Aprendizaje de máquinas: Discriminación

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

29 de junio de 2020

Contenido

Outline

Introducción

- El aprendizaje de maquinas 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 maquinas 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 esta 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.
- Equilibrio: las predicciones desconocen que pueden existir interacciones estratégicas.
- Privacidad: con suficiente información es posible identificar personas o atributos privados.
- Este es uno de los grandes retos en las aplicaciones de las matemáticas y aprendizaje de máquinas al mundo real.

Outline

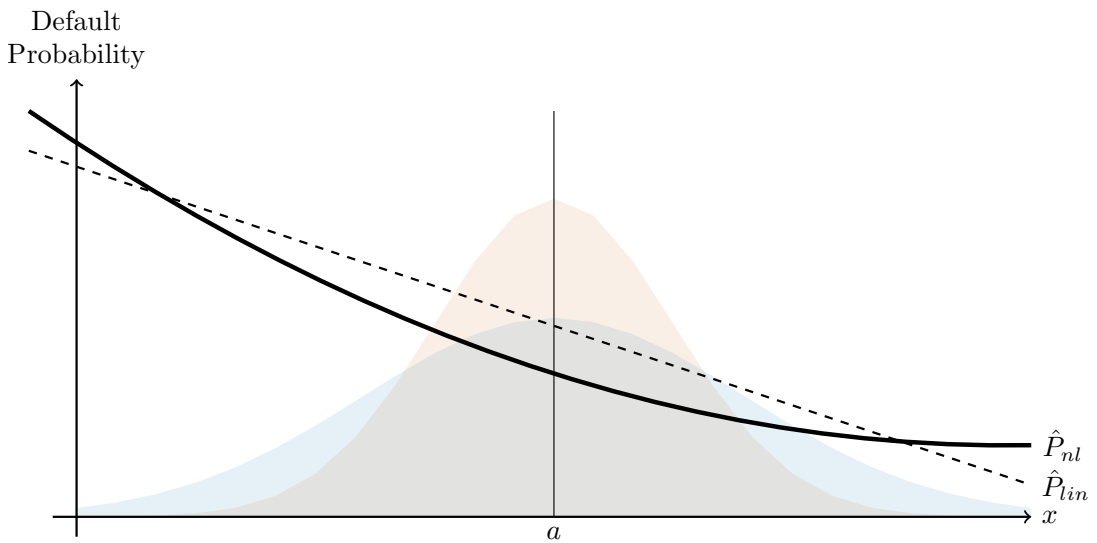
Introducción

- Usan una base de datos de 10 millones de hipotecas.
- Los autores encuentran que los negros y e hispanos ganan menor con la introducción de ML en el sentido de que sus probabilidades de incumplimiento no disminuyen.
- Los modelos de ML aumentan el otrogamiento 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 felxibiliad) por encima de fenómeno de triangulación.

