University of Michigan

Project Report Ver2.0

A Locally-aware Fairness Calibration Model Prediction as well as its application on pre-trial crime prediction task

Reporter: Naicheng Wu

1

[1] This project is heavily based on the application and discussion of the paper *Toward Group-wise Calibration: A Locally-aware Calibration Method for Proba-bilistic Classifiers*. So the reference should also includes all reference made in that paper.

Contents

1 Introduction

These days, with the fast development of Machine Learning Technology, the usage of algorithm prediction/detection is gradually expanding to the field of human action prediction, rather than objects like image or natural language. For example, courts are now widely using software, like COMPAS , to predict the defendant's possibility of recidivism or escape from trail. Such predictions are normally used to decide the bail amount/ opportunity in the pretrial hearing, however, some areas like Wisconsin started to use COMPAS for the official sentencing purpose. [1]

Such rapid increasing of usage in those critical area, often related to the fairness and future among races, surely worth more careful exam. Unfortunately, most scholars in ML fairness area believe that COMPAS is obviously biased against black people, compare to the white. For example, Angwin et al. state that "blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend" in 2016 [2], which is also confirmed in my project.

Considering the potential impact of those algorithms, as well as their critical fairness defect, this project firstly confirms existing (individual) unfairness, then analysis the cause of them and finally offers constructive improvement method for COMPAS's recidivism prediction, as an example. More importantly, this project also proposed a uniform and general solution for (data-caused) individual bias, especially the one which applies a novel fairness metric "Local Calibration" as well as its Probability Estimator and Calibrator.

2 Related Works

2.1 COMPAS's Similarity with Simple Logistic Regression

Although COMPAS's detailed algorithm is business secret, scholars like Julia Dressel et al. successfully shows that COMPAS's performance has no statistical difference with a simple Logistic Regression among those important benchmark like accuracy and individual fairness (among different races).[3] More importantly, such a Logistic Regression only uses 2 of the 137 features used by COMPAS: Age and Adulthood Crime counts. There major results are shown below.

Performances	LR_2	COMPAS
Accuracy% (Overall)	66.8	65.4
Accuracy% (Black)	66.7	63.8
Accuracy% (White)	66.4	67.0
False Positive% (Black)	45.6	44.8
False Positive% (White)	25.3	23.5
False Negative% (Black)	21.6	28.0
False Negative% (White)	46.1	47.7

So based on this results, we can focus on the analysis and improvement on the logistic regression model.

2.2 Fairness Measurement and Improvement Insight

Jon Kleinberg et al. mentioned three common measurement of fairness among groups.[4] They argue that to achieve perfect fairness among different groups (in this project, the group means race), one binary classification algorithm must assure the following three rates are correspondingly the same among different groups: (TP = True Positive, so as TN, FP, FN)

$$Calibration : TP/(TP + FP) \tag{1}$$

$$False\ Negative\ Rate : FN/(TP + FN) \tag{2}$$

$$False\ Positive\ Rate : FP/(TN + FP) \tag{3}$$

Actually the first one measure the (positive) accuracy while the remaining two means the statistically, people from different groups will be treated similarly if they are similar (reflected as their ground truth y).

In this project, we focus on term(2) and term(3) since they are directly linked to the fairness performance. For the recidivism prediction, "Positive" result means you are more likely to conduct a crime again, which means higher false positive rate for a certain group means a punishment or discrimination to such group while the false negative rate means the opposite thing. So based on previous table, black people do bear unfair discrimination from this algorithm.

Also, Jon Kleinberg et al. argue that the three rates can't be the same between two groups unless one of the following requirements are met:

$$The\, predictor\ has\ 100\%\ success\, rate. \tag{4}$$

$$The\ Positive/\ Negative\, rates\, are\, the\, same\, among\, two\, groups. \tag{5}$$

Although they are both quite obvious (the first one is trivial), it gives me some insight in the data-balance issue.

2.3 Local Calibration

This is the key idea in this project, which is recently introduced by Dr.Arya Farahi as well as Prof.Danai Koutra from The Michigan Institute for Data Science (MIDAS) of the University of Michigan.

In their paper *Toward Group-wise Calibration: A Locally-aware Calibration Method for Probabilistic Classifiers*[5], they introduced a new metric of fairness called local calibration.

