17° JORNADAS DE

DATA MINING

4, 5 y 6 DE OCTUBRE

UNIVERSIDAD AUSTRAL | INGENIERÍA Posgrados

Equidad en modelos de aprendizaje automático para la predicción de desempleo

M.G. Beiró [1,2], *K. Kalimeri* [3]

1 Facultad de Ingeniería, Universidad de Buenos Aires, Buenos Aires, Argentina
2 INTECIN (UBA-CONICET), Buenos Aires, Argentina
3 ISI Foundation, Turín, Italia

1. CONTEXTO

* ¿Es legítimo utilizar **modelos de aprendizaje automático** para tomar decisiones que afectan a las personas? ¿Existen dilemas éticos al respecto?

* En caso afirmativo, ¿qué precauciones se deben tomar para evitar situaciones de discriminación? ¿Dependen estas precauciones de cada escenario?

* En este trabajo analizamos una situación en que un algoritmo debe seleccionar en forma automática a personas potencialmente desempleadas para asistirlas en la búsqueda de empleo.

2. INTRODUCCIÓN Y NOTACIÓN

* En un problema de **clasificación** buscamos predecir una variable Y que puede tomar un conjunto de valores finito $C = \{c_1, c_2, ..., c_M\}$ a partir de un conjunto de características (*features*) X.

* Si aplicamos un enfoque probabilístico, podemos asumir que X e Y son variables aleatorias con una **distribución conjunta** (X, Y).

* Entrenar un modelo de clasificación es hallar una función f que brinde una estimación $\hat{Y}$ a partir de X, intentando minimizar cierta **función de pérdida** $l(\hat{Y}, Y)$:

$$\hat{Y} = f(X)$$

¡$\hat{Y}$ es también una variable aleatoria!

2. INTRODUCCIÓN Y NOTACIÓN

- Existen distintas funciones de pérdida, como:
  - **Error de clasificación:** $l(\hat{Y}, Y) = \mathbb{P}\{\hat{Y} \neq Y\}$
  - **Error cuadrático:** $l(\hat{Y}, Y) = \mathbb{E}[(\hat{Y} - Y)^2]$
  - **Entropía cruzada:** $l(\hat{Y}, Y) = H(\hat{Y}, Y)$

- Cada función se adapta mejor a ciertos casos. Por ejemplo, en modelos de *deep learning* para clasificación multiclase es común utilizar la entropía cruzada.

- Cuando utilizamos el error de clasificación como función de pérdida, el clasificador óptimo es siempre el **máximo a posteriori** (MAP):

$$f^*(X) = \arg_{c \in C} \max \left(\mathbb{P}\{ Y = c \mid X \} \right)$$

2. INTRODUCCIÓN Y NOTACIÓN

* Para el caso particular en que sólo hay 2 clases, $C = \{0, 1\}$, dicho clasificador óptimo f^* se puede calcular como un *threshold* aplicado a la siguiente función:

$$r(X) = \mathbb{P}\{ Y = 1 \mid X \} = \mathbb{E}\{ Y \mid X \}$$

* Si $r(X) \geq 0.5$ predecimos $f^*(X)=1$, y en caso contrario predecimos $f^*(X)=0$.

* A dicha función $r(X)$ la llamamos **función de riesgo** *(risk score)*.

* Esto explica por qué es frecuente atacar un problema de clasificación como un problema de regresión en que se busca estimar $\mathbb{P}\{ Y = 1 \mid X \}$, seguido de un threshold.

3. FORMULACIÓN DEL PROBLEMA

* Consideremos el caso en que las **muestras** de nuestro problema de clasificación representan **personas**.
* Y supongamos que dentro de nuestro esquema existen ciertos atributos, A, que codifican la pertenencia de una persona a una **categoría protegida**.

> *Una categoría protegida es aquella que alguna vez ha servido de base para tratar a las personas en forma adversa sistemáticamente y sin justificación.*

* **Ejemplos** de categorías protegidas: *grupo etáreo, género, etnia, creencias religiosas, orientación sexual, lugar de residencia, lugar de nacimiento, discapacidad*.

