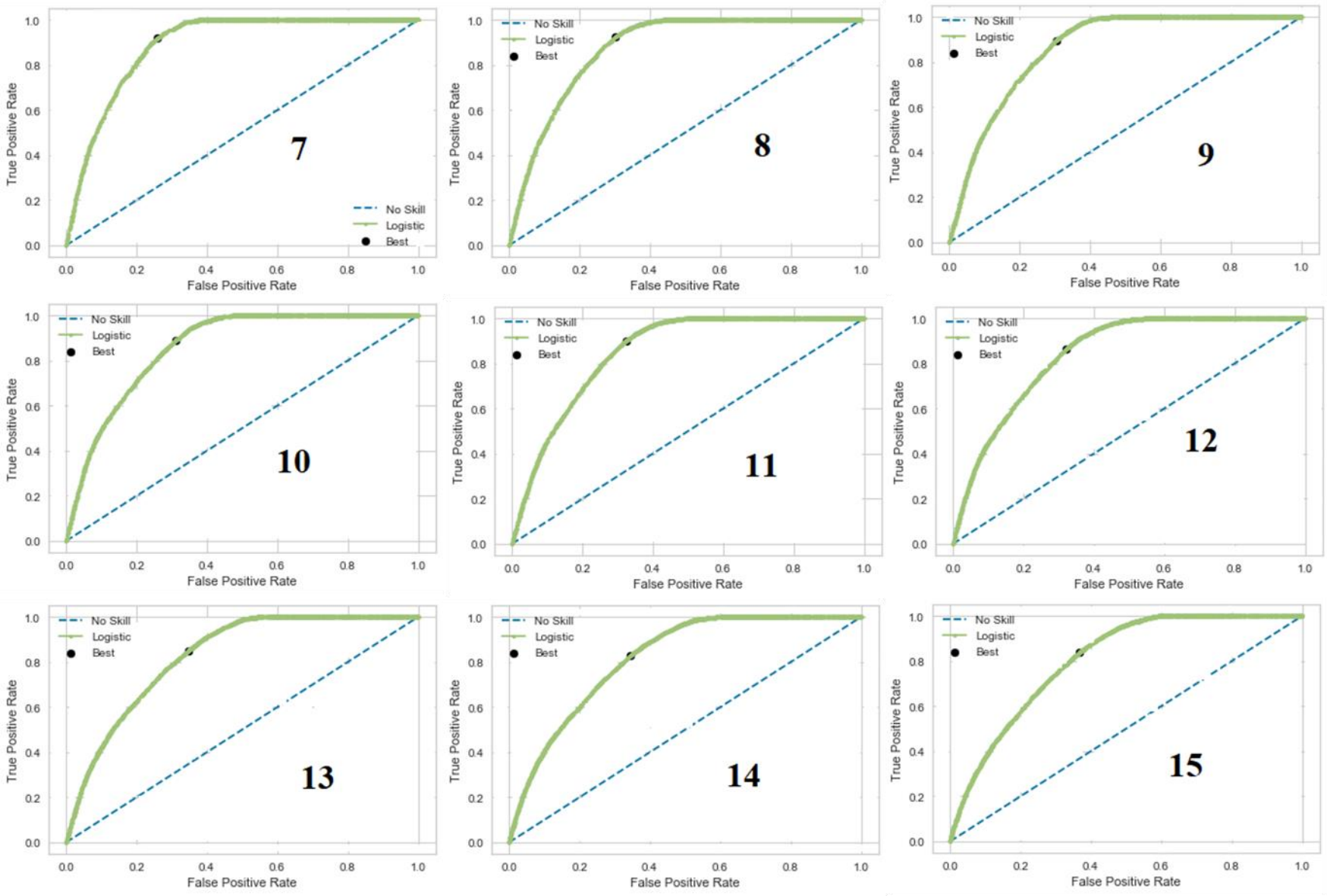
Stop16 : $\begin{bmatrix} 31663 & 8082 \\ 8175 & 10826 \end{bmatrix}$, Stop17: $\begin{bmatrix} 27088 & 8497 \\ 8520 & 14641 \end{bmatrix}$, Stop18: $\begin{bmatrix} 20566 & 7641 \\ 7782 & 22757 \end{bmatrix}$, Stop19: $\begin{bmatrix} 18434 & 7073 \\ 7092 & 26147 \end{bmatrix}$, Stop20: $\begin{bmatrix} 14528 & 5734 \\ 5792 & 32692 \end{bmatrix}$

Figures

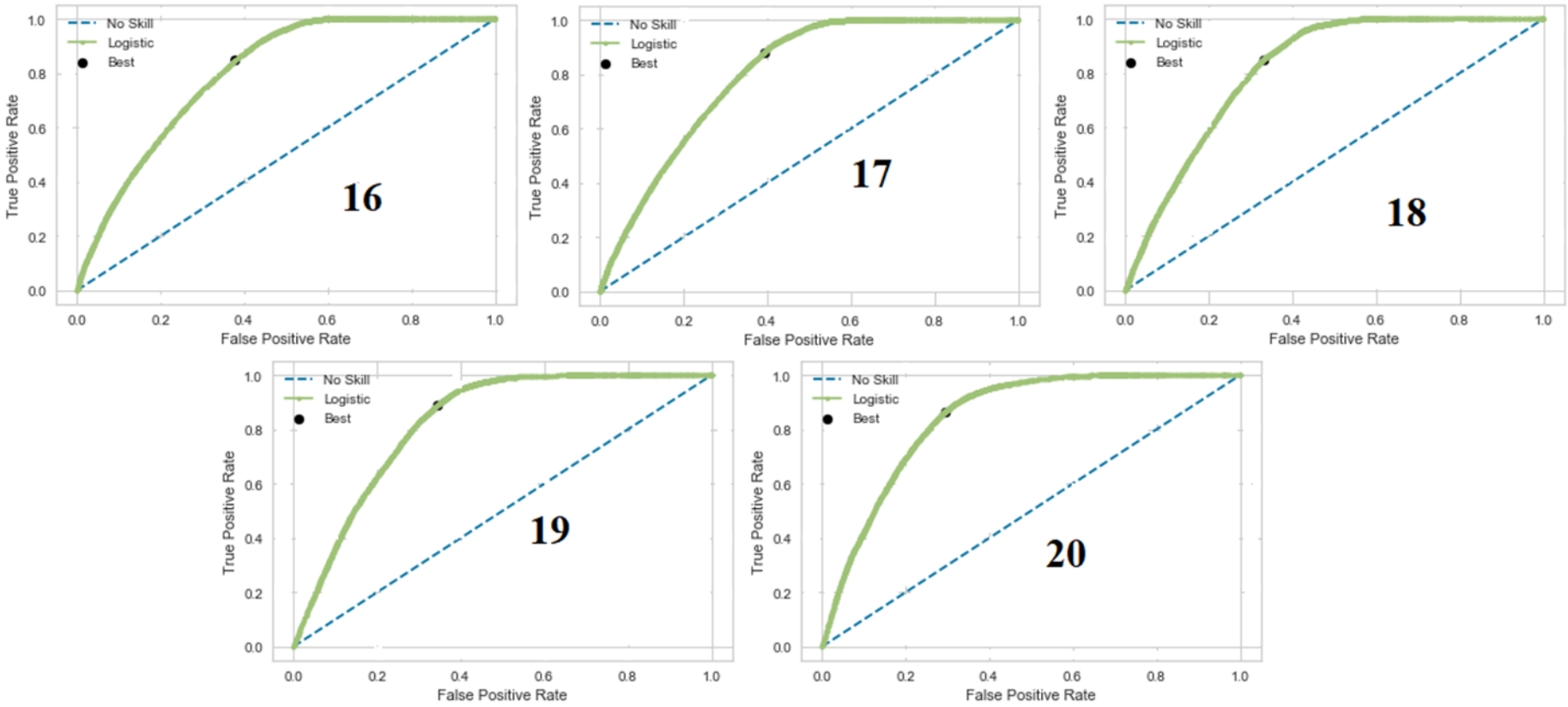
Figures F1 (continued)



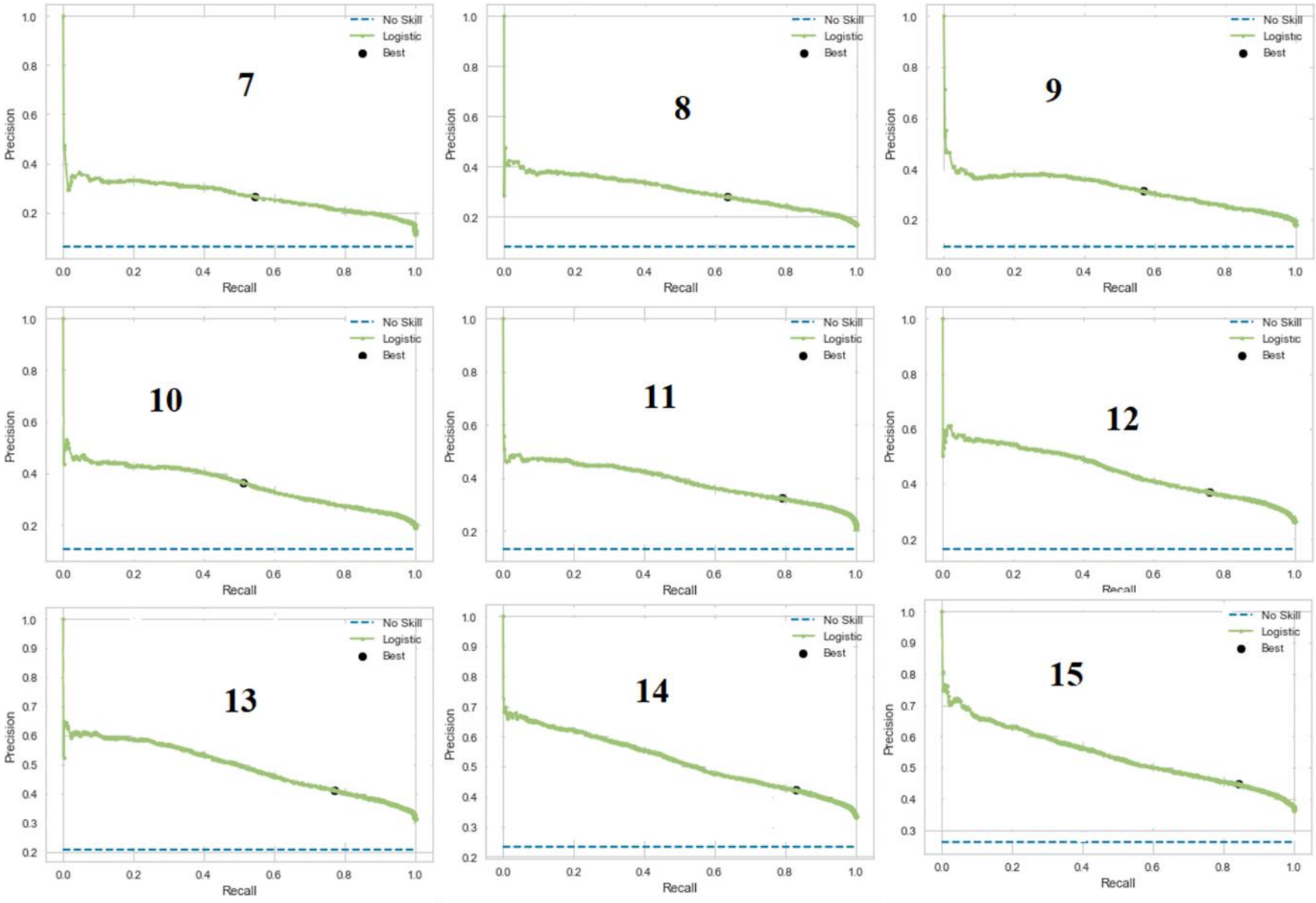
Figures F1. Bar plots for the number of members in each class in the train data set



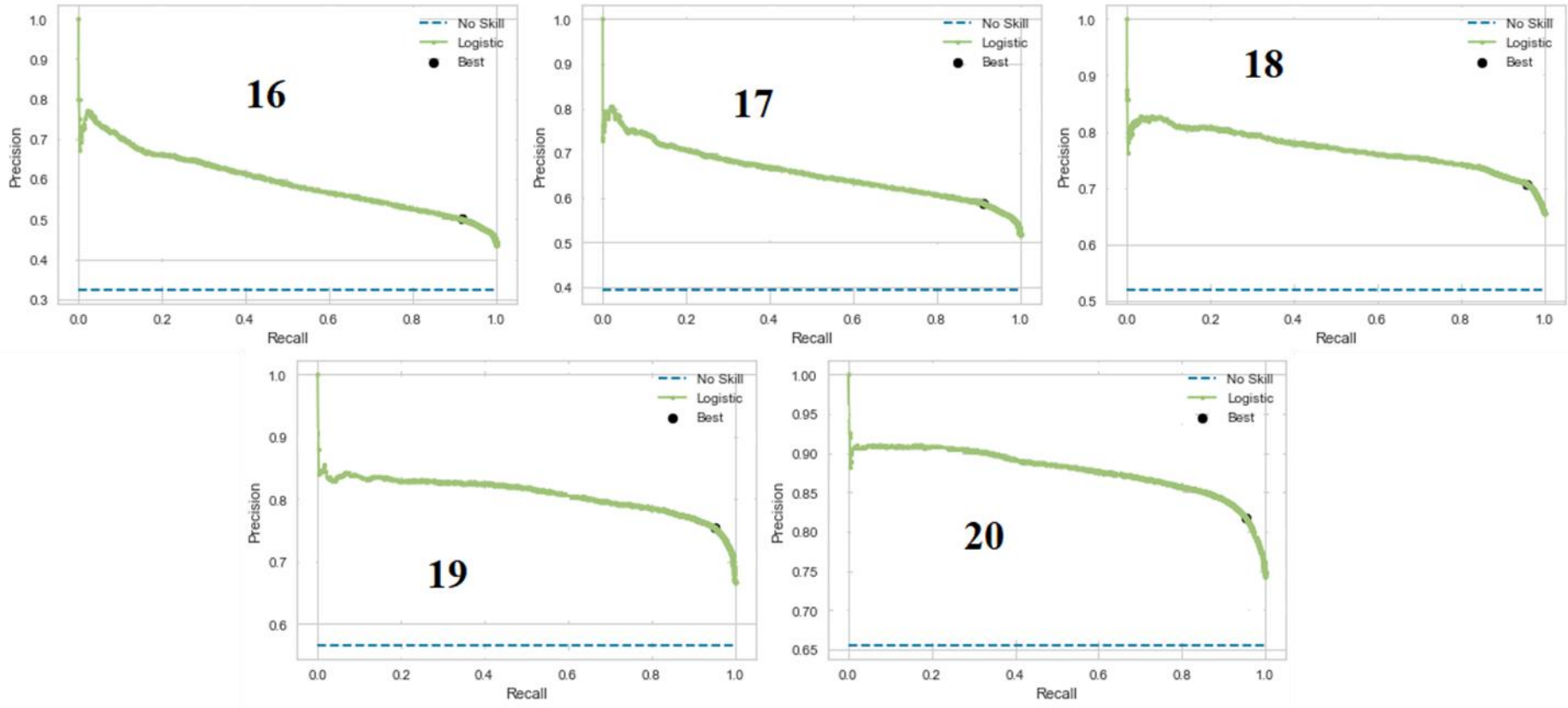
Figures F2 (continued)



Figures F2. ROC curve line plots for each model, with the optimal threshold



Figures F3 (continued)



Figures F3. Precision-Recall curves for each model, with the optimal threshold