Suppose $p(\mathbf{y} = 1 \mid \mathbf{x}, \widehat{f} = \alpha)$ is the conditional true frequency of a data instance $\mathbf{x}$ belongs to class $\mathbf{y} = 1$ at probability level α. $p(\mathbf{x})$ is the population join density over domain space $\mathcal{X}$. Finally, let α be the probability level corresponds to all instances of data that the model $\widehat{f}$ predicts probability of α. Then the model $\widehat{f}$ is locally calibrated if satisfies

$$p(\mathbf{y} = 1 \mid \mathbf{x}, \widehat{f} = \alpha) = \alpha. \tag{6}$$

for all $\mathbf{x} \in \mathcal{X}$ and $\alpha \in [0, 1]$.

And they also mention that if the above conditional true frequency of the data is not conditioned in $\mathbf{x}$, but uniformly on the whole dataset, namely:

$$p(\mathbf{y} = 1 \mid \widehat{f} = \alpha) = \alpha. \tag{7}$$

then this model is called globally calibrated. So apparently this is a weaker calibration status. The authors argued that most current research on calibration fields only focus on the latter, however, they believed this is not enough.

They argued that the meaning of the novel metric local calibration is to focus on group-wise fairness. For example, if group A's samples are all positive while group B's are all negative. If we mixed them into the same set, due to the limitation of the model's expression ability, if it is only globally calibrated, fairness issue among groups may happen. (like huge false positives in B but huge false positives in A)

They also introduced an estimator for local-calibration called Expected Local Calibration Error (ELCE):

Firstly define a helper random variable:

$$\mathbb{Y}_\alpha := \mathbb{I}(\mathbf{y} = 1 \mid \widehat{f} = \alpha), \tag{8}$$

where $\mathbb{I}[\cdot]$ is an indicator function that takes the value of 1/0 if its argument is true/false. Thus, $\mathbb{Y}_\alpha$ is a binary random variable with value 1 or 0.

$$\mathrm{ELCE}^2[\mathcal{G}, \widehat{f}, p] := \left[\sup_{g \in \mathcal{G}} \mathbb{E}_{\mathbf{x} \sim p}\left[(\mathbb{Y}_\alpha - \alpha) g(\mathbf{x}, \alpha)\right]\right]^2. \tag{9}$$

where $\mathcal{G}$ is the domain of non-zero functions g. $\widehat{f}$ is locally calibrated if and only if $\mathrm{ELCE}^2[\cdots] = 0$ for every bounded function g defined on the domain $\mathcal{X} \times \mathcal{R}$, which they offered their proof.

2.4 Calibrator

Naturally, with an effective estimator of local calibration, they introduced two calibrator. First we assume the calibrated model(classifier) can be modeled as the original classifier adding a bias which produced by the calibrator, namely:

$$\widehat{f_c}(\mathbf{x}) = \widehat{f}(\mathbf{x}) + b(\mathbf{x}). \tag{10}$$

The first calibrator is directly from the ELCE, called EWF(error witness function). EWF is actually a closed form solution of ELCE that determines the discrepancy between the observed class labels and predicted probabilities.

$$\mathrm{EWF}[K, \widehat{f}, \mathbf{x}'] := \mathbb{E}_{\mathbf{x} \sim q}\left[(\mathbf{y} - \widehat{f}(\mathbf{x}))\, K(\mathbf{x}', \mathbf{x})\right], \tag{11}$$

where $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j) \times l(\widehat{f_i}, \widehat{f_j})$ (k, l are kernel functions, in this project, we choose RBF with $\sigma = 1$).

After properly adapt the above EWF to our interested binary classification case with normalization, we get the following calibrator:

$$b_{\mathrm{EWF}}(\mathbf{x}') = \mathbb{E}_{\mathbf{x} \sim q}\left[(\mathbf{y} - \widehat{f}(\mathbf{x}))\, K(\mathbf{x}, \mathbf{x}'))\right]. \tag{12}$$

The second model take advantage of the kernel space and employ a linear model to estimate the additive bias. However, since the dimensionallity may be very large, imposing sparsity improves the performance of our model. They, therefore, take kernel Ridge Regression model to construct the second estimator.

$$\widehat{b}_{\mathrm{KRR}}(\mathbf{x}) = \mathbf{b}(K + \lambda\mathbb{I})^{-1}\kappa^\top(\mathbf{x}) \tag{13}$$

where $\mathbf{b} = [(\mathbf{y}_1 - \widehat{f_1}), \cdots, (\mathbf{y}_n - \widehat{f_n})]$, $K_{i,j}$ is a kernel function, $\mathbb{I}$ is an identity matrix of size $n \times n$, and $\kappa(\mathbf{x}) = [k(\mathbf{x}_0, \mathbf{x}), \cdots, k(\mathbf{x}_n, \mathbf{x})]$. is a regularization penalty coefficient which determines the smoothness of learned re-calibration model and is set by the user.

