Trabaja práctico 1: Predicción de Ingresos

Mendoza Franco[1]

I. Introducción.

Las prácticas fraudulentas relacionadas con la subdeclaración de ingresos es un comportamiento habitual en muchas economías del mundo. En Estados Unidos, se estima que alrededor del 85,8% de los impuestos se pagaron voluntariamente entre 2011 y 2013 y el porcentaje restante se explican en parte por la falta de declaración de ingresos (IRS, 2019). En este sentido, la capacidad de prever los ingresos desempeña un papel crucial en la toma de decisiones. En el ámbito público, comprender los ingresos de la sociedad facilita el cálculo de tasas impositivas y optimiza la implementación de políticas redistributivas. Por otro lado, en el sector privado, la segmentación de mercados según los ingresos se erige como una estrategia fundamental en muchas industrias, mientras que en el sector financiero, la predicción de ingresos posibilita una evaluación más precisa de posibles riesgos crediticios (Matkowski, M., 2021).

La econometría ha proporcionado modelos de estimación de ingresos con resultados tanto comprensibles como robustos. Sin embargo, debido a la presencia de supuestos rigurosos, su capacidad predictiva con datos "fuera de la muestra" a menudo es limitada. Bajo esta limitación de la econometría clásica y con el aumento en la disponibilidad de datos, las prácticas de aprendizaje automático (Machine Learning, ML) han irrumpido en el campo de la predicción y con resultados satisfactorios para diversas áreas de la economía.

En el ámbito de la predicción de ingresos mediante técnicas de aprendizaje automático, se destaca el trabajo de Matkowski, M. (2021), quien utilizó datos del Current Population Survey 2017-2020 para prever los ingresos de los estadounidenses mediante la estimación de siete métodos de aprendizaje automático. El autor llegó a la conclusión de que los modelos de aprendizaje automático superan el rendimiento de las predicciones tradicionales en la estimación del ingreso individual total. De manera análoga, Wang, J. (2022) aplicó técnicas de aprendizaje automático para predecir los ingresos anuales basándose en diversos atributos, concluyendo que ciertos algoritmos de aprendizaje automático pueden utilizarse con un alto grado de precisión para prever el nivel de ingresos de un individuo. Por otro lado, Gomez-Cravioto, D. et al. (2022) encontraron que los modelos de aprendizaje automático superaron a los modelos paramétricos de regresión lineal y logística en la predicción del ingreso actual de los exalumnos.

[1] Mail: mendozaantoniofranco@gmail.com
Link a GitHub: https://github.com/FranMendozaAntonio/Problem-Set-1-Predicting-Income

Este estudio sigue la línea de las investigaciones previamente mencionadas. Utilizando datos censales de ingresos de 9,785 individuos provenientes de la Gran Encuesta Integrada de Hogares (GEIH) del año 2018, llevada a cabo por el Departamento Administrativo Nacional de Estadística (DANE), el objetivo principal es prever el nivel de ingresos por hora de un individuo basándose en diversos atributos personales (edad, sexo, categoría ocupacional, horas trabajadas y nivel educativo máximo alcanzado).

Se crearon diez modelos con diferentes niveles de complejidad, y su capacidad predictiva se evaluó mediante la partición de la muestra. Se empleó una submuestra no utilizada en el entrenamiento del modelo, conocida como muestra de prueba, para evaluar el rendimiento predictivo. De esta manera, se identificaron como los modelos de menor error predictivo aquellos que incorporaron controles para todos los atributos mencionados anteriormente y sus combinaciones polinómicas de grado 2 y 4, respectivamente. En tanto, los de mayor error predictivo fueron los de grado 5 y 6 dado que a mayor complejidad, menor es el poder de predicción por mayor variabilidad.

Finalmente, se aplicó la técnica de Leave-One-Out Cross-Validation (LOOCV) a los dos modelos con mejor predicción. Esta estrategia aborda algunas limitaciones del mecanismo anterior, donde la elección aleatoria de la partición de la muestra podía presentar desafíos. Una vez más, los resultados indicaron que el modelo con menor error predictivo es aquel que incorpora combinaciones polinómicas de grado 2.

