Problem Set N°1 - Machine Learning
UNLP

Cantero, Lara Sofia
Temossi, Francisco
Vollmer, Candela

Repositorio de GitHub

1)

CONTEXTO

En el sector público, un reporte preciso de los ingresos de los individuos es indispensable para el cálculo de los impuestos. Sin embargo, el fraude fiscal siempre ha sido un problema importante. De acuerdo con el IRS, un 83,6 % de los impuestos son pagados voluntariamente y a tiempo en Estados Unidos. Una de las causas de esa brecha es la subdeclaración de ingresos por parte de los individuos.

Un modelo de predicción de ingresos podría ayudar potencialmente a detectar casos de fraude que podrían conducir a la reducción de la brecha. Además, un modelo de predicción de ingresos puede ayudar a identificar a personas y familias vulnerables que puedan necesitar más asistencia.

El objetivo principal del trabajo es construir un modelo de salario horario individual:

$$w = f(X) + u$$

donde X es una matriz que incluye potenciales variables explicativas/predictoras.

INTRODUCCIÓN

La obtención de información precisa sobre los ingresos es un desafío, especialmente en economías en desarrollo como las de América Latina. Los impuestos sobre la renta constituyen una fuente importante de recursos para los países; sin embargo, es común que algunos individuos subdeclaren sus ingresos para evadir parte de sus obligaciones fiscales. Aunque existen agencias especializadas en detectar estos casos de evasión, los procesos de auditoría suelen ser costosos. En este contexto, han comenzado a implementarse técnicas de machine learning, las cuales son eficaces para predecir un resultado sin requerir una comprensión detallada de la relación causal entre la variable dependiente, como ingresos o beneficios declarados, y las variables utilizadas para la predicción, como características socioeconómicas. En términos generales, los modelos predictivos pueden contribuir a una recaudación más eficiente.

En cuanto al uso de machine learning para la predicción de ingresos, la evidencia aún es incipiente. No obstante, estudios recientes, como los de Jo (2024), Matkowski (2021) y Wang (2022), han mostrado ciertos avances en este ámbito.

Para el caso colombiano, De Roux et al. (2018) utilizaron datos de declaraciones de impuestos de la ciudad de Bogotá y aplicaron una técnica de aprendizaje no supervisado para identificar subdeclaraciones en los impuestos de construcción. Por su parte, Battiston et al. (2024) emplearon métodos de clasificación supervisada para predecir el subreporte de ingresos por parte de trabajadores independientes en Italia. Battaglini et al.(2024) utilizaron datos administrativos para identificar qué casos de evasión permiten maximizar los ingresos de la agencia tributaria tras una auditoría.

Sobre la adopción de estas técnicas por los organismos tributarios, la OCDE (2021) estima que aproximadamente el 75 % de las administraciones fiscales ya implementan machine learning con el fin de mejorar la eficiencia y optimizar la asignación de recursos.

Los datos utilizados se obtuvieron del reporte "Medición de Pobreza Monetaria y Desigualdad" que extrae información de las bases de microdatos de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 que elabora el Departamento Administrativo Nacional de Estadística (DANE). La GEIH es una integración de las tres más importantes Encuestas a Hogares de Colombia: la Encuesta Continua de Hogares (ECH), la Encuesta Nacional de Ingresos y Gastos (ENIG) y la Encuesta de Calidad de Vida (ECV). Se propuso en 2005 con el propósito de ampliar el alcance temático de la investigación y reducir el costo de la aplicación.

Una de las principales ventajas de esta fuente de datos es la amplitud de los aspectos relevados, que abarcan una variedad de temáticas demográficas y socioeconómicas. Esto la convierte en una herramienta particularmente valiosa para analizar cuestiones relacionadas con la estructura del mercado laboral y la distribución del ingreso, lo que la hace especialmente adecuada para la construcción de un modelo de predicción de salarios.

