

Fair Prediction Sets Through Multi-objective Hyperparameter Optimization

Alberto García-Galindo, Marcos López-De-Castro, Rubén Armañanzas

Supplementary Material

A. Analysis of efficiency-fairness trade-off ($\alpha = 0.20$)

Similar to the procedure used in the main manuscript, we report the overall results of the produced Pareto set of optimal conformal predictors in terms of the proposed efficiency and fairness metrics averaged across 10 independent runs. This time, we set to 0.20 the significance level.

Table 1 Efficiency-fairness trade-off for the Adult Income dataset.

		SetSize _{val}	CovGap _{val}	SetSize _{test}	CovGap _{test}
logit	min	1.494 _(0.018)	1.822 _(0.996)	1.492 _(0.020)	1.903 _(0.554)
	Q1	1.503 _(0.024)	1.022 _(0.953)	1.501 _(0.026)	1.263 _(0.554)
	Q2	1.509 _(0.027)	0.624 _(0.955)	1.507 _(0.028)	0.917 _(0.602)
	Q3	1.517 _(0.031)	0.367 _(0.715)	1.516 _(0.035)	0.564 _(0.311)
	max	1.529 _(0.036)	0.143 _(0.372)	1.528 _(0.038)	0.319 _(0.225)
	Standard	1.565 _(0.014)	3.396 _(1.178)	1.566 _(0.012)	3.427 _(1.075)
	Mondrian	1.555 _(0.013)	3.451 _(1.192)	1.556 _(0.011)	3.603 _(1.085)
decision tree	min	1.275 _(0.013)	5.356 _(0.893)	1.273 _(0.016)	5.127 _(1.142)
	Q1	1.287 _(0.019)	4.261 _(0.988)	1.283 _(0.017)	4.022 _(0.975)
	Q2	1.301 _(0.022)	3.401 _(0.811)	1.297 _(0.023)	3.464 _(1.065)
	Q3	1.317 _(0.032)	2.671 _(0.779)	1.314 _(0.034)	2.880 _(1.098)
	max	1.426 _(0.090)	1.407 _(0.715)	1.426 _(0.092)	1.545 _(1.077)
	Standard	1.325 _(0.033)	4.969 _(1.709)	1.325 _(0.030)	4.318 _(1.881)
	Mondrian	1.311 _(0.028)	5.380 _(1.329)	1.311 _(0.026)	5.043 _(1.652)
random forest	min	1.213 _(0.008)	7.102 _(1.948)	1.213 _(0.007)	6.809 _(1.916)
	Q1	1.219 _(0.010)	5.608 _(0.684)	1.221 _(0.009)	5.093 _(0.600)
	Q2	1.230 _(0.012)	5.221 _(0.556)	1.231 _(0.012)	4.870 _(0.633)
	Q3	1.297 _(0.098)	4.517 _(0.971)	1.294 _(0.092)	4.361 _(0.942)
	max	1.392 _(0.127)	3.247 _(1.893)	1.393 _(0.127)	3.251 _(1.783)
	Standard	1.235 _(0.008)	7.937 _(0.982)	1.235 _(0.007)	7.793 _(0.772)
	Mondrian	1.232 _(0.011)	7.781 _(0.589)	1.232 _(0.011)	7.453 _(0.850)

Average distribution of efficiency in terms of average prediction set size (f_1), and fairness in terms of equalized coverage gap (f_2) for the validation and test sets. Equalized coverage gap is displayed as a percentage. Bold numbers highlight percentiles that were significantly different with respect to both the standard and Mondrian conformal predictors (Friedman test of differences and Nemenyi test to assess pairwise differences).

Table 2 Efficiency-fairness trade-off for the COMPAS dataset.