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**



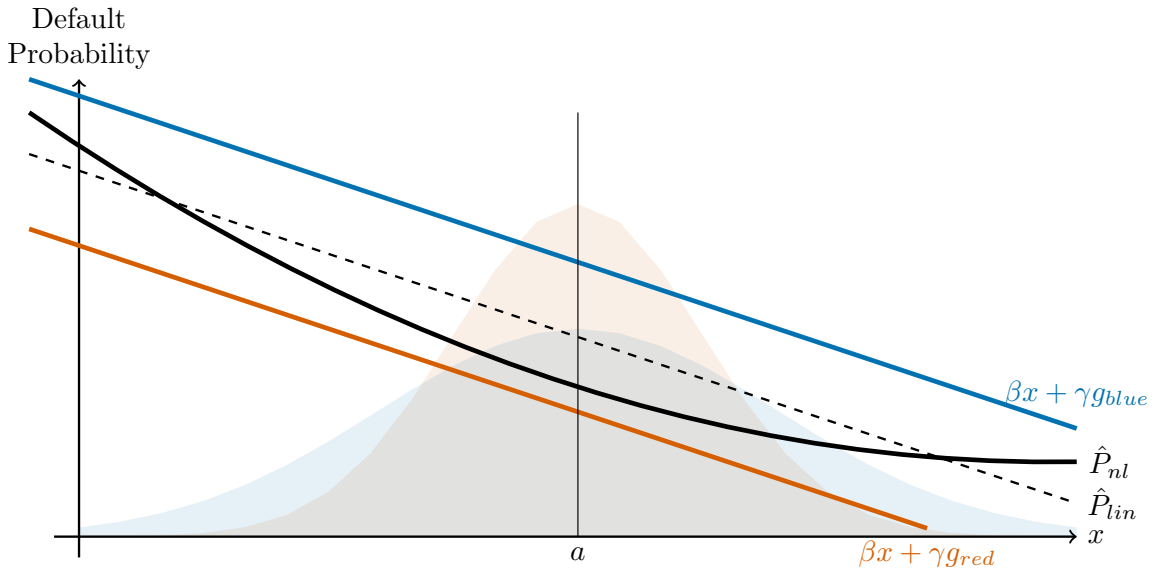
Triangulación

- Ahora suponga que existe una carcaterística (i.e., raza) que no puede ser utilizada para discriminar y tiene correlacion cero con la variable observada.
- Supongamos que el verdadero modelo es de la forma $y = \beta x + \gamma g$.
- Una tecnología no lineal no puede, por definición, mejorar el anterior modelo.
- Sin embargo, en asuencia de la variable protegida en el modelo, sí puede ser mejor descubriendo el papel de la variable protegida.

Triangulación

- Una mejor tecnología tiene efectos negativos sobre el grupo azul (los más dispersos).

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

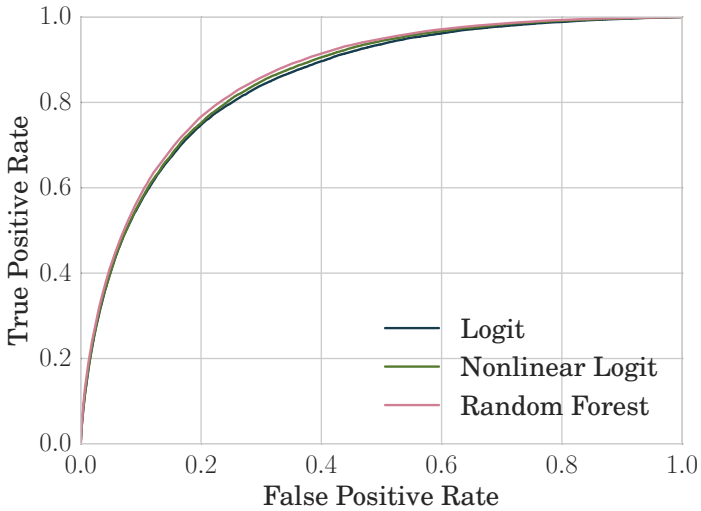
Table 2: **Variable List**

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)	

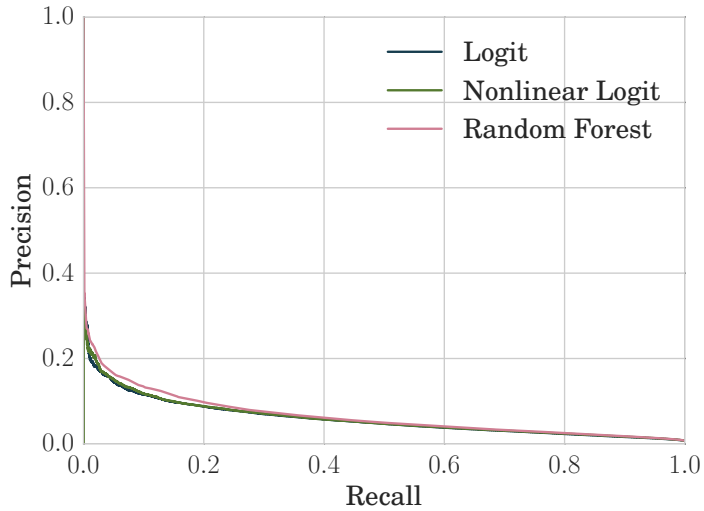
Note: Variables used in the models. Data source: HMDA-McDash matched dataset of conventional fixed-rate mortgages.

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

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

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.
- En particular, si los positivos son muy poco (incumplidos) los falsos positivos son aún menos (piense en un modelo perfecto en donde los falsos positivos serían cero) y FPR es cercano a cero.