AGREGAR preview of the results and main takeaways

DATOS

La Gran Encuesta Integrada de Hogares es una encuesta que recolecta información sobre las condiciones de empleo de las personas en Colombia (si trabajan, en qué trabajan, cuánto ganan, si tienen seguridad social en salud o si están buscando empleo), además de las características generales de la población como sexo, edad, estado civil y nivel educativo. La GEIH proporciona información a nivel nacional, cabecera - resto, regional, departamental, y para cada una de las capitales de los departamentos. La recolección de la GEIH empezó el 7 de agosto de 2006 en su módulo central de mercado laboral e ingresos y, a partir del 11 de septiembre, con su módulo de gastos de los hogares.

Las características más importantes de esta encuesta son:

- Unidad de observación: Hogares.
- Muestra (total anual): 240.000 hogares, aproximadamente.
- Periodicidad: trimestral y anual.
- Temas: algunos son ingresos, propiedad, ahorro/inversión, empleo y desempleo, situación económica, economía rural, vivienda, fertilidad y uso de servicios públicos.
- Cobertura: cobertura nacional que permite obtener resultados por zona urbana y rural, cinco grandes regiones (Región Atlántica, Región Oriental, Región Central, Región Pacífica y Región Bogotá) y total por departamentos: 23 departamentos, las 13 grandes ciudades con sus áreas metropolitanas y 11 ciudades intermedias.

En el presente trabajo se seleccionaron 9 variables que se creyeron pertinentes para poder formular un modelo de predicción de salarios. Las variables numéricas consideradas son la edad, la cantidad de horas trabajadas[1] y la antigüedad laboral. También se han incluido variables categóricas como el sexo, la condición respecto al sistema de pensiones, el máximo nivel de educación, el tipo de ocupación y dos variables relativas al tamaño de la firma en la cual trabaja el individuo. Por un lado, la variable MicroEmpresa toma valor 1 si la empresa tiene 5 empleados o menos o el trabajador es cuentapropista y 0 en caso contrario. Por el otro, consideramos la variable sizeFirm que contempla más desagregaciones respecto al tamaño de la empresa, considerando 5 tamaños posibles.

Los datos fueron obtenidos mediante la técnica de web scraping de una página web que contiene información detallada sobre 32,177 observaciones correspondientes a la Gran Encuesta Integrada de Hogares (GEIH) del año 2018. Como el acceso a los microdatos anonimizados de uso público es de carácter gratuito y está disponible en la página Web del DANE, no se nos presentó ninguna restricción para su acceso.

La base de datos final se obtuvo filtrando a los individuos que cumplen las siguientes condiciones: tener ingresos horarios superiores a cero, estar empleados y ser mayores de 18 años. Además, solo se han conservado las observaciones que no presentan valores faltantes en las variables seleccionadas para la predicción del ingreso.

Tabla 1: Estadísticas descriptivas de la submuestra seleccionada

	n	Media	S.E.	Mínimo	25 %	50 %	75 %	Máximo
Ingreso horario	14764	8541.9	13866.1	0.5	3796.5	4837.5	7899.31	350583.3
Edad	14764	38.9	13.2	18.0	28.0	37.0	49.0	91.0
Sexo	14764	0.5	0.5	0.0	0.0	1.0	1.0	1.0
Horas trabajadas	14764	47.6	15.2	1.0	40.0	48.0	50.0	130.0
MicroEmpresa	14764	0.4	0.5	0.0	0.0	0.0	1.0	1.0
Antigüedad	14764	62.0	88.0	0.0	7.0	24.0	72.0	720.0
Formal	14477	0.6	0.5	0.0	0.0	1.0	1.0	1.0

Nota: elaboración propia en base a GEIH 2018 (DANE)

En la tabla 1 se pueden observar las estadísticas descriptivas de algunas variables seleccionadas. La submuestra seleccionada, compuesta por personas empleadas mayores de 18 años con salario horario positivo, incluye 14,764 observaciones, representando aproximadamente el 46 % de la muestra original. La media y la mediana del ingreso horario confirman la esperada asimetría positiva (o hacia la derecha) de la distribución. La edad promedio de la población analizada es de 39 años, y el 53 % de los individuos son hombres. En cuanto a las horas trabajadas, el promedio semanal asciende a 47 horas, lo cual sugiere una jornada laboral promedio de 8 horas diarias durante 6 días a la semana, superando el estándar de 40 horas semanales.