3. FORMULACIÓN DEL PROBLEMA

* A los atributos *A* de nuestro problema que codifican la pertenencia a una categoría protegida los llamamos atributos sensibles *(sensitive attributes)*.

* ¿Cómo evitamos que nuestro modelo tome decisiones que discriminen a las personas en función de su pertenencia a una categoría protegida?

> **No fairness through unawareness**
> *Remover los atributos sensibles de un conjunto de datos no necesariamente evitará situaciones de discriminación en nuestro modelo.*

* **Ejemplo:** un conjunto de datos de compras en un supermercado que incluya las marcas o productos que una persona compra puede ser altamente indicativo de su género o su nivel socio-económico.

4. CASO DE ESTUDIO: PREDICCIÓN DE DESEMPLEO

- **Escenario:** Se quiere realizar una campaña a través de Facebook para invitar a usuarios desempleados a una feria laboral.

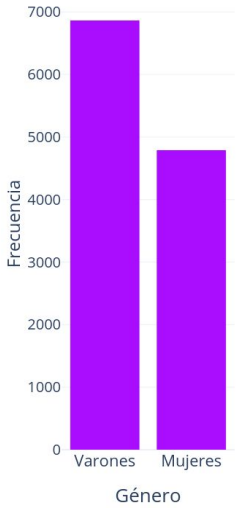
- **Conjunto de datos:** Likes en Facebook de un conjunto de 64.000 usuarios mayoritariamente jóvenes en Italia para los que conocemos:
  - Las páginas de Facebook a las cuales dieron Like.
  - Atributos demográficos: género, edad, región.
  - Condición laboral (sólo para 11.000 usuarios).

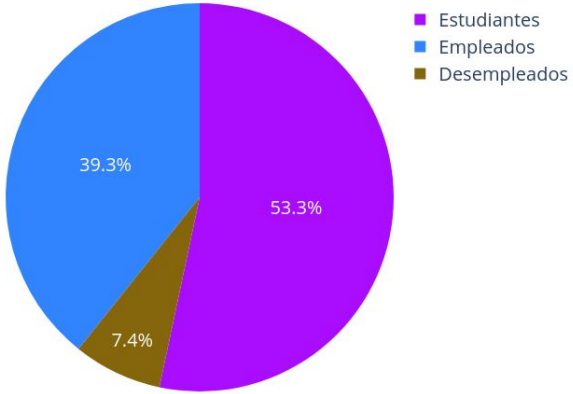
- **Problema:** Predecir la condición laboral de las personas (desempleado / no desempleado) a fin de contactar a usuarios desempleados. Por razones de presupuesto, se prefiere tener alta *precision* (que la mayoría de los seleccionados sean desempleados) aún a costa de baja *recall* (que pocos desempleados sean seleccionados).

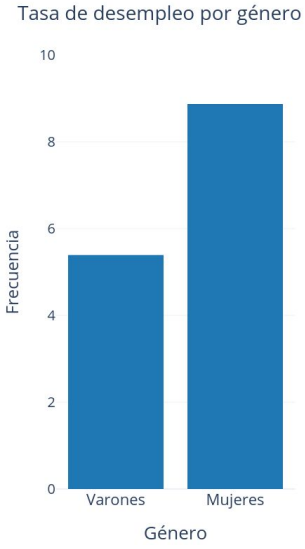
4. CASO DE ESTUDIO: PREDICCIÓN DE DESEMPLEO

* **Nos concentraremos en estudiar las disparidades de género en el modelo**

* **Estadísticas generales:**







4. CASO DE ESTUDIO: PREDICCIÓN DE DESEMPLEO

- **Modelo:** Clasificador LightGBM, 10-fold cross-validation + grid search*.
  - Para cada muestra se obtiene un *score* $r \in [0,1]$. Si $r > \vartheta$ (*threshold*), entonces $\hat{Y}=1$ (desempleado). De lo contrario, $\hat{Y}=0$ (no desempleado).

