Maestría en Economía

Problem Set 1: Predicting Income

Basile, Carolina

Coloma Conte-Grand, Carolina

Onofri, Milagros

Profesor: Ignacio Sarmiento Barbieri

Materia: Machine Learning

24 de Noviembre de 2024

1. Repositorio y aclaraciones

Realizamos el trabajo en el siguiente repositorio de GitHub: https://github.com/milagrosonofri/Problem_Set_1

Vale aclarar que los envíos al repositorio de las figuras asociadas al código fueron hechos de forma automática, como puede verse en el historial del repositorio y en el cuaderno de Jupyter llamado `PS1_gráficos.ipynb`. Esta parte del código, sin embargo, no pudo ser subida al Colab `PS1.ipynb` que utilizamos para la mayor parte del trabajo debido a un error de guardado de GitHub que no logramos resolver.

2. Introducción

La predicción de ingresos individuales es una herramienta clave tanto en el sector público como en el privado, dado su impacto en áreas críticas como la determinación de impuestos y la identificación de vulnerabilidades socioeconómicas. En el ámbito fiscal, la subdeclaración de ingresos representa un desafío significativo, contribuyendo a la brecha fiscal que enfrenta el gobierno. Según el Servicio de Impuestos Internos de los Estados Unidos (IRS), cerca del 83.6 % de los impuestos se pagan voluntariamente, dejando un 16.4 % de ingresos fiscales potenciales sin registrar.

La magnitud de este problema resalta la importancia de desarrollar modelos predictivos de ingresos que puedan ser utilizados como una herramienta de monitoreo y permitan identificar posibles casos de incumplimiento. Al hacerlo, no solo se contribuye a reducir la brecha fiscal, sino que se mejora la equidad del sistema tributario al redistribuir de manera más justa la carga fiscal, ya que pequeñas disminuciones en el cumplimiento tienen un gran costo en términos de ingresos perdidos y trasladan la carga tributaria de aquellos que no pagan sus impuestos a aquellos sí lo hacen (IRS, 2023).

Más allá del ámbito tributario, los modelos predictivos de ingresos tienen aplicaciones relevantes en la identificación de individuos y hogares vulnerables que podrían beneficiarse de políticas públicas específicas, entendiendo que del mismo modo que existen subdeclaraciones de ingresos en la cola superior, también las hay en la cola inferior de la distribución (Albina et al., 2024).

En este trabajo, se construirá un modelo predictivo de salarios horarios individuales utilizando datos provenientes de la Gran Encuesta Integrada de Hogares (GEIH) del año 2018 para Bogotá. Esta base de datos, elaborada por el Departamento Administrativo Nacional de Estadística (DANE), ofrece información sociodemográfica y laboral que resulta adecuada para abordar la problemática planteada. La GEIH incluye información detallada sobre aspectos laborales, ingresos y características demográficas de los individuos.

En base a esto, el objetivo principal será el de estimar la siguiente función,

$$w = X\beta + u$$

donde w representa el salario horario, X es un conjunto de variables explicativas, β los coeficientes asociados a las mismas y u un término de error. En función de los resultados de los modelos planteados se buscará predecir información vinculada a los ingresos de la población de referencia.

3. Datos

Los datos utilizados en este trabajo provienen de la Gran Encuesta Integrada de Hogares (GEIH) del año 2018, específicamente de la Medición de Pobreza Monetaria y Desigualdad en Bogotá. Este informe es elaborado por el Departamento Administrativo Nacional de Estadística (DANE) y contiene información detallada sobre aspectos sociodemográficos, laborales y de ingresos de la población.

En el contexto de este trabajo, los datos serán utilizados para construir un modelo predictivo de salarios horarios individuales que permita estudiar cómo distintas características explicativas afectan el ingreso laboral.

3.1. Extracción de datos

El conjunto de datos fue obtenido mediante **web scraping** desde la página web GEIH 2018 Sample, (Sarmiento, 2018). La extracción de datos se realizó utilizando herramientas de Python como `requests` y `BeautifulSoup`, dado que los datos estaban organizados en tablas distribuidas en 10 páginas consecutivas.

3.1.1. Pasos del proceso

Estructura de las páginas web: Las tablas con datos estaban distribuidas en 10 páginas HTML con URLs que seguían un patrón sistemático. Cada tabla contenía filas correspondientes a individuos, con columnas que representaban características sociodemográficas, laborales y de ingresos.

Extracción de datos: Usando la biblioteca `requests`, se enviaron solicitudes HTTP `GET` para obtener el contenido HTML de cada página. Con la biblioteca `BeautifulSoup`, se localizaron las tablas en el código HTML. En la primera iteración, se extrajeron los encabezados de las columnas, mientras que en todas las iteraciones se capturaron las filas de datos.

Consolidación de datos: Todas las filas de datos fueron almacenadas en una lista de Python y posteriormente convertidas a un `DataFrame` de *pandas*, utilizando los encabezados de las columnas como nombres de variables.

3.1.2. Restricciones y desafíos encontrados

En cuanto al acceso a los datos, no se encontraron restricciones de acceso en la página web, ya que los datos son públicos y accesibles sin autenticación.

Luego, los datos estaban en formato de strings, lo que requirió un procesamiento posterior para convertirlos a tipos de datos numéricos en variables clave, como salarios y horas trabajadas.

Por último, dado que las tablas estaban distribuidas en múltiples páginas, fue necesario implementar un sistema automatizado para recopilar toda la información de manera eficiente.

3.2. Limpieza de datos

Una vez obtenidos los datos, se realizó un proceso exhaustivo de limpieza y filtrado para garantizar la calidad del conjunto de datos y su adecuación al análisis. Este proceso incluyó las siguientes etapas:

- **Restricción de la población de interés**: el análisis se centró únicamente en individuos empleados mayores de 18 años. Para ello, se aplicaron los siguientes filtros:

- La variable `age` debía ser mayor o igual a 18.
- La variable `ocu`, que indica si el individuo está empleado, debía ser igual a 1.

