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	14763	8542.18	13866.55	0.47	3796.53	4837.83	7899.31	350583.34
Edad	14763	38.89	13.2	18.0	28.0	37.0	49.0	91.0
Sexo	14763	0.53	0.50	0.00	0.00	1.00	1.00	1.00
Horas trabajadas	14763	47.20	15.06	1.00	40.0	48.00	50.00	130.00
Antigüedad	14763	61.97	88.03	0.00	7.00	24.00	72.00	720.00
Formal	14476	0.60	0.49	0.00	0.00	1.00	1.00	1.00

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.

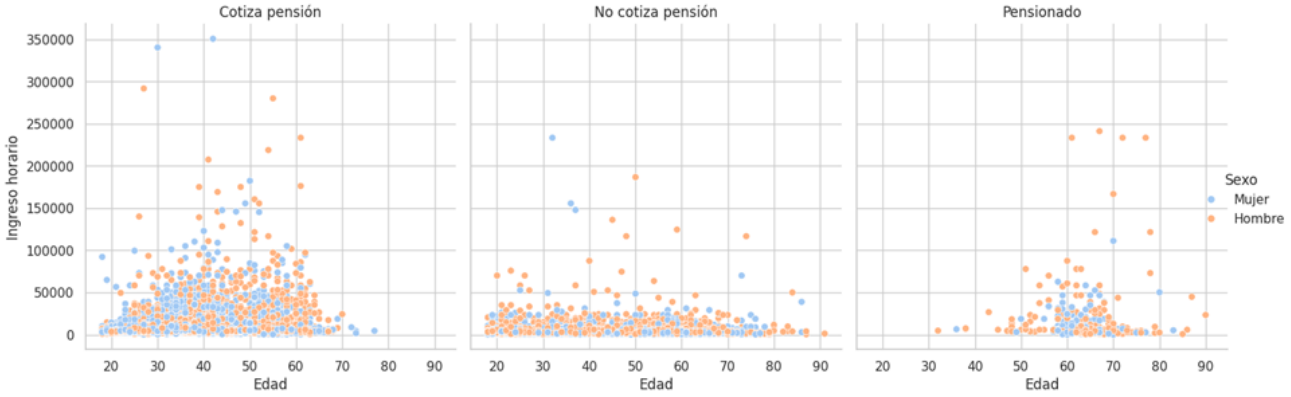
En relación a la antigüedad, esta presenta una gran dispersión, con un promedio de cinco años. Cabe destacar que el 25 % de los encuestados tiene una permanencia de siete meses o menos en su empresa actual. Por último, se ha creado una variable "formal" que toma valor 1 si el individuo cotiza pensión y 0 en caso contrario (excluyendo a los pensionados), lo cual permite aproximar el nivel de informalidad de la economía. Al considerar la

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

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).

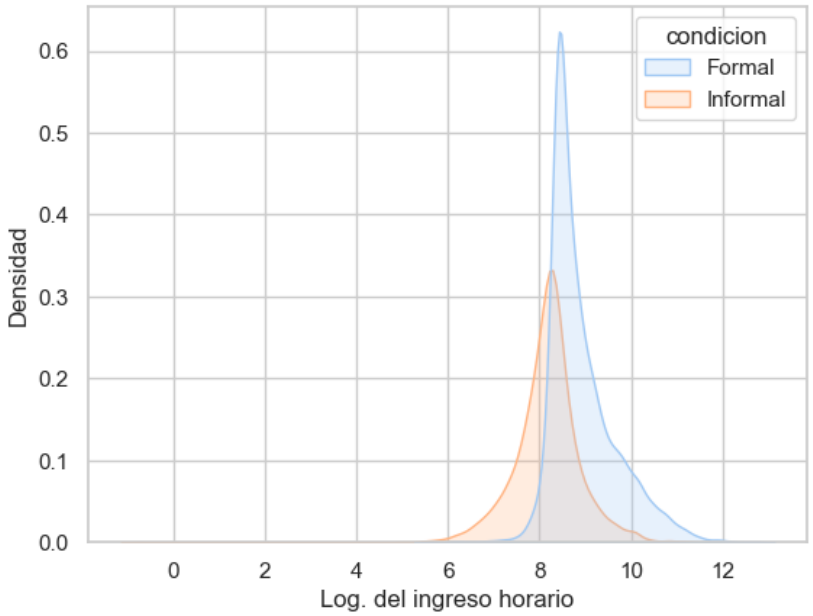
Figura 1: Ingreso por sexo, edad y condición de pensión



