Aprendizaje de máquinas: Discriminación

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

16 de febrero de 2026

Contenido

Outline

Introducción

- El aprendizaje de máquinas se utiliza de forma generalizada para resolver problemas de predicción en muchas circunstancias de la vida humana: otorgamiento de crédito, predicción de crimen, alertas de desempeño académico, transacciones atípicas en el sector financiero, etc.

Introducción

- El aprendizaje de máquinas se utiliza de forma generalizada para resolver problemas de predicción en muchas circunstancias de la vida humana: otorgamiento de crédito, predicción de crimen, alertas de desempeño académico, transacciones atípicas en el sector financiero, etc.
- En todos estos casos surgen problemas con la recolección de datos o uso de los modelos:
  - La predicción está correlacionada con el instrumento de recolección de información (input del modelo). Este fenómeno se conoce como retroalimentación circular o *feeback loop*: genera potenciales sesgos y resultados disciminatorios en una población.
  - Los modelos de ML infieren información de variables no observadas que normativamente no deben ser utilizadas en los modelos. Este fenómeno se llama triangulación: potencialmente genera resultados discriminatorios.

Otros problemas

- Sub reporte: los datos son una muestra sesgada de la realidad (e.g., procesos sesgados, seres humanos que intervienen en el proceso).
- Equilibrio: las predicciones desconocen que pueden existir interacciones estratégicas.
- Privacidad: con suficiente información es posible identificar personas o atributos privados.
- En la actualidad esté es uno de los grandes retos de las aplicaciones de las matemáticas y aprendizaje de máquinas al mundo real.

Outline

Efectos heterogéneos

- Usan una base de datos de 10 millones de hipotecas.
- Los autores encuentran que los negros e hispanos ganan menos con la introducción de ML en el sentido de que sus probabilidades de incumplimiento no disminuyen.
- En equilibrio, cuando consideramos las consecuencias de estas predicciones en el mercado, los modelos de ML aumentan el otorgamiento de todos los grupos raciales pero incrementa la disparidad en los grupos y entre grupos.
- El responsable de estos efectos es principlamente el cambio tecnológico (mayor flexibilidad) por encima del fenómeno de triangulación.

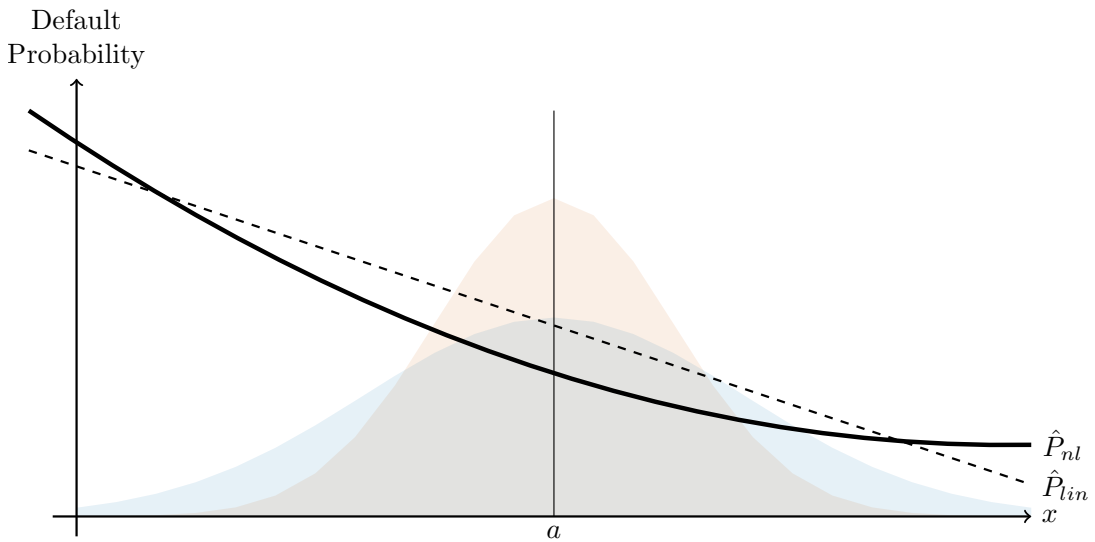
Introducción

- Intuición básica.
  - Una mejor tecnología, en el sentio de disminuir el error cuadrático medio, es más dispersa.
  - Mayor dispersión implica que esa tecnología tiene efectos diferenciales en la poblacion.
- Ejemplo: Si se usa un modelo ingenuo para estimar probabilidad de incumplimiento (i.e., probabilidad igual a la proporción de incumplidos) y se compara con un modelo lineal en cualquier característica (por ejemplo sexo), habrá ganadores y perdedores.

Efectos de una mejor tecnología

- Dos tipos de deudores: azul y rojo.
- Una característica observable, eje x en la figura.
- Supongamos que **la verdadera relación entre ingreso e incumplimiento es cuadrática**.
- Una mejor tecnología (modelo no lineal) tiene efectos negativos sobre parte de la población de los azules (los más dispersos)

Figure 1: **Unequal Effects of Better Technology**



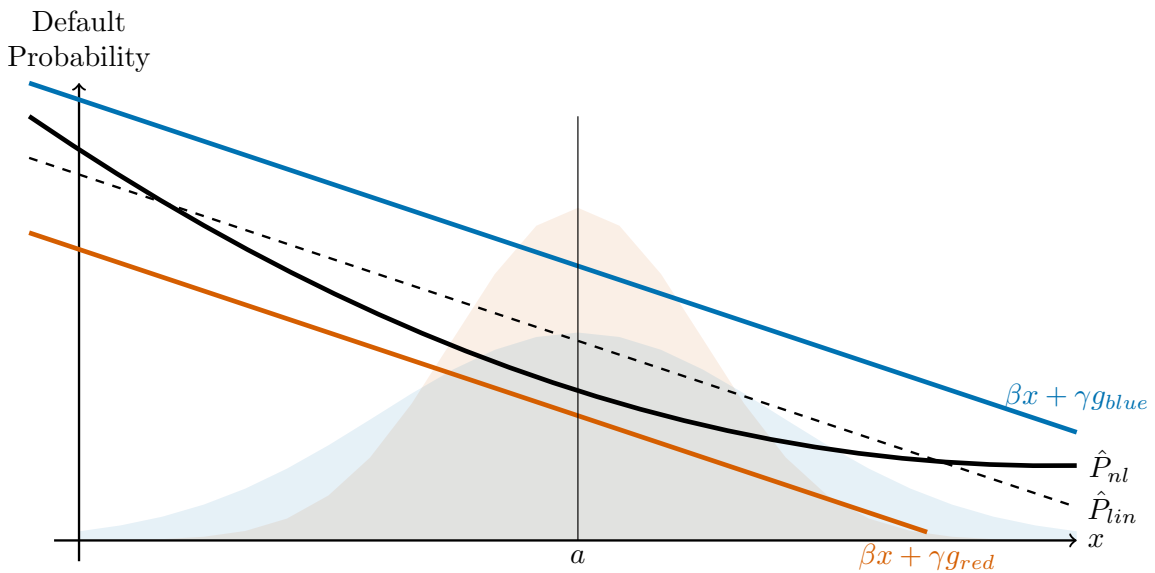
Triangulación

- Ahora suponga que existe una carcaterística (i.e., raza) que no puede ser utilizada para discriminar.
- **Supongamos que el verdadero modelo** es de la forma $y = \beta x + \gamma g + \epsilon$, donde ϵ no se correlaciona con x, g (x, g son exógenas).
- Una tecnología no lineal no puede, por definición, mejorar el anterior modelo (en el que se observan todas las variables).
- Sin embargo, en asuencia de la variable protegida, el modelo no lineal si puede hacer un mejor ajuste (ser mejor descubriendo el papel de la variable protegida).

Triangulación

- En este caso una mejor tecnología tiene efectos negativos sobre el grupo azul (los más dispersos): aproxima mejor el verdadero modelo cuando no se observa la variable protegida.