- Al observar los resultados de la predicción, encontramos que el modelo tiende a predecir más mujeres como desempleadas que varones desempleados.
  - Esto ocurre por desigualdades históricas que hacen que la tasa de desempleo en Italia sea mayor en mujeres que en varones.
  - En otros problemas podría deberse a sesgos en los datos (p.ej., que hubiera igual tasa de desempleo en cada género, pero que por algún motivo las mujeres desempleadas fueran más propensas a usar Facebook que los varones desempleados).

* *(learning rate, lambda, n_estimators, max_depth).*

5. NOCIONES DE EQUIDAD EN MODELOS DE CLASIFICACIÓN

- **Entonces, ¿bajo qué criterio decidimos si nuestro modelo es o no es justo?**

- Existen distintas nociones de **fairness (equidad)** en el aprendizaje automático:
  - Independencia (*paridad demográfica*)
  - Separación (*igualdad de oportunidades*)
  - Suficiencia (*calibración*)

- A continuación exploraremos cada una de estas tres.

5. NOCIONES DE EQUIDAD EN MODELOS DE CLASIFICACIÓN

INDEPENDENCIA

* Un modelo tiene independencia respecto a un atributo sensible A cuando la predicción $\hat{Y}$ es independiente del atributo sensible A.

* En términos de probabilidades:

$$\hat{Y} \perp A$$

* La independencia implica que la distribución condicional de $\hat{Y}$ no depende del valor de A (p. ej., del nivel socio-económico de la persona). Por eso se conoce también como **paridad demográfica**.

$$P\{\hat{Y} = c \mid A = a_1\} = P\{\hat{Y} = c \mid A = a_2\}$$

5. NOCIONES DE EQUIDAD EN MODELOS DE CLASIFICACIÓN

INDEPENDENCIA

* En otras palabras, todos los grupos tienen la misma tasa de aceptación (acceptance rate). La probabilidad de ser seleccionado no depende de la clase.

* Este criterio no siempre es razonable. Ejemplo: si intentamos predecir la probabilidad de que una persona tenga una enfermedad a partir de una imágen médica, y determinado grupo de la población tiene mayores chances de padecer la enfermedad, no sería sensato pretender que todos los grupos tengan igual proporción de enfermos predichos.

  ➢ Nos llevaría a considerar enfermas a personas de otros grupos que están sanas, o a no detectar la enfermedad en personas del grupo vulnerable, a fin de igualar las tasas de aceptación.

5. NOCIONES DE EQUIDAD EN MODELOS DE CLASIFICACIÓN

SEPARACIÓN

* Un modelo ofrece separación cuando, condicionado a personas en una clase Y en particular, el valor predicho es independiente del atributo sensible.

$$\hat{Y} \perp A \mid Y$$

* Es fácil probar que esta definición implica que la tasa de error en cada grupo debe ser la misma. Es decir, dentro de cada grupo debe haber igual FPR (*false positive rate*), FNR, TPR y TNR.

* También se conoce como criterio de **igualdad de oportunidades** (*equality of opportunity*).

5. NOCIONES DE EQUIDAD EN MODELOS DE CLASIFICACIÓN

SUFICIENCIA

* Un modelo satisface la suficiencia cuando el valor predicho tiene toda la información sobre el atributo sensible que puede afectar a la clase de la persona.

* Es decir, condicionada al valor predicho, la verdadera clase es independiente del atributo sensible:

$$Y \perp A \mid \hat{Y}$$

* Un clasificador no restringido intentará por defecto capturar esta noción. En particular, el clasificador óptimo $f^* = \mathbb{E}\{Y \mid X\}$ satisface la suficiencia.

5. NOCIONES DE EQUIDAD EN MODELOS DE CLASIFICACIÓN

Resultados de imposibilidad y criterio escogido

- Existen resultados teóricos que muestran que es imposible satisfacer varias nociones de equidad simultáneamente.