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

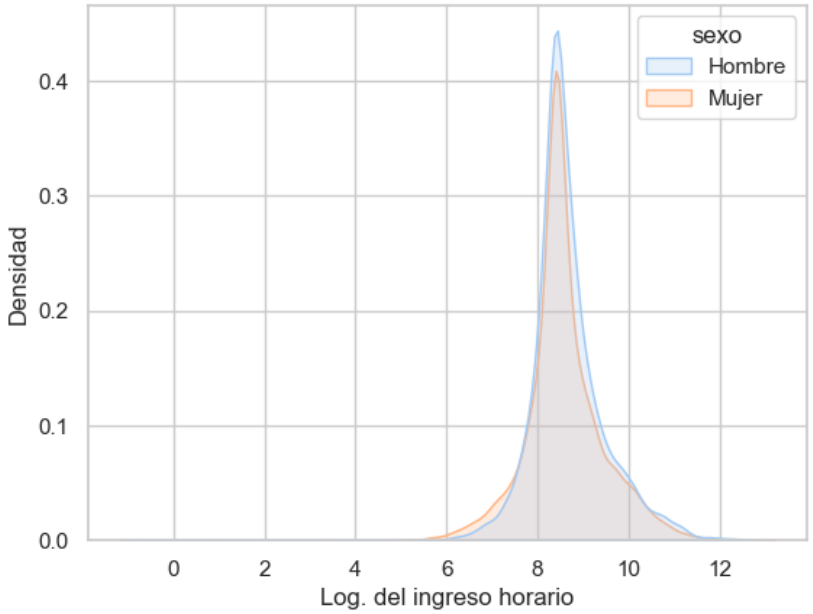
Las Figuras 2 y 3 muestran la densidad del logaritmo del ingreso horario, desglosada por condición laboral (formal o informal) y por sexo. En la primera se observa una diferencia notable entre trabajadores formales e informales. Es plausible que los trabajadores formales se vean influenciados por una 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)

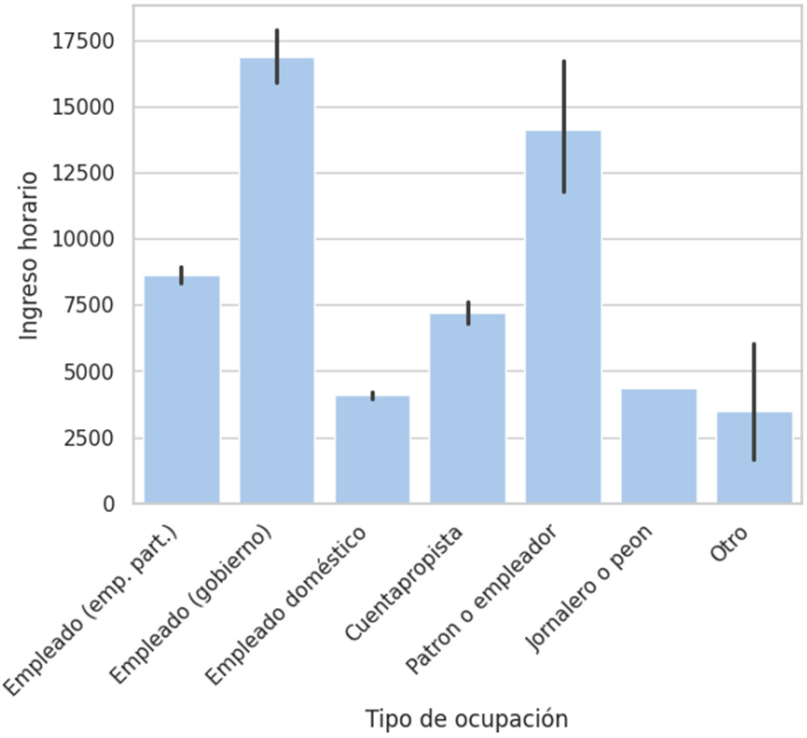
Figura 3: Distribución del ingreso horario (en log.) de hombres y mujeres



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

a)

En esta sección del trabajo, llevamos a cabo el modelo de predicción de salarios.

Para ello se divide la muestra en dos partes, una para entrenamiento (70 %) y otra para prueba (30 %). Dividir los datos en conjuntos de entrenamiento y prueba permite evaluar el desempeño del modelo en datos no vistos, lo cual sirve para estimar que tan bien se comportará el modelo.

Luego, se fija un estado aleatorio (123 en este caso) lo cual significa que cada vez que se ejecute el código con el mismo estado aleatorio, se obtendrá la misma división de entrenamiento y prueba. Esto permite comparar diferentes modelos o especificaciones bajo las mismas condiciones.