		SetSize _{val}	CovGap _{val}	SetSize _{test}	CovGap _{test}
logit	min	1.252 _(0.014)	2.123 _(1.304)	1.259 _(0.023)	2.368 _(0.780)
	Q1	1.261 _(0.013)	1.812 _(1.181)	1.268 _(0.024)	1.939 _(0.726)
	Q2	1.268 _(0.014)	1.575 _(1.163)	1.275 _(0.024)	1.682 _(0.844)
	Q3	1.275 _(0.012)	1.392 _(1.186)	1.283 _(0.022)	1.382 _(0.840)
	max	1.282 _(0.014)	1.254 _(1.190)	1.290 _(0.023)	1.199 _(0.885)
	Standard	1.272 _(0.017)	1.737 _(1.295)	1.280 _(0.022)	1.688 _(0.995)
	Mondrian	1.278 _(0.017)	1.748 _(1.502)	1.285 _(0.025)	1.757 _(0.795)
decision tree	min	1.266 _(0.022)	2.017 _(1.076)	1.274 _(0.027)	2.625 _(1.640)
	Q1	1.273 _(0.021)	1.045 _(0.571)	1.283 _(0.021)	2.174 _(1.465)
	Q2	1.285 _(0.019)	0.489 _(0.322)	1.292 _(0.019)	1.962 _(1.462)
	Q3	1.296 _(0.021)	0.207 _(0.243)	1.302 _(0.021)	1.255 _(1.004)
	max	1.319 _(0.026)	0.125 _(0.227)	1.325 _(0.026)	0.755 _(0.679)
	Standard	1.403 _(0.029)	2.858 _(1.682)	1.403 _(0.027)	1.516 _(1.772)
	Mondrian	1.403 _(0.031)	3.115 _(1.694)	1.404 _(0.027)	2.099 _(1.831)
random forest	min	1.319 _(0.229)	1.600 _(0.853)	1.319 _(0.229)	2.384 _(1.201)
	Q1	1.324 _(0.227)	1.191 _(0.636)	1.324 _(0.227)	1.873 _(1.034)
	Q2	1.329 _(0.226)	0.821 _(0.524)	1.330 _(0.226)	1.587 _(1.033)
	Q3	1.332 _(0.225)	0.561 _(0.418)	1.333 _(0.225)	1.395 _(0.989)
	max	1.340 _(0.222)	0.316 _(0.306)	1.340 _(0.222)	1.153 _(0.992)
	Standard	1.288 _(0.030)	2.277 _(1.290)	1.291 _(0.035)	2.777 _(1.581)
	Mondrian	1.298 _(0.043)	2.177 _(1.253)	1.296 _(0.041)	2.580 _(1.682)

Average distribution of efficiency in terms of average prediction set size (f_1), and fairness in terms of equalized coverage gap (f_2) for the validation and test sets. Equalized coverage gap is displayed as a percentage. Bold numbers highlight percentiles that were significantly different with respect to both the standard and Mondrian conformal predictors (Friedman test of differences and Nemenyi test to assess pairwise differences).

Table 3 Efficiency-fairness trade-off for the Diabetes dataset.

		SetSize _{val}	CovGap _{val}	SetSize _{test}	CovGap _{test}
logit	min	1.620 _(0.005)	1.491 _(0.411)	1.618 _(0.006)	1.802 _(0.584)
	Q1	1.622 _(0.005)	1.164 _(0.436)	1.621 _(0.006)	1.552 _(0.579)
	Q2	1.625 _(0.005)	1.020 _(0.365)	1.623 _(0.005)	1.403 _(0.558)
	Q3	1.630 _(0.005)	0.901 _(0.341)	1.628 _(0.005)	1.262 _(0.544)
	max	1.706 _(0.016)	0.753 _(0.318)	1.705 _(0.014)	1.071 _(0.566)
	Standard	1.623 _(0.006)	1.248 _(0.401)	1.620 _(0.007)	1.278 _(0.492)
	Mondrian	1.623 _(0.006)	1.181 _(0.432)	1.621 _(0.006)	1.475 _(0.528)
decision tree	min	1.639(0.005)	1.736 _(0.496)	1.636(0.008)	1.914 _(0.518)
	Q1	1.643(0.006)	1.244 _(0.380)	1.641(0.007)	1.659 _(0.433)
	Q2	1.658(0.029)	1.087 _(0.293)	1.656(0.030)	1.462 _(0.403)
	Q3	1.674 _(0.061)	0.953(0.221)	1.672 _(0.061)	1.303 _(0.373)
	max	1.697 _(0.066)	0.745(0.196)	1.695 _(0.068)	1.022(0.350)
	Standard	1.699 _(0.013)	1.694 _(0.503)	1.697 _(0.014)	1.734 _(0.530)
	Mondrian	1.701 _(0.013)	1.729 _(0.564)	1.699 _(0.013)	1.826 _(0.498)
random forest	min	1.620 _(0.009)	1.491 _(0.239)	1.618 _(0.008)	1.802 _(0.372)
	Q1	1.622 _(0.010)	1.164 _(0.186)	1.621 _(0.008)	1.552 _(0.363)
	Q2	1.625 _(0.010)	1.020 _(0.238)	1.623 _(0.008)	1.403 _(0.355)
	Q3	1.630 _(0.010)	0.901 _(0.246)	1.628 _(0.009)	1.262 _(0.340)
	max	1.706 _(0.148)	0.753(0.237)	1.705 _(0.148)	1.071 _(0.351)
	Standard	1.623 _(0.010)	1.248 _(0.323)	1.620 _(0.008)	1.278 _(0.379)
	Mondrian	1.623 _(0.009)	1.181 _(0.285)	1.621 _(0.008)	1.475 _(0.404)