La variable "microempresa" permite observar que el 43 % de los trabajadores están empleados en establecimientos con cinco trabajadores o menos, incluyendo a los cuentapropistas. En relación con la antigüedad, esta presenta una gran dispersión, con un promedio

[1]Se considera el total de las horas trabajadas la semana previa a la encuesta.

de cinco años. Cabe destacar que el 25 % de los encuestados tiene una permanencia de siete meses o menos en su empresa actual.

Por último, se ha creado una variable "formal" que toma valor 1 si el individuo cotiza pensión y 0 en caso contrario (excluyendo a los pensionados), lo cual permite aproximar el nivel de informalidad de la economía. Al considerar la submuestra completa, la informalidad alcanza aproximadamente el 40 %, pero disminuye al 23 % al excluir a cuentapropistas y empleadores.

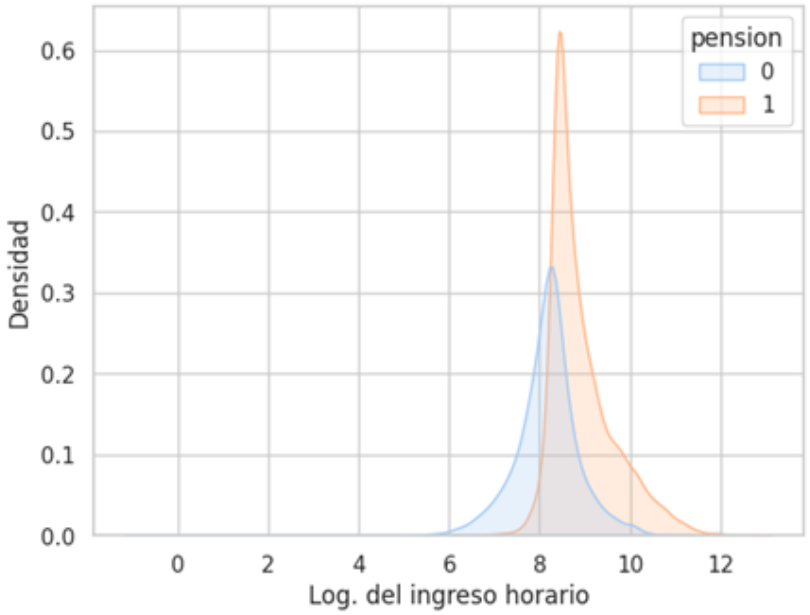
La Figura 1 ilustra la relación entre ingreso horario y edad, desagregada por sexo y condición laboral (formal, informal o pensionado).

Figura 1: Ingreso por sexo, edad y condición de pensión

Nota: elaboración propia en base a GEIH 2018 (DANE)

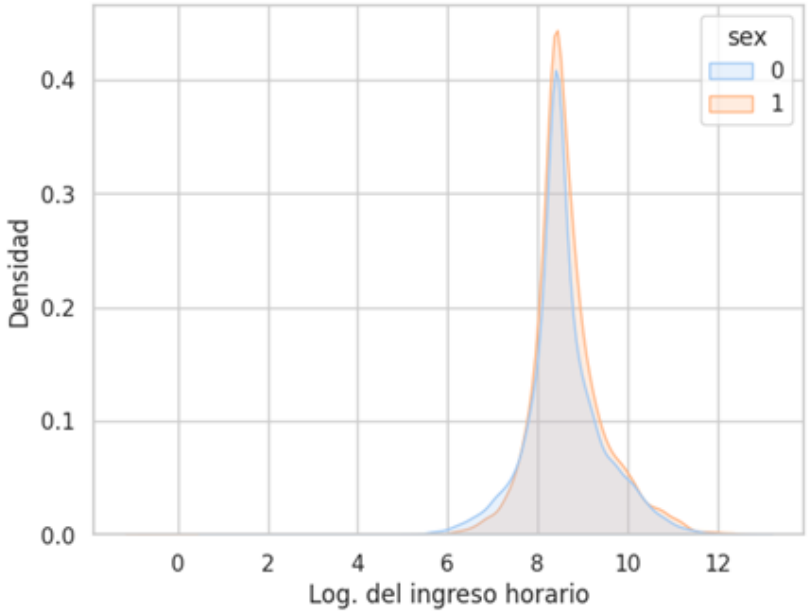
Las Figuras 2 y 3 muestran la densidad del logaritmo del ingreso horario, desglosada por condición laboral (formal o informal) y por sexo. En la primera se observa una diferencia notable entre trabajadores formales e informales. Es plausible que los trabajadores formales se vean influenciados por una medida de salario mínimo, lo que contribuye a que sus ingresos sean, en promedio, más altos que los de los trabajadores informales. En cuanto a las diferencias de género, aunque ambos grupos presentan distribuciones similares, los hombres alcanzan un pico más alto en la densidad de ingresos y acumulan menor densidad en los niveles de ingreso más bajos.