Curvas: ROC y Precision-Recall

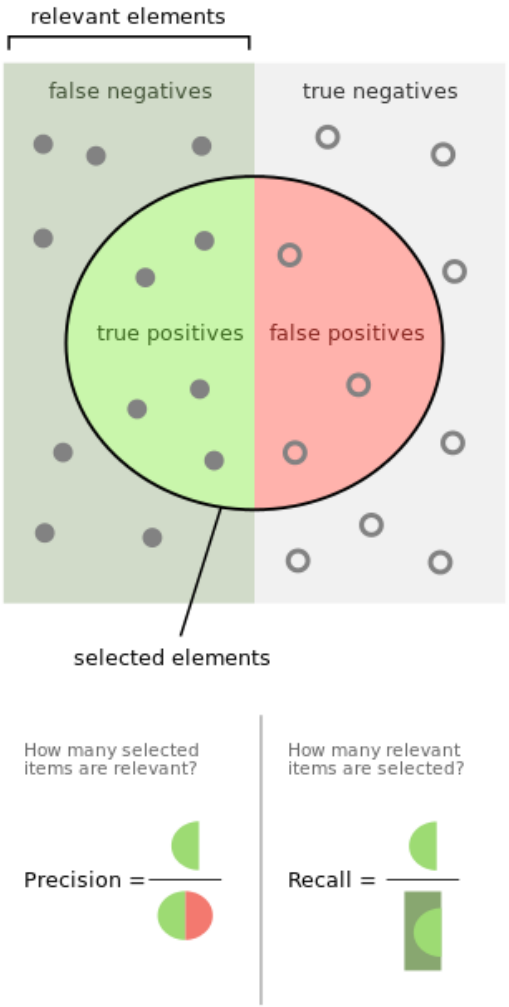
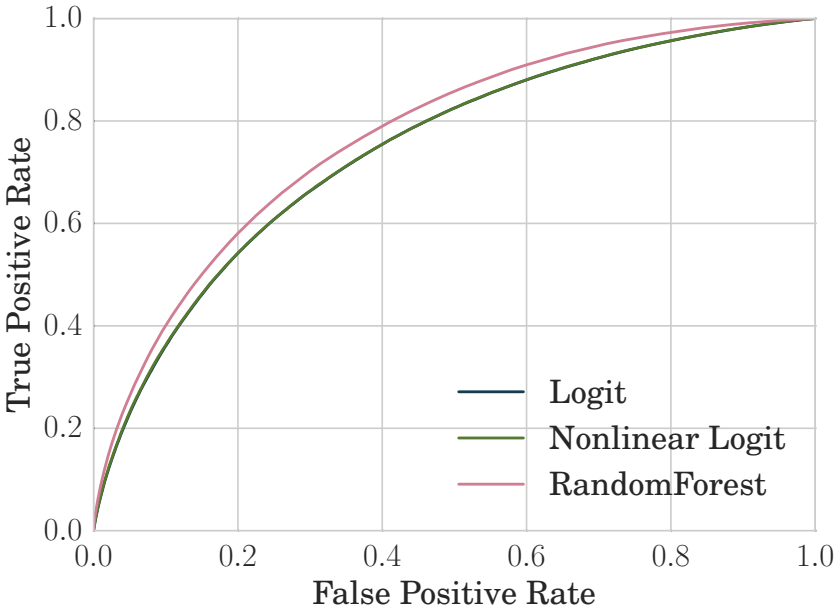


Figura 1: By Walber - Own work, CC BY-SA 4.0, `https://commons.wikimedia.org/w/index.php?curid=36926283`