II. Datos

Los datos empleados en este estudio se originan en la Gran Encuesta Integrada de Hogares (GEIH), un relevamiento llevado a cabo por el Departamento Administrativo Nacional de Estadística (DANE) de Colombia. Esta encuesta recopila información detallada sobre las condiciones laborales de los individuos en el país, abordando aspectos como la ocupación, el ingreso, la afiliación a seguridad social en salud, y la búsqueda de empleo. Además, indaga sobre características generales de la población, tales como género, edad, estado civil, nivel educativo, y fuentes de ingresos. La GEIH ofrece información a nivel nacional, urbano-rural, regional, departamental, e individual para cada capital departamental. Los datos abarcan alrededor de 32,187 observaciones correspondientes a 12 períodos mensuales del año 2018.

Los datos están disponibles al público a través de la página oficial del DANE y están fragmentados para los diferentes períodos mensuales. En este estudio, se accedió a esta información mediante 10 bases de datos que contienen variables originales y otras reconstruidas, disponibles en la siguiente página web: https://ignaciomsarmiento.github.io/GEIH2018_sample/. Se emplearon técnicas de web scraping, sin encontrar ningún inconveniente al construir la base de datos.

Para la construcción de los modelos predictivos, se seleccionó una muestra compuesta por individuos mayores de 18 años cuyos ingresos laborales por hora sean superiores a cero. Asimismo, se decidió quitar aquellas observaciones que no tenían computado ningún nivel de ingreso. De esta manera, la muestra se redujo a un total de 9,785 personas. Cabe destacar, que los ingresos informados son en términos reales. No obstante, se considera que el efecto de los precios puede no tener un gran impacto en el análisis, dado que los datos abarcan un período de 12 meses consecutivos. A continuación se muestran las estadísticas descriptivas básicas para la variable a predecir.

Tabla 1. Estadística básica de la variable de ingresos horario laborales en pesos colombianos

	Observaciones	Media	Mínimo	25%	50%	75%	Máximo
Ingreso horario laboral	9785	$7984.26	$151.9	$3797.7	$4520.8	$7291.6	$291666.6

Fuente. Elaboración propia con base a DANE (2018)

Se muestran también las estadísticas descriptivas de otras variables que se consideran importantes a la hora de predecir el salario horario tomando a la literatura sobre determinantes de los ingresos. Entre ellas la edad, la educación, el grado de formalidad del puesto de trabajo, la cantidad de horas trabajadas usualmente por el individuo y el sexo.[2]

Tabla 2. Estadística básica de variables de interés para la predicción

Variables	Valor
Promedio de edad	36
Cantidad de hombres en la muestra (%)	50.1
Cantidad de ocupados formales (%)	77.3
Promedio de horas trabajadas semanalmente	48
Dummy de educación alcanzada	
Sin educación (%)	0.45
Primario (%)	10.3
Secundario (%)	9.4
Media (%)	34.3
Universitario (%)	45.4
Sin datos (%)	0

Nota. Haber alcanzado la educación primaria implica alrededor de 5 años de estudio, la secundaria 9 años, la media 13 años y la universitaria más de 14.
Fuente. Elaboración propia con base a DANE (2018)

[2] La elección de las variables fue siguiendo un estricto relevamiento de la literatura y comparando con los datos que la DANE ofrecía. Cabe destacar, que en muchos estudios con técnica de ML la elección "arbitraria" de los predictores, y los posibles problemas asociados a ella, se solucionan con herramientas estadísticas sofisticadas que en este trabajo no abordaremos.

Al emplear técnicas econométricas tradicionales, el enfoque para predecir ingresos implica la minimización de una función de pérdida esperada. En un contexto de regresión, una función de pérdida comúnmente utilizada es la pérdida cuadrática $L(d) = d^2$, y bajo esta función, la pérdida esperada se traduce en el error cuadrático medio (MSE, por sus siglas en inglés). En este contexto, el objetivo es estimar una función que relacione X con Y y que tenga una pérdida esperada baja en la predicción. Esta función corresponde a la media condicional E[y|X=x]. Para aproximar dicha media, podemos estimar el siguiente modelo de regresión (1) para el caso de los salarios, utilizando las covariables descritas en la tabla 2.

$$f(x) = \beta_0 + \beta_1 age + \beta_2 age2 + \beta_3 sex + \beta_4 formal + \beta_5 hourworkusual + \beta_6 educ_{primary} + \beta_7 educ_{secondary} + \beta_8 educ_{media} + \beta_7 educ_{univ} + \beta_7 educ_{nodata} + u$$