Figura 2: Distribución del ingreso horario (en log.) de formales e informales



Nota: elaboración propia en base a GEIH 2018 (DANE)

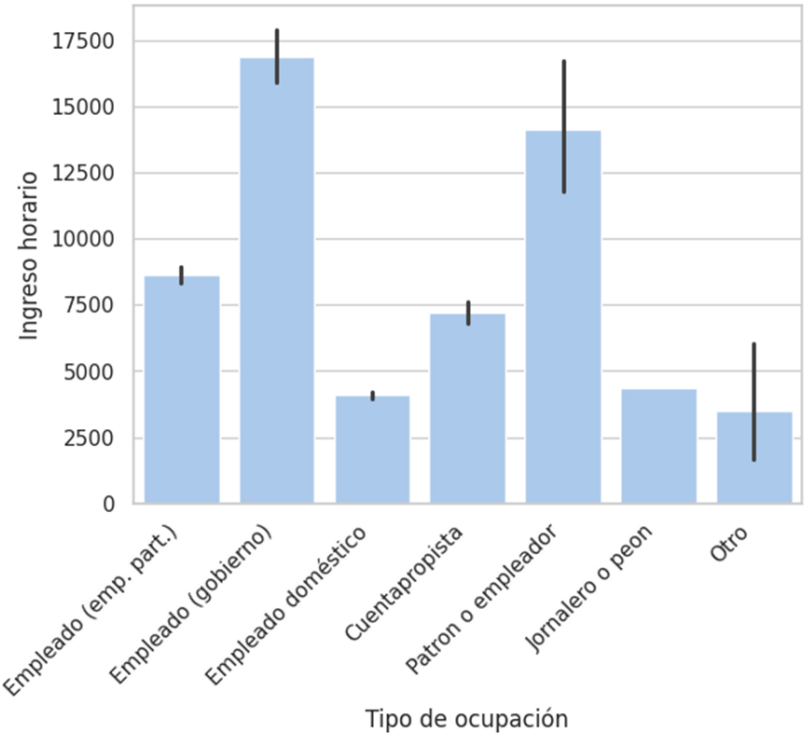
Figura 3: Distribución del ingreso horario (en log.) de hombres y mujeres



Nota: elaboración propia en base a GEIH 2018 (DANE)

La Figura 4 muestra el ingreso promedio reportado por individuos según el tipo de ocupación, junto con sus respectivos intervalos de confianza. Esto sugiere que el tipo de ocupación puede ser un predictor relevante del nivel de ingresos.

Figura 4: Ingreso horario por tipo de ocupación



Nota: elaboración propia en base a GEIH 2018 (DANE)

PREDICCIÓN DE SALARIOS

b)

Modelo	**RMSE**
Regresión Lineal Simple	11571.6712
Regresión Lineal con Interacciones	11369.8616
Regresión Ridge	11571.6226
Regresión Lasso	11571.6520
Árbol de Decisión (Sin Regularización)	16584.2103
Árbol de Decisión (Con Regularización)	11363.7877
Random Forest	11689.5497
Gradient Boosting	10983.7115
SVR	12944.7715
K-Vecinos Más Cercanos	11806.9751

Tabla 2: Resultados de RMSE para diferentes modelos

c)

i. Sobre el rendimiento general de los modelos:

Al comparar los RMSE de los modelos, queda claro que hay diferencias importantes en su capacidad predictiva. Los modelos más básicos, como la regresión lineal simple y la regresión lineal con interacciones, tienen RMSE de 11571.6712 y 11369.8616, respectivamente. O sea, estos modelos capturan las tendencias generales en los datos, pero no son lo suficientemente sofisticados para atrapar relaciones más complejas.

