Introducción (problema de investigación y el objetivo)

En el contexto del mercado laboral estadounidense, cada éxito económico revela una desigualdad racial significativa, lo que ha suscitado un debate sobre si los empleadores practican la discriminación racial (Bertrand y Mullainathan, 2004). Numerosos estudios han documentado la persistencia de la discriminación racial en los procesos de contratación. En este sentido, Bertrand y Mullainathan (2004) llevaron a cabo un influyente experimento de campo en el que enviaron currículos ficticios a ofertas de empleo reales, modificando únicamente los nombres de los candidatos para que se percibieran como pertenecientes a determinados grupos raciales. Sus conclusiones revelaron que los candidatos cuyos nombres se asociaban con personas blancas tenían muchas más posibilidades de ser contactados para una entrevista que aquellos cuyos nombres se percibían como afroamericanos, incluso si sus perfiles eran equivalentes. Este trabajo tiene como objetivo reproducir y analizar empíricamente este fenómeno utilizando esta base de datos simulada construida. Mediante modelos de regresión logística, buscamos estudiar el efecto de las diferentes variables del CV en la probabilidad de recibir una llamada para entrevista, con el fin de identificar posibles sesgos discriminatorios en el proceso de selección de candidatos.

Descripción de los datos

La base de datos utilizada en este análisis proviene de la experiencia sobre el terreno diseñada por Bertrand y Mullainathan (2004) y contiene información sobre 4870 currículos ficticios enviados a ofertas de empleo reales en las ciudades de Boston y Chicago (Estados Unidos). Las variables incluidas comprenden información sobre la experiencia profesional, el nivel de estudios, los premios recibidos, la presencia de un correo electrónico, la experiencia militar, la ciudad del empleo, así como el origen étnico y el sexo deducidos del nombre. La variable de interés es “recibir llamada”, que indica si el currículum ha dado lugar a una llamada para una entrevista (1 = sí, 0 = no).

Variables numéricas

En promedio, los candidatos tienen 7.84 años de experiencia laboral con una desviación estándar de 5.04 años, lo que sugiere una gran variabilidad en los años de experiencia laboral (cuadro 1). En cuanto a la educación, en promedio, 3.62 años de estudios universitarios con una desviación estándar de 0.72 años (Cuadro 1). Entonces, no hay una grande variabilidad

entre los años de estudios universitarios, lo que indica que los candidatos tienen niveles educativos similares.

Cuadro 1.- Estadísticas de los candidatos según los años de educación universitaria y los años de experiencia

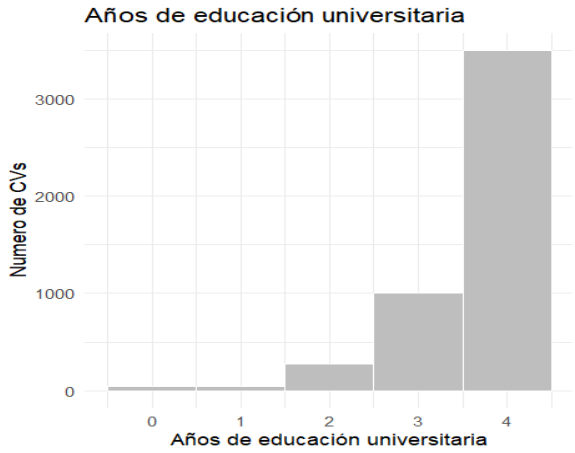
Variables	Promedio	Desviación estándar
Experiencia Laboral	7.84	5.04
Años_universitaria	3.62	0.72

Fuente: Elaboración propia con datos ficticios de Bertrand y Mullainathan (2004).

Gráfica 1.- Reparticiones de los años de experiencia laboral.

