Predicciones salariales en Colombia: Un análisis utilizando la GEIH*

Emiliano Bohorquez

Brayan A. Condori Luque

24 de noviembre de 2024

*Este trabajo corresponde al Trabajo Práctico 1 del curso de Machine Learning de la Maestría en Economía de la UNLP.

1. Introduction

Para tablas:
Para figuras

2. Data

El conjunto de datos utilizados proviene de la Gran Encuesta Integrada de Hogares (GEIH) de Colombia, la cual tiene por finalidad brindar información socioeconómica y sobre la situación laboral del país a través de una muestra representativa de dicha población (DANE-DIMPE, 2023).

Esta base es el insumo principal para nuestro modelo de predicción salarial, ya que cuenta con diferentes variables que representan la remuneración de un trabajador (ingreso laboral mensual, ingreso laboral mensