

Data subset	ECE				MCE				Brier
	Equidistant bins		Equisized bins		Equidistant bins		Equisized bins		
	15 bins	50 bins	15 bins	50 bins	15 bins	50 bins	15 bins	50 bins	
DeepJIT with OPENSTACK									
Val. folds Avg	0.3509	0.3510	0.3509	0.3509	0.6197	0.7052	0.5302	0.6303	0.2378
Val. folds Min	0.3231	0.3231	0.3231	0.3231	0.5237	0.6127	0.4970	0.5659	0.2189
Val. folds Max	0.3788	0.3788	0.3788	0.3788	0.8704	0.8641	0.5720	0.7021	0.2599
Test	0.3470	0.3472	0.3470	0.3470	0.6823	0.5608	0.5628	0.6570	0.2362
DeepJIT with QT									
Val. folds Avg	0.3330	0.3330	0.3330	0.3330	0.6177	0.7226	0.5784	0.6495	0.1923
Val. folds Min	0.3157	0.3157	0.3157	0.3157	0.6440	0.5354	0.6000	0.1806	
Val. folds Max	0.3529	0.3529	0.3529	0.3529	0.6919	0.9420	0.6294	0.7140	0.2104
Test	0.3264	0.3264	0.3264	0.3264	0.6633	0.7443	0.6168	0.6665	0.1927

Table 1: ECE, MCE, Brier, and ROC-AUC results for different validation folds and binning methods -full table

Validation Fold	ECE				MCE				Brier
	Equal-interval bins		Equal-weight bins		Equal-interval bins		Equal-weight bins		
	15 bins	50 bins	15 bins	50 bins	15 bins	50 bins	15 bins	50 bins	
Average	0.3509	0.3510	0.3509	0.3509	0.6197	0.7052	0.5302	0.6303	0.2378
Val.Min	0.0140	0.0244	0.0214	0.0428	0.0970	0.2090	0.0458	0.1506	0.0875
Val. Max	0.0335	0.0439	0.0433	0.0596	0.0596	0.6127	0.1475	0.2563	0.1080
Test set result	0.3470	0.3472	0.3470	0.3470	0.6823	0.5608	0.5628	0.6570	0.2362

Table 2: Platt scaled DeepJIT