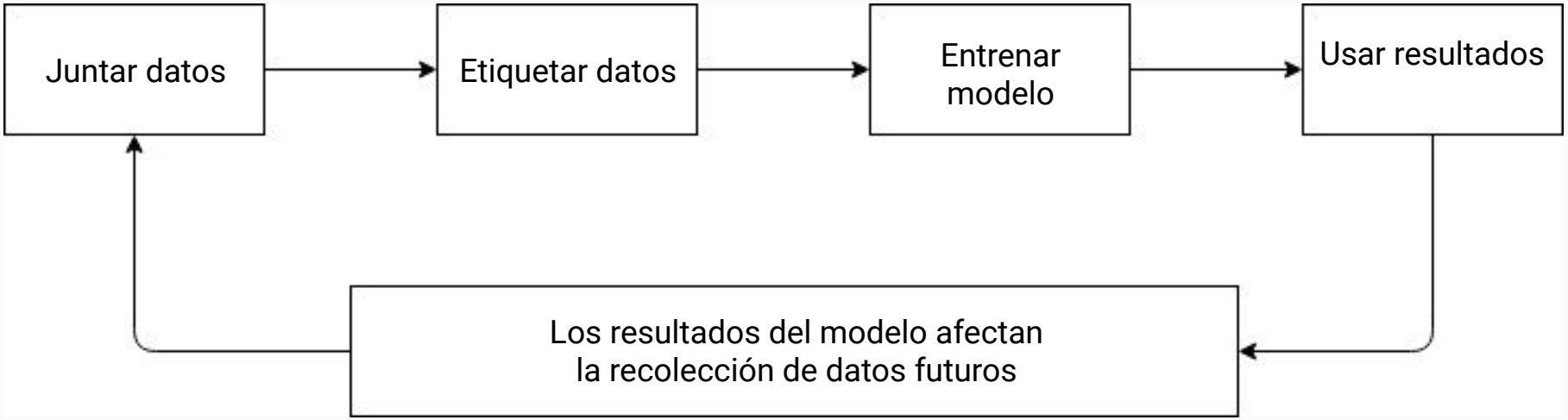
All Computers Are Bastards

Introducción a *fairness* en *machine learning*

Sebastián Waisbrot

Machine learning

Flujo de *machine learning*



Sysarmy - Encuesta de sueldos

Cada seis meses hacen un relevamiento de sueldos.

Preguntan muchas cosas:

- Información personal (género, edad, orientación sexual, discapacidades)
- Información laboral (experiencia, años en la empresa, capacitaciones)
- Información del empleador (cantidad de empleados, lenguajes de programación, beneficios)
- Sueldo

Sebastian Ariel Waisbrot
@seppo0011

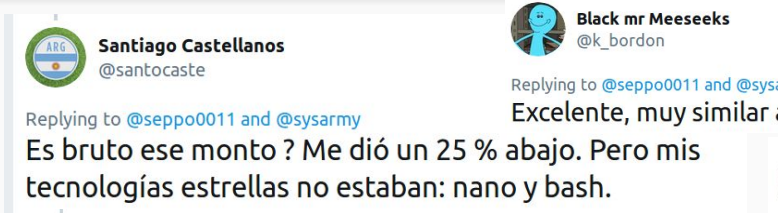
Usé las respuestas de la encuesta de @sysarmy para armar un modelo de sueldos. Pueden probarlo y ver cómo lo armé acá seppo0010.github.io/sysarmy-sueldo…

Translate Tweet
5:23 PM · Mar 12, 2019 · Twitter Web Client

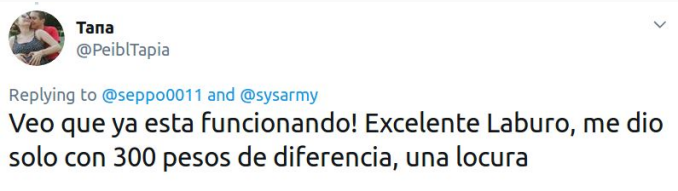
View Tweet activity

70 Retweets **225** Likes

Resultados


Santiago Castellanos
@santocaste
Replying to @seppo0011 and @sysarmy
Es bruto ese monto ? Me dió un 25 % abajo. Pero mis tecnologías estrellas no estaban: nano y bash.


Black mr Meeseeks
@k_bordon
Replying to @seppo0011 and @sysarmy
Excelente, muy similar a lo que estoy cobrando


Tana
@PeiblTapia
Replying to @seppo0011 and @sysarmy
Veo que ya esta funcionando! Excelente Laburo, me dio solo con 300 pesos de diferencia, una locura


Juan Soto
@juansoto86
Replying to @seppo0011 and @sysarmy
Muy bien me dio!


Patricia
@Luthoriel_
Replying to @seppo0011 and @sysarmy
La pegó de una manera muy exacta. Para mí que me espió el recibo de sueldo.


Sebastian Cipolat
@seba_cipolat
Replying to @seppo0011 and @sysarmy
Le error un toque pero como idea sirve


Tamar Anush
@tamar_moz
Replying to @seppo0011 and @sysarmy
Muy bueno! Pero a mí me dio que debería tener un sueldo que es más del doble del actual

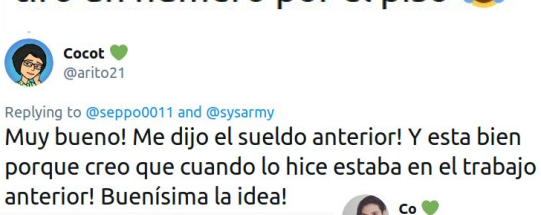

Santiago Basulto
@santiagobasulto


SeedVicious
@SeedVicious
Replying to @seppo0011 and @sysarmy
200 $ de diferencia

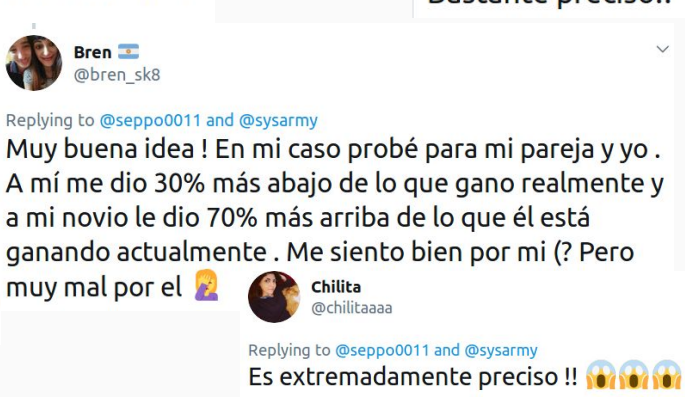

Wertygol
@wertygol
Replying to @seppo0011 and @sysarmy
Bastante preciso!!


Pink Whale
@PinkyWhale
Replying to @seppo0011 and @sysarmy
Le pifió por 7k


Replying to @seppo0011 and @sysarmy
Se ve que debo ser un gran cebador de mates, pq me tiró un número por el piso


Cocot
@arito21
Replying to @seppo0011 and @sysarmy
Muy bueno! Me dijo el sueldo anterior! Y esta bien porque creo que cuando lo hice estaba en el trabajo anterior! Buenísima la idea!