Figure 2: **Triangulation**



Predicción de incumplimiento

Table 1: **Descriptive Statistics, 2009-2013 Originations**

Group		FICO	Income	LoanAmt	Rate (%)	SATO (%)	Default (%)
	Mean	764	122	277	4.24	-0.07	0.42
Asian	Median	775	105	251	4.25	-0.05	0.00
(N=574,812)	SD	40	74	149	0.71	0.45	6.49
	Mean	735	91	173	4.42	0.11	1.88
Black	Median	744	76	146	4.50	0.12	0.00
(N=235,673)	SD	58	61	109	0.71	0.48	13.57
	Mean	746	90	187	4.36	0.07	0.99
White Hispanic	Median	757	73	159	4.38	0.07	0.00
(N= 381,702)	SD	52	63	115	0.71	0.47	9.91
	Mean	761	110	208	4.33	-0.00	0.71
White Non-Hispanic	Median	774	92	178	4.38	0.02	0.00
(N=7,134,038)	SD	45	73	126	0.69	0.44	8.37
Native Am, Alaska,	Mean	749	97	204	4.39	0.04	1.12
Hawaii/Pac Isl	Median	761	82	175	4.45	0.04	0.00
(N=59,450)	SD	51	65	123	0.70	0.46	10.52
	Mean	760	119	229	4.38	0.00	0.79
Unknown	Median	773	100	197	4.50	0.02	0.00
(N=984,310)	SD	46	78	141	0.68	0.44	8.85

Note: Income and loan amount are measured in thousands of USD. SATO stands for "spread at origination" and is defined as the difference between a loan's interest rate and the average interest rate of loans originated in the same calendar quarter. Default is defined as being 90 or more days delinquent at some point over the first three years after origination. Data source: HMDA-McDash matched dataset of fixed-rate mortgages originated over 2009-2013.

Predicción de incumplimiento

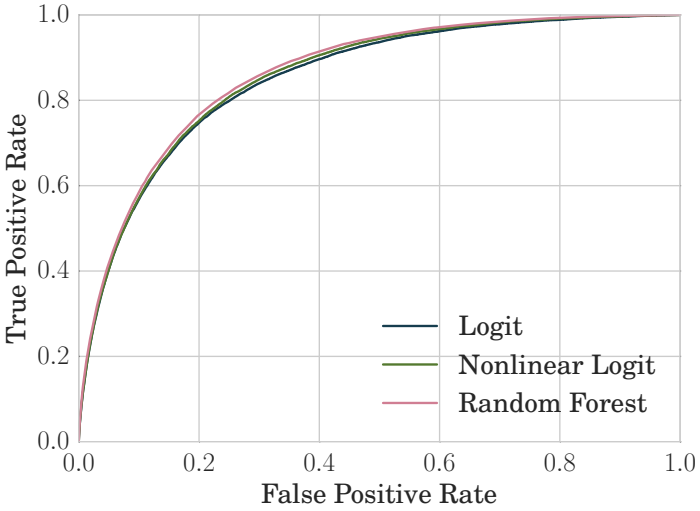
Table 2: **Variable List**

Logit	*Nonlinear Logit*
Applicant Income (linear)	Applicant Income (25k bins, from 0-500k)
LTV Ratio (linear)	LTV Ratio (5-point bins, from 20 to 100%; separate dummy for LTV=80%)
FICO (linear)	FICO (20-point bins, from 600 to 850;) separate dummy for FICO<600)
(with dummy variables for missing values)	
Common Covariates	
Spread at Origination "SATO" (linear)	
Origination Amount (linear and log)	
Documentation Type (dummies for full/low/no/unknown documentation)	
Occupancy Type (dummies for vacation/investment property)	
Jumbo Loan (dummy)	
Coapplicant Present (dummy)	
Loan Purpose (dummies for purchase, refinance, home improvement)	
Loan Term (dummies for 10, 15, 20, 30 year terms)	
Funding Source (dummies for portfolio, Fannie Mae, Freddie Mac, other)	
Mortgage Insurance (dummy)	
State (dummies)	
Year of Origination (dummies)	

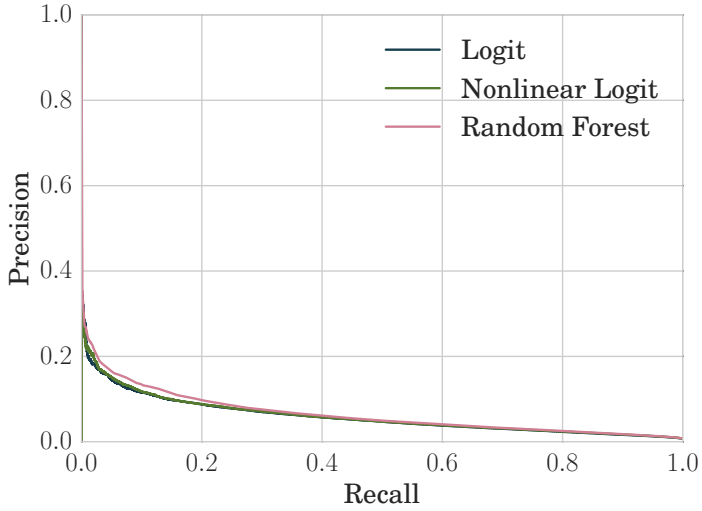
Note: Variables used in the models. Data source: HMDA-McDash matched dataset of conventional fixed-rate mortgages.

Predicción de incumplimiento

Figure 3: **ROC and Precision-Recall Curves**



Panel A: ROC



Panel B: Precision-Recall

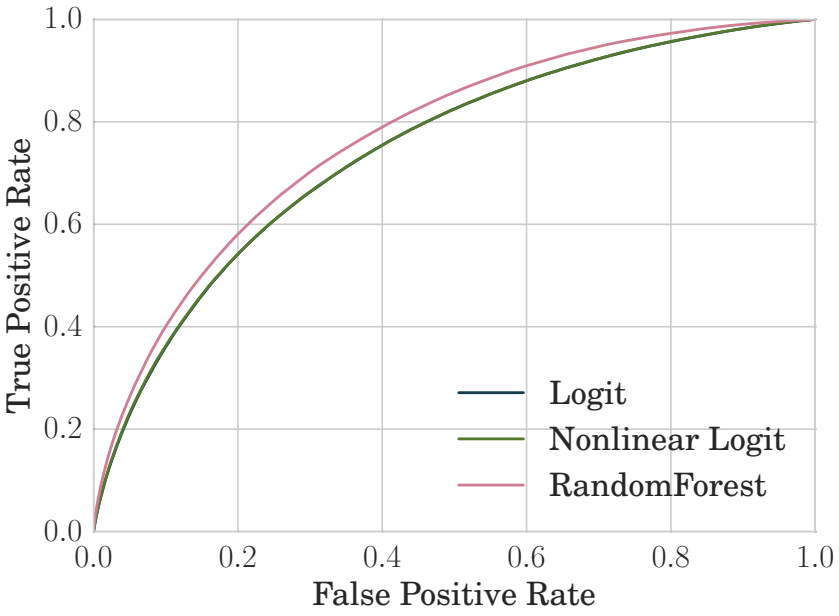
Predicción de incumplimiento

Table 3: **Performance of Different Statistical Technologies Predicting Default**

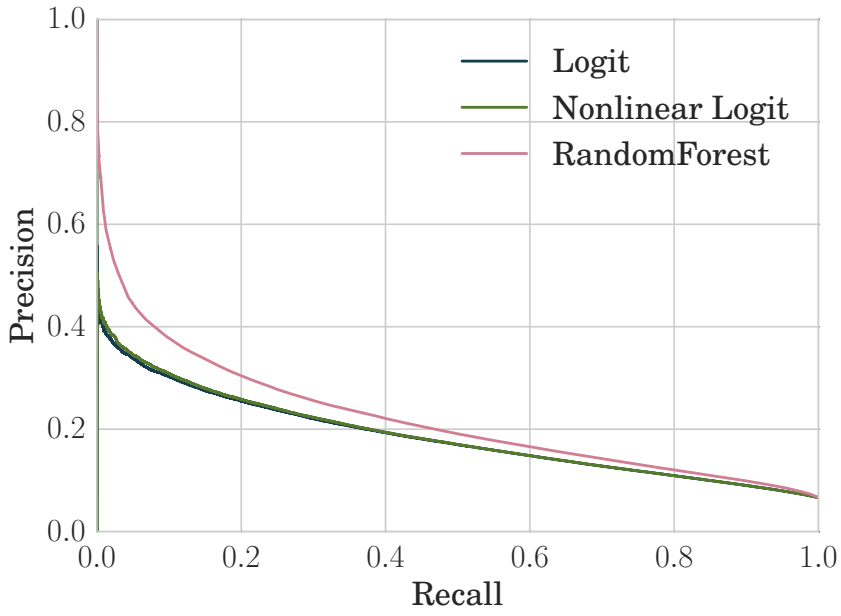
	ROC AUC		Precision Score		Brier Score × 100		R^2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Model	No Race	Race	No Race	Race	No Race	Race	No Race	Race
Logit	0.8522	0.8526	0.0589	0.0592	0.7172	0.7171	0.0245	0.0246
Nonlinear Logit	0.8569	0.8573	0.0598	0.0601	0.7146	0.7145	0.0280	0.0281
Random Forest	0.8634	0.8641	0.0630	0.0641	0.7114	0.7110	0.0323	0.0329