- El criterio que utilizaremos en nuestro caso de estudio es el de *igualdad de oportunidades* (separación): queremos que la tasa de error con que mujeres desempleadas son predichas como empleadas (falso negativo) sea la misma que aquella con que varones desempleados son predichos como empleados, y viceversa.

- Utilizaremos como métrica de equidad el cociente $FNR_{mujeres}/FNR_{varones}$.

6. METODOLOGÍA

Al entrenar el modelo sin restricciones encontramos los siguientes resultados:

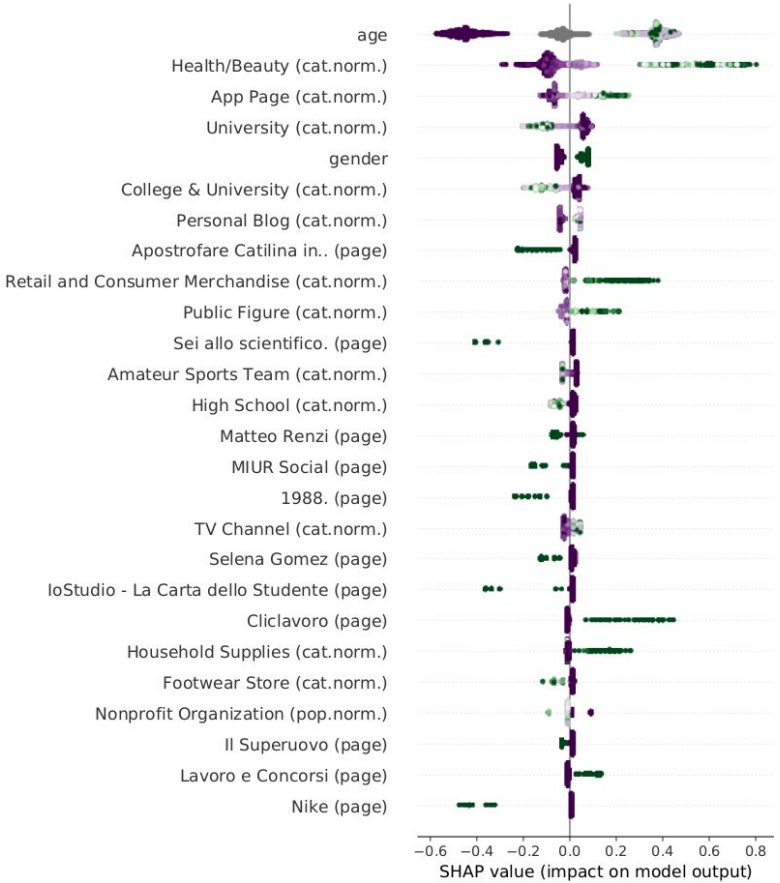
Métrica	Valor
AUC (Area Under the ROC Curve)	0.74
Recall	0.56
Precision	0.16
Fairness (FNR_{muj}/FNR_{var})	**0.47**

Precision y recall dependen del threshold del modelo. Estos resultados corresponden a un threshold de 0.5.

6. METODOLOGÍA

¿Cuáles fueron los principales predictores de desempleo?

Respondemos esta pregunta utilizando la herramienta SHAP para observar el impacto del valor de cada *feature* en la salida del modelo →