Donde:

β_0 es el intercepto del modelo
age es una variable numérica que indica los años del individuo
$age2$ es el cuadrado de la edad del individuo
sex es una variable dummy que vale 1 si el individuo es hombre
$formal$ es una variable dummy que vale 1 si el individuo trabaja en el sector formal
$hourworkusual$ es una variable numérica que indica las horas usuales de trabajo semanal del individuo
$educ$ es una variable que indica el nivel educativo máximo alcanzado por el individuo. Para este caso se utilizó como variable omitida aquella que indica que el individuo no tuvo educación.
u es el término de error del modelo que suponemos se comporta de manera normal

Haciendo la regresión lineal a través de mínimos cuadrados ordinarios (OLS) nos quedan los siguientes resultados de la tabla 3

Tabla 3. Resultados del modelo de regresión 1

Variable	Coeficiente
Intercept	7.017***
	(0.102)
age	0.050***
	(0.003)
age2	-0.000***
	(0.000)
sex	0.169***
	(0.012)
formal	0.366***
	(0.014)
hourworkusual	-0.011***
	(0.000)
educ_primary	0.2093**
	(0.086)
educ_secundary	0.258***
	(0.086)
educ_media	0.387***
	(0.085)
educ_univ	1.002***
	(0.085)
educ_nodata	0.02
	(0.568)
Observations	9785
R-squared	0.395

Standard error in parentheses
***p<0.001, **p<0.05, *p<0.1

Nota. Para las variables de educación, la omitida fue "sin educación".

Todas los coeficientes arrojan los signos esperados y en su gran mayoría con resultados estadísticamente significativos, con la excepción de la variable que sugiere la no existencia de datos para los niveles de educación. Solo por mencionar algunos resultados de inferencia obtenidos, cuando aumenta en un año la educación, el salario horario aumenta en promedio un 5% manteniendo constante las demás variables[3]. En tanto, las variables de educación mostraron signos positivos y creciente, esto quiere decir que, por ejemplo, las personas con educación universitaria tienen un aumento del 171.8% en sus salarios en comparación con aquellos sin educación, manteniendo el resto de las variables constantes[4].

En un contexto predictivo, lo que se hace es reemplazar los coeficientes estimados en el modelo (1) y de esa manera se obtiene un modelo predictivo del ingreso. No obstante, este tipo de modelos, si bien suelen ser insesgados con un fuerte poder predictivo dentro de la muestra, podría no presentar un buen comportamiento a la hora de reducir el error de predicción con datos

[3] Se trata de un modelo log-level.
[4] Se realizó la transformación necesaria del coeficiente para interpretar los resultados del modelo log-level

por fuera de la muestra. Al buscar reducir el sesgo complejizando el modelo, se incrementa la varianza. Esta es una de las limitaciones de la econometría tradicional. En ML, al interesarnos la predicción se comienza asumir la idea de resignar insesgadez para ganar en proder predictivo. En la siguiente sección abordaremos este punto.

III. Predicción de ingresos

Se evaluaron diez modelos diferentes en este estudio, abarcando diversos niveles de complejidad y linealidad. La muestra se dividió en un 70% para llevar a cabo el entrenamiento del modelo y un 30% restante para realizar las pruebas de predicción y calcular los errores cuadráticos medios (MSE, por sus siglas en inglés). El propósito fue comparar el MSE de los diferentes modelos para evaluar su rendimiento en la predicción. A continuación, se detallan cada uno de los modelos.

Modelo 1. Es el de menor complejidad en donde solo se tiene en cuenta un modelo lineal con intercepto sin covariables. En este caso, la predicción para el salario logarítmico horarios es el promedio muestral promedio de la muestra de entrenamiento.

Modelo 2. Al modelo 1 se le agrega dos covariable que son la edad y la edad al cuadrado de los individuos para capturar la idea de que a mayor edad, mayor es el ingreso del individuo hasta que atraviesa cierto umbral de vejez.

Modelo 3. Al modelo dos se le suma una variable que captura la cantidad usuales de horas trabajadas por semana por el trabajador.

Modelo 4. Se adiciona al anterior modelo una dummy que identifica el sexo del individuo. Esta incorporación se sustenta en el hecho de que existen brechas de género en el mercado laboral que son necesarias controlar para hacer las predicciones de ingreso.