Este paso eliminó observaciones irrelevantes para el análisis, como menores de edad, individuos no empleados o con información faltante en estas variables.

Luego, para facilitar la interpretación, se renombraron variables clave como:

- `p6500` a `salario_empleo_principal`
- `p7070` a `salario_empleo_secundario`
- `p6426` a `antig`
- `sex` a `sexo`
- `age` a `edad`

Además, se convirtieron las variables relacionadas con salarios y horas trabajadas a formato numérico.

Por otro lado, se definió un diccionario de categorías para agrupar los diferentes códigos de la variable `oficio` en categorías más agregadas. Llevamos adelante este proceso con el objetivo de poder considerar los efectos del oficio pero sin restringir demasiado los grados de libertad de los modelos. Se creó una nueva variable, `oficio_new`, que asigna cada registro a su respectiva categoría.

Por último, las variables categóricas seleccionadas (`relab`, `sizeFirm`, `maxEducLevel` y `oficio_new`) fueron transformadas en *dummies* para incluirlas como variables explicativas en el modelo predictivo.

De esta forma, el conjunto de datos limpio y procesado incluye ahora todas las variables relevantes en el formato adecuado para el análisis.

3.3. Análisis descriptivo

3.3.1. Elección de variables explicativas

Se seleccionaron las variables explicativas más relevantes para modelar los ingresos laborales, considerando tanto la teoría económica como la disponibilidad de datos en la base utilizada. En esta sección, describiremos las variables y justificaremos su inclusión en el modelo.

En primer lugar, en el Cuadro 1 se presenta una descripción detallada de cada variable seleccionada, junto con sus categorías o codificaciones principales. Es importante destacar que, como se mencionó anteriormente, algunas variables categóricas, como el tipo de ocupación laboral (`relab`), el tamaño de la firma (`sizeFirm`), el nivel educativo máximo alcanzado (`maxEducLevel`) y el oficio (`oficio_new`), fueron transformadas en variables *dummies* para capturar correctamente sus efectos en el modelo.

Cuadro 1: Variables utilizadas en el modelo predictivo

Variable	Descripción	Categorías o Codificación
`edad`	Edad del individuo (años).	Numérica
`informal`	Condición laboral según afiliación a seguridad social.	0: Formal, 1: Informal
`relab`[a]	Tipo de ocupación laboral.	1: Obrero/Empleado empresa privada, 2: Obrero/Empleado gobierno, 3: Empleado doméstico, 8: Jornalero/Peón
`sexo`	Sexo del individuo.	0: Mujer, 1: Hombre
`sizeFirm`[a]	Tamaño de la firma.	1: Autoempleo, 2: 2-5 empleados, 3: 6-10 empleados, 4: 11-50 empleados, 5: Más de 50 empleados
`maxEducLevel`[a]	Nivel educativo máximo alcanzado.	1: Sin educación, 2: Preescolar, 3: Primaria incompleta, 4: Primaria completa, 5: Secundaria incompleta, 6: Secundaria completa, 7: Terciario, 9: No aplica
`antig`	Antigüedad en la empresa o negocio (años).	Numérica
`oficio`[a]	Categoría del oficio según áreas ocupacionales.	Ver Nota[b]

[a] **Nota**: Las variables categóricas fueron codificadas numéricamente para facilitar el análisis. Las categorías y sus codificaciones se describen en la columna correspondiente.
[b] **Nota**: Las categorías del oficio se agruparon en áreas ocupacionales. Detalle completo disponible en el Anexo 4.3.
Fuente: Elaboración propia en base a datos de la Gran Encuesta Integrada de Hogares (GEIH) 2018.

Las variables seleccionadas para el modelo predictivo de salarios horarios individuales se eligieron con base en su relevancia teórica y empírica en la literatura económica.

Los gráficos de dispersión presentados en el Anexo 2 permiten examinar las relaciones entre las variables explicativas y el salario horario. Estas visualizaciones ofrecen una primera aproximación a las correlaciones y patrones identificados en la literatura. Además, los *scatterplots* son útiles para detectar posibles no linealidades y valores atípicos, proporcionando una base sólida para validar las hipótesis teóricas y guiar el modelado predictivo posterior. A continuación, se describen en detalle estas variables y sus tendencias en los datos.

- **Edad (`edad`)**: La edad es un indicador de la experiencia laboral acumulada, la cual tiene un impacto significativo en los ingresos laborales. Se espera una relación positiva entre la edad y el salario en las etapas iniciales de la vida laboral, ya que el aumento de la experiencia suele asociarse con mayores ingresos. Sin embargo, más allá de cierto punto, esta tendencia podría estabilizarse o incluso revertirse, reflejando un posible estancamiento o disminución en los ingresos con el envejecimiento.

  Si observamos la Figura 2a), donde se grafica el **salario horario** en función de la **edad**, podemos observar que esta relación parece ser efectivamente cuadrática. Se observa que los salarios aumentan hasta cierto rango etario, probablemente asociado con una mayor experiencia laboral, y luego tienden a estabilizarse o incluso disminuir en edades más avanzadas. Además, se identifican algunos valores atípicos de salarios significativamente más altos que el promedio, atados a *outliers* en la distribución del ingreso.

- **Condición laboral (`informal`)**: La formalidad está directamente vinculada con mejores condiciones de trabajo, estabilidad y acceso a beneficios de seguridad social. Asimismo, suele estar asociada con el cumplimiento de regulaciones laborales como el pago del salario mínimo, lo cual representa un piso en los ingresos que no necesariamente se observa en empleos informales. De esta forma, distinguir entre trabajadores formales e informales resulta esencial para capturar estas heterogeneidades

que podrían estar asociadas a un mayor salario para los trabajadores formales.

Podemos observar la relación entre el **salario horario** y el **trabajo informal** en la Figura 2b). Efectivamente, se refleja una clara distinción por grado de formalidad. Los trabajadores formales (0 en el eje horizontal) tienden a reportar, en promedio, mayores salarios horarios y mayor dispersión en los mismos en comparación con los trabajadores informales (1 en el eje horizontal).