Fuente: Elaboración propia con datos ficticios

Gráfica 2.- Reparticiones de los anos universitarios



Fuente: Elaboración propia con datos ficticios

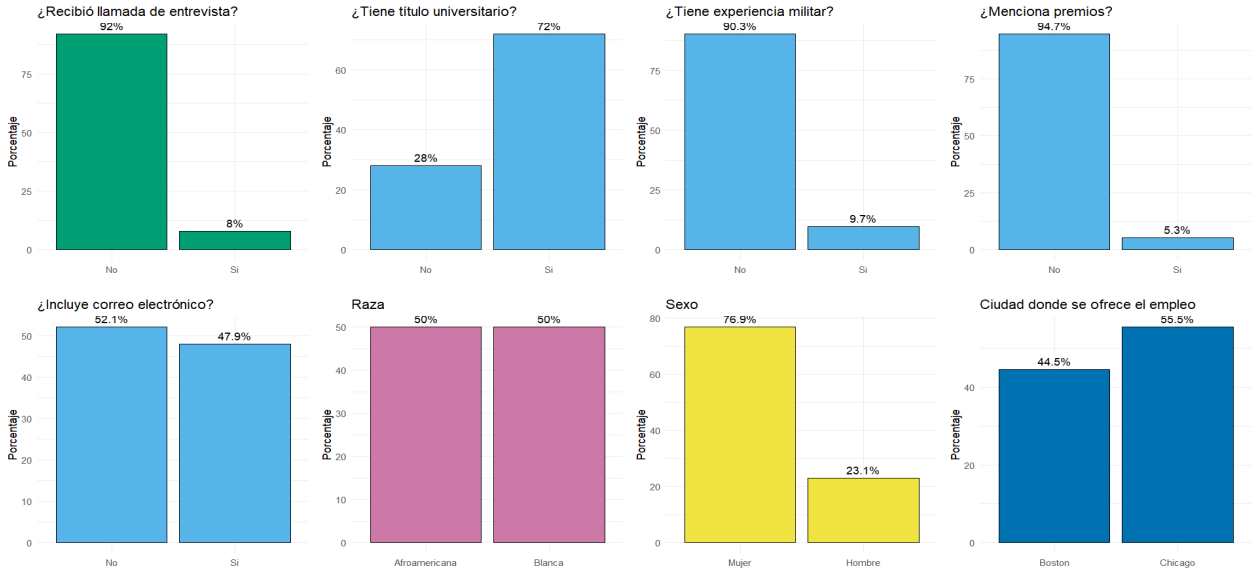
En la gráfica 2, la mayoría de los CVs (4510 CVs que representan 92.6%) reportan que los candidatos pasaron 3 y 4 años de educación universitaria y casi de 1% de los candidatos (46 CVs, 0.94%) no tiene año de estudio universitario (0 años). Y de otro lado, los años de experiencia laboral presentan una mayor dispersión, con candidatos que van desde poca hasta mucha experiencia. En la mayoría de los CVs, los candidatos tienen entre 5 y 10 años de experiencia laboral y tenemos un dato atípico donde un candidato ya tiene 44 años de experiencia en el mercado laboral (gráfica 1).

Variables categóricas

En la gráfica 3, se ve que, hay el mismo peso de afroamericanos que personas de raza blanca, de hecho, 50% de los candidatos son Afroamericanos y 50% son de raza blanca. Más de tres cuartos de los candidatos son mujeres (76.9% de los candidatos son femeninos y 23.1% son masculinos). En cuanto a la ciudad donde se ofrece el empleo, 44.48% fueron en Boston y 55.52% en Chicago.

Al reportar experiencia militar, por cada 10 candidatos, casi 9 no tiene ninguna experiencia (90.3%). Casi la mitad (48.0%) de los candidatos tienen un correo electrónico. Y, 72 % de los candidatos reportaron que tienen un título universitario. Por cada 10 candidatos, más de 9 (94.7%) no recibieron premios o reconocimientos según su CV. Y, por último, 92% de los candidatos no recibieron llamada a entrevista.

Gráfica 3.- Reparticiones de las variables categóricas.



Fuente: Elaboración propia con datos ficticios de Bertrand y Mullainathan (2004).