Modelo 5. Dada la existencia de mercados informales en Colombia, rasgo que se materializa en toda la región latinoamericana, es necesario incorporar una dummy que controle por la formalidad de los individuos ocupados. En este sentido, al modelo cuatro se le incorporó esta variable.

Modelo 6. En este modelo se controla por todas las variables mencionadas anteriores más distintas dummy por educación. Se trata del modelo (1) expuesto en la sección anterior.

Modelo 7. Este modelo se encuentra en el umbral de los más complejos de los 10 modelos dado que muestra las distintas combinaciones de polinomios de las características con grado menor o igual a 2. Esto quiere decir, por ejemplo, que si tengo

un modelo inicial con dos covariables $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ el Modelo 7 sería $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + u$. En nuestro caso particular, vamos a complejizar el Modelo 6.

Modelo 8, 9 y 10. El Modelo 8 es igual al modelo 7 pero en vez de un grado menor o igual a 2 es un grado menor o igual a 4. En tanto el Modelo 9 sigue la lógica del 7 y 8 pero con una grado menor o igual a 5 y el Modelo 10 con un grado menor o igual a 6.

A continuación, en la tabla 4 se presentan los rendimientos predictivos de cada uno de estos modelos.

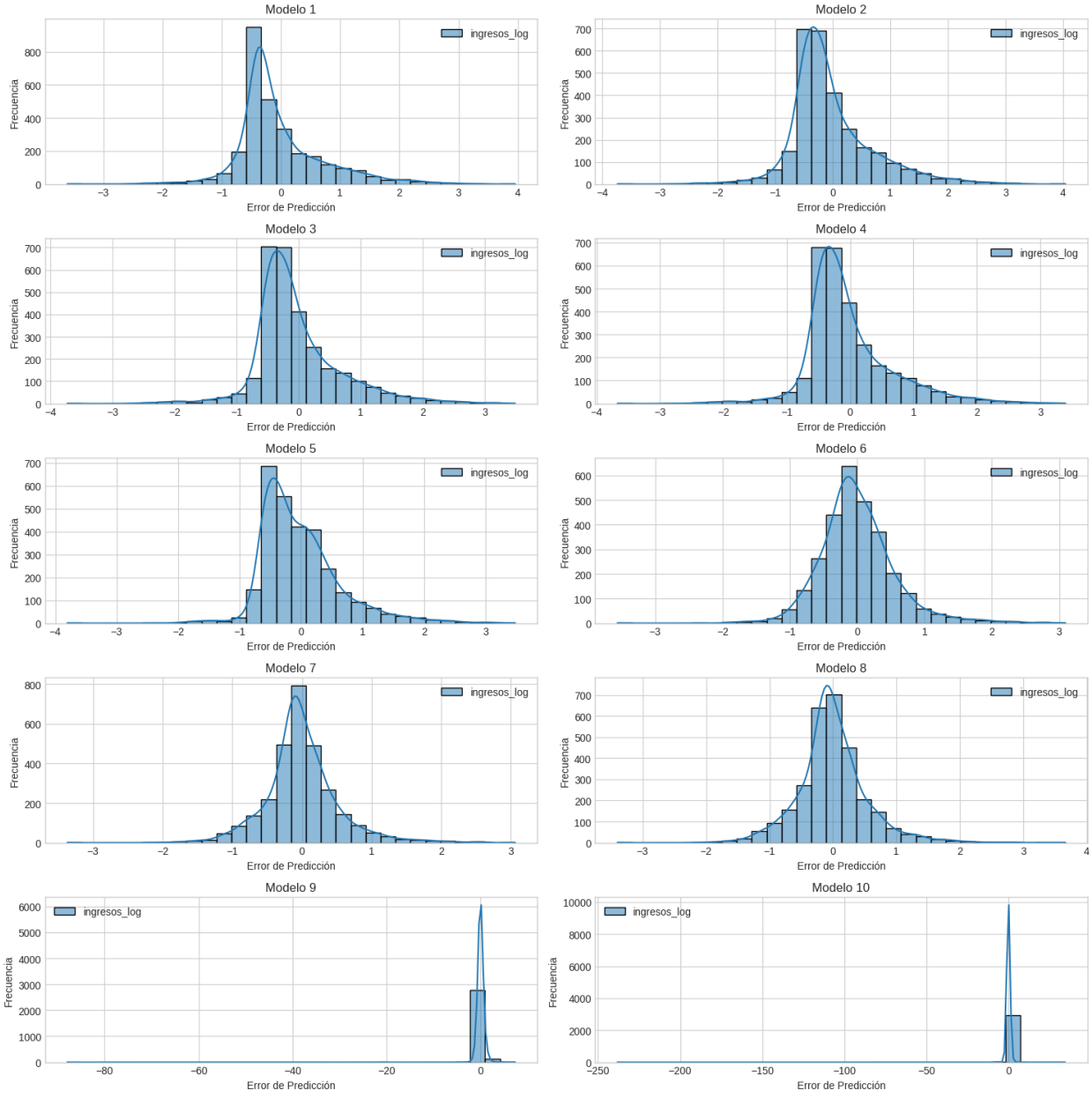
Tabla 4. Poder predictivo fuera de muestra de los modelos