They then produced the following experiment to show both of their ELCE estimator and two calibrators are effective.

The following generative model is employed to produce pairs of $\{\mathbf{x}, \mathbf{y}\}$.

$$\mathbf{x}_1 \sim \mathcal{N}(\mu = 0, \sigma^2 = 1), \tag{14}$$

$$\mathbf{x}_2 \sim \mathcal{N}(\mu = 0, \sigma^2 = 1), \tag{15}$$

$$\mathbf{y} \sim \mathrm{Bernoulli}(\bar{p}). \tag{16}$$

where $\bar{p} = \mathrm{sigmoid}(\mathbf{x}_1 + \mathbf{x}_2)$. $f(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{sigmoid}(\mathbf{x}_1 + \mathbf{x}_2)$ is the true (properly calibrated) model. Suppose the following classification models: $\widehat{f_1}(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{sigmoid}(\mathbf{x}_1)$ and $\widehat{f_2}(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{sigmoid}(0.5+$

$1.3\mathbf{x}_1$). $\widehat{f}_1$ and $\widehat{f}_2$ are respectively locally and globally+locally mis-calibrated. A sample of size 10,000 is drawn from this generative model. A random subset of size 5,000 is employed for post-processing re-calibration and the rest for test.

The result is shown in the following table:

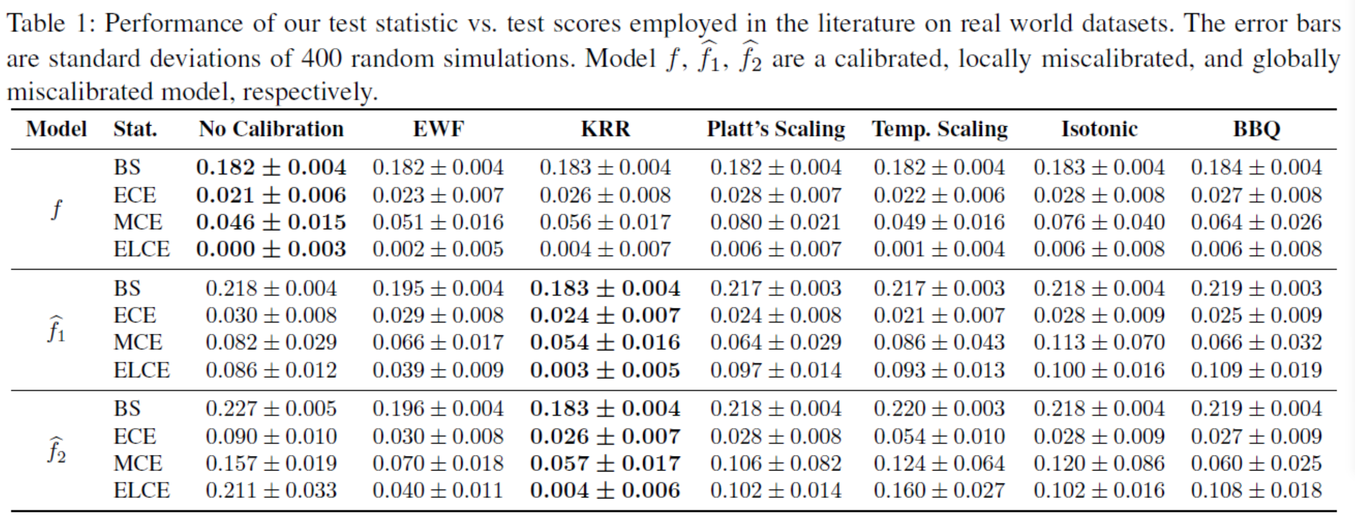


Table 1: Performance of our test statistic vs. test scores employed in the literature on real world datasets. The error bars are standard deviations of 400 random simulations. Model f, $\widehat{f}_1$, $\widehat{f}_2$ are a calibrated, locally miscalibrated, and globally miscalibrated model, respectively.