Co
@co_constanza
Replying to @seppo0011 @ramblinn_ and @sysarmy
¡Genial! Muy, muy aproximado


Bren
@bren_sk8
Replying to @seppo0011 and @sysarmy
Muy buena idea ! En mi caso probé para mi pareja y yo . A mí me dio 30% más abajo de lo que gano realmente y a mi novio le dio 70% más arriba de lo que él está ganando actualmente . Me siento bien por mi (? Pero muy mal por el
Chilita
@chilitaaaa
Replying to @seppo0011 and @sysarmy
Es extremadamente preciso !!


Facundo
@FacundSua
Replying to @seppo0011 and @sysarmy
Bueno no le vamos a pasar ésto a mí empleador porque, en papel, me estoy choreando 10 lucas casi


agustin
@Agusttiny
Replying to @seppo0011 and @sysarmy
Le pegó bastante bien, esto también te sirve como par
orientarte hacia donde ir $:p

Resultados

Malena Rey 💚
@malerey_

Replying to @malerey_ @seppo0011 and @sysarmy

Lo depresivo es que muestra muy claro que el mismo perfil cobra menos solo cambiando la variable hombre/mujer

Gabriel Patiño 💚🧡
@gepatino

Replying to @seppo0011 and @sysarmy

Te baja el sueldo casi un 10% por cambiar de hombre a mujer. Sigue pasando en nuestra industria?
Mal ahí...

Fairness

Doctrinas de discriminación

Tratamiento desparejo

Impacto desparejo

Clases

Ejemplos de casos de *fairness*

No hay una única definición. Vamos a elegir una definición.

Resultados son independientes de una variable que consideramos sensible y no relacionada (por ejemplo género, grupo étnico, orientación sexual)

Veamos ejemplos del mundo legal.

Ricci vs DeStefano

Bomberos de Connecticut tomaron un examen para ser promovidos.

Entre los que rendían el examen eran 57% blancos, 23% negros, y 20% latinos. Los resultados que dieron promocionarían 88% blancos y el resto, latinos.

Ricci vs DeStefano

Doctrinas de discriminación

Tratamiento desparejo

(ignorar resultados)

Impacto desparejo

(usar resultados)

Grupos étnicos

Universidad pública en Texas

Problema: Necesita satisfacer tratamiento parejo e impacto parejo simultáneamente.

Solución: *El Plan del Diez Por Ciento de Texas* provee admisión automática del **10 por ciento superior de cada secundario** a cualquier universidad pública del estado.

Larry P. vs Riles

Niñes de California eran ubicados en clases para les "*educable mentally retarded*" basado en tests de Coeficiente Intelectual.

25% de las personas que fallaban la prueba eran personas negras, cuando sólo representaban al 10% del total.

> *While many think of the **I.Q. as an objective measure of innate, fixed intelligence**, the testimony of the experts overwhelmingly demonstrated that this conception of I.Q. **is erroneous**.*

Larry P. vs Riles

***Defendants are enjoined from utilizing**, permitting the use of, or approving the use **of any standardized intelligence tests**, including those now approved pursuant to Cal.Admin.Code § 3401, **for the identification of black E.M.R. children** or their placement into E.M.R. classes, without securing prior approval by this court.*

¿Sesgos de algoritmos? Un estudio empírico de discriminación aparente basada en geńero al mostrar publicidades de STEM

Publicidades para promocionar las oportunidades de trabajo en *STEM* (*Science, technology, engineering, and mathematics*).

Sin preferencia de género en el diseño y publicación.

Más impresiones en hombres que en mujeres en todas las plataformas.

Tratamiento parejo, impacto disparejo.


GDPR

General Data Protection Regulation

*Processing of personal data revealing **racial or ethnic origin**, **political opinions**, **religious or philosophical beliefs**, or trade union membership, and the processing of **genetic data**, **biometric data** for the purpose of uniquely identifying a natural person, data concerning **health** or data concerning a natural person's **sex life or sexual orientation** shall be **prohibited**.*

General Data Protection Regulation

*The data subject shall have the right **not to be subject to a decision based solely on automated processing**, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*

(...)

*In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the **right to obtain human intervention**, to express his or her point of view, **to obtain an explanation of the decision** reached after such assessment and **to challenge the decision**.*

Leyes Argentina

I AM NOT A LAWYER

GRAPHIC DESIGN IS MY PASSION

*Legislar y promover **medidas de acción positiva** que garanticen la **igualdad real** de oportunidades y de trato, y el pleno goce y ejercicio de los derechos reconocidos por esta Constitución y por los tratados internacionales vigentes sobre derechos humanos, **en particular respecto de los niños, las mujeres, los ancianos y las personas con discapacidad**.*

Constitución Nacional - Artículo 75 Inciso 23

*Legislar y promover **medidas de acción positiva** que garanticen la **igualdad real** de oportunidades y de trato, y el pleno goce y ejercicio de los derechos reconocidos por esta Constitución y por los tratados internacionales vigentes sobre derechos humanos, **en particular respecto de los niños, las mujeres, los ancianos y las personas con discapacidad**.*

Debate Asamblea Constituyente 1994

No se trata de crear vocaciones artificiales sino de recrear el medio para que las auténticas vocaciones políticas puedan manifestarse espontáneamente.
Las mujeres somos más de la mitad de la población del país. Aun con esa superioridad numérica a favor soy testigo de ***cuán complicado es participar políticamente en un medio hasta ahora dominado por los hombres.***
La Constitución que estamos forjando no puede dejar de tener en cuenta este hecho para ser tan justa y democrática como la queremos.

Nilda Romero

Debate Asamblea Constituyente 1994

No puede haber democracia sin igualdad de derecho entre los sexos. Discriminar a la mujer es una forma de violencia encubierta al ejercer por la fuerza el veto a la igualdad. Sin embargo, ***las pautas culturales aceptadas en nuestra sociedad se traducen perpetuando la discriminación****. Por eso no sólo es importante que nuestros derechos tengan rango constitucional sino que nosotras sepamos concientizar tanto a hombres como a mujeres acerca de la justicia de nuestros logros. Debemos aprender y enseñar a participar con sentido solidario junto al hombre, pero conservando nuestra condición de mujer, difundiendo estos derechos para que sean —como deben serlo— interpretados como un reconocimiento justo. Porque nada se nos dio ni se nos concedió, sino que sólo se tacharon de la historia de la mujer oscuros olvidos e incomprensibles negaciones.*

Dora Sachs de Repetto

*Más adelante, el texto agrega algo que parece un parche. Dice así: "...y el pleno goce y ejercicio de los derechos reconocidos por esta Constitución y por los tratados internacionales sobre derechos humanos vigentes, respecto de los niños, las mujeres, los ancianos y las personas con discapacidad." **Me siento excluido** porque no reúno ninguna de estas condiciones. (Risas) Tal vez pueda estarlo por mi edad.*

Álvaro Alsogaray

Debate Asamblea Constituyente 1994

Debate Asamblea Constituyente 1994

No quiero que mis palabras se interpreten como una falta de amor al prójimo, pero considero que los niños, por su inmadurez psíquica y física, los discapacitados por su situación de discapacidad, y los ancianos, que por estar en una etapa biológica natural no tienen el pleno ejercicio de sus aptitudes y facultades, no están en la misma condición que las mujeres. Entiendo que la redacción que se propone de algún modo implicaría aceptar una situación de inferioridad que no es justa respecto de la mujer.