Average distribution of efficiency in terms of average prediction set size (f_1), and fairness in terms of equalized coverage gap (f_2) for the validation and test sets. Equalized coverage gap is displayed as a percentage. Bold numbers highlight percentiles that were significantly different with respect to both the standard and Mondrian conformal predictors (Friedman test of differences and Nemenyi test to assess pairwise differences).

Table 4 Efficiency-fairness trade-off for the Nursery dataset.

		SetSize _{val}	CovGap _{val}	SetSize _{test}	CovGap _{test}
logit	min	0.954 (0.023)	1.436 _(1.397)	0.956 (0.020)	2.648 _(1.088)
	Q1	0.968 (0.024)	0.645 _(0.971)	0.970 (0.022)	1.953 _(1.092)
	Q2	0.976 (0.024)	0.234 _(0.401)	0.978 (0.022)	1.552 _(0.974)
	Q3	0.985 (0.025)	0.091 _(0.122)	0.985 (0.022)	1.245 _(0.993)
	max	1.007 _(0.025)	0.026 _(0.031)	1.006 _(0.027)	0.894 _(0.949)
	Standard	1.030 _(0.020)	1.691 _(1.027)	1.030 _(0.021)	2.473 _(1.085)
	Mondrian	1.030 _(0.022)	1.744 _(1.001)	1.030 _(0.024)	2.393 _(1.191)
decision tree	min	0.800 (0.013)	1.455 _(1.200)	0.800 (0.012)	2.526 _(1.410)
	Q1	0.809 (0.014)	0.687 _(0.696)	0.810 (0.012)	2.053 _(1.077)
	Q2	0.812 (0.014)	0.371 _(0.429)	0.814 (0.013)	1.567 _(0.808)
	Q3	0.819 (0.016)	0.102 _(0.078)	0.820 (0.014)	1.252 _(0.745)
	max	0.835 (0.033)	0.029 _(0.023)	0.835 (0.029)	0.957 _(0.776)
	Standard	0.932 _(0.007)	2.036 _(1.129)	0.928 _(0.006)	1.952 _(1.159)
	Mondrian	0.933 _(0.007)	1.938 _(1.096)	0.928 _(0.006)	1.878 _(1.163)
random forest	min	0.792 _(0.011)	3.513 _(4.128)	0.797 _(0.011)	3.907 _(2.855)
	Q1	0.797 _(0.011)	1.573 (1.284)	0.801 _(0.012)	1.879 (0.789)
	Q2	0.800 _(0.012)	0.992 (0.904)	0.804 _(0.013)	1.432 (0.745)
	Q3	0.849 _(0.095)	0.736 (0.848)	0.853 _(0.099)	1.067 (0.731)
	max	0.922 _(0.143)	0.440 (0.633)	0.924 (0.144)	0.527 (0.532)
	Standard	0.798 _(0.011)	4.192 _(1.184)	0.801 _(0.011)	3.723 _(1.007)
	Mondrian	0.801 _(0.012)	4.211 _(0.892)	0.803 _(0.013)	4.014 _(0.981)

Average distribution of efficiency in terms of average prediction set size (f_1), and fairness in terms of equalized coverage gap (f_2) for the validation and test sets. Equalized coverage gap is displayed as a percentage. Bold numbers highlight percentiles that were significantly different with respect to both the standard and Mondrian conformal predictors (Friedman test of differences and Nemenyi test to assess pairwise differences).