Las variables predictoras del modelo serán:

- *age*: Edad del individuo
- *sex*: Género del individuo
- *clase*: Si es un trabajador urbano o rural
- *cotPension*: Si cotiza en un fondo de pensiones
- *totalHoursWorked*: Cantidad de horas trabajadas en la última semana

- *maxEducLevel*: Máximo nivel de educación alcanzado
- *sizeFirm*: Tamaño de la empresa por categorías
- *p6426*: Tiempo trabajado en dicha empresa
- *relab*: Tipo de ocupación

Por último, la variable objetivo es *y_total_m_ha* (salario horario tanto de asalariados como de independientes, construido a partir de las variables de ingreso y horas trabajadas)

b)

A continuación, se muestra una tabla con el RMSE (raíz del error cuadrático medio) para distintas especificaciones del modelo. A partir de este dato, se puede evaluar la capacidad predictiva de cada especificación.

Modelo	**RMSE**
Modelo Simple	11570.8477
Modelo con Interacciones (Grado 2)	11360.3387
Modelo Polinómico (Grado 2)	11151.5381
Modelo Polinómico (Grado 3)	11091.4099
Modelo Polinómico (Grado 4)	11484.1186
Modelo con Selección de Variables (age, sex)	12522.8203
Modelo con Selección de Variables (clase, p6426, sizeFirm)	12225.7520
Modelo con Selección de Variables (cotPension, totalHoursWorked)	12219.6146
Modelo con Sexo, Edad y Edad al Cuadrado	12493.9406
Regresión Ridge	11565.6524
Regresión Lasso	11570.8475
Modelo Completo	11554.7253

Tabla 2: Resultados de RMSE para diferentes modelos

c)

i. Sobre el rendimiento general de los modelos:

Los RMSE obtenidos muestran diferencias claras en la capacidad predictiva de los modelos evaluados. La **regresión lineal simple** tiene un RMSE de 11570.85, un valor alto comparado al resto de especificaciones dado que que este modelo captura las tendencias generales, pero no logra captar relaciones más complejas en los datos. La **regresión con interacciones (grado 2)** tiene un RMSE de 11360.34, un poco menor, ya que este modelo puede capturar relaciones más complejas al tener en cuenta cómo se combinan las variables, lo que le permite ajustarse mejor a los datos y mejorar las predicciones.

Los **modelos polinómicos** muestran un mejor desempeño a medida que el grado aumenta hasta el grado 3. El **modelo polinómico de grado 2** tiene un RMSE de 11151.54, mientras que el **modelo polinómico de grado 3** obtiene el segundo RMSE más bajo, 11091.41. Sin embargo, el **polinomio de grado 4** (RMSE = 11484.12) parece sobreajustar los datos, lo que empeora el rendimiento.

Los modelos con selección de variables, como el que incluye solo **"age" y "sex"** o **"cotPension" y "totalHoursWorked"**, tienen RMSE relativamente altos (12522.82 y 12219.61, respectivamente). Esto sugiere que excluir variables importantes limita la capacidad de los modelos para predecir con precisión.

Los modelos de regularización, **Ridge** (11565.65) y **Lasso** (11570.85), tienen un rendimiento muy similar al de la regresión lineal simple. Esto indica que la penalización al sobreajuste que realizan estos modelos no tiene un gran impacto, probablemente porque no hay un problema serio de multicolinealidad o sobreajuste en los datos.

Finalmente, el **modelo completo**, que incluye todas las variables y la edad al cuadrado, obtiene un RMSE ligeramente mejor que los modelos lineales básicos (11554.73). Esto muestra que incluir todas las variables y agregar la edad de forma cuadrática puede mejorar el ajuste un poco, pero sin grandes mejoras en comparación con modelos polinómicos más complejos.

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

El modelo con el menor RMSE es el **modelo polinómico de grado 3** (11091.41). Esto se debe a su capacidad para capturar relaciones no lineales y combinaciones de variables que los modelos lineales simples y con regularización no lograron captar.

Sin embargo, este modelo puede ser más costoso computacionalmente y más propenso a sobreajustar. Por eso, en situaciones donde se busque simplicidad en la interpretación, los modelos de grado 2 o el modelo con interacciones podrían ser una mejor opción, ya que ofrecen un buen equilibrio entre complejidad y precisión. La elección del modelo dependerá de cual es el objetivo final, si maximizar la precisión o mantener la simplicidad y la interpretabilidad.

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

Para estudiar 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. Además, se muestran estadísticas sobre los errores extremos, los cuales son aquellos errores de predicción cuyo valor absoluto es mayor que dos veces el desvío estandar de todos los errores de predicción.