- **Tipo de ocupación laboral (`relab`)**: El tipo de ocupación influye significativamente en los ingresos percibidos, ya que ciertas actividades laborales, como el trabajo por cuenta propia, suelen presentar una mayor variabilidad en los ingresos en comparación con empleos más estructurados, como los del sector gubernamental, que suele presentar mayor estabilidad en los ingresos.

  La asociación entre el **salario horario** y la **relación laboral** (ver Figura 2c)) revela una notable heterogeneidad en los ingresos según el tipo de ocupación. Cabe destacar que la ausencia de observaciones en ciertas categorías laborales se debe a los criterios de selección aplicados en el análisis, donde únicamente se incluyeron personas que perciben ingresos laborales.

- **Antigüedad (`antig`)**: La antigüedad mide la experiencia específica en un empleo, lo cual puede reflejar estabilidad, así como una mayor productividad debido a acumulación de capital humano especifico, factores que suelen estar asociados con mayores ingresos. Por otra parte, a niveles altos de antigüedad también empiezan, posiblemente, a ser considerados efectos como la edad de la persona.

  Por su parte, la Figura 2d), que relaciona el **salario horario** y la **antigüedad** indica que existe una dispersión considerable y una tendencia ligeramente negativa. De esta forma, si bien la experiencia laboral acumulada podría influir positivamente en los ingresos, existen otros factores determinantes que generan variabilidad en los salarios.

- **Sexo (`sexo`)**: Incorporar el sexo como variable permite analizar disparidades salariales de género, un tema ampliamente estudiado en la literatura.

  En el análisis de la relación entre **salario horario** y **sexo** (ver Figura 2e)) se observa una mayor dispersión en los salarios horarios de los hombres (valor 1 en el eje horizontal) en comparación con las mujeres (valor 0). Además, los hombres tienden a reportar, en promedio, salarios horarios superiores.

- **Tamaño de la firma (`sizeFirm`)**: El tamaño de la firma suele estar relacionado positivamente con los niveles salariales debido a factores como la capacidad financiera, economías de escala, entre otras.

  En la Figura 2f) se muestra esto claramente. Los empleados de empresas más grandes suelen percibir salarios horarios más altos en promedio, mientras que en las firmas más pequeñas se concentra una mayor proporción de ingresos bajos. Además, se observa mayor dispersión en firmas más grandes.

- **Máximo nivel educativo alcanzado (`maxEducLevel`)**: La educación es uno de los determinantes más importantes y estudiados en la literatura del ingreso laboral. Se espera que exista una relación positiva y significativa entre el nivel educativo y el salario.

  La Figura 2g) confirma esto. Los individuos con niveles educativos más altos tienden a tener, en promedio, salarios horarios significativamente mayores en comparación con aquellos con niveles educativos más bajos o sin educación formal.

- **Oficio, categoría amplia** (`oficio_new`): Es lógico pensar que existirá una amplia variabilidad en los ingresos entre sectores de ocupación. Esto se ve evidenciado en la Figura 2h), que relaciona el **salario horario** con el **tipo de oficio**.

Las ocupaciones relacionadas con ciencias, ingeniería y tecnología, administración y gestión, y ciencias sociales y humanidades valores significativamente más altos en comparación con otros sectores. Esto sugiere que estos sectores pueden estar asociados con mejores oportunidades salariales o empleos altamente remunerados.

A su vez, sectores como servicios personales y comunitarios, agricultura, pesca y explotación de recursos naturales y otros oficios y trabajos no clasificados tienden a concentrar ingresos horarios más bajos.

3.3.2. Estadísticas descriptivas para variables de ingresos

A continuación, se presentan el Cuadro 2 y el Cuadro 3, que muestran estadísticas descriptivas de las variables de interés. En el Anexo 3 se presentan histogramas para las distintas categorías de horas trabajadas e ingresos tanto en su escala original como en escala logarítmica.

Cuadro 2: Estadísticas descriptivas: Observaciones, Media, Desvío Estándar, Mínimo y Máximo

Variable	Observaciones	Media	Desvío Estándar	Mínimo	Máximo
Horas trabajadas	16.397	47	15	1	130
Horas trabajadas en empleo secundario	556	11	9	1	50
Salario del empleo principal	9.785	1.566.233	2.258.106	10.000	34.000.000
Salario del empleo secundario	453	697.084	1.129.732	6.000	10.000.000
Ingreso laboral nominal mensual	9.785	1.757.076	2.413.728	30.000	60.100.000
Ingreso laboral nominal horario	9.785	8.868	12.917	326	350.583

Nota: Los salarios están expresados en pesos colombianos de 2018.
Fuente: Elaboración propia en base a datos de la Gran Encuesta Integrada de Hogares (GEIH) 2018.

Cuadro 3: Estadísticas descriptivas: Observaciones y Percentiles

Variable	Observaciones	Percentil 25	Percentil 50	Percentil 75	
Horas trabajadas	16.397	40	48	50	
Horas trabajadas en empleo secundario	556	5	10	16	
Salario del empleo principal	9.785	781.242	900.000	1.500.000	
Salario del empleo secundario	453	120.000	300.000	0	750.000
Ingreso laboral nominal mensual	9.785	869.453	1.040.000	1.625.000	
Ingreso laboral nominal horario	9.785	4.226	5.071	8.101	

Nota: Los salarios están expresados en pesos colombianos. La variable *Salario del empleo secundario* fue eliminada del análisis debido a su falta de información útil, ya que en los tres percentiles analizados contiene ceros.
Fuente: Elaboración propia en base a datos de la Gran Encuesta Integrada de Hogares (GEIH) 2018.

En cuanto a las **horas trabajadas totales**, los trabajadores reportan en promedio 47 horas semanales, con una desviación estándar de 15 horas, mientras que la mediana es de 48 horas y el percentil 75 de 50. Aunque esto está alineado con la jornada laboral estándar en Colombia, se observan valores extremos, como jornadas que alcanzan las 130 horas semanales, lo cual podría indicar semanas atípicas, situaciones de empleo informal o errores en el registro de los datos. Podemos observar en el histograma (ver Figuras 3a) y 3b)) que existe un pequeño porcentaje que excede las 60 horas, y muy pocos casos con valores extremos.