Más que un informe lo que estoy expresando es una inquietud que formulo a los integrantes de la comisión que elaboró el dictamen a efectos de que traten de aclarar ***por qué a las mujeres nos han ubicado*** *—trato de tener mucho cuidado con el término a utilizar—* ***en esta categorización junto a los niños, los ancianos y los discapacitados****. Lo hago —reitero— sin querer que mis palabras se interpreten como una falta de caridad o de amor al prójimo puesto que nada está más alejado de mis verdaderos sentimientos. Simplemente se trata de algo que no considero justo.*

María Teresita Colombo

Leyes Argentina

Criterio de razonabilidad: la medida tiene que satisfacer que los medios son adecuados y proporcionados al fin.

Criterio estricto: los medios son efectivos para los fines sustanciales que persigue y no hay medios alternativos para lograrlo.

El criterio estricto se ha usado de acuerdo a la vulnerabilidad de los grupos.

Sisnero, Mirtha Graciela y otros c/Taldelva SRL y otros s/amparo

Transporte público urbano.

Sisnero no es contratada.

No hay mujeres choferes.

Sisnero, Mirtha Graciela y otros c/Taldelva SRL y otros s/amparo

Si el reclamante puede acreditar la existencia de hechos de los que pueda presumirse su carácter discriminatorio, corresponderá al demandado la prueba de su inexistencia.

Corte Suprema de Justicia de la Nación

Sisnero, Mirtha Graciela y otros c/Taldelva SRL y otros s/amparo

Un claro ejemplo en esta dirección, por cierto, lo constituyen las manifestaciones de ***uno de los empresarios*** *demandados ante un medio periodístico, quien, con relación a este juicio, señaló sin ambages y "entre risas" que "esto es Salta Turística, y* ***las mujeres deberían demostrar sus artes culinarias*** *[...]* ***Esas manos son para acariciar****, no para estar llenas de callos [...] Se debe ordenar el tránsito de la ciudad, y [...]* ***no es tiempo de que una mujer maneje colectivos*** *[_] (cf. entrevista agregada a fs. 564).*

Machine learning y leyes

Al aplicar *machine learning* el criterio razonable está automáticamente cumplido porque usar *machine learning* para tomar una decisión es razonable.

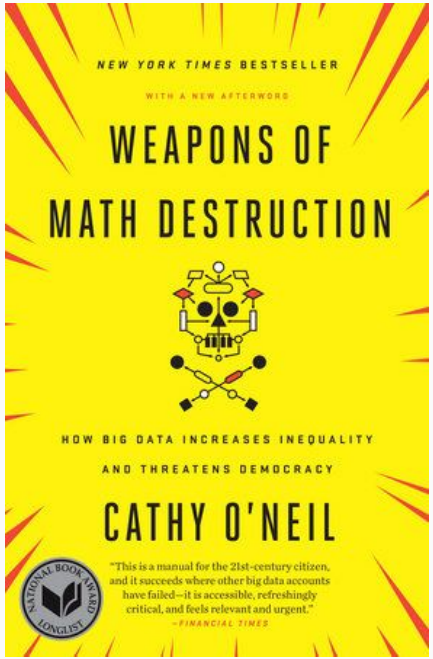
El criterio estricto va a ser difícil de argumentar porque requiere demostrar que no había una forma alternativa y es difícil discutirle a una computadora.

Machine learning es inevitable

Cada vez se usa más *machine learning*.

Se pueden tomar mejores decisiones o a menor costo.

Weapons of Math Destruction



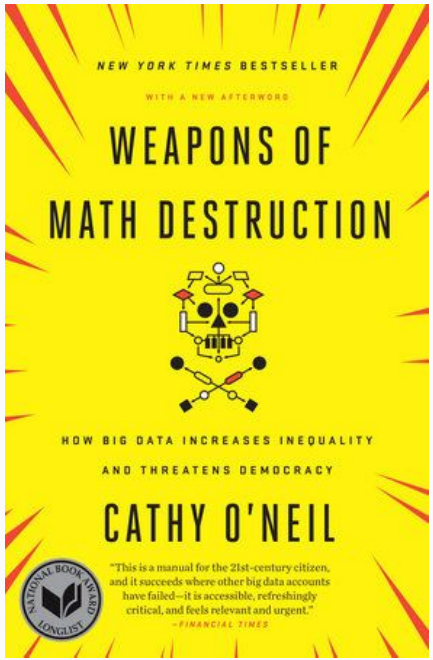
*The models being used today are **opaque, unregulated, and uncontestable**, even when they're wrong. Most troubling, they reinforce discrimination: If a poor student can't get a loan because a lending model deems him too risky (by virtue of his zip code), he's then cut off from the kind of education that could pull him out of poverty, and a **vicious spiral** ensues.*

Weapons of Math Destruction

I've been trying to convince, say, large companies who use hiring algorithms to hire me to audit those algorithms, with the argument that they're the ones on the hook if something goes wrong. Unfortunately they simply don't think they really would be on the hook. So ***we need some legal or regulatory standards*** *to be established to get this business model to work.*

Cathy O'Neil

Weapons of Math Destruction



I've been trying to convince, say, large companies who use hiring algorithms to hire me to audit those algorithms, with the argument that they're the ones on the hook if something goes wrong. Unfortunately they simply don't think they really would be on the hook. So ***we need some legal or regulatory standards*** *to be established to get this business model to work.*

Cathy O'Neil

Fairness

Considerar el tratamiento e impacto sobre distintas clases.

Cuando aplicamos *machine learning* no parece haber obligación legal de considerarlo.

Como llegan los sesgos a *machine learning*

Muestras sesgadas

Datos recolectados de un grupo de personas que no representan al resto.

Ejemplos:

- Encuestas en Twitter
- Reportes de crímenes

Ejemplos contaminados

Usar información sesgada generada en el pasado.

Ejemplos:

- Contrataciones
- Puntajes de jefes

Limited features

Usar datos que varíen en calidad para distintos grupos.

Ejemplo:

- Combinar encuestas telefónicas y online

Sample size disparity

El peso de una clase en las métricas de resultados es proporcional a la cantidad de observaciones.

Proxies

Correlación entre clase protegida y features.

Ejemplo:

- Género y redes sociales

Paradoja de Simpson

Una tendencia que se ve en distintos grupos desaparece o se revierte en el agregado.

Admisiones a UC Berkeley en 1973.

	Men		Women	
	Applicants	**Admitted**	**Applicants**	**Admitted**
Total	8442	**44%**	4321	35%

Department	**Men**		**Women**	
	Applicants	**Admitted**	**Applicants**	**Admitted**
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**

Corto plazo vs largo plazo

Los algoritmos en general están entrenados con métricas de corto plazo sin tener en cuenta la influencia que pueden tener en el largo.

Ejemplos:

- Paperclip Maximizer
- Youtube, the Great Radicalizer

Retroalimentación

Las decisiones tomadas por un modelo afectan los datos que serán recolectados para un nuevo modelo.

Ejemplos:

- Pre-filtro de candidates

Herramientas para mejorar *fairness*

Explorar/Explotar

Alternar entre usar el modelo (explotar) o buscar datos nuevos (explorar).