Predicción de raza

Figure 4: **ROC and Precision-Recall Curves of Predicting Race**



Panel A



Panel B

Predicción de raza

Table 4: **Performance of Different Statistical Technologies Predicting Race**

Model	ROC AUC	Precision Score	Brier Score × 10	R^2
Logit	0.7478	0.1948	0.5791	0.0609
Nonlinear Logit	0.7485	0.1974	0.5783	0.0622
Random Forest	0.7527	0.2110	0.5665	0.0813

Mejor tecnología vrs triangulación

Table 5: **Decomposition of Performance Improvement**

	Race	Technology
ROC-AUC	5.88	94.12
Precision	7.90	92.10
Brier	3.25	96.75
R^2	2.04	97.96

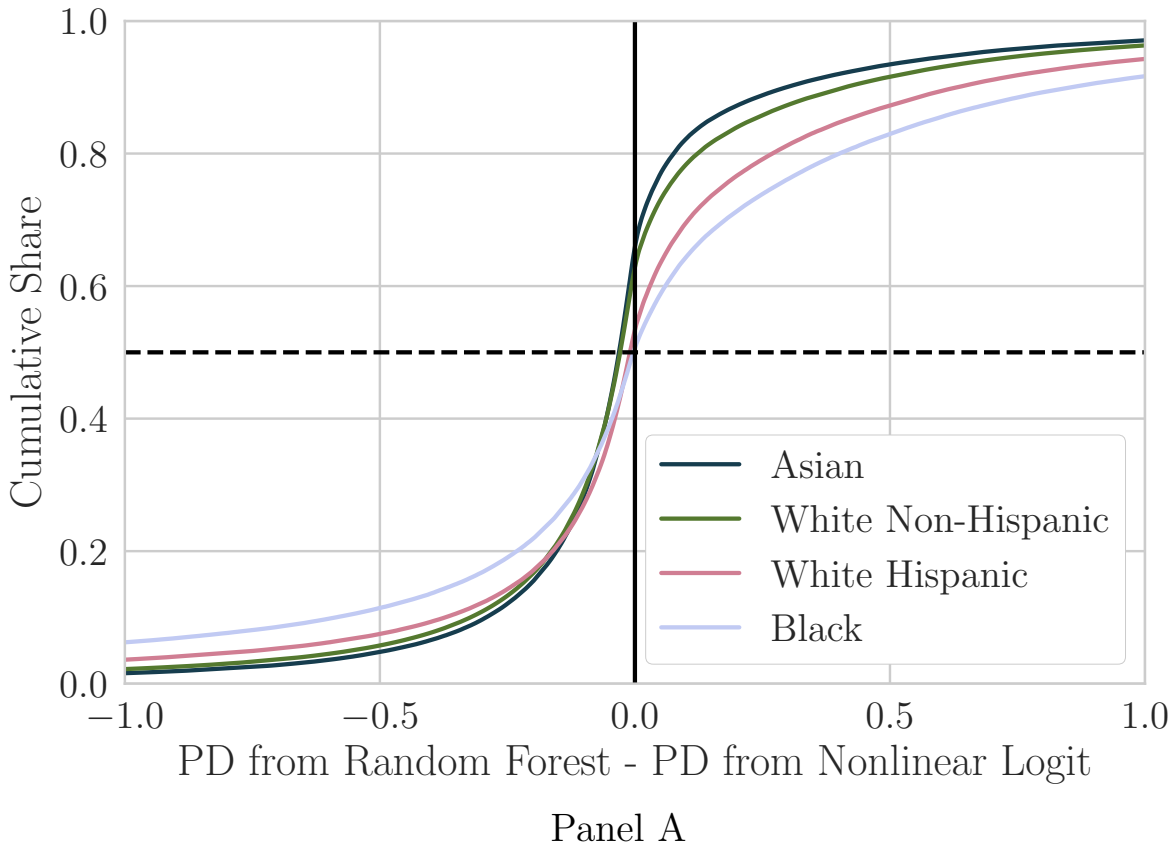
Panel A: Race Controls First

	Technology	Race
ROC-AUC	91.16	8.84
Precision	77.21	22.79
Brier	90.63	9.37
R^2	87.75	12.25

Panel B: New Technology First

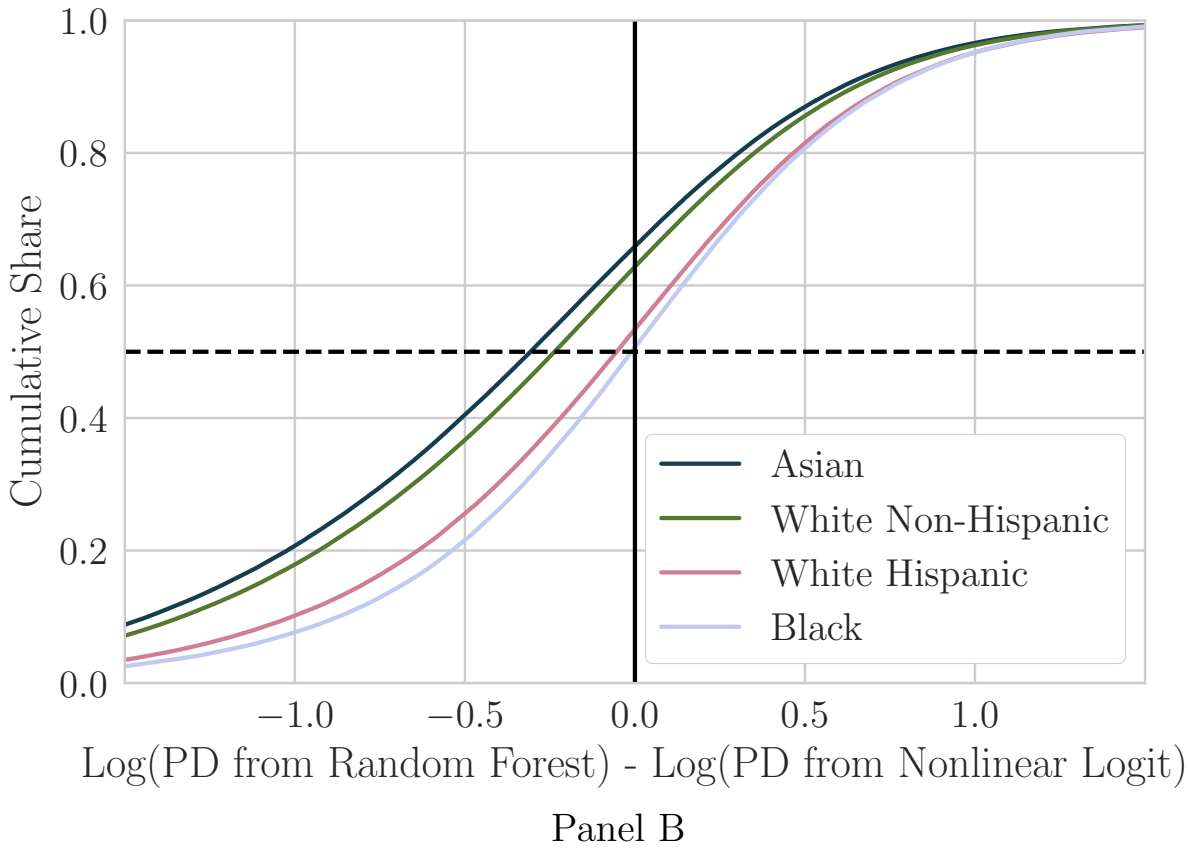
Efectos discriminatorios

Figure 6: **Comparison of Predicted Default Probabilities Across Models, by Race Groups**



Panel A

Efectos discriminatorios



Panel B

Curvas: ROC y Precision-Recall

- La curva ROC (TPR o sensitivity vrs. FPR o 1-especificidad) no es una muy buena medida cuando las clases están muy desbalanceadas.
- Si los positivos son muy pocos, TPR va a tender a ser alto.