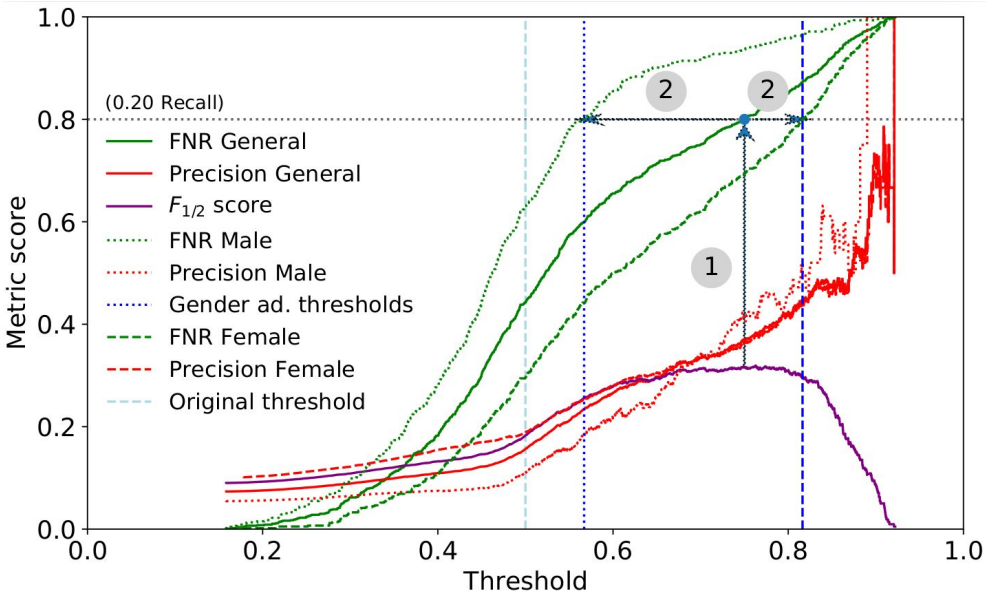
6. METODOLOGÍA

Para equiparar la tasa de error de distintos grupos, aplicaremos un threshold adaptable al clasificador en función del grupo de pertenencia de la persona.

Como queremos privilegiar la precision sobre el recall, maximizamos el $F_{1/2}$ score (que otorga más peso a la *precision*) para encontrar el threshold global óptimo.

A partir de este threshold global determinamos el threshold a aplicar a cada género.

$$F_{1/2} = \frac{5 \cdot precision \cdot recall}{precision + 4 \cdot recall}$$



7. RESULTADOS

Con estos thresholds adaptados encontramos estos siguientes resultados:

Métrica	Valor
AUC (Area Under the ROC Curve)	0.74
Recall	~~0.56~~ 0.22
Precision	~~0.16~~ 0.25
Fairness (FNR_{muj}/FNR_{var})	**~~0.47~~ 1.00**

Detectamos al 22% de todos los desempleados.

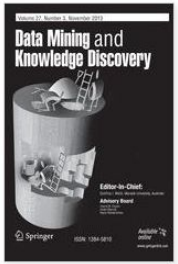
La cuarta parte de los seleccionados son desempleados.

7. RESULTADOS

Resumen de resultados hallados:

- Eliminar los atributos vulnerables no logró evitar la discriminación por género.
- El uso de thresholds adaptados permitió cumplir con la equidad según el criterio de igualdad de oportunidades para cada género.
- En nuestro **artículo** mostramos que el uso de thresholds por género también mejoró por sí solo la equidad frente a otros dos atributos vulnerables: la edad y la región de pertenencia.

Fairness in vulnerable attribute prediction on social media
Mariano G. Beiró & Kyriaki Kalimeri
Data Mining and Knowledge Discovery (2022)



8. CONCLUSIONES

* La aplicación de tecnologías de aprendizaje automático está transformando distintos aspectos de nuestra sociedad y plantea discusiones y dilemas éticos.

* Su utilización en forma sistemática para la toma de decisiones podría acabar con situaciones de discriminación hacia determinados grupos sociales. Sin embargo, también requiere tomar precauciones para que los algoritmos no terminen reforzando discriminaciones históricas.

* La métrica de equidad a utilizar en cada caso debe ser escogida en forma conveniente. Los resultados de imposibilidad muestran que no es posible en general satisfacer varios criterios de equidad en forma simultánea.

* Es posible diseñar estrategias para cumplir con requisitos de equidad en escenarios reales como el caso de estudio analizado.

PREGUNTAS


17° JORNADAS DE
DATA MINING
4, 5 y 6 DE OCTUBRE
UNIVERSIDAD AUSTRAL | INGENIERÍA Posgrados