Model	Stat.	No Calibration	EWF	KRR	Platt's Scaling	Temp. Scaling	Isotonic	BBQ
f	BS	$\mathbf{0.182 \pm 0.004}$	0.182 ± 0.004	0.183 ± 0.004	0.182 ± 0.004	0.182 ± 0.004	0.183 ± 0.004	0.184 ± 0.004
	ECE	$\mathbf{0.021 \pm 0.006}$	0.023 ± 0.007	0.026 ± 0.008	0.028 ± 0.007	0.022 ± 0.006	0.028 ± 0.008	0.027 ± 0.008
	MCE	$\mathbf{0.046 \pm 0.015}$	0.051 ± 0.016	0.056 ± 0.017	0.080 ± 0.021	0.049 ± 0.016	0.076 ± 0.040	0.064 ± 0.026
	ELCE	$\mathbf{0.000 \pm 0.003}$	0.002 ± 0.005	0.004 ± 0.007	0.006 ± 0.007	0.001 ± 0.004	0.006 ± 0.008	0.006 ± 0.008
$\widehat{f}_1$	BS	0.218 ± 0.004	0.195 ± 0.004	$\mathbf{0.183 \pm 0.004}$	0.217 ± 0.003	0.217 ± 0.003	0.218 ± 0.004	0.219 ± 0.003
	ECE	0.030 ± 0.008	0.029 ± 0.008	$\mathbf{0.024 \pm 0.007}$	0.024 ± 0.008	0.021 ± 0.007	0.028 ± 0.009	0.025 ± 0.009
	MCE	0.082 ± 0.029	0.066 ± 0.017	$\mathbf{0.054 \pm 0.016}$	0.064 ± 0.029	0.086 ± 0.043	0.113 ± 0.070	0.066 ± 0.032
	ELCE	0.086 ± 0.012	0.039 ± 0.009	$\mathbf{0.003 \pm 0.005}$	0.097 ± 0.014	0.093 ± 0.013	0.100 ± 0.016	0.109 ± 0.019
$\widehat{f}_2$	BS	0.227 ± 0.005	0.196 ± 0.004	$\mathbf{0.183 \pm 0.004}$	0.218 ± 0.004	0.220 ± 0.003	0.218 ± 0.004	0.219 ± 0.004
	ECE	0.090 ± 0.010	0.030 ± 0.008	$\mathbf{0.026 \pm 0.007}$	0.028 ± 0.008	0.054 ± 0.010	0.028 ± 0.009	0.027 ± 0.009
	MCE	0.157 ± 0.019	0.070 ± 0.018	$\mathbf{0.057 \pm 0.017}$	0.106 ± 0.082	0.124 ± 0.064	0.120 ± 0.086	0.060 ± 0.025
	ELCE	0.211 ± 0.033	0.040 ± 0.011	$\mathbf{0.004 \pm 0.006}$	0.102 ± 0.014	0.160 ± 0.027	0.102 ± 0.016	0.108 ± 0.018

Columns represent different calibration methods and rows represent models. BS, ECE, MCE are all currently major error estimation functions in the ML-fairness area. It is clear that only ELCE can effectively differential the status of Local Calibration and KRR, an improved calibrator based on EWF, achieve the best performance with regards to the local calibration. ELCE, EWF and KRR will be the tools this project implement and the KRR will be used to solve the mentioned pre-trial prediction problem.

3 Data Set and Basic Data Processing

For this project, We use the ”Broward County's 2013 – 2014 COMPAS prediction data - official release”. This dataset contains 7215 entries (defendants). Except for the ground truth y feature (1 indicates did recidivism) and the COMPAS's prediction result (as given in 2.1 section), We only extract 7 relevant feature: age, race, sex, juvenile felony count, juvenile misconduct count, adult felony count, current charge type. However, for most of the logistic regression in my experiment, We only use age and adult felony count as my predictor, just like the LR_2 implemented in section 2.1 by Julia Dressel et al..

For simplicity, we only use entries with race ”white” or ”black”. And since this dataset is also heavily imbalanced, not only black defendants are over 60% overall, but over 51% of black entries are positive entries while less than 38% of white are positive. we know many linear classifier, like the logistic regression We examine here, are sensitive to the balance situation of its training dataset, so We simply use random SMOTE (over-sampling) to balance the data. After the balancing, black and white share same amount and positive rates and the total entry amount is 7344. For any dataset mentioned in the remaining report, if it is not originally balanced, we manually balance it to an ideal status because data balancing is not my focus in this project.