Curvas: ROC y Precision-Recall

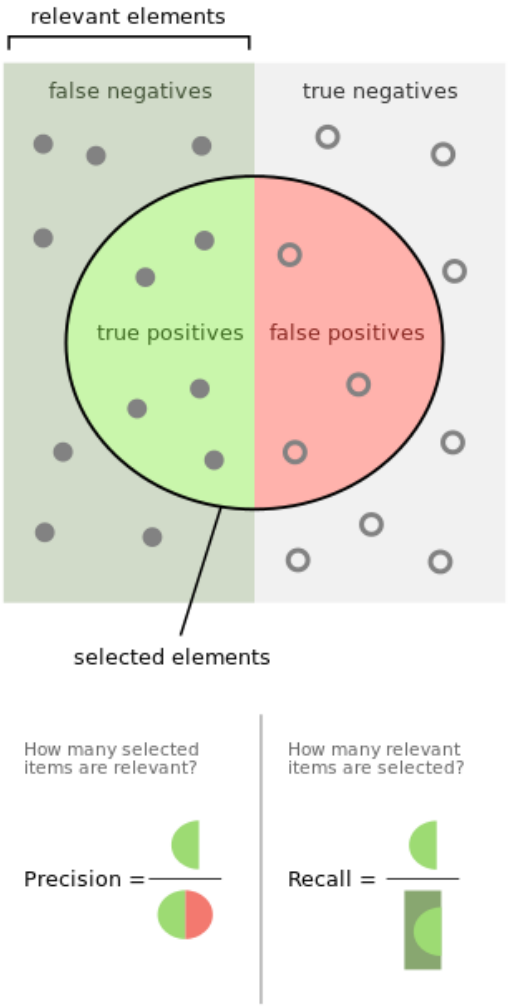


Figura 1: By Walber - Own work, CC BY-SA 4.0, `https://commons.wikimedia.org/w/index.php?curid=36926283`

Obsérvese que TPR = Recall

Brie Score y R^2

- El Brie Score se define como: $\frac{1}{n}\sum_n(\hat{P}(x_i) - y_i)^2$
- Este se puede descomponer entre Reliability, Resolucion e Incertidumbre.

Outline

Introducción

- Existen muchas definiciones formales de discriminación algortimica.
- Vamos a estudiar formalmente:
  1. Paridad demográfica.
  2. *Equalized odds*.
  3. Calibración y calibración condicional.
  4. Balance de clases positivas (igualdad de oportunidad).
  5. Balance de clases negativas.

Motivación

- El grado, o evidencia de discriminación puede depender de la medida cuantitativa que se use como definición de discriminación.
- El siguiente ejemplo es uno de los principales que inicio y motivo mucho el interés en las implicaciones éticas de usar modelos de aprendizaje de máquinas para tomar decisiones en el ámbito público.

El Caso de COMPAS

- COMPAS: Correctional Ofender Management Profiling for Alternative Sanctions.

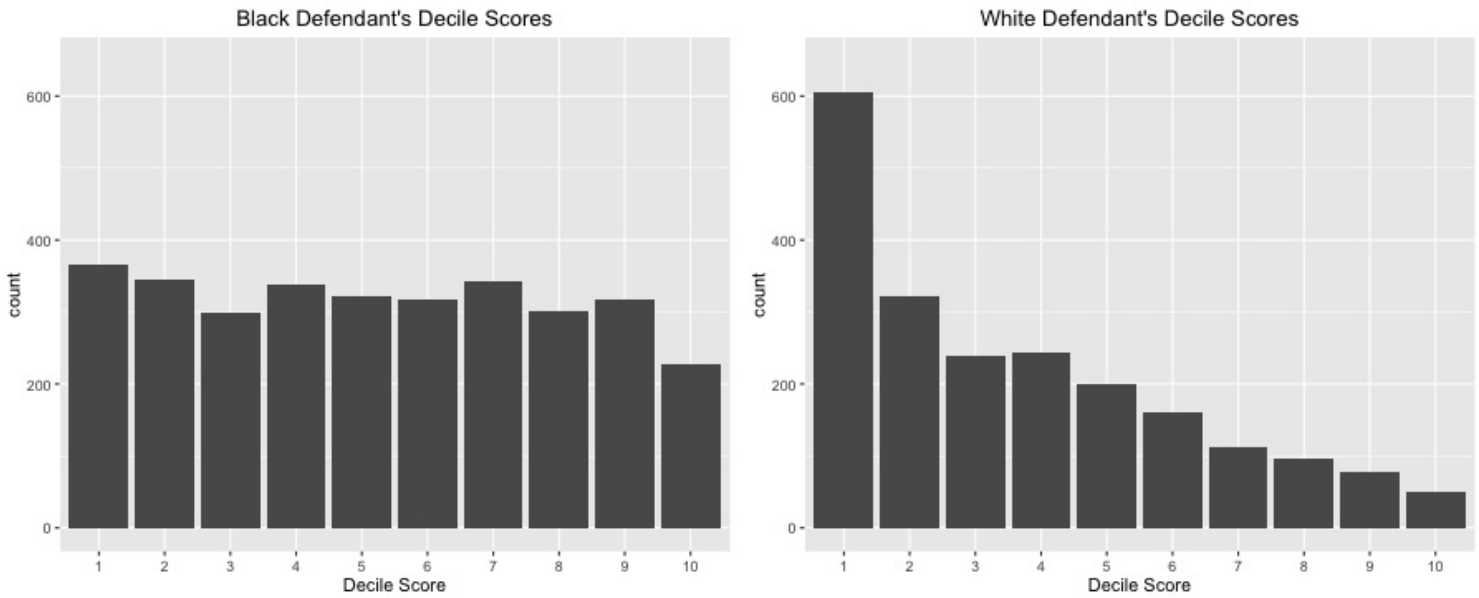


Figura 2: Histograma de los que no reincidieron

El Caso de COMPAS

- ¿Qué explica el score de COMPAS?

Risk of General Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	0.221*** (0.080)
Age: Greater than 45	-1.356*** (0.099)
Age: Less than 25	1.308*** (0.076)
Black	0.477*** (0.069)
Asian	-0.254 (0.478)
Hispanic	-0.428*** (0.128)
Native American	1.394* (0.766)
Other	-0.826*** (0.162)
Number of Priors	0.269*** (0.011)
Misdemeanor	-0.311*** (0.067)
Two year Recidivism	0.686*** (0.064)
Constant	-1.526*** (0.079)
Observations	6,172
Akaike Inf. Crit.	6,192.402

*Note: *p<0.1; **p<0.05; ***p<0.01*

Figura 3: Logit

El Caso de COMPAS

- The accuracy, fairness, and limits of predicting recidivism. Julia Dressel and Hany Farid. 2018.
- El accuracy es 67 % y 63.8 % (aciertos positivos y negativos sobre el total de ejemplo) para blanco y negros respectivamente.
- Los negros que no reincidieron, COMPAS predice que en promedio van a reincidir con probabilidad 44.9 % mientras que el grupo análogo de blancos 23.5 %.
- Sin embargo COMPAS score satisface: predictive parity, AUC equity y calibration.
- Este estudio se concentra en entender si COMPAS realmente es más preciso o más sesgado que lo que un grupo de expertos haría: encuentran que no es mas accurate y que con solo 5 variables y un modelo lineal se puede tener la misma capacidad rerdictiva que COMPAS que usa 137 variables.

El Caso de COMPAS

Table 1. Human versus COMPAS algorithmic predictions from 1000 defendants. Overall accuracy is specified as percent correct, AUC-ROC, and criterion sensitivity (d') and bias (β). See also Fig. 1.

	(A) Human (no race)	(B) Human (race)	(C) COMPAS
Accuracy (overall)	67.0%	66.5%	65.2%
AUC-ROC (overall)	0.71	0.71	0.70
d'/β (overall)	0.86/1.02	0.83/1.03	0.77/1.08
Accuracy (black)	68.2%	66.2%	64.9%
Accuracy (white)	67.6%	67.6%	65.7%
False positive (black)	37.1%	40.0%	40.4%
False positive (white)	27.2%	26.2%	25.4%
False negative (black)	29.2%	30.1%	30.9%
False negative (white)	40.3%	42.1%	47.9%

Figura 4: Humanos

Formalización

1. Y es una variable binaria que queremos predecir.
2. $\hat{Y}$ es la prediccion.
3. La clase 1 se va considerar la decisión social preferida (otorgar un crédito, ser beneficiario de un programa social, etc.).
4. X son las variables observadas de cada individuo.
5. A es una variable binaria que representa la clase protegida (hombres, mujeres, blanco, negro, religioso, ateo, etc.)