B. Analysis of efficiency-fairness trade-off ($\alpha = 0.05$)

Similarly, we report the overall results of the produced Pareto set of optimal conformal predictors in terms of the proposed efficiency and fairness metrics averaged across 10 independent runs. Here, we set to 0.05 the significance level.

Table 5 Efficiency-fairness trade-off for the Adult Income dataset.

		SetSize _{val}	CovGap _{val}	SetSize _{test}	CovGap _{test}
logit	min	2.199 _(0.031)	3.368 _(0.666)	2.197 _(0.029)	3.338 _(0.500)
	Q1	2.223 _(0.031)	2.309 _(0.532)	2.224 _(0.032)	2.394 _(0.310)
	Q2	2.244 _(0.029)	1.868 _(0.363)	2.244 _(0.030)	1.827 _(0.232)
	Q3	2.261 _(0.027)	1.549 _(0.266)	2.260 _(0.030)	1.579 _(0.213)
	max	2.288 _(0.037)	1.223 _(0.318)	2.287 _(0.038)	1.227 _(0.320)
	Standard	2.270 _(0.029)	2.029 _(0.465)	2.269 _(0.029)	2.033 _(0.416)
	Mondrian	2.266 _(0.034)	2.198 _(0.577)	2.266 _(0.034)	2.176 _(0.564)
decision tree	min	1.884 _(0.018)	1.173 _(0.817)	1.886 _(0.015)	1.458 _(0.532)
	Q1	1.898 _(0.021)	0.667 _(0.305)	1.898 _(0.022)	0.966 _(0.275)
	Q2	1.956 _(0.081)	0.302 _(0.245)	1.953 _(0.082)	0.569 _(0.229)
	Q3	1.970 _(0.083)	0.222 _(0.206)	1.967 _(0.085)	0.368 _(0.224)
	max	2.004 _(0.104)	0.034 _(0.046)	2.003 _(0.105)	0.196 _(0.181)
	Standard	1.978 _(0.031)	0.621 _(0.402)	1.981 _(0.029)	0.812 _(0.587)
	Mondrian	1.980 _(0.027)	0.590 _(0.438)	1.981 _(0.029)	0.712 _(0.550)
random forest	min	1.827 _(0.019)	1.846 _(0.871)	1.828 _(0.018)	1.553 _(0.479)
	Q1	1.855 _(0.027)	1.174 _(0.272)	1.853 _(0.026)	1.233 _(0.393)
	Q2	1.886 _(0.034)	0.814 _(0.218)	1.886 _(0.032)	0.992 _(0.381)
	Q3	1.930 _(0.030)	0.531 _(0.199)	1.926 _(0.025)	0.635 _(0.354)
	max	2.072 _(0.314)	0.143 _(0.123)	2.073 _(0.314)	0.191 _(0.273)
	Standard	1.850 _(0.022)	1.777 _(0.394)	1.847 _(0.019)	1.756 _(0.439)
	Mondrian	1.844 _(0.025)	1.799 _(0.368)	1.844 _(0.021)	1.641 _(0.329)

Average distribution of efficiency in terms of average prediction set size (f_1), and fairness in terms of equalized coverage gap (f_2) for the validation and test sets. Equalized coverage gap is displayed as a percentage. Bold numbers highlight percentiles that were significantly different with respect to both the standard and Mondrian conformal predictors (Friedman test of differences and Nemenyi test to assess pairwise differences).

Table 6 Efficiency-fairness trade-off for the COMPAS dataset.