Metodología

Elegimos 10 variables de esta base de datos, 8 de ellas categóricas y 2 numéricas de la base de datos para crear un modelo de regresión logística binaria adecuada para estudiar nuestra variable dependiente categórica: si el candidato recibió o no una llamada para una entrevista (recibir _llamada). En la primera etapa, realizamos una descripción general de los datos para

explorar nuestra base de datos; en la segunda etapa, se implementó una estrategia de selección de variables en dos fases: primero, se ajustó un modelo completo con todas las variables; luego, se descartaron progresivamente aquellas que no eran significativas. También se evaluaron modelos con interacciones y un modelo con selección automática "backward" usando el criterio de información de Akaike (AIC); en la tercera etapa, utilizaremos el modelo de regresión logística y comprobaremos si es un buen clasificador con una división aleatoria de los datos en 80% para los datos de entrenamiento y 20% para los de prueba. Con el modelo entrenado, se estimaron probabilidades y se clasificaron las observaciones del conjunto de prueba usando un umbral de 0.5. Finalmente, se construyó una matriz de confusión para evaluar el desempeño del clasificador. Para llevar a cabo este trabajo, realizaremos los análisis estadísticos utilizando el lenguaje informático R (versión 4.5.0), utilizando las librerías "tidyverse" , "graphics", "readxl", "class", "MASS", "ggeffects", "caret", y "gridExtra".

Resultados

a) Estimación de un modelo logístico.

Considerando al principio todas las variables de la base, nos salió en R en modelo que tiene todas sus variables significativas. Se estimó también un modelo con interacciones entre las variables significativas y ninguna de las interacciones es significativa y el valor del AIC fue más grande (2687.4). Entonces, no vale la pena incluir las interacciones en el modelo. En cuento al modelo ajustado mediante selección automática usando el criterio de información de Akaike, este modelo presenta un menor AIC (2644.9) que el modelo 2 (AIC = 2677.6). Sin embargo, también presenta una estructura más compleja y menos intuitiva para interpretar individualmente los efectos. El modelo 2 es más parsimonioso, incluye solo efectos principales, y todas las variables que lo componen son significativas. Por ello, se optó por conservar el modelo 2 y con todos los predictores significativos.

Entonces, nuestro modelo optimo es lo siguiente:

$$\log\left(\frac{P(X)}{1-P(X)}\right) = -2.77 - 0.35 * trabajo\_ciudadChicago + 0.03 * experiencia\_laboral + 0.79 * premiosSi + 0.44 * razaBlanca$$

Cuadro 2.- Resultados de la regresión logística (el modelo optimo)

Coefficients:	estimate	std.error	statistic	p.value
(Intercept)	-2.77	0.134395	-20.6307	<2.22E-16
trabajo_ciudadChicago	-0.350	0.108876	-3.21834	0.001289
experiencia_laboral	0.0264	0.009585	2.755845	0.005854
premiosSi	0.793	0.182863	4.339293	1.43E-05
razaBlanca	0.440	0.107878	4.07753	4.55E-05
AIC= 2677.6				

Fuente: Elaboración propia con datos ficticios

Según el cuadro 2, sin considerar las variables regresoras, en promedio, los momios logarítmicos de ser llamado a entrevista son de -2.77. Considerando constantes las otras variables regresoras, en promedio, ser en Chicago tiene un efecto negativo en los momios logarítmicos de ser llamado a entrevista de 0.350; por cada año que aumenta el número de años de experiencia laboral de un candidato, en promedio, los momios logarítmicos aumentan de 0.0264; al reportar en el CV que recibió premios o reconocimientos aumentan de 0.793 en los momios logarítmicos, en promedio; y por último, ser de raza blanca aumenta de 0.44 en los momios logarítmicos, en promedio.

b) Predicciones con el modelo considerado

Según modelo final elegido, la probabilidad de una persona blanca, en Chicago, sin anos de experiencia laboral y sin premios reportados de ser llamada a entrevista es de 6.40%. Para una persona afroamericana que tiene las mismas otras características, la probabilidad de ser llamada a entrevista es de 4.22%.