El promedio de **horas trabajadas en el empleo secundario** es de 11 horas semanales, con un desvío estándar de 9 horas y una mediana de 10. A su vez, el bajo número de observaciones para esta variable (556) sugiere que muy pocos individuos tienen un segundo empleo. La distribución del histograma puede verse en las Figuras 3c) y 3d) y exhibe una dispersión considerable, denotando que la cantidad de horas dedicada a esta ocupación extra varía mucho según el individuo.

Para el **salario del empleo principal**, observamos que este presenta una media de 1.467.828 pesos colombianos, con una desviación estándar elevada (2.123.494) y una mediana de 880.000. Esto evidencia

la existencia de *outliers* con salarios extremadamente altos. El rango de valores llega hasta los 34.000.000 pesos, lo cual evidencia que hay individuos con ingresos significativamente superiores al promedio que los ingresos derivados del empleo principal. El histograma (ver Figura 3e) muestra esta concentración y la existencia de outliers, que puede verse de mejor forma introduciendo una escala logarítmica en la Figura 3f).

El **salario del empleo secundario** tiene una media significativamente más baja, de 19.294 pesos, y un desvío estándar también elevado, de 219.840 pesos. Esto tiene sentido, ya que los empleos secundarios por definición aportan ingresos inferiores a los de la ocupación principal.

En el histograma (Figura 3g)) observamos que la distribución es enormemente asimétrica, con la mayoría de los salarios concentrados cerca del cero. Esto sugiere que muchas personas no tienen ingresos de un empleo secundario, y que las que lo tienen, perciben ingresos bajos. De manera similar, en el histograma en escala logarítmica (Figura 3h)), la transformación logarítmica hace que la distribución sea más informativa, evidenciando que aunque la mayoría de los individuos tienen ingresos cercanos a cero, la fracción con ingresos secundarios que le sigue presenta una distribución normal bastante marcada.

El **ingreso laboral nominal mensual** tiene una media de 1.757.076 pesos, algo superior al salario del empleo principal, reforzando el hecho de que algunos trabajadores complementan sus ingresos con un empleo secundario. De todas formas, los histogramas (ver Figuras 3i) y 3j)) exhiben una forma muy similar a los analizados previamente, con una dinámica de *outliers* de ingresos análoga. Lógicamente, lo mismo ocurrirá con el **ingreso laboral nominal horario**, que cuenta con una media de 8.868 pesos y cuya distribución puede verse en las Figuras 3k) y 3l).

4. Predicción de salarios

4.1. Set-up

Definimos nuestra variable Y como el logaritmo de la variable `y_ingLab_m_ha`, el ingreso laboral horario nominal de todas las ocupaciones de la persona (incluyendo propinas y comisiones). Tomamos el logaritmo debido a que la gran varibilidad de los salarios horarios, al haber un número pequeño de observaciones en la cola superior con salarios muy grandes respecto al valor de la media. Esta dispersión es suavizada al tomar logaritmo y esto facilita llegar a predicciones más certeras. A su vez, definimos nuestra X como todas aquellas variables previamente seleccionadas, excluyendo `y_ingLab_m_ha`.

Seguimos el enfoque de partición de datos conocido como de *Validation Set Approach*, que permite evaluar el desempeño del modelo fuera de la muestra (*out-of-sample*). La muestra se divide en dos subconjuntos:

- **Muestra de entrenamiento:** Representa el 70 % de los datos y se utiliza para construir y ajustar el modelo predictivo.
- **Muestra de prueba (test):** Representa el 30 % de los datos y se utiliza para evaluar el desempeño del modelo sobre datos no vistos.

Para garantizar la reproducibilidad, utilizamos un estado aleatorio fijo (`random_state = 123`) en Python. La partición se implementó utilizando la función `train_test_split` de `scikit-learn`.

Para poder utilizar las variables categóricas halladas de una forma más sencilla en los modelos, creamos la función `related_columns`, que devuelve todas aquellas columnas asociadas a un tipo específico de dato.

4.2. Modelos

En total corrimos 14 modelos con diferentes grados de complejidad. En aquellos modelos lineales, el ajuste fue mejorando a medida que se elevaba el nivel de complejidad, hasta llegar a un RMSE de 0.4984 en el Modelo 6, que incorpora todas las variables seleccionadas previamente como relevantes.

Al incorporar interacciones y no linealidades de segundo grado, estas redujeron el RMSE para todos los modelos estudiados. Esto indica que las variables cuadráticas y las interacciones entre variables parecen ser relevantes para la explicación del logaritmo del salario horario.

No sucede lo mismo al incorporar interacciones y no linealidades de tercer grado. En este caso, el ajuste del modelo mejora para los Modelos 3 y 4, pero empeora en el caso de los Modelos 5 y 6. Entendemos que esto probablemente se deba a un problema de *overfitting* generado por el exceso de variables explicativas. Esto se ve específicamente para el Modelo 14, que añade el componente de tercer grado al Modelo 6, el más complejo de los lineales. Las 11 variables categóricas de `oficio_new` que sumamos en este caso generan una cantidad enorme de variables adicionales a la hora de incluir interacciones de tercer grado, lo que explica que el RMSE se haya disparado, más que triplicando el del Modelo 6.

La especificación con el mejor ajuste es la del Modelo 13, que incorpora no linealidades e interacciones de segundo grado y utiliza como variables explicativas al intercepto, el máximo nivel educativo alcanzado, la edad, el sexo, la condición de formalidad, la relación laboral, la antigüedad, el tamaño de la firma y el tipo de oficio. El RMSE en este caso es de 0.4861.