Paridad Demográfica

Definition

Decimos que un algoritmo satisface paridad demogáfica si:

$$P(\hat{Y} = 1 \mid A) = P(\hat{Y} = 1) \tag{1}$$

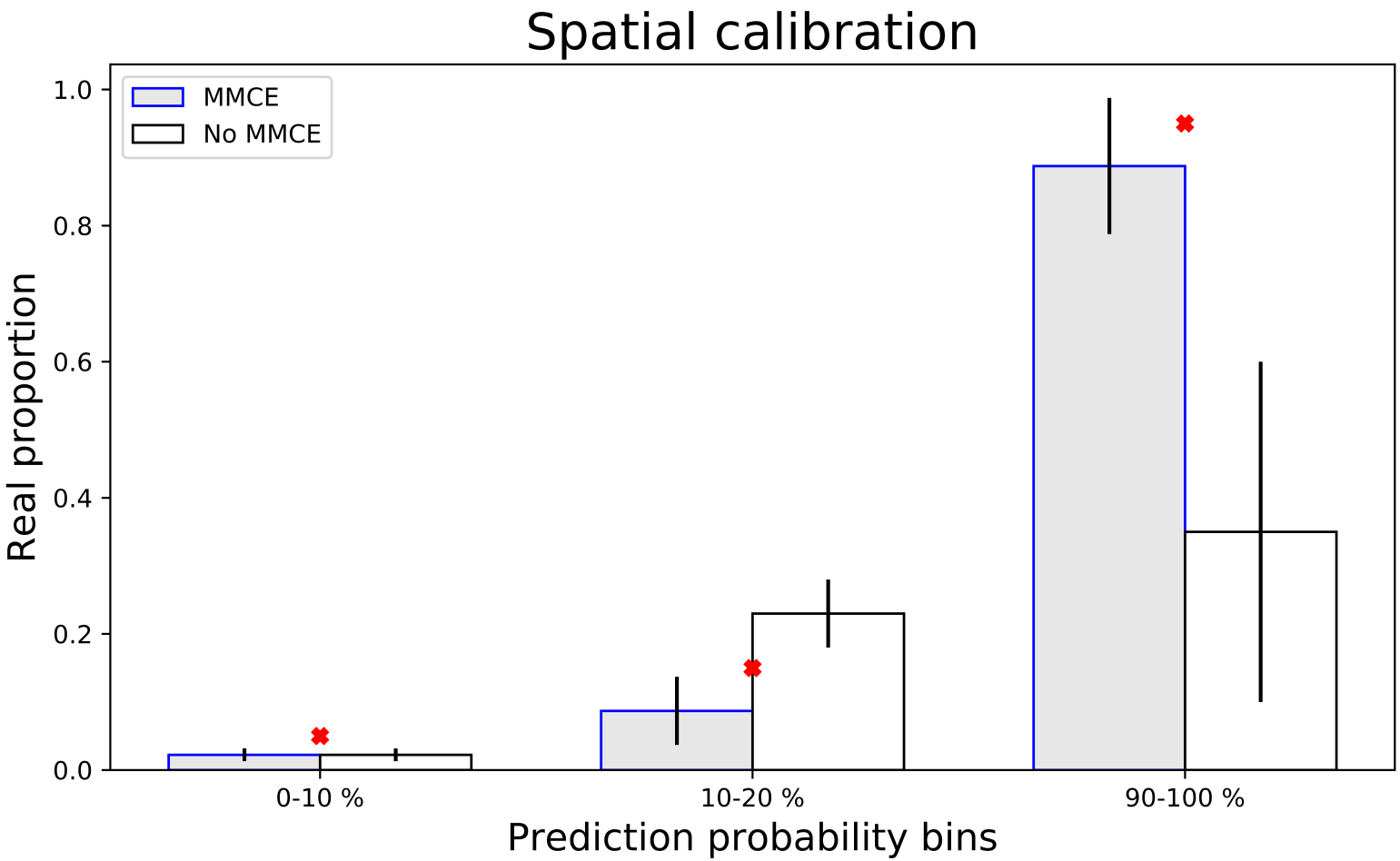
Calibración

Definition

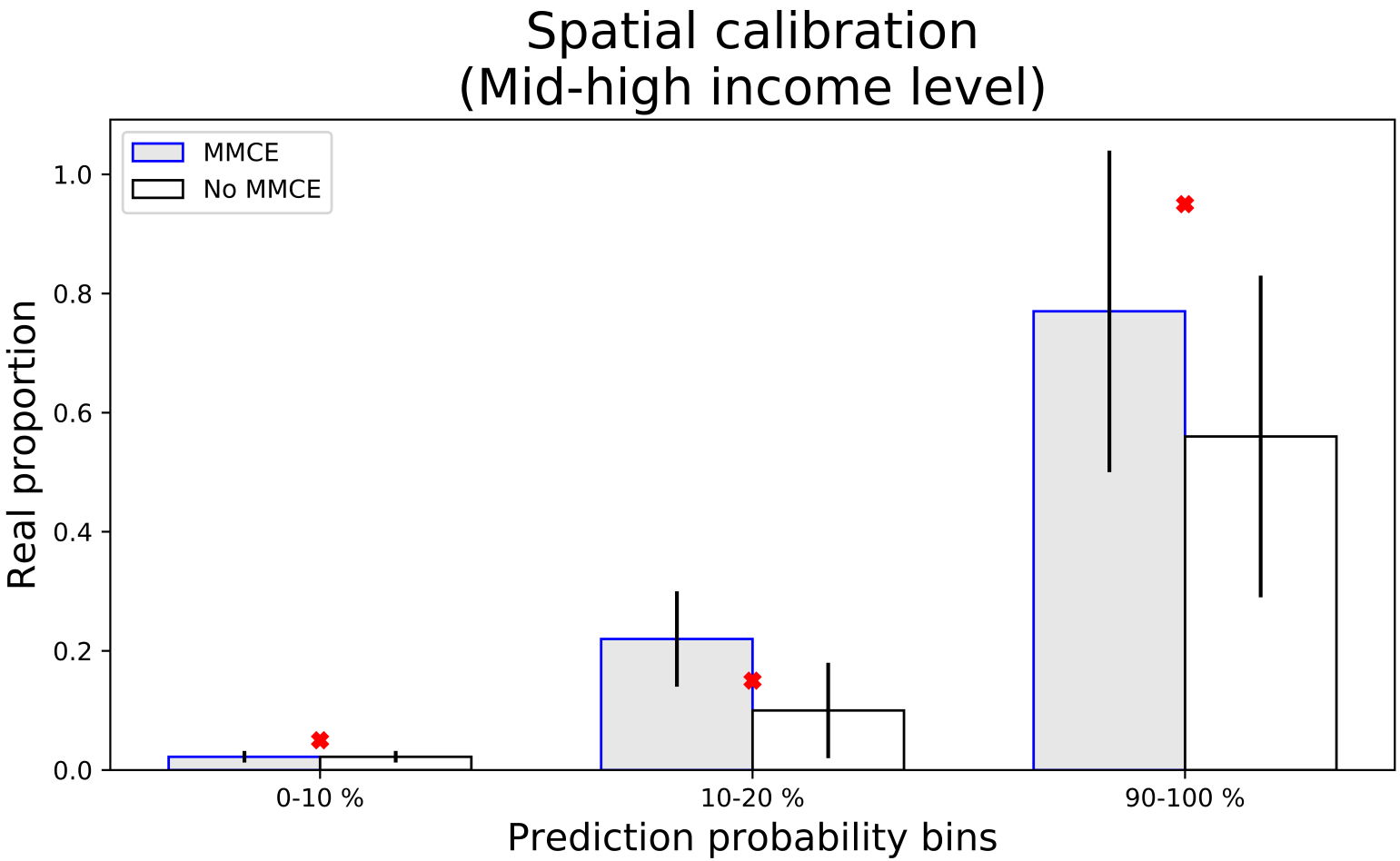
Decimos que un algoritmo satisface la propiedad de calibracion con respecto a la variable protegida si:

1. El pronóstico está bien calibrado.
2. Condicional a la variable protegida el pronóstico está bien calibrado.

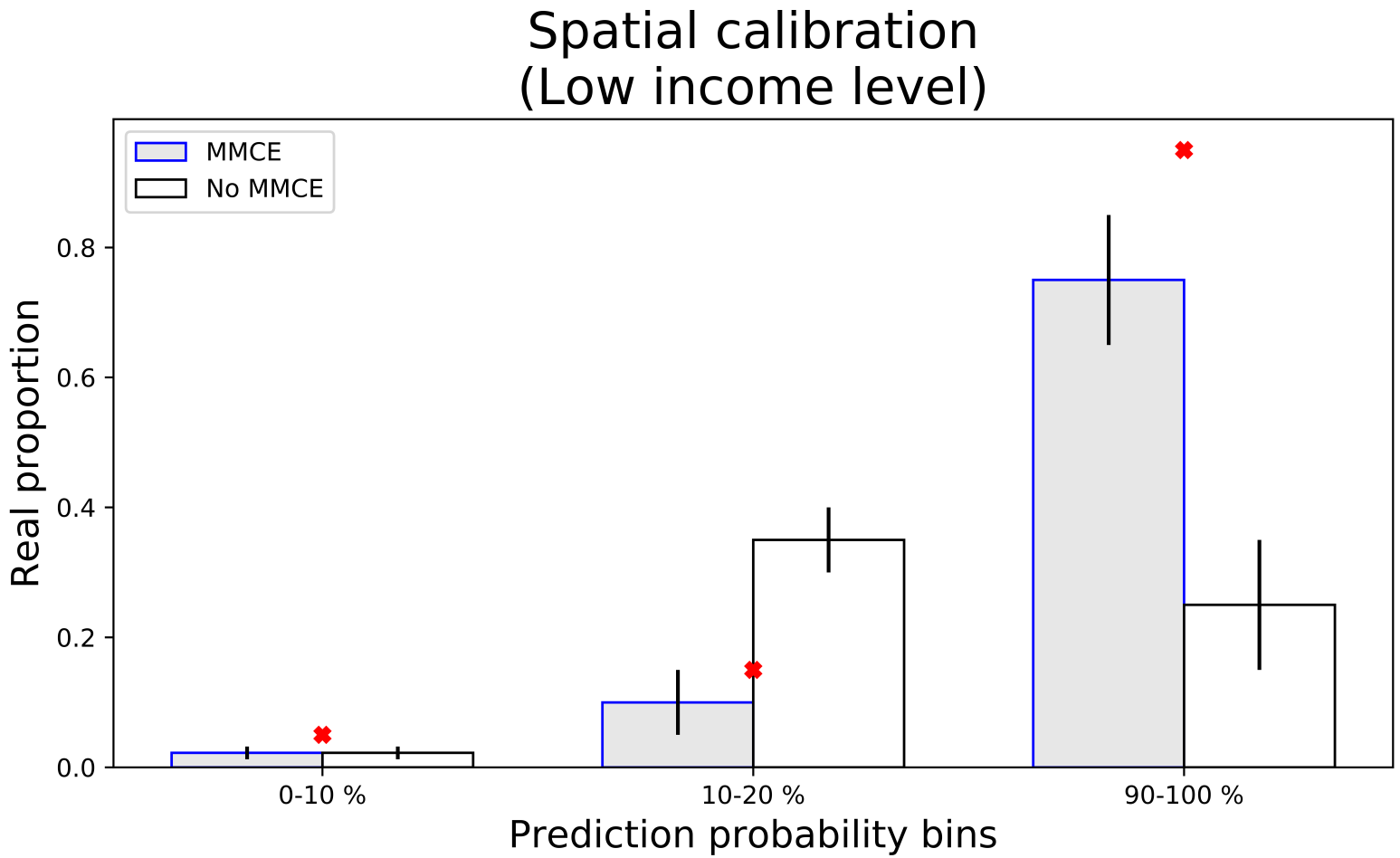
Calibración: Ejemplo Predicción de Crimen



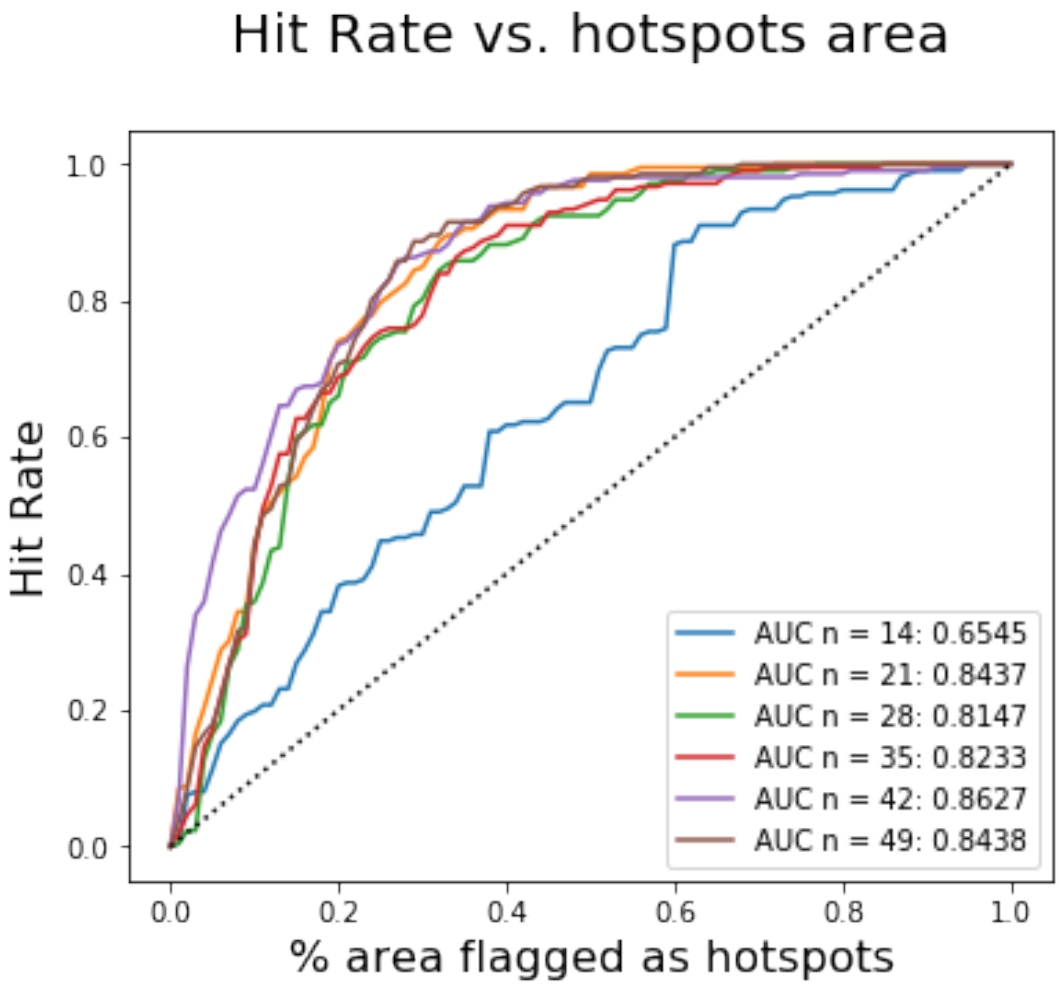
Calibración: Ejemplo Predicción de Crimen