Ejemplo:

- Pre-filtro de candidates

Los modelos son contextuales

La salida de los modelos no se usa en un estado puro. Lo que sea que vaya a consumirlo puede considerar que esto tendrá errores y sesgos y ver cómo mitigarlo.

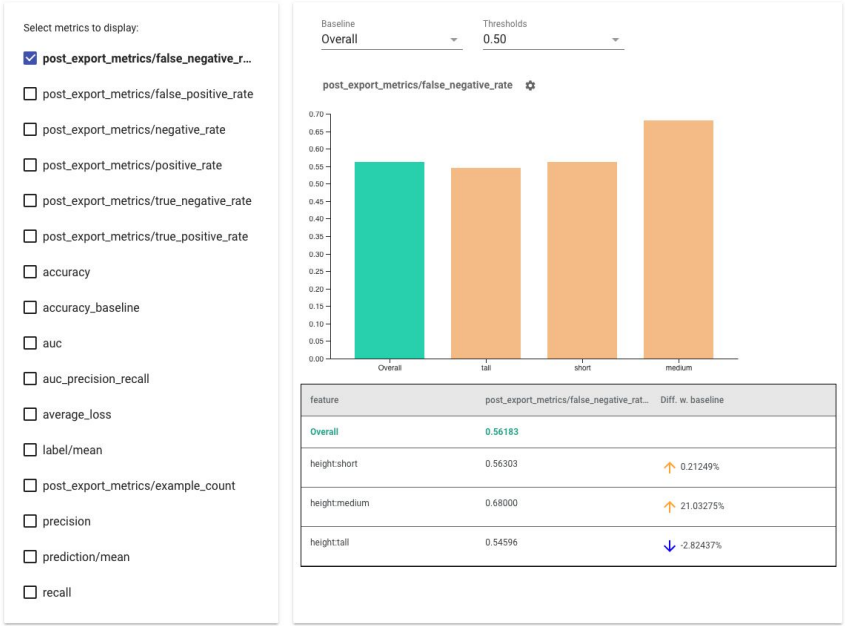
Ejemplo:

- Género en Google Translate.

fairness-indicators

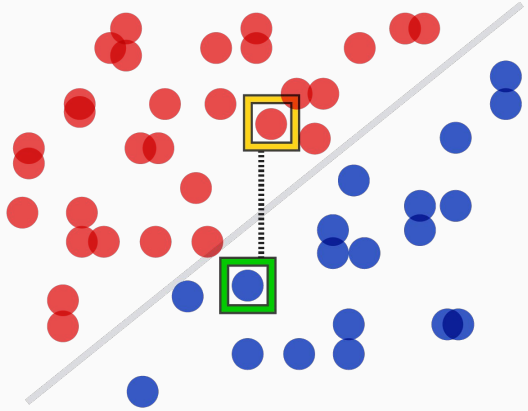
Fairness Indicators includes the ability to:

- *Evaluate the distribution of datasets*
- *Evaluate model performance, sliced* ***across defined groups of users***
- *Dive deep into individual slices to explore root causes and* ***opportunities for improvement***

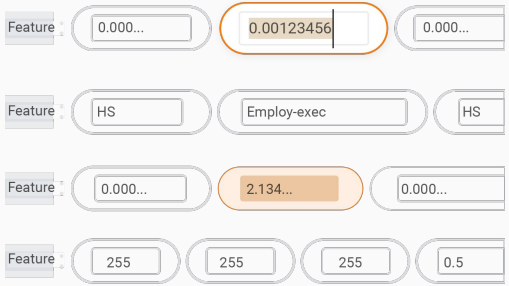


What-If Tool

For any selected datapoint, find the most similar datapoint of a different classification.

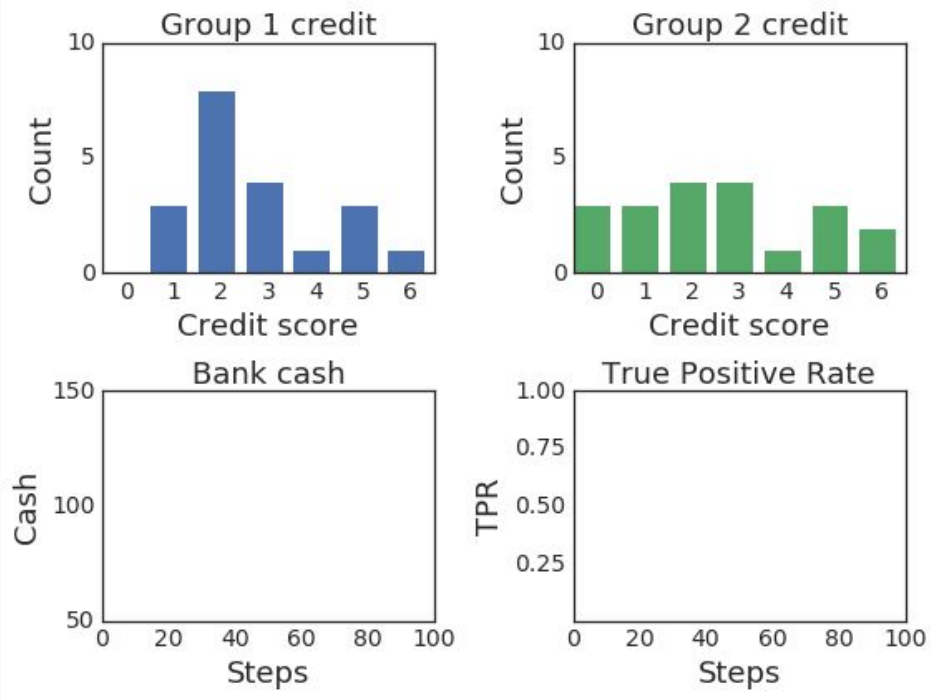


Edit, add or remove features or feature values for any selected datapoint and then run inference to test model performance.