Cuadro 4: Resumen de modelos

Modelo	**Root Mean Squared Error**
Modelo 1: Intercepto	0.7280
Modelo 2: Intercepto y `maxEducLevel`	0.6234
Modelo 3: Intercepto, `maxEducLevel`, `edad` y `sexo`	0.5661
Modelo 4: Intercepto, `maxEducLevel`, `edad`, `sexo`, `informal` y `relab`	0.5572
Modelo 5: Intercepto, `maxEducLevel`, `edad`, `sexo`, `informal`, `relab`, `antig` y `sizeFirm`	0.5449
Modelo 6: Intercepto, `maxEducLevel`, `edad`, `sexo`, `informal`, `relab`, `antig`, `sizeFirm` y `oficio_new`	0.4984
Modelo 7: Modelo 3 con interacciones y no linealidades de segundo grado	0.5415
Modelo 8: Modelo 3 con interacciones y no linealidades de tercer grado	0.5388
Modelo 9: Modelo 4 con interacciones y no linealidades de segundo grado	0.5334
Modelo 10: Modelo 4 con interacciones y no linealidades de tercer grado	0.5310
Modelo 11: Modelo 5 con interacciones y no linealidades de segundo grado	0.5197
Modelo 12: Modelo 5 con interacciones y no linealidades de tercer grado	0.5526
Modelo 13: Modelo 6 con interacciones y no linealidades de segundo grado	0.4861
Modelo 14: Modelo 6 con interacciones y no linealidades de tercer grado	1.5639

Fuente: Elaboración propia en base a datos de la Gran Encuesta Integrada de Hogares (GEIH) 2018.

A continuación, tomamos los dos modelos con mayor poder predictivo y analizamos sus errores de predicción a través de los gráficos ubicados en el Anexo 3.

Podemos observar que los errores de predicción de ambos modelos poseen una media cercana al cero y una tendencia a la sub-estimación del salario horario de los individuos. Esto implica que los modelos tienen una gran capacidad predictiva para salario medios pero pierde cierta capacidad predictiva al considerar valores extremos de la variable de interés.

Este punto se ve claramente en los *scatterplot*, donde las estimaciones se posicionan en gran medida en la zona media del salario horario, generando grandes errores de predicción al considerar salarios extremadamente altos o bajos.

A priori, es posible observar que el Modelo 13 captura mucho mejor las particularidades de los salarios altos y ligeramente mejor la de los salarios bajos, a diferencia del Modelo 6 que prácticamente no tiene predicciones que reflejen la variabilidad que existe en los datos por fuera de los valores medios.

Dada la complejidad de los modelos analizados y la gran capacidad predictiva de los mismos en la gran mayoría de los casos, tiene sentido pensar que los *outliers* surgen de casos aislados que podrían ser analizados por el DANE. En particular, y dada la tendencia a la subestimación del salario horario, los errores podrían deberse a la subdeclaración de ingresos por parte de los individuos encuestados.

4.3. *Leave One Out Cross-Validation* (*LOOCV*)

Para el caso del Modelo 6 y el Modelo 13, los dos con el menor error cuadrático medio, los estimamos nuevamente utilizando el método de *Leave One Out Cross-Validation* (*LOOCV*). Este método recorre las observaciones y utiliza cada una de ellas una vez como el set de validación, empleando al resto como el set de entrenamiento.

Para el caso del Modelo 6, el análisis utilizando el *LOOCV* arroja un RMSE de 0.5052, muy similar al valor de 0.4984 al cual habíamos llegado a partir del método de *Validation Set* con 70 % entrenamiento y 30 % test.

Tiene sentido que este valor sea algo más grande por dos motivos. En primer lugar, cuando una de las observaciones de la cola superior es elegida como el set de testeo, habrá un error de subestimación muy grande. En segundo lugar, en el caso de elegir una de las observaciones que no forman parte de la cola superior como testeo, habrá también una sobrestimación considerable. Asumiendo una distribución pareja entre las muestras del *Validation Set Approach*, este segundo problema estará presente, mientras que el primero no lo estará.

Para el Modelo 13, en cambio, el análisis utilizando el *LOOCV* arroja un RMSE de 7961.95, muy significativamente mayor al 0.4861 al cual habíamos llegado a partir del *Validation Set Approach*. Siguiendo con la línea de análisis previa, podemos pensar que los errores más grandes con los que contábamos se vieron multiplicados muchas veces debido a las interacciones nuevas de segundo grado que incorpora este nuevo modelo.

Si miramos distintos percentiles de la distribución de los RMSE calculados en las iteraciones realizadas, encontramos que los valores de RMSE son bajos en casi toda la distribución hasta llegar al percentil 99.99, lo cual es consistente con la idea de que esto se debe a un ajuste pobre en observaciones específicas donde se testea sobre valores extremos y su posterior expansión a través de las interacciones.

Viendo tanta cantidad de casos que arrojaron RMSE altos en el LOOCV, nos resultó llamativo que el RMSE hubiera dado un valor original tan pequeño, por lo que condujimos un nuevo análisis del Modelo 13 cambiando el *random state* a 1. Aquí encontramos un RMSE de 7680.6, también mucho mayor al original. Esto nos lleva a pensar que el RMSE bajo fue solo un golpe de suerte que se debió a que el *random state 123* arrojó muestras balanceadas entre si al hacer la distribución 70-30. Esto, sin embargo, podría perfectamente no ocurrir y el ajuste cambiar completamente, exacerbado por el *overfitting* del modelo.

Como chequeo adicional de robustez realizamos un *cross-validation K-fold* para este modelo y, nuevamente, encontramos un RMSE muy elevado respecto al original.

Todos estos hallazgos nos llevan a pensar que el Modelo 13 incurre en overfitting de variables, y que es preferible quedarnos con el Modelo 6 lineal y no incorporar interacciones adicionales.

Referencias

Albina, I., Laguinge, L. A., Gasparini, L. C., Tornarolli, L., Cruces, G. A., y Afonso, S. (2024). Ajustando la imagen de la distribución del ingreso en argentina: encuestas y registros administrativos. *Documentos de Trabajo del CEDLAS*.

IRS. (2023). *The tax gap*. https://www.irs.gov/newsroom/the-tax-gap.

Sarmiento, I. M. (2018). *Geih 2018 sample*. https://ignaciomsarmiento.github.io/GEIH2018_sample/.