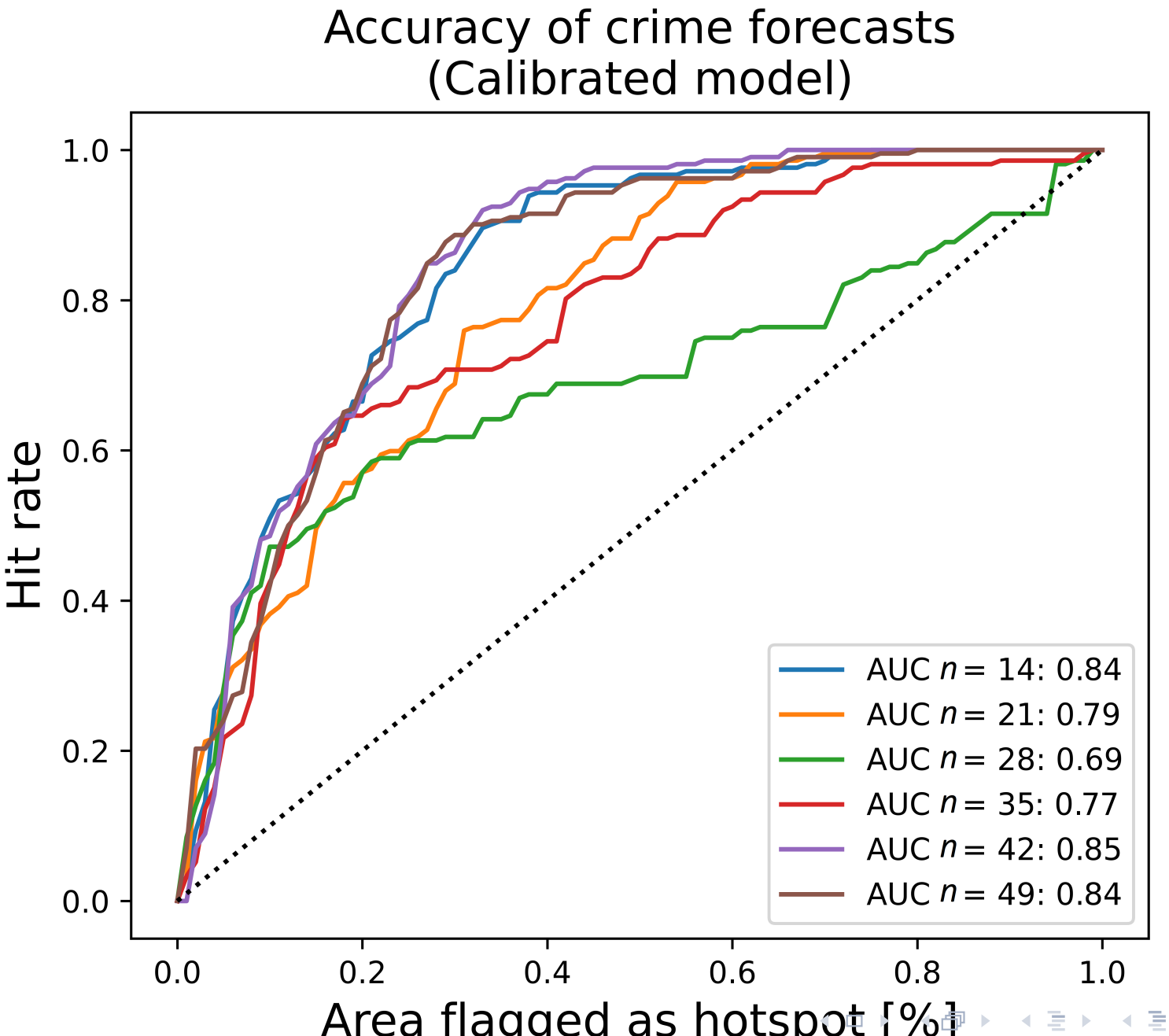
Calibración: Ejemplo Predicción de Crimen



Calibración: Ejemplo Predicción de Crimen



Calibración: Ejemplo Predicción de Crimen



Igualdad de Oportunidad (Balance clases positivas)

Definition

Decimos que un algoritmo satisface igualdad de oportunidad si:

$$P(\hat{Y} = 1 \mid Y = 1, A) = P(\hat{Y} = 1 \mid Y = 1) \tag{2}$$

Balance de Clases Negativas

Definition

Decimos que un algoritmo satisface balance de clases negativas:

$$P(\hat{Y} = 1 \mid Y = 0, A) = P(\hat{Y} = 1 \mid Y = 0) \tag{3}$$

Igualdad de Probabilidades (*equalized odds*)

Definition

Decimos que un algoritmo satisface igualdad de probabilidades si:

1. Balance de clases positivas.
2. Balance de clases negativas.

Outline

Predictor derivado

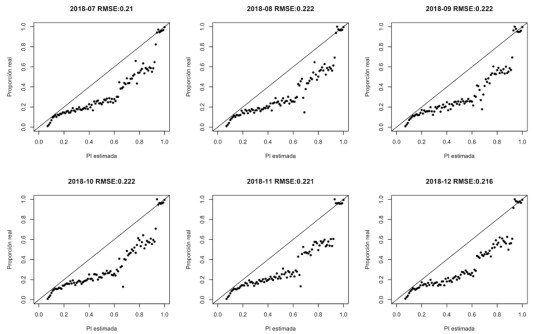
Definition (Predictor derivado)

Un predictor $\tilde{Y}$ es *derivado de la variable aleatoria R (e.g., score) y atributo A* si es una función de unicamente de (R, A) En particular, $\tilde{Y}$ es independiente de los covariantes X (que se usaron para construir R) condicional en (R, A).

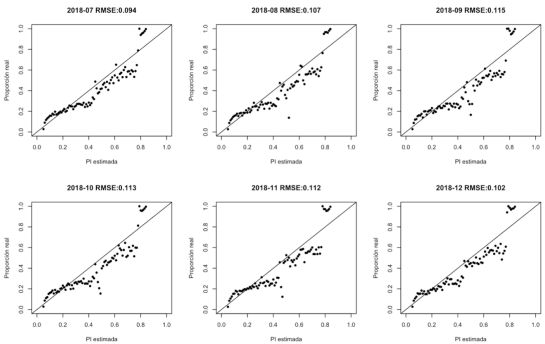
- $\tilde{Y}$ puede incluir una aleatorización adicional.

Calibración

- Platt scaling: $P(\tilde{Y}_i = 1) = \frac{1}{1+Ae^{\beta_1 P(\hat{Y}_i=1)+\beta_0}}$