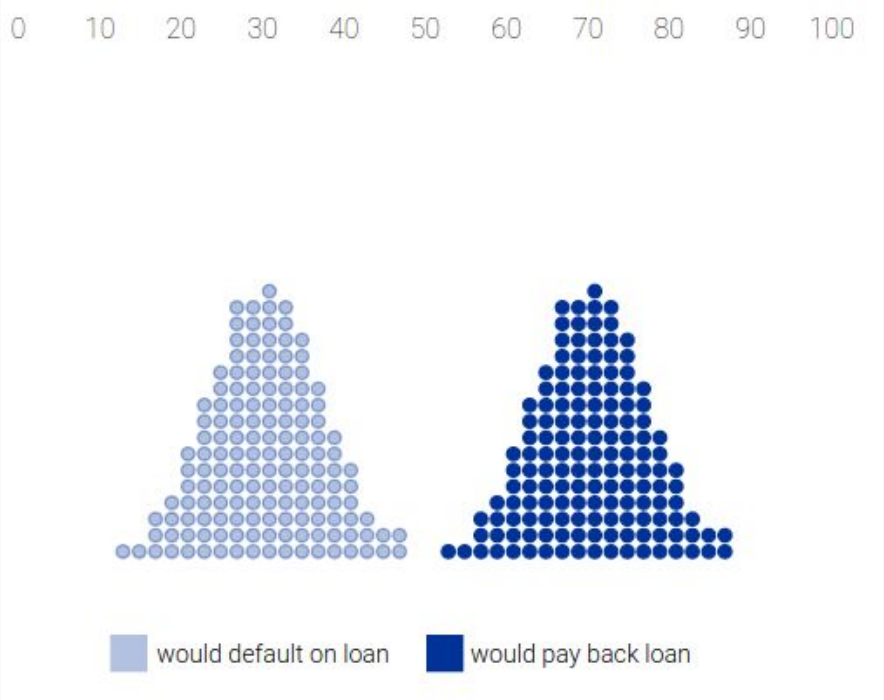


ML-fairness-gym

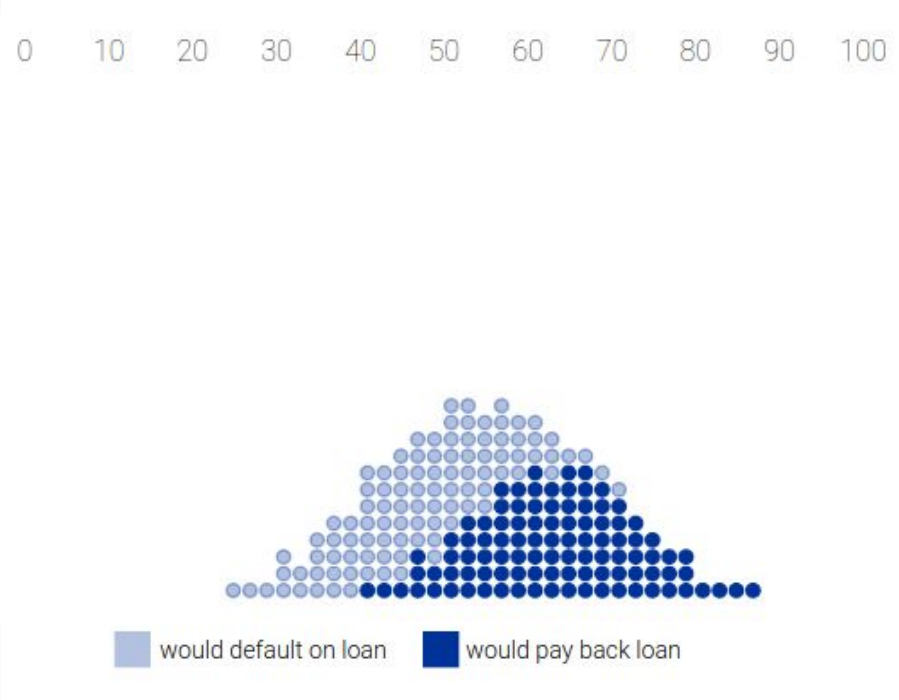
*A set of components for building simple simulations that explore potential **long-run impacts** of deploying machine learning-based decision systems in social environments.*



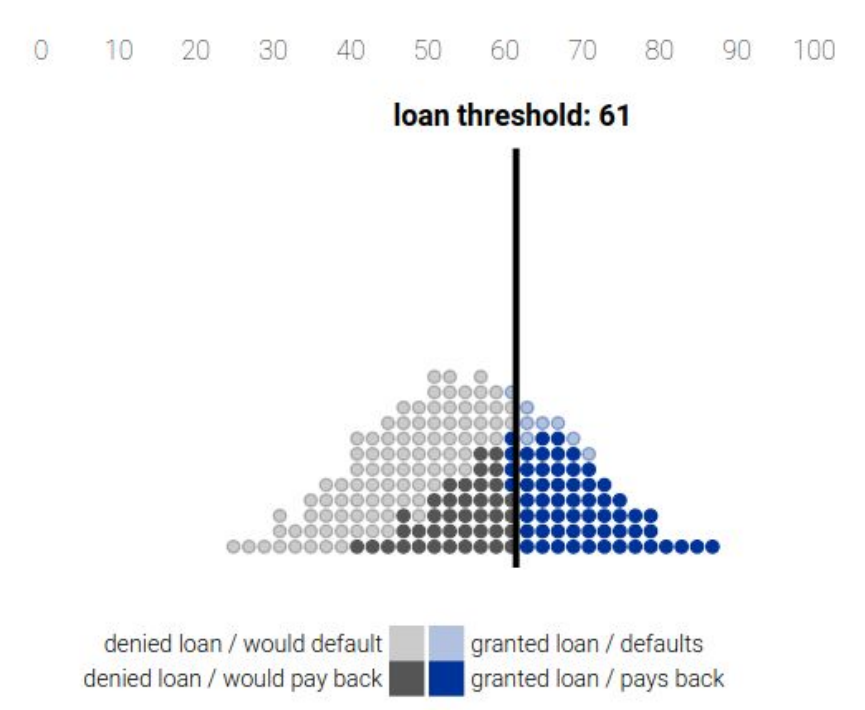
Ejemplo de modelo *machine learning* Caso ideal



Ejemplo de modelo *machine learning*
Caso realista



Ejemplo de modelo *machine learning* Caso realista



Ejemplo de modelo *machine learning*
Matriz de confusión

	Denied loan	Granted loan	Total
Defaults	A	B	A+B
Pays back	C	D	C+D
Total	A+C	B+D	A+B+C+D

Ejemplo de modelo *machine learning*
Matriz de confusión

(th=61)	Denied loan	Granted loan	Total
Defaults	91	8	99
Pays back	40	59	99
Total	131	67	198