Con los modelos con regularización, como Ridge y Lasso, los RMSE son prácticamente iguales al de la regresión simple (11571.6226 y 11571.6520, respectivamente). Esto significa que, en este caso, la regularización no está ayudando mucho, tal vez porque no hay un problema fuerte de sobreajuste en el dataset.

El árbol de decisión sin regularización tiene el peor RMSE (16584.2103), lo que deja claro que sobreajusta los datos. Pero, al meterle regularización, el árbol mejora bastante y queda con un RMSE de 11363.7877, similar a los modelos lineales.

Por otro lado, los modelos más complejos, como Gradient Boosting, Random Forest y K-Vecinos, son los que mejor se desempeñan. El campeón es Gradient Boosting con el RMSE más bajo (10983.7115), seguido de los vecinos más cercanos (11806.9751) y Random Forest (11689.5497). Esto tiene sentido porque estos modelos suelen ser mejores para detectar patrones complejos.

El SVR no tuvo un buen rendimiento (12944.7715). Probablemente se deba a que es un modelo más delicado y depende mucho de los hiperparámetros y la estructura de los datos.

ii. Sobre la especificación con el menor error de predicción:

El modelo de **gradient boosting** sobresale con el menor RMSE (10983.7115). Este modelo combina múltiples árboles de decisión ajustados de forma secuencial, enfocándose en corregir errores de los árboles previos, lo que lo hace particularmente efectivo para capturar patrones complejos y no lineales.

Este rendimiento es particularmente destacable considerando que no requiere una gran cantidad de parámetros ajustables comparado con modelos como los árboles de decisión o random forest, y parece estar equilibrando correctamente el ajuste a los datos sin sobreajustar. Sin embargo, es importante considerar el costo computacional asociado en comparación con modelos lineales más simples, que tienen un RMSE competitivo en este caso.

iii. Exploración de las observaciones con el mayor error de predicción:

Para explorar las observaciones con mayor error de predicción, calculamos los errores de predicción en la muestra de prueba y analizamos su distribución. El error de predicción para cada observación es la diferencia entre el valor observado y el valor predicho. A continuación, se muestra el análisis de la distribución de estos errores.