Considerando una persona blanca, en Boston, con 10 años de experiencia laboral y con premios reportados de ser llamada a entrevista es de 21.84% y para una persona afroamericana con las mismas características, la probabilidad es de 15.25%.

c) Características de que arroja la mayor probabilidad y la menor

Con este modelo, las características del que arroja la mayor probabilidad es una persona blanca sea en Boston, que tiene 26 años de experiencia, y que reporta que recibió premios. La probabilidad estimada de recibir una llamada es de 29.89%. Y, las características del que arroja la menor probabilidad es una persona afroamericana que sea en Chicago, que tiene 1 año de experiencia y que reporta que no recibió premios con una probabilidad de 4.32% de ser llamado.

d) Clasificación

Con una regresión logística

Con un clasificador de una regresión logística (cuadro 3), el clasificador predicó correctamente la posibilidad de que un candidato reciba o no una llamada por entrevista en 91.98% del tiempo. Y se equivoque en 8.02% de los casos. Pero, este clasificador no es un buen clasificador porque clasifica solamente a los negativos, entonces la sensibilidad es de 0.

Cuadro 3.- Matriz de confusión con el modelo logístico como clasificador.

		Recibir_llamada
Predicciones	**No**	**Sí**
No	895	78
Sí	0	0
	Precisión = 0.9198	

Fuente: Elaboración propia con datos ficticios.

Discusión

Parte 1: Comparación entre los modelos

En la comparación de modelos se consideraron dos criterios principales: la significancia estadística de los predictores y el valor del criterio de información de Akaike (AIC). El modelo 2, ajustado en el punto 3, incluye únicamente efectos principales (sin interacciones) y todas las variables que lo componen resultaron estadísticamente significativas (trabajo_ciudad, experiencia_laboral, premios y raza). Su valor de AIC fue de 2677.6.

Por otro lado, el modelo AIC, obtenido mediante selección automática por criterio de Akaike, incluye más variables e interacciones. Si bien este modelo presentó un valor más bajo de AIC (2644.9), no todos sus predictores principales fueron estadísticamente significativos, lo que complica la interpretación. Además, muchas de sus interacciones tampoco mostraron significancia.

Desde una perspectiva de predicción, el modelo AIC podría tener un desempeño levemente mejor debido a su menor AIC. Sin embargo, desde la perspectiva de la inferencia estadística, el modelo 2 es preferible: es más parsimonioso, todas las variables son interpretables y

significativas, y facilita extraer conclusiones claras sobre los efectos de cada variable en la probabilidad de recibir una llamada.

Por último, se detectó un problema de desbalance en los datos: solo una fracción menor de los candidatos recibió una llamada (la mayoría tienen valor 0). Este desbalance puede afectar el desempeño del clasificador, en especial en términos de sensibilidad y precisión.

Parte 2: Resultados e interpretación de los modelos

El modelo 2, elegido por su simplicidad y me permite comprender con mayor claridad los patrones y asociaciones entre las variables, permite identificar las variables que importan son los que aumentan mas en los momios logarítmicos del modelo logístico. Entonces, podemos hablar de recibir premios, la raza y la ciudad donde se anunció el empleo.

Al utilizar CV ficticios asociados a nombres que sugerían el sexo y la etnia de los candidatos, los resultados de la regresión de nuestro modelo final mostraron claramente que, en función de su raza, algunas personas tenían más probabilidades de recibir una llamada para una entrevista que otras. Al calcular las probabilidades considerando las mismas características para una persona blanca y una afroamericana, observamos que la probabilidad de que una persona blanca reciba una llamada para una entrevista es mayor (6.40%) que la de una afroamericana (4.22%).

Bibliografía

Bertrand, M., & Mullainathan, S. (2003). *Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination* (NBER Working Paper No. 9873). National Bureau of Economic Research. https://doi.org/10.3386/w9873