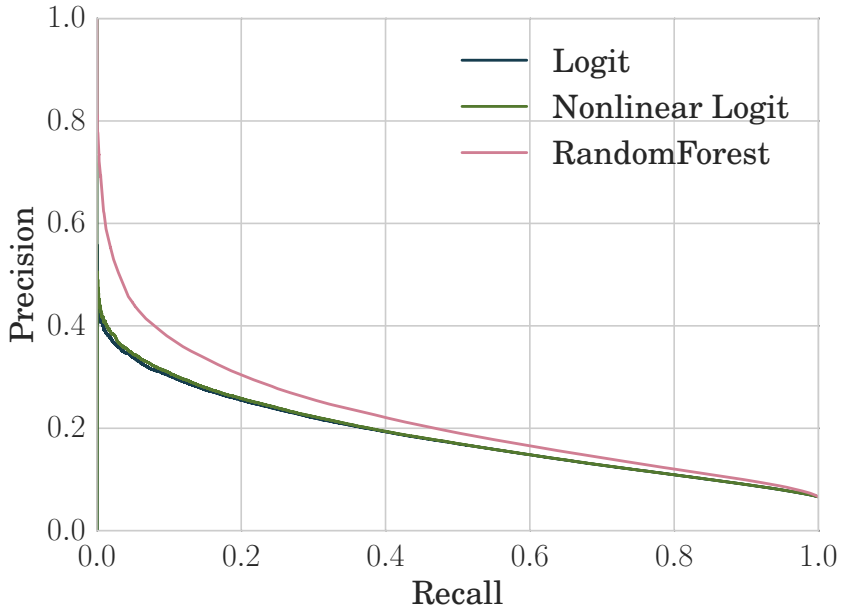
Obsérvese que TPR = Recall

Predicción de raza

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



Panel A



Panel B

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

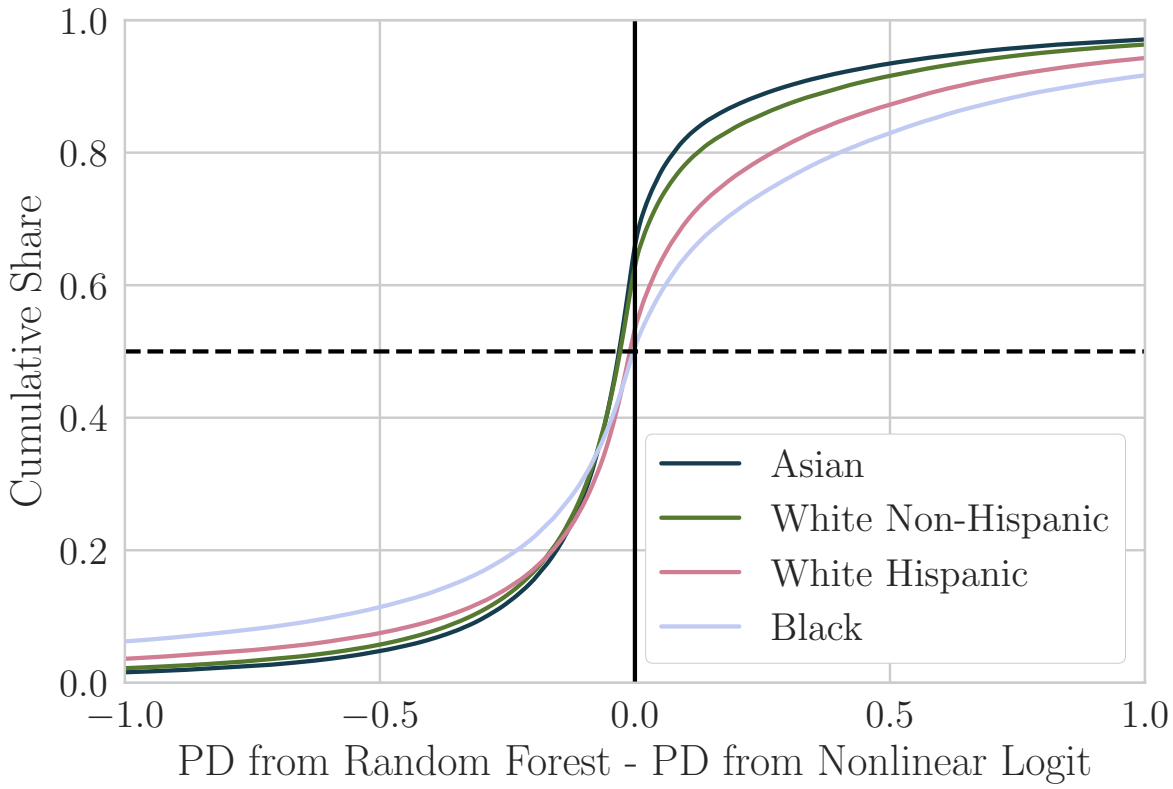
Panel A: Race Controls First

	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