		SetSize _{val}	CovGap _{val}	SetSize _{test}	CovGap _{test}
logit	min	1.728 _(0.017)	1.959 _(0.891)	1.735 _(0.017)	1.184 _(0.761)
	Q1	1.730 _(0.017)	1.885 _(0.908)	1.738 _(0.017)	1.115 _(0.768)
	Q2	1.733 _(0.018)	1.735 _(0.920)	1.741 _(0.016)	1.050 _(0.767)
	Q3	1.736 _(0.017)	1.678 _(0.947)	1.744 _(0.015)	1.012 _(0.775)
	max	1.739 _(0.015)	1.614 _(0.951)	1.746 _(0.014)	0.955 _(0.766)
	Standard	1.736 _(0.021)	2.040 _(0.908)	1.741 _(0.019)	1.046 _(0.746)
	Mondrian	1.743 _(0.022)	1.802 _(0.889)	1.749 _(0.019)	1.294 _(0.635)
decision tree	min	1.734 _(0.036)	1.000 _(0.819)	1.748 _(0.034)	1.978 _(0.813)
	Q1	1.746 _(0.036)	0.637 _(0.601)	1.760 _(0.034)	1.487 _(0.835)
	Q2	1.755 _(0.040)	0.284 _(0.127)	1.770 _(0.036)	1.222 _(0.849)
	Q3	1.768 _(0.038)	0.102 _(0.047)	1.782 _(0.036)	0.979 _(0.749)
	max	1.782 _(0.041)	0.018 _(0.021)	1.799 _(0.041)	0.600 _(0.731)
	Standard	2.000 _(0.000)	0.000 _(0.000)	2.000 _(0.000)	0.000 _(0.000)
	Mondrian	2.000 _(0.000)	0.000 _(0.000)	2.000 _(0.000)	0.000 _(0.000)
random forest	min	1.701 _(0.031)	2.096 _(1.063)	1.702 _(0.028)	2.034 _(0.667)
	Q1	1.714 _(0.032)	1.605 _(0.936)	1.713 _(0.029)	1.655 _(0.643)
	Q2	1.721 _(0.035)	1.365 _(0.883)	1.721 _(0.030)	1.396 _(0.608)
	Q3	1.732 _(0.033)	1.112 _(0.760)	1.732 _(0.028)	1.135 _(0.645)
	max	1.749 _(0.030)	0.603 _(0.490)	1.744 _(0.029)	0.680 _(0.625)
	Standard	1.756 _(0.030)	1.488 _(0.734)	1.754 _(0.024)	1.149 _(0.772)
	Mondrian	1.759 _(0.038)	1.639 _(1.025)	1.756 _(0.032)	0.957 _(0.492)

Average distribution of efficiency in terms of average prediction set size (f_1), and fairness in terms of equalized coverage gap (f_2) for the validation and test sets. Equalized coverage gap is displayed as a percentage. Bold numbers highlight percentiles that were significantly different with respect to both the standard and Mondrian conformal predictors (Friedman test of differences and Nemenyi test to assess pairwise differences).

Table 7 Efficiency-fairness trade-off for the Diabetes dataset.

		SetSize _{val}	CovGap _{val}	SetSize _{test}	CovGap _{test}
logit	min	2.386 _(0.006)	0.859 _(0.198)	2.385 _(0.005)	0.581 _(0.207)
	Q1	2.387 _(0.006)	0.806 _(0.187)	2.386 _(0.005)	0.540 _(0.161)
	Q2	2.390 _(0.005)	0.746 _(0.154)	2.388 _(0.006)	0.511 _(0.155)
	Q3	2.392 _(0.007)	0.730 _(0.148)	2.389 _(0.006)	0.504 _(0.203)
	max	2.394 _(0.006)	0.696 _(0.136)	2.393 _(0.005)	0.486 _(0.178)
	Standard	2.399 _(0.006)	0.863 _(0.186)	2.398 _(0.008)	0.616 _(0.173)
	Mondrian	2.401 _(0.006)	0.733 _(0.172)	2.401 _(0.008)	0.730 _(0.161)
decision tree	min	2.384 _(0.012)	0.646 _(0.236)	2.384 _(0.013)	0.869 _(0.231)
	Q1	2.390 _(0.012)	0.500 _(0.148)	2.391 _(0.013)	0.765 _(0.245)
	Q2	2.396 _(0.013)	0.457 _(0.143)	2.395 _(0.014)	0.672 _(0.228)
	Q3	2.408 _(0.023)	0.408 _(0.125)	2.408 _(0.025)	0.569 _(0.230)
	max	2.458 _(0.108)	0.381 _(0.135)	2.458 _(0.108)	0.478 _(0.226)
	Standard	2.567 _(0.033)	1.628 _(0.565)	2.569 _(0.032)	1.505 _(0.760)
	Mondrian	2.567 _(0.033)	1.442 _(0.498)	2.567 _(0.027)	1.505 _(0.647)
random forest	min	2.375 _(0.007)	0.728 _(0.236)	2.375 _(0.008)	0.841 _(0.305)
	Q1	2.378 _(0.008)	0.641 _(0.220)	2.379 _(0.008)	0.740 _(0.236)
	Q2	2.380 _(0.008)	0.562 _(0.202)	2.380 _(0.008)	0.683 _(0.204)
	Q3	2.381 _(0.008)	0.518 _(0.189)	2.382 _(0.008)	0.633 _(0.200)
	max	2.453 _(0.194)	0.373 _(0.204)	2.454 _(0.192)	0.503 _(0.270)
	Standard	2.383 _(0.010)	0.714 _(0.265)	2.383 _(0.011)	0.675 _(0.244)
	Mondrian	2.384 _(0.010)	0.582 _(0.280)	2.385 _(0.010)	0.696 _(0.307)