Anexo 1: Detalle de la variable oficio

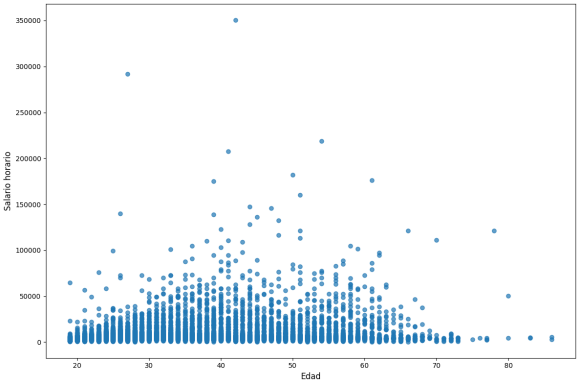
La variable `oficio` fue agrupada en las siguientes áreas ocupacionales, basándose en códigos numéricos:

Cuadro 5: Agrupaciones para la variable `oficio`

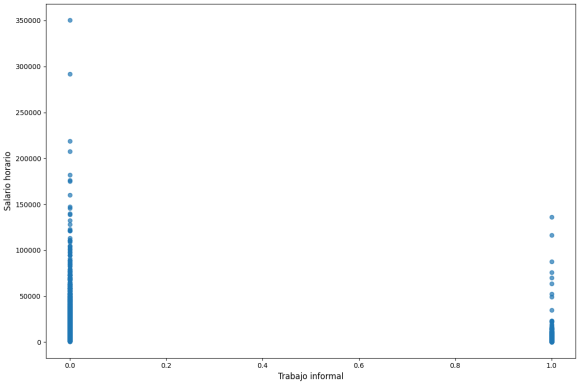
Área Ocupacional	**Códigos Asociados**
Ciencias, Ingeniería y Tecnología	1, 2, 3, 4, 8
Ciencias Biológicas y de la Salud	5, 6, 7
Ciencias Sociales y Humanidades	9, 11, 12, 13, 19
Arte, Cultura y Medios	15, 16, 17, 18, 92
Administración y Gestión	20, 21, 30
Oficios Administrativos y Comerciales	31, 32, 33, 34, 38, 39, 40, 41, 42, 43, 44, 45, 49
Servicios Personales y Comunitarios	14, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59
Agricultura, Pesca y Explotación de Recursos Naturales	60, 61, 62, 63, 64, 71, 72, 73
Industria, Manufactura y Construcción	70, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 93, 94, 95, 96
Transporte y Logística	35, 36, 37, 97, 98
Otros Oficios y Trabajos No Clasificados	99

Anexo 2: Gráficos de dispersión

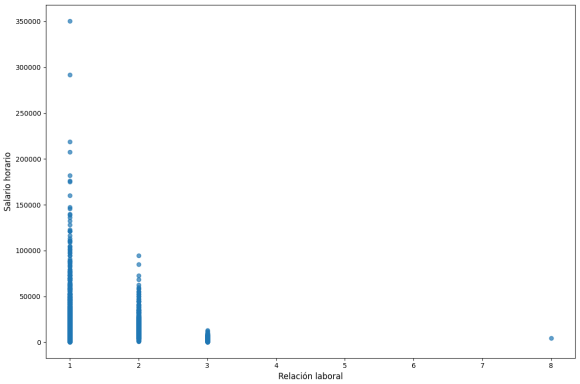
2a) Relación entre salario horario y edad.



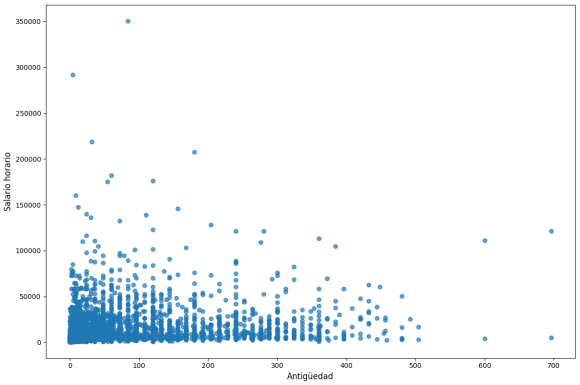
2b) Relación entre salario horario y trabajo informal.



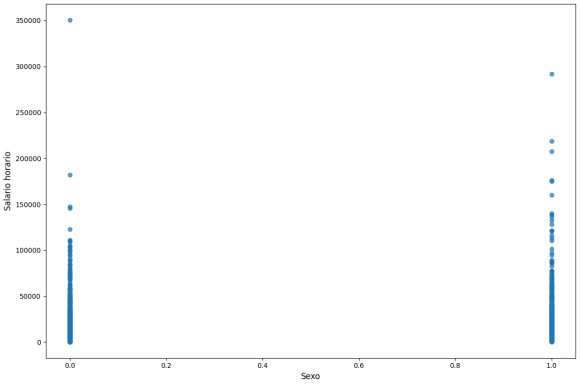
2c) Relación entre salario horario y relación laboral.



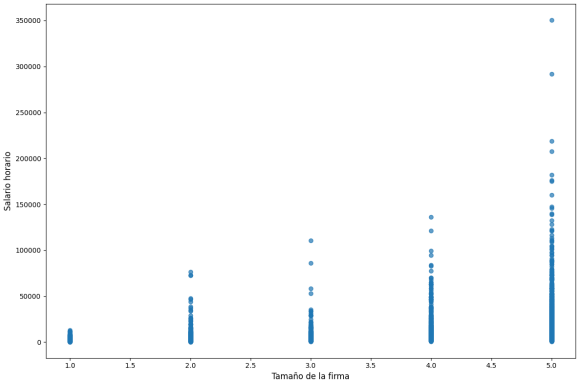
2d) Relación entre salario horario y antigüedad.