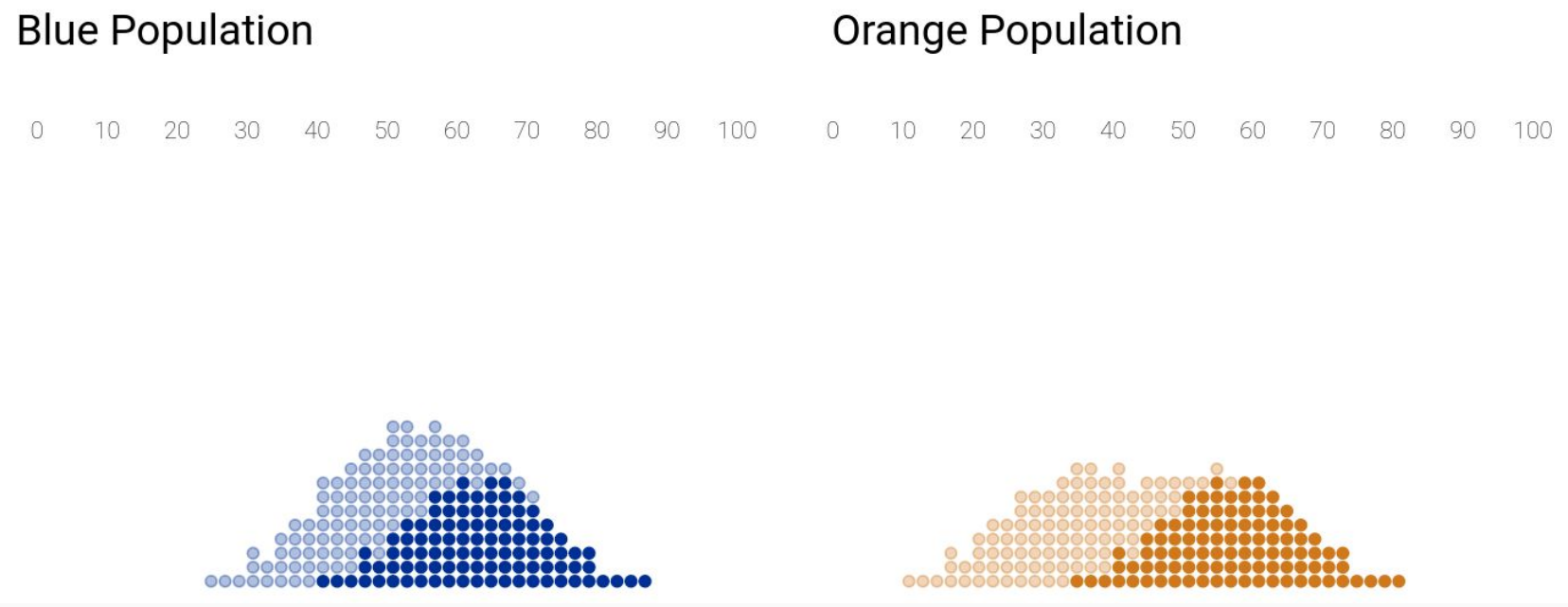
Métricas de Matriz de confusión

Accuracy: (A+D)/(A+B+C+D) = **0.76**
Aciertos sobre el total de lo evaluado.

Positive Rate: (A+C)/(A+B+C+D) = **0.66**
Rechazos sobre el total de lo evaluado.

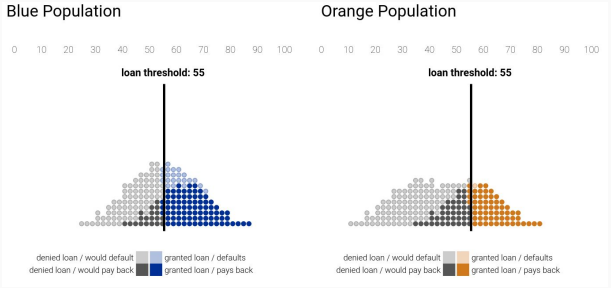
True Positive Rate: A/(A+C) = **0.60**
Rechazado sobre los que no hubiesen devuelto el préstamo.

Ejemplo de modelo *machine learning*

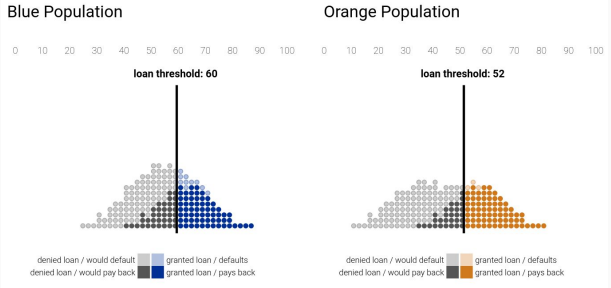


Ejemplo de Modelo Machine Learning

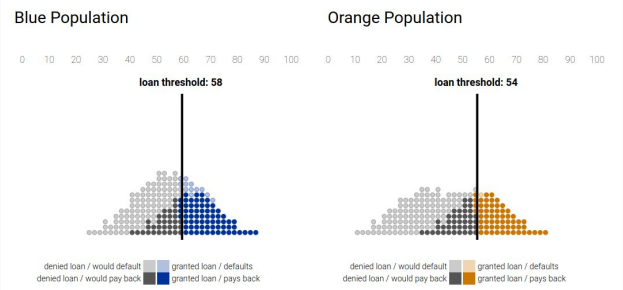
Group Unaware



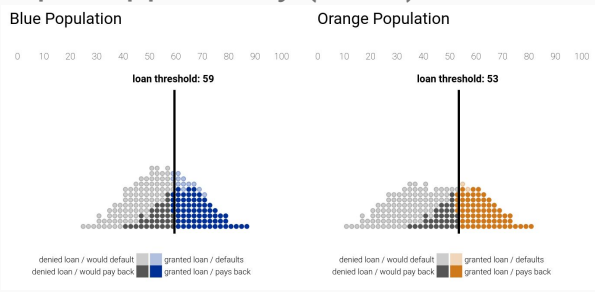
Demographic parity (=PR)



Overall Accuracy Equality (=Acc)



Equal opportunity (=TPR)



Ejemplo de Modelo Machine Learning

Group Unaware

(th=55, azul)	Denied loan	Granted loan	(th=55, naranja)	Denied loan	Granted loan
Defaults	77	22	Defaults	98	1
Pays back	19	80	Pays back	40	59

Demographic parity (=PR)

(th=60, azul)	Denied loan	Granted loan	(th=52, naranja)	Denied loan	Granted loan
Defaults	89	10	Defaults	96	3
Pays back	36	63	Pays back	29	70

Overall Accuracy Equality (=Acc)

(th=58, azul)	Denied loan	Granted loan	(th=55 naranja)	Denied loan	Granted loan
Defaults	85	14	Defaults	98	1
Pays back	29	70	Pays back	40	59

Equal opportunity (=TPR)

(th=59, azul)	Denied loan	Granted loan	(th=53, naranja)	Denied loan	Granted loan
Defaults	87	12	Defaults	97	2
Pays back	32	67	Pays back	32	67

Ejemplo de modelo *machine learning*

Group Unaware

	Azul	Naranja
Accuracy	0.79	0.79
Positive Rate	0.48	0.70
True positive rate	0.81	0.60

Demographic parity (=PR)

	Azul	Naranja
Accuracy	0.76	0.83
Positive Rate	**0.67**	**0.65**
True positive rate	0.60	0.68

Overall Accuracy Equality (=Acc)

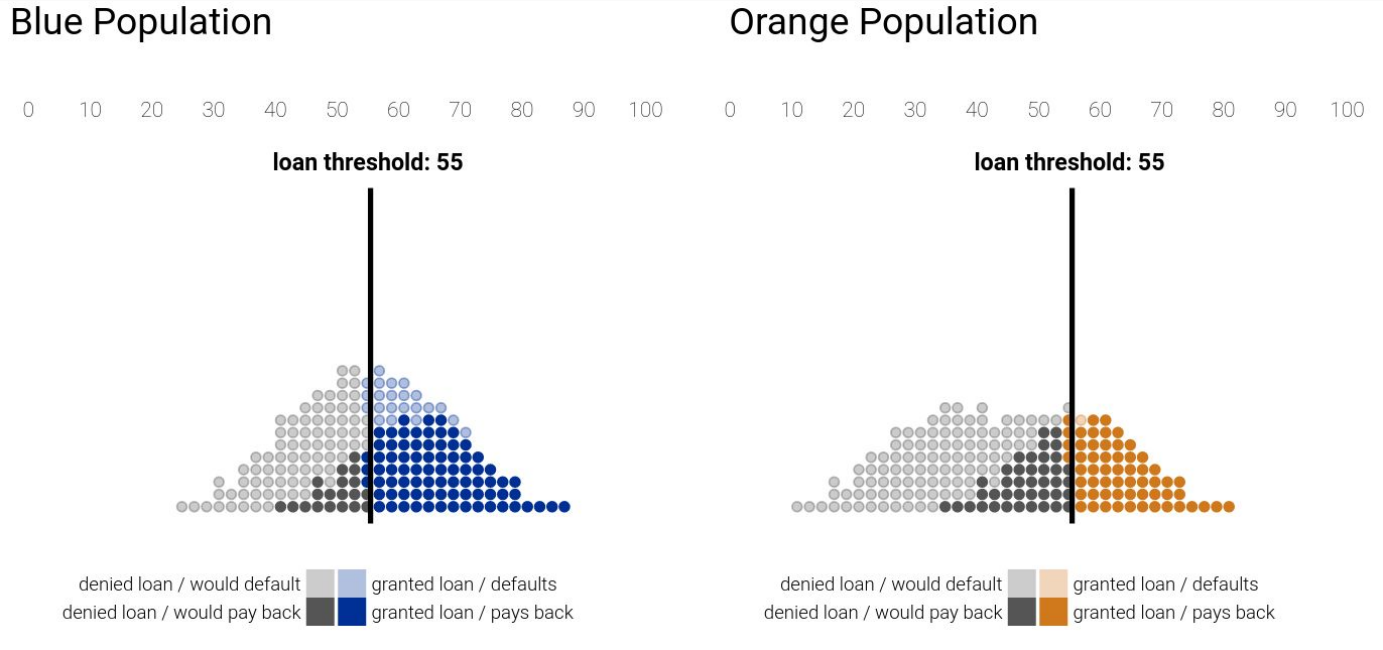
	Azul	Naranja
Accuracy	**0.78**	**0.79**
Positive Rate	0.58	0.70
True positive rate	0.71	0.60