(a) Modelo no calibrado

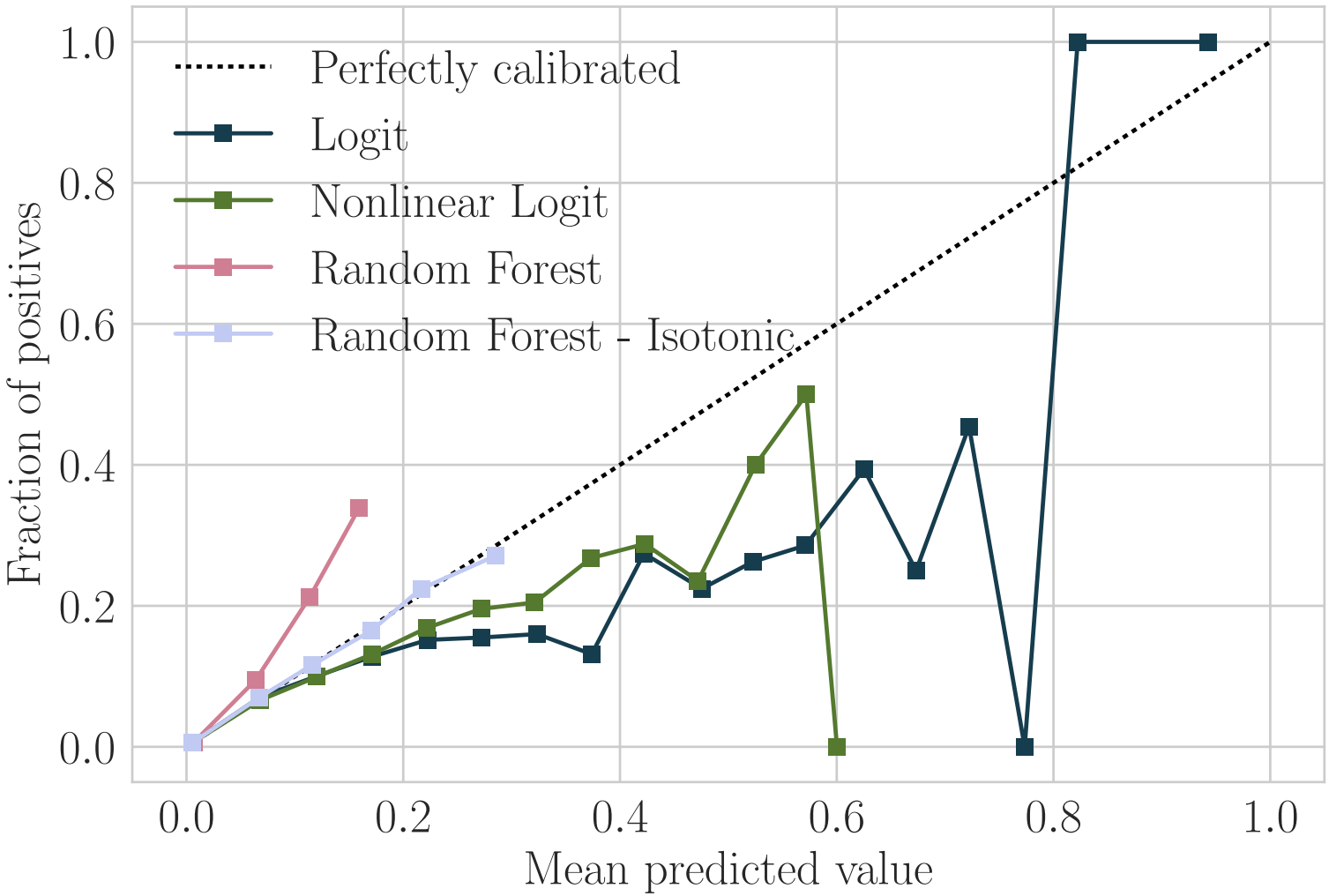


(b) Modelo calibrado con Platt Scaling

Figura 13: Curvas de calibración para las PI estimadas los meses de Agosto a Diciembre 2018 para (a) el modelo no calibrado de PI, y (b) el modelo con Platt Scaling, así como la raíz del error medio cuadrado (RMSE) de la diagonal, como medida de falta de calibración (entre menor sea este valor, mejor calibrado el modelo)

Calibrando Modelos: Transfomración Isotónica

Figure A-1: **Calibration Curve.**



Igualdad de Oportunidades

- Equality of Opportunity in Supervised Learning. 2016. Moritz Hardt, Eric Price, Nathan Srebro.
- Definamos:

$$\gamma_a(\hat{Y}) = \Big(P(\hat{Y} = 1 \mid A = a, Y = 0), P(\hat{Y} = 1 \mid A = a, Y = 1)\Big) \tag{4}$$

- La primera componente $\gamma_a(\hat{Y})$ es el *false positive rate* de $\hat{Y}$ en el grupo $A = a$.
- La segunda componente es *true positive rate* de $\hat{Y}$ en el grupo $A = a$.
- Obsérvece que $\gamma_a(\hat{Y})$ se puede calcular de la distribución conjunta de $(\hat{Y}, A, Y)$.

Igualdad de Oportunidades

- Igualdad de oportunidades e igualdad de probabilidades se pueden expresar en términos de $\gamma_a(\hat{Y})$
- Definamos:

$$P_a(\hat{Y}) = \text{convhull}\{(0,0), \gamma_a(\hat{Y}), \gamma_a(1-\hat{Y}), (1,1)\} \tag{5}$$

- Un predictor $\tilde{Y}$ es derivado de (R, A) si y solo si para todo a, $\gamma_a(\hat{Y}) \in P_a(\hat{Y})$.

Igualdad de Oportunidades

- Ajuste óptimo:

$$\min_{\tilde{Y}} E[L(\tilde{Y}, Y)] \tag{6}$$

$$\forall a : \gamma_a(\tilde{Y}) \in P_a(\hat{Y}) \tag{7}$$

$$\gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y}) \tag{8}$$

Igualdad de Oportunidades

- Umbrales específicos (Rodolfa, Lamba y Ghani [2021]. Empirical observation of negligible fairness-accuracy trade-offs in ML for public policy).

Igualdad de Oportunidades

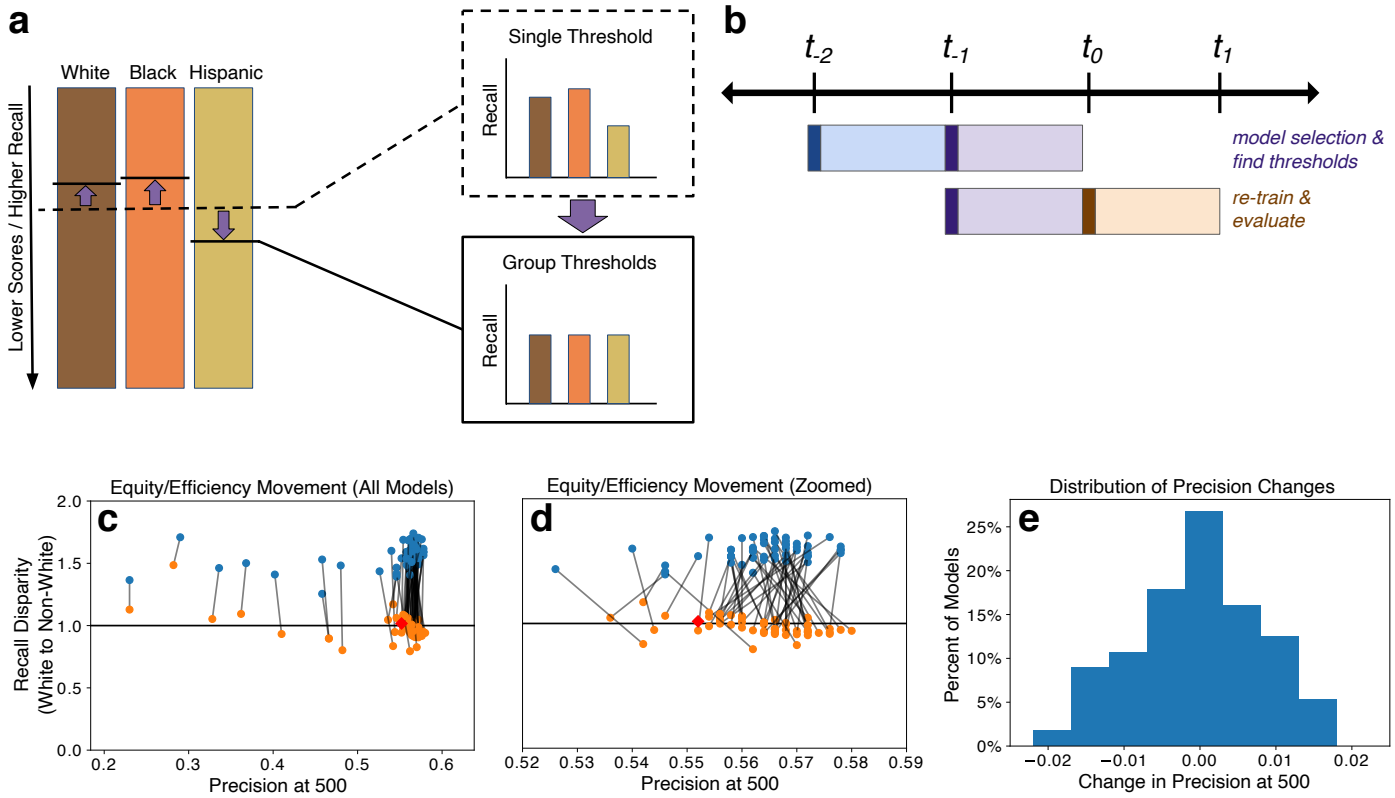


Illustration of the methods used and motivating results. (a) Subgroup-specific thresholds are applied to a modeled risk score to improve the recall equity among individuals chosen for intervention while maintaining a desired overall list size. (b) Temporal validation strategy: a grid of models is trained using examples as of t_{-2} (dark blue, with labels derived from the time shown in light blue) and predictions on a cohort as of t_{-1} (dark purple with labels derived from the time shown in light purple) are used to determine the equity-balancing thresholds described in (a). Models are then re-trained on this cohort for "current day" predictions as of t_0 (dark tan, with labels in light tan) used to evaluate model performance with equity adjustments. (c and d): Changes in race/ethnicity recall disparities before (blue) and after (orange) making post-hoc score adjustments for fairness in the Inmate Mental Health context. (c) shows all model specifications and (d) shows the cluster of well-performing models. The red diamond reflects the performance of a composite model combining the best-performing model for each subgroup. (e) Distribution of precision changes after adjusting for disparities for the models shown in (c), relative to the precision attained by the same model specification without adjustment (that is, the difference along the x-axis of the blue and orange dots).

Outline

Kleinber, Mullainathan y Raghavan

- Decimos que un algoritmo satisface *equalized odds* si satisface balance de clase positivas más balance de clase negativas.
- Teorema de Imposibilidad: Es imposible que, excepto en casos triviales, un algoritmo satisfaga: equalized odds y la propiedad de calibración.