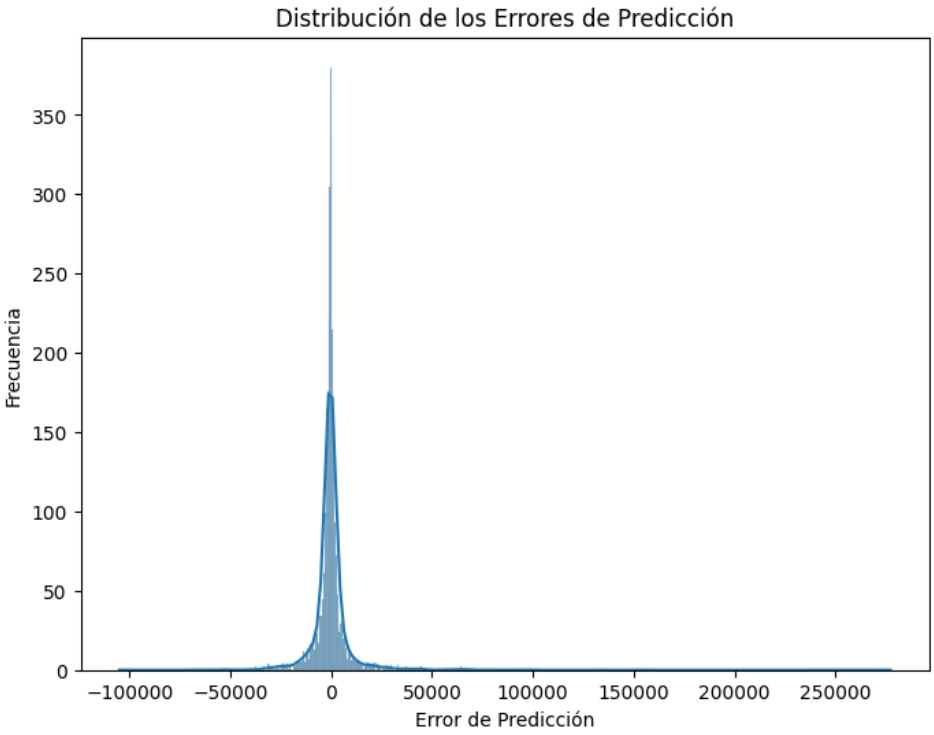


Figura 5: Distribución de los errores de predicción

Estadística	**Valor**
Cantidad de errores extremos	132
Media	29393.47
Desvío estándar	48208.50
Mínimo	-65857.53
Mediana (50 %)	30214.46
Máximo	254676.78

Tabla 3: Resumen de errores extremos para el Modelo Polinómico (Grado 3)

Al analizar la distribución de los errores de predicción, encontramos que tiene una forma muy similar a una distribución normal. Esto podría sugerir que, en su mayoría, el modelo tiene un rendimiento razonablemente bueno. Sin embargo, las estadísticas indican que hay una considerable variabilidad en los errores de predicción, con una cantidad significativa de errores extremos (132 sobre 4429 observaciones), cuya media y desvío estandar son relativamente altos (29393.47 y 48208.50 respectivamente). Además, los valores extremos son muy grandes (mínimo de -65857.53 y máximo de 254676.78).

Dichos errores extremos o valores atípicos se encuentran en las colas de la distribución. 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 DIAN debería examinar más de cerca, ya que los ingresos reportados podrían ser inusuales

y podrían ayudar a detectar evasión fiscal o identificar a personas que realmente necesitan asistencia del estado.

- **El modelo no está bien ajustado para estos casos atípicos,** lo que podría sugerir que el modelo no está capturando adecuadamente ciertos factores relevantes. Si los errores extremos se distribuyen de manera aleatoria sin relación con características específicas de las personas, el modelo podría estar sobreajustando o no considerando correctamente algunas de las características importantes de estas observaciones, por lo que podría necesitar ajustes adicionales.

Dado que hay una cantidad significativa de errores extremos, sería prudente investigar estas observaciones más detalladamente. El gobierno colombiano podría enfocarse en los casos con errores muy grandes (tanto positivos como negativos) para determinar si se deben a ingresos atípicos no reportados correctamente. Si consideramos que estamos frente a un modelo defectuoso, se podrían recolectar más datos, sumar técnicas de machine learning más avanzadas o añadir la validación cruzada para detectar y corregir problemas de sobreajuste o subajuste.

d)

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

- **Modelo Polinómico (Grado 3)**: De 11091.41 a 5362.55
- **Modelo Polinómico (Grado 2)**: De 11151.54 a 5287.62

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

Llama la atención cómo mediante este método, el modelo polinómico de grado 2 pasa a tener un menor RMSE que el de grado 3, sugiriendo que su utilización por encima de este último podría ser adecuada.

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.techfore.2023.123080.
- 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/6d01c9v7.
- Matkowski, M. (2021). Prediction of Individual Level Income: A Machine Learning Approach [Honors Thesis, Bryant University]. Archivo digital.
https://digitalcommons.bryant.edu/honors_economics/39/.
- 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.