Equal opportunity (=TPR)

	Azul	Naranja
Accuracy	0.78	0.83
Positive Rate	0.60	0.65
True positive rate	**0.68**	**0.68**

Ejemplo de modelo *machine learning*

Group Unaware



Ejemplo de modelo *machine learning*

Group Unaware

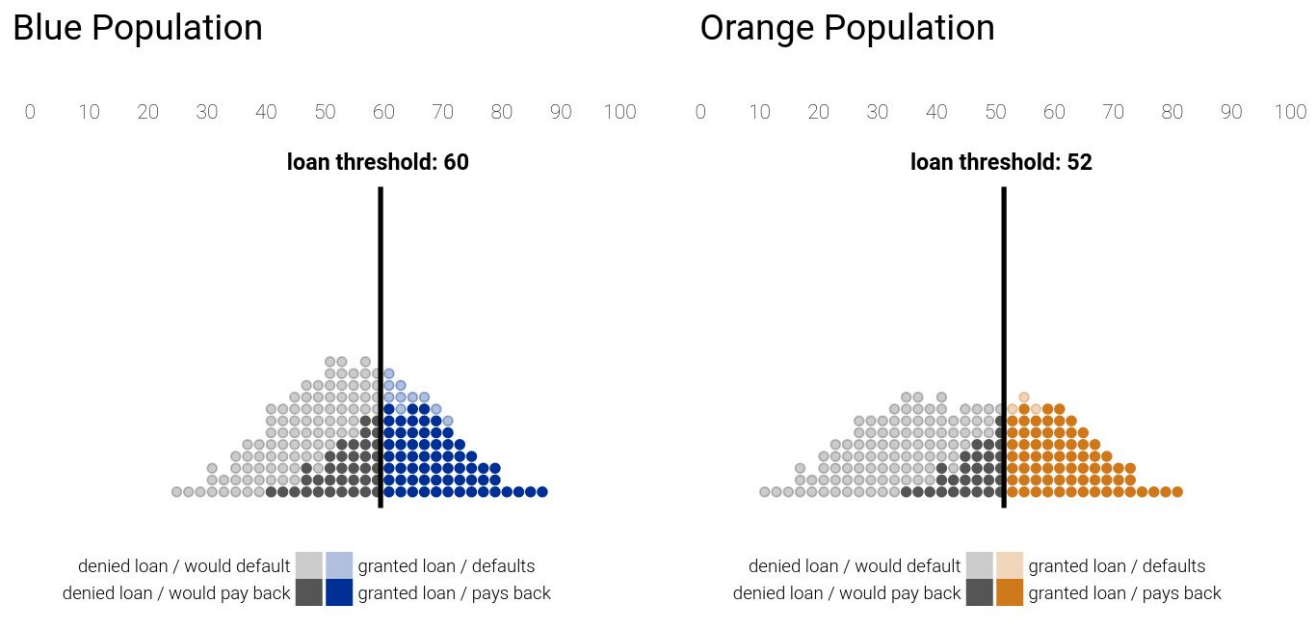
(th=55, azul)	Denied loan	Granted loan
Defaults	77	22
Pays back	19	80

(th=55, naranja)	Denied loan	Granted loan
Defaults	98	1
Pays back	40	59

	Azul	Naranja
Accuracy	0.79	0.79
Positive Rate	0.48	0.70
True positive rate	0.81	0.60

Ejemplo de modelo *machine learning*

Demographic parity (=PR)



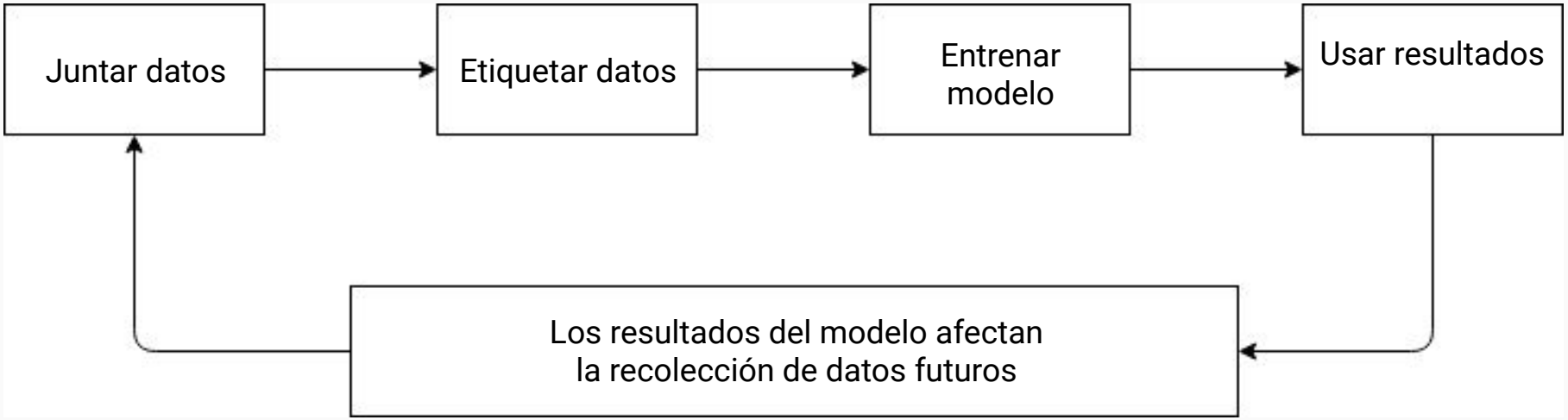
Ejemplo de modelo *machine learning*

Demographic parity (=PR)