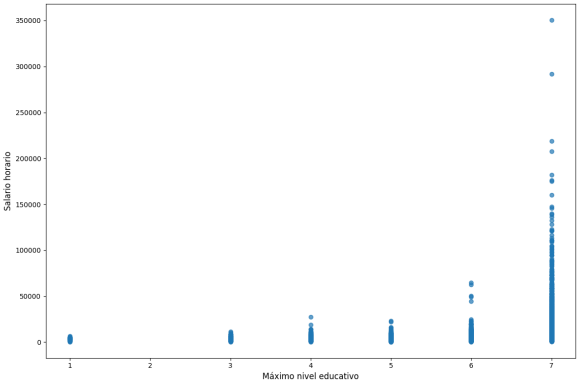
2e) Relación entre salario horario y sexo.



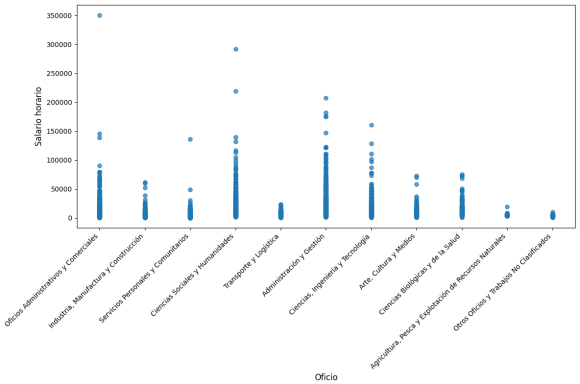
2f) Relación entre salario horario y tamaño de la firma.



2g) Relación entre salario horario y nivel educativo máximo.

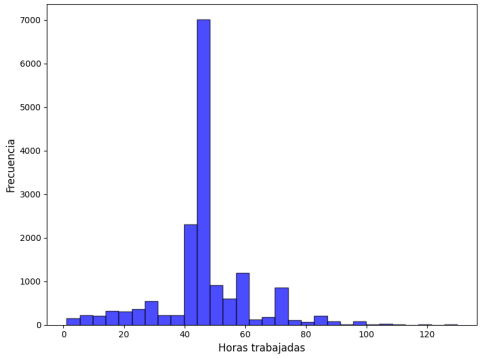


2h) Relación entre salario horario y oficio.

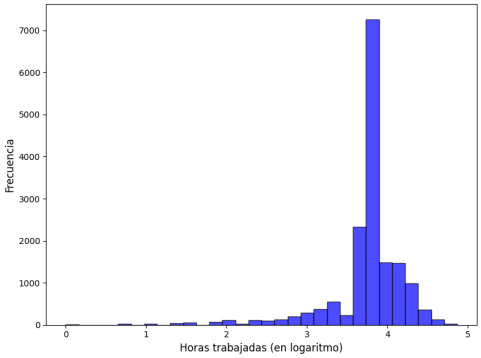


Anexo 3: Histogramas

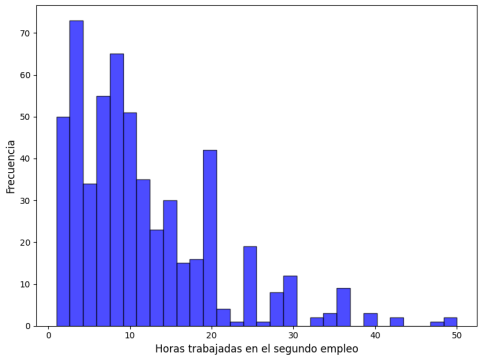
3a) Histograma de horas trabajadas.



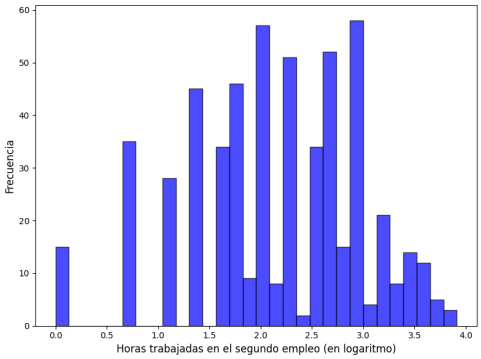
3b) Histograma de horas trabajadas (log).



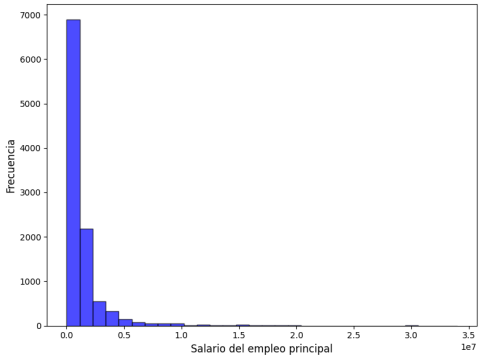
3c) Histograma de horas en empleo secundario.



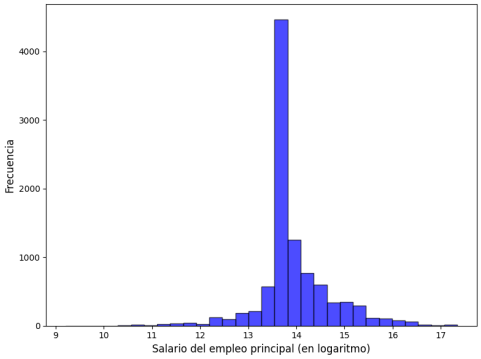
3d) Histograma de horas trabajadas en empleo secundario (log).



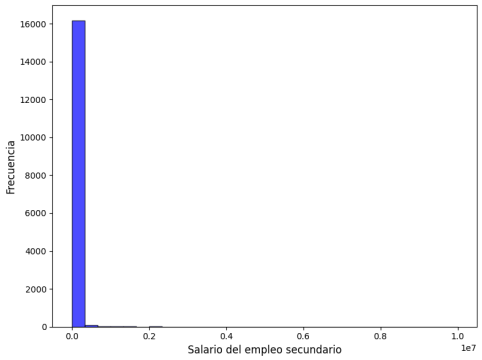
3e) Histograma de salario del empleo principal.