Average distribution of efficiency in terms of average prediction set size (f_1), and fairness in terms of equalized coverage gap (f_2) for the validation and test sets. Equalized coverage gap is displayed as a percentage. Bold numbers highlight percentiles that were significantly different with respect to both the standard and Mondrian conformal predictors (Friedman test of differences and Nemenyi test to assess pairwise differences).

Table 8 Efficiency-fairness trade-off for the Nursery dataset.

		SetSize _{val}	CovGap _{val}	SetSize _{test}	CovGap _{test}
logit	min	1.329 _(0.021)	1.843 _(0.926)	1.332 _(0.021)	2.392 _(0.717)
	Q1	1.346 _(0.017)	1.280 _(0.636)	1.349 _(0.015)	1.659 _(0.537)
	Q2	1.362 _(0.022)	0.841 _(0.552)	1.367 _(0.021)	1.361 _(0.568)
	Q3	1.375 _(0.022)	0.570 _(0.509)	1.379 _(0.023)	1.087 _(0.570)
	max	1.551 _(0.446)	0.204 _(0.304)	1.557 _(0.444)	0.553 _(0.451)
	Standard	1.482 _(0.111)	2.140 _(0.575)	1.482 _(0.109)	2.448 _(0.507)
	Mondrian	1.493 _(0.129)	2.269 _(0.518)	1.494 _(0.126)	2.362 _(0.349)
decision tree	min	0.967 _(0.016)	1.973 _(0.810)	0.969 _(0.015)	1.931 _(0.680)
	Q1	0.974 _(0.016)	1.301 _(0.562)	0.976 _(0.015)	1.371 _(0.443)
	Q2	0.993 _(0.043)	0.685 _(0.611)	0.994 _(0.043)	1.019 _(0.364)
	Q3	1.019 _(0.073)	0.262 _(0.324)	1.020 _(0.074)	0.714 _(0.276)
	max	1.087 _(0.112)	0.027 _(0.074)	1.088 _(0.116)	0.320 _(0.296)
	Standard	0.977 _(0.010)	1.985 _(0.974)	0.976 _(0.009)	1.930 _(0.684)
	Mondrian	0.981 _(0.012)	2.374 _(0.622)	0.980 _(0.012)	2.158 _(0.611)
random forest	min	0.953 _(0.008)	1.447 _(0.681)	0.954 _(0.009)	1.770 _(0.568)
	Q1	0.958 _(0.007)	0.897 _(0.499)	0.959 _(0.008)	1.434 _(0.474)
	Q2	0.961 _(0.007)	0.633 _(0.435)	0.962 _(0.008)	1.136 _(0.474)
	Q3	0.966 _(0.009)	0.383 _(0.192)	0.968 _(0.009)	0.835 _(0.451)
	max	1.032 _(0.163)	0.106 _(0.111)	1.032 _(0.161)	0.504 _(0.491)
	Standard	0.998 _(0.015)	1.497 _(0.475)	0.998 _(0.015)	1.139 _(0.543)
	Mondrian	1.004 _(0.017)	1.078 _(0.578)	1.005 _(0.018)	0.984 _(0.580)

Average distribution of efficiency in terms of average prediction set size (f_1), and fairness in terms of equalized coverage gap (f_2) for the validation and test sets. Equalized coverage gap is displayed as a percentage. Bold numbers highlight percentiles that were significantly different with respect to both the standard and Mondrian conformal predictors (Friedman test of differences and Nemenyi test to assess pairwise differences).