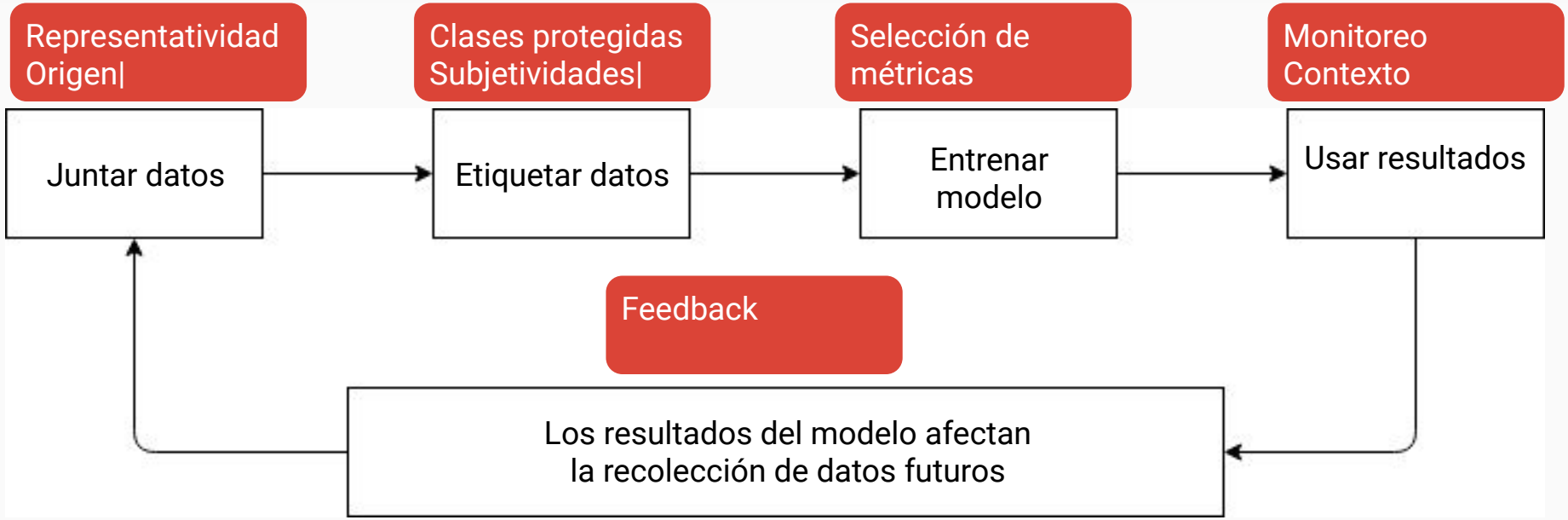
(th=60, azul)	Denied loan	Granted loan	(th=52, naranja)	Denied loan	Granted loan
Defaults	89	10	Defaults	96	3
Pays back	36	63	Pays back	29	70

	Azul	Naranja
Accuracy	0.76	0.83
Positive Rate	0.67	0.65
True positive rate	0.60	0.68

Flujo de *machine learning*



Flujo de machine learning - fairness



Conclusión

Hoy en día nadie nos obliga a considerar *fairness*.

Las empresas no tienen grandes incentivos para considerar *fairness*.

Las personas que trabajamos con *machine learning* deberíamos considerar *fairness*.

Existen herramientas que podemos usar para que los modelos sean más justos.

Referencias

Sysarmy - Encuesta IT https://twitter.com/sysarmy/status/1082676532403937280

Modelo de sueldos https://github.com/seppo0010/sysarmy-sueldos

Fairness in Machine Learning, Solon Barocas and Moritz Hardt https://vimeo.com/248490141

Ricci vs DeStefano https://supreme.justia.com/cases/federal/us/557/557/

Texas 10 percent rule
https://www.educationnext.org/texas-ten-percent-plans-impact-college-enrollment/
https://capitol.texas.gov/billlookup/text.aspx?LegSess=75R&Bill=HB588

Larry P. vs Riles https://law.justia.com/cases/federal/district-courts/FSupp/495/926/2007878/

Anja Lambrecht, Catherine Tucker (2019) Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. Management Science
https://pubsonline.informs.org/doi/10.1287/mnsc.2018.3093

Referencias

Categorías sospechosas y control de constitucionalidad, Guillermo F. Treacy
http://www.derecho.uba.ar/publicaciones/lye/revistas/89/treacy-guillermo-f-categorias-sospechosas-y-control-de-constitucionalidad.pdf

Caso Calvo y Pesini, Fallos 321:194
http://www.saij.gob.ar/corte-suprema-justicia-nacion-federal-ciudad-autonoma-buenos-aires-calvo-pesini-rocio-cordoba-provincia-amparo-fa98000022-1998-02-24/123456789-220-0008-9ots-eupmocsollaf?q=%20fecha-rango%3A%5B19980224%20TO%2019980224%5D&o=15&f=Total%7CTipo%20de%20Documento/Jurisprudencia/Fallo%7CFecha%7COrganismo%7CTribunal%7CPublicaci%F3n%7CTema%7CEstado%20de%20Vigencia%7CAutor%7CJurisdicci%F3n&t=105#

Convención sobre la eliminación de todas las formas de discriminación contra la mujer
https://www.ohchr.org/sp/professionalinterest/pages/cedaw.aspx

Constitución Nacional http://servicios.infoleg.gob.ar/infolegInternet/anexos/0-4999/804/norma.htm

Transcripción Debate Constituyente 1994
https://hcdcorrientes.gov.ar/DCN-1994/Indice%20-Debate-constituyente.htm

Referencias

Sisnero, Mirtha Graciela y otros c. Tadelva SRL y otros s/ amparo, Corte Suprema de Justicia de la Nación http://www.defensoria.org.ar/sisnero-mirtha-graciela-y-otros-c-tadelva-srl-y-otros-s-amparo/

Weapons of Math Destruction https://www.goodreads.com/book/show/28186015-weapons-of-math-destruction

Entrevista a Cathy O'Neil https://algorithmwatch.org/en/story/cathy-oneil-orcaa/

To predict and serve?, Kristian Lum and William Isaac https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1740-9713.2016.00960.x

Fairness in Machine Learning, NIPS 2017 Tutorial — Part I, Solon Barocas and Moritz Hardt https://vimeo.com/248490141 https://mrtz.org/nips17/#/

A Survey on Bias and Fairness in Machine Learning, Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan https://arxiv.org/abs/1908.09635

Sex Bias in Graduate Admissions: Data from Berkeley, P. J. Bickel, E. A. Hammel, J. W. O'Conne https://homepage.stat.uiowa.edu/~mbognar/1030/Bickel-Berkeley.pdf

Referencias

YouTube, the Great Radicalizer, Zeynep Tufekci
https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Fairness in Criminal Justice Risk Assessments: The State of the Art, Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth https://arxiv.org/abs/1703.09207

The Frontiers of Fairness in Machine Learning, Alexandra Chouldechova, Aaron Roth
https://arxiv.org/abs/1810.08810

Attacking discrimination with smarter machine learning
https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Repositorio donde calculé los números de la charla
https://github.com/seppo0010/fairness-talk/blob/master/Untitled.ipynb

Machine Learning Fairness: Lessons Learned (Google I/O'19)
https://www.youtube.com/watch?v=6CwzDoE8J4M

Referencias

fairness-indicators https://pypi.org/project/fairness-indicators/

What-If Tool https://pair-code.github.io/what-if-tool/

ML-fairness-gym https://ai.googleblog.com/2020/02/ml-fairness-gym-tool-for-exploring-long.html

GDPR https://gdpr-info.eu/

¿Preguntas?

¡Gracias!