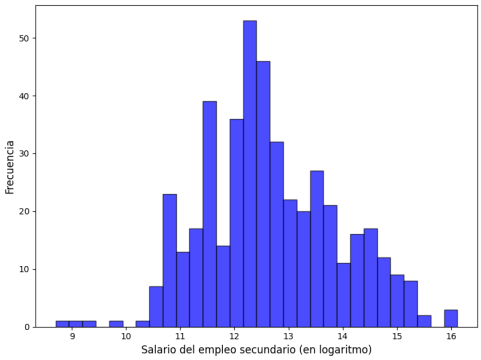
3f) Histograma de salario del empleo principal (log).



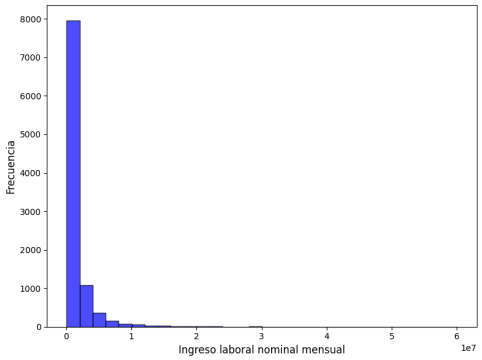
3g) Histograma de salario de empleo secundario.



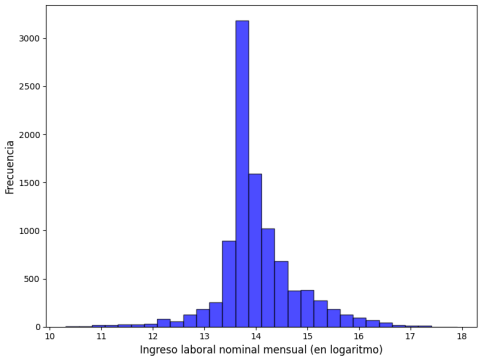
3h) Histograma de salario de empleo secundario (log)



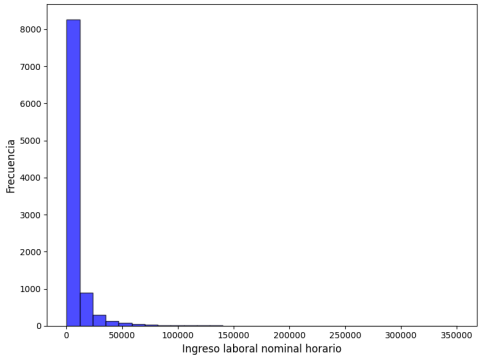
3i) Histograma de ingreso laboral mensual.



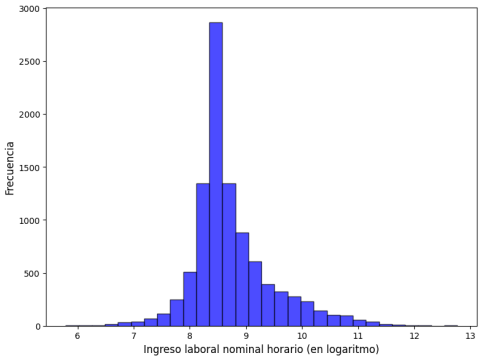
3j) Histograma de ingreso laboral mensual (log).



3k) Histograma de ingreso laboral horario.

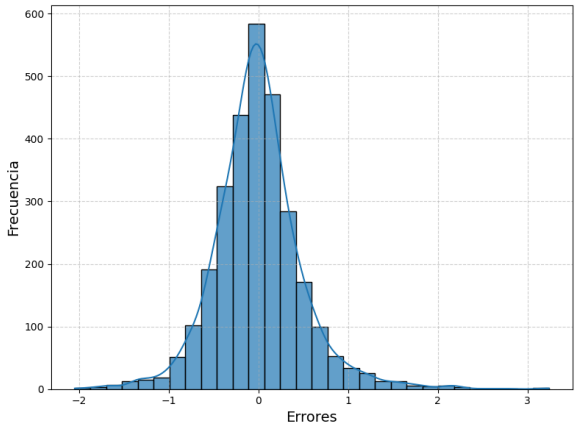


3l) Histograma de ingreso laboral horario (en log).

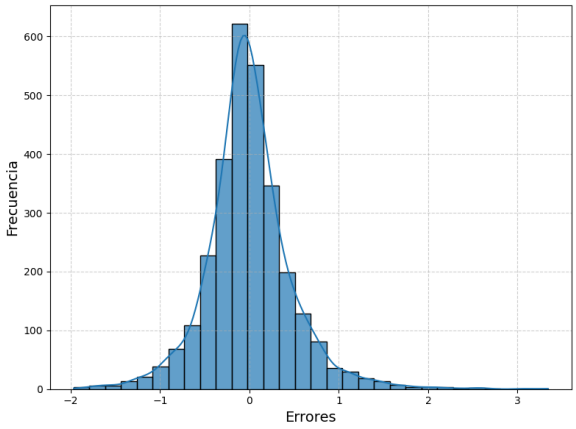


Anexo 4: Errores de predicción de los modelo 6 y 13

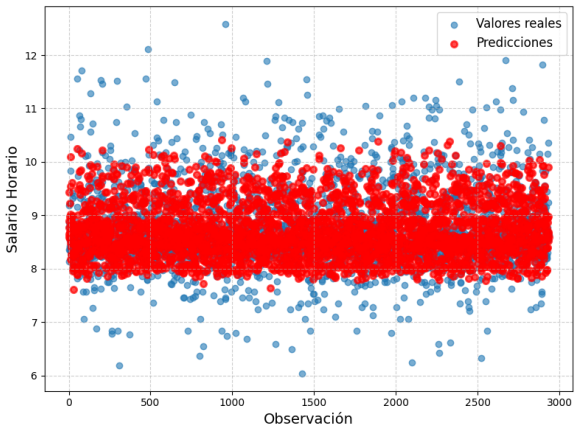
4a) Modelo 6: Distribución de los errores.



4b) Modelo 13: Distribución de los errores.



4c) Modelo 6: Salario horario observado y predicho.



(a) Modelo 13: Salario horario observado y predicho.