Modelo	Covariables	MSE
Modelo 1	-	0.5244
Modelo 2	2	0.507
Modelo 3	3	0.4784
Modelo 4	4	0.4737
Modelo 5	5	0.4111
Modelo 6	10	0.3106
Modelo 7	**66**	**0.2767**
Modelo 8	1001	0.3095
Modelo 9	3003	3.969
Modelo 10	8008	20.34

Lo primero que se observa es que efectivamente a medida que se complejiza el modelo, el error predictivo fuera de muestra se va reduciendo hasta cierto punto en donde vuelve a incrementarse. En este caso, a partir del Modelo 8 con 1000 covariables comienza a aumentar el MSE con un fuerte salto en el Modelo 9 y 10. Asimismo, modelos de muy baja complejidad como el 1 y 2 también presentan un nivel alto de MSE. Tomando los 10 modelos estimados, el de menor error predictivo es el modelo 7, seguido por el modelo 8.

Graficar la distribución de los errores predictivos es una manera ágil para intuir el poder predictivo de los modelos. En términos visuales lo ideal sería contar con una campana lo más cercana a cero posible. Esto indicaría que para la mayoría de las observaciones, el valor que se predijo por el modelo fue muy cercano al verdadero valor. No obstante también necesitamos que exista cierta asimetría de la distribución lo que insinúa cierta "estabilidad" del modelo a la hora de predecir. En otras palabras, una baja varianza. En este sentido, la ilustración 1 muestra las distintas distribución de los errores predictivos.

Ilustración 1. Distribución de los errores predictivos para los distintos modelos analizados



Un aspecto a resaltar, y que va en línea con la problemática de la subdeclaración de ingresos mencionada en la introducción, es que la distribución de los errores predictivos del modelo 7 presenta una leve cola más larga hacia la izquierda. Esto podría significar que el valor predicho para el logaritmo del ingreso horario fue menor al observado. Notar que para estos modelos, no nos interesa realizar interpretaciones causales, sino más bien hacer predicciones. Por eso se presentan los cálculos del MSE.

A. Leave-One-Out Cross-Validation (LOOCV)

Si bien la partición de la muestra es una técnica útil y necesaria para medir el error predictivo de un modelo por fuera de muestra, pueden suscitar dos problemas derivados en parte de la existencia de cierta arbitrariedad. Esto tiene que ver sobre todo en la decisión sobre qué datos se utilizarán para entrenar el modelo y cuáles para probarlo. La división de la muestra puede cambiar fuertemente el poder predictivo para un mismo modelo.

Bajo este problema surge la técnica de Leave-One-Out Cross-Validation (LOOCV). Lo que hace es usar 1 observación para hacer el test y n-1 para hacer el entrenamiento. Este ejercicio lo voy a repetir para cada una de las observaciones hasta la última. De esta manera se partió la muestra "n" veces y todas las observaciones pasaron a ser train y test. Esto quita el problema de la arbitrariedad a la hora de particionar. En este sentido, si tomamos los dos modelos con menor error predictivo con la técnica de partición vemos que tomando LOOCV obtenemos que el modelo 7 sigue siendo el que cuenta con menor error predictivo tal como se muestra en la tabla 5

Tabla 5. Poder predictivo fuera de muestra de los modelos

Modelos	LOOCV
Modelo 7	0.285
Modelo 8	1.64

No obstante el Modelo 8 aumentó significativamente su error de predicción por fuera de muestra, lo que da cuenta del componente arbitrario que presenta la partición de la muestra para calcular los MSE.