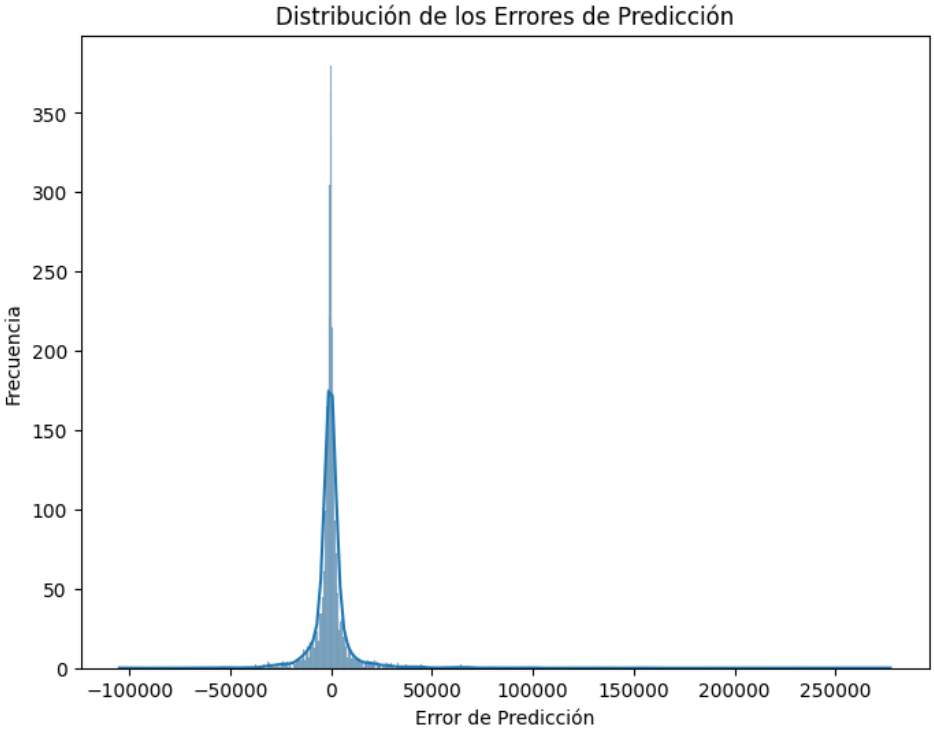


Figura 5: Distribución de los errores de predicción

Al analizar la distribución de los errores de predicción, encontramos que tiene una forma muy similar a una distribución normal. Esto sugiere que, en su mayoría, el modelo tiene un rendimiento razonablemente bueno, con pocos errores grandes. Sin embargo, observamos que existen algunas observaciones en las colas de la distribución, especialmente en la cola izquierda, que se alejan significativamente de la media. Estas observaciones indican que el modelo presenta algunas fallas en predicciones muy bajas, lo cual es un indicio de que hay casos atípicos que el modelo no captura correctamente.

Las observaciones que se encuentran en las colas de la distribución (especialmente aquellas que presentan errores negativos significativos) son consideradas valores atípicos. Estas observaciones pueden tener varias causas posibles:

- **Personas con características únicas que el modelo no está capturando bien,** tales como ingresos inusualmente bajos, comportamientos laborales atípicos o situaciones excepcionales. En este caso, podrían ser individuos con características que la DANE debería examinar más de cerca, ya que los ingresos reportados podrían ser inusuales o indicar un posible fraude fiscal.
- **El modelo no está bien ajustado para estos casos atípicos,** lo que podría sugerir que el modelo no está capturando adecuadamente ciertos factores relevantes. En este caso, el modelo podría estar sobreajustando o no considerando correctamente algunas de las características importantes de estas observaciones, por lo que podría necesitar ajustes adicionales.

*d)
Al usar validación cruzada *leave-one-out* (LOOCV), se ve que los errores bajaron bastante:

- **Gradient Boosting**: De 10983.7115 a 4847.63
- **Árbol de decisión (con regularización)**: De 11689.54 a 4906.71

Esto pasa porque con LOOCV el modelo usa casi todos los datos para entrenar y deja fuera solo un punto en cada iteración. Esto suele mejorar la generalización, especialmente si el conjunto de datos no es tan grande.

A pesar de la mejora, esto no siempre significa que el modelo será mejor fuera de la muestra. LOOCV puede ser sensible a ciertas características del conjunto de datos y, si el modelo tiene un poco de sobreajuste, puede no funcionar tan bien en datos nuevos.

Referencias

- Battaglini, M; Guiso, L.; Lacava, C.; Miller, D. L.; Patacchini, E. (2024). Refining public policies with machine learning: The case of tax auditing. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2024.105847.
- Battiston, P.; Gamba, S.; Santoro, A. (2024). Machine learning and the optimization of prediction-based policies. *Technological Forecasting and Social Change* 199. https://doi.org/10.1016/j.techfo
- de Roux, D.; Perez, B.; Moreno, A.; Villamil, M.; Figueroa, C. (2018). Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (KDD '18). Association for Computing Machinery, New York, NY, USA, 215–222. https://doi.org/10.1145/3219819.3219878.
- Jo, K. (2024). Income Prediction Using Machine Learning Techniques. UCLA. ProQuest ID: Jo_ucla_0031N_22838. Merritt ID: ark:/13030/m54c47kf. https://escholarship.org/uc/item/6d
- Matkowski, M. (2021). Prediction of Individual Level Income: A Machine Learning Approach [Honors Thesis, Bryant University]. Archivo digital. https://digitalcommons.bryant.edu/hon
- OCDE (2021), Tax Administration 2021: Comparative Information on OECD and other Advanced and Emerging Economies, OECD Publishing, Paris, https://doi.org/10.1787/cef472b9-en.
- Wang, J. (2022). Research on Income Forecasting based on Machine Learning Methods and the Importance of Features. ICIDC – EAI. https://eudl.eu/doi/10.4108/eai.17-6-2022.2322745.