One major problem about the discussed paper is that they do not link their calibrator with real world problem since ELCE or local calibration alone do not guarantee anything valuable directly. After directly contacting the authors, this concern is confirmed. So in this project, we will show such connection.

4 Improvement and Experiment

4.1 Predicting with one Classifier

The first thing we do is to directly make a simple classifier with the same method as mentioned in 2.1. The results are as follows. (All results shown here are from 100 times random 10 fold.)

Performances	My LR_2	COMPAS
Accuracy% (Overall)	66.2	65.4
Accuracy% (Black)	66.3	63.8
Accuracy% (White)	66.1	67.0
False Positive% (Black)	45.2	44.8
False Positive% (White)	24.9	23.5
False Negative% (Black)	23.1	28.0
False Negative% (White)	45.3	47.7

As shown above, there is no statistical different fairness performance between my direct LR_2 and the COMPAS. Because both Julia and we only use 2 features as predictor, by directly investigate the features "age" and "adult felony count". About the "adult felony count", we find average count for actual positive black is 5.8 but such average count is only 3.9 for white. Likewise, for actual negative black, the average count is 3.0 while 1.7 for white. This is already indicate that "similar person" reflected by the data actually does not present the "similar person" in reality. So if compared with reality, bias will be inevitable.

Be more specific, both the Spearman Correlation and the coefficient of my trained linear model indicate the count has a strong positive correlation with the final positive label. (1 = being predicted as future recidivist) So based on the different mean account between "similar" black and white people, it becomes obvious why my linear model discriminate black people.

Generally speak, for a binary model, as long as its predicted mean/ actual mean differs from each group, such unfair will always happen.

One possible solution is to add back the "race" feature and hope it can correct the overall bias. After doing so, with three feature, the results are shown below:

Performances	My LR_3	COMPAS
Accuracy% (Overall)	66.5	65.4
Accuracy% (Black)	69.7	63.8
Accuracy% (White)	63.2	67.0
False Positive% (Black)	27.1	44.8
False Positive% (White)	41.7	23.5
False Negative% (Black)	38.2	28.0
False Negative% (White)	32.6	47.7

So, apparently reversing the unfairness is not a success and it also shows that adding the race feature is not sufficient to balance the group prediction mean following the default cost function. But actually, if we are allowed to use/consider the feature "race", there are better way which can guarantee such fairness among different races.

4.2 Predicting with multiple Classifiers

One nature idea is to train two different model for two different race (black and white). By doing so, it is trivial that mathematically, the fairness among groups will be solved as long as the overall accuracy are similar. For each classifier, since the data is balanced, the cost (for default cost function which cares only about the correctness of classification) of increasing false positive rate is the same as increasing false negative rate. Hence, firstly the FPR and FNR among any group should be similar and thus if the accuracy among each group is similar, the FPR and FNR among different groups should be similar as well. While the accuracy simply relies on the overall correlation between predictors and the labels. The results by doing so are as follows:

Performances	My LR_2m	COMPAS
Accuracy% (Overall)	67.0	65.4
Accuracy% (Black)	67.6	63.8
Accuracy% (White)	66.5	67.0
False Positive% (Black)	31.4	44.8
False Positive% (White)	34.0	23.5
False Negative% (Black)	33.5	28.0
False Negative% (White)	33.7	47.7

This is just as we expected. However, this may still far from solving this question because although this is the simplest way to solve such unfairness issue and actually without hurting the accuracy, there may be law and other obstruction which forbid us by using the "race" in anyway. So how to solve, improve this problem without consideration of "race"?

In fact, this question is how to solve the unfairness among groups without knowing the groups at all. From logic, it is somehow self-contradiction.

One way is to use some clustering algorithm to divide the dataset into some smaller "virtual" groups and hope that the existing features can lead data entries into their "actual groups".

For example, by using the K-means with features "adult felony count", "juvenile felony count" and "age", as well as K = 3, we get the following results:

Performances	My LR_2c	COMPAS
Accuracy% (Overall)	67.0	65.4
Accuracy% (Black)	67.6	63.8
Accuracy% (White)	66.5	67.0
False Positive% (Black)	33.3	44.8
False Positive% (White)	21.1	23.5
False Negative% (Black)	32.6	28.0
False Negative% (White)	49.8	47.7

Although it does not correct the white group's "privilege", it does help with the black group's discrimination. Such method heavily depends on the features you choose and their relationship with the target group. For example, the Spearman Correlation between "adult felony count" and "race" is only 0.16, while 0.14 for "juvenile felony count" and -0.011 for "age". Such weak connection explains the relatively weak performance.

However, the actual world data should be able to offer more information and connection among our targeted groups, especially if such groups are considered by public to have significant difference and meaning. For example, the average US's household income (as of March 2018) is about $90000 for white and $60000 for black. Although the dataset we use do not contain income information, but we can do experiment by adding random household income based on the normal distribution to check the statistical meaning. (Mean as mentioned above and Std as $ 15000)

After adding "household income" into the K-Mean's feature and repeat the above experiment, we can get results as follows:

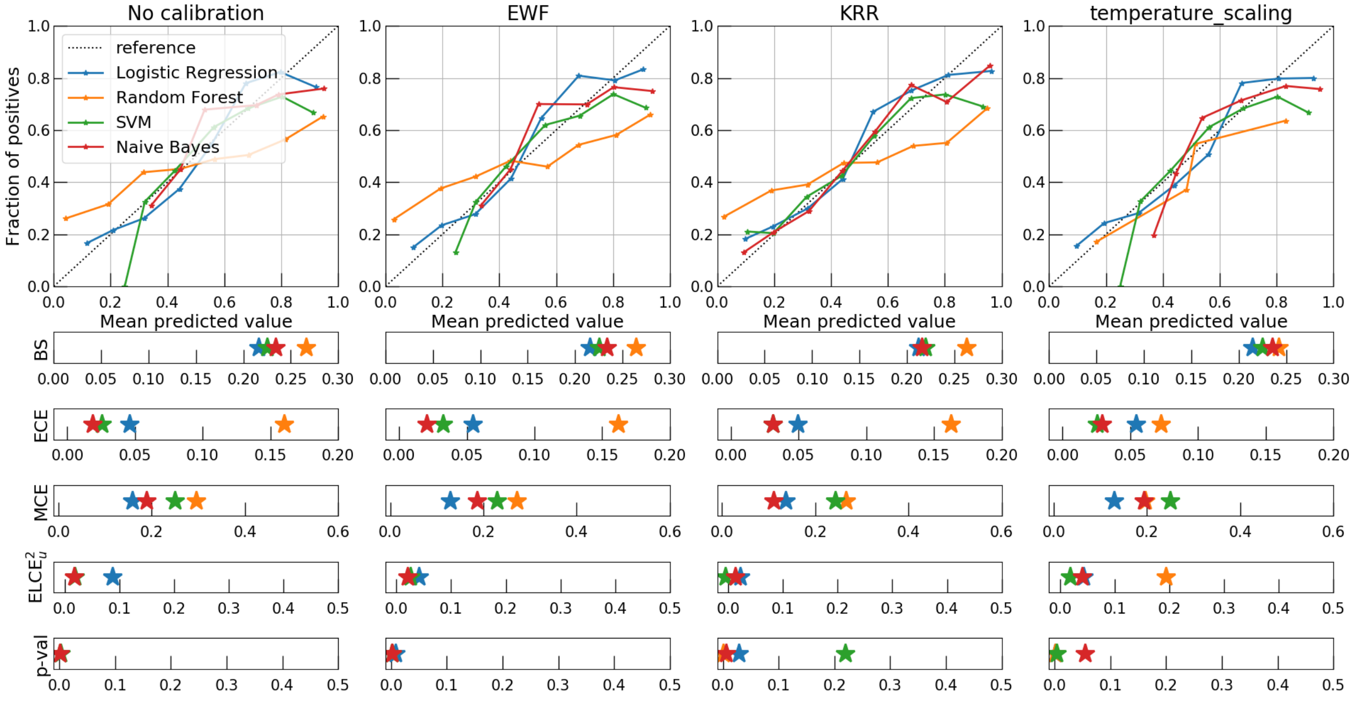
Performances	My LR_2c	COMPAS
Accuracy% (Overall)	67.2	65.4
Accuracy% (Black)	68.0	63.8
Accuracy% (White)	66.4	67.0
False Positive% (Black)	34.4	44.8
False Positive% (White)	31.1	23.5
False Negative% (Black)	32.1	28.0
False Negative% (White)	38.8	47.7

This shows that as long as the targeted unknown group really has some significance and the overall dataset can reflect this, it is quite hopeful to use this simple Cluster - Classification method to solve similar fairness concern.

4.3 Prediction with the KRR

Now, we implement the ELCE, EWF and KRR mentioned before, but before we directly apply KRR to the original question and measure the results by the FP/FN rate among groups, we will firstly consider if they all works similarly in my real-world dataset rather than previous human-created data.

Although we must use logistic regression as my original model in later fairness performance check, we will apply the local calibration tools on additional 3 binary classification models, just for comparison. The results are as follows:



The top panel shows the situation of global calibration: predicted possibility value v.s. fraction of positives.

It is clear that KRR performs best with regards to the ELCE, aka the local calibration in this pre-trial prediction task in all 4 classification models and even improves the global calibration significantly.

So after we get this similarity with the previous experiment conducted by the inventor of KRR calibrator, we apply such method to my task and get the following results:

Performances	KRR	My LR_2c	COMPAS
Accuracy% (Overall)	72.7	67.0	65.4
Accuracy% (Black)	70.1	67.6	63.8
Accuracy% (White)	74.6	66.5	67.0
False Positive% (Black)	31.2	33.3	44.8
False Positive% (White)	22.1	21.1	23.5
False Negative% (Black)	25.3	32.6	28.0
False Negative% (White)	38.9	49.8	47.7

The My LR_2c is the previous trail which achieves the best general fairness performance with regards to group-wise FP/FN, without using the protect feature: race(group). It is clear that the KRR, using same information, outperforms the previous best method in almost all important fairness metric used in this project and almost reachs an ideal fairness status among black and white people. More importantly, it significantly improves the accuracy, which all previous methods can not do.

So we here successfully accomplish my original goal: to show a link between the local calibration and other real-world fairness metric, for further study.

5 Further Thinking and Plans

One thing we must remind the reader is it is never my goal to consider separately based on their race. This is simply because, just like the cases above, being black is not the reason for you to have averagely more crime records then a white person with similar recidivism chance, it is issues like household income, family structure, educational level, living area, etc, which actually decide those issue and the race is nothing but a efficient indicator of those issue. However, it is just an indicator based on statistical meaning. For example, if a white has all those "hidden features" as same as a black person, their scores should simply on the "same scale".

However, it is also shown previous that in many occasion, group itself is an important indicator. In other words, individuals in different groups may actually behave differently. So the idea or principle to consider different groups' people differently may be very valuable. If the group feature is unknown or protected, certain unsupervised (with respect to the group feature) calibration method will be helpful. Naive ways may includes K-Mean, which does help a little bit in this project. The KRR calibrator, with similar "group-wise calibration" concept, solve the COMPAS unfairness problem successfully.

One important issue is that, just like the K-mean, this local calibration method also heavily depends on the assumption that the $\mathbf{x}$ can reflect, in some degree, their own group, which means every instant contains their group information. This is obvious since it actually use kernel to define similarity of instance. Thus it also implies the importance of the data quality.

However, even if we show a real-world application with this novel method, it lacks certain math proof, just like the mentioned paper whose proof ignore many important assumptions. After contacting with the original author of this method, he agrees that the next step is to make some valid proof on the connection between the local calibration with real world metric, like group-wise FP/FN rate. Such connection will likely be vague or estimated, hence we believe the further abstract math modeling and proof will be needed and should be a good topic for a series of academic papers.

References

[1] Kirkpatrick, Keith. *It's not the algorithm, it's the data.* Communications of the ACM. 60 (2): 21–23.

[2] Angwin, Julia; Larson, Jeff. *Machine Bias.* ProPublica.

[3] Julia Dressel; Hany Farid. *The accuracy, fairness, and limits of predicting recidivism.* Science Advances.

[4] Jon Kleinberg; Sendhil Mullainathan; Manish Raghavan. *Inherent Trade-Offs in the Fair Determination of Risk Scores.* ACM SIGMETRICS (PER).

[5] Farahi, Koutra. *Toward Group-wise Calibration: A Locally-aware Calibration Method for Probabilistic Classifiers.* To be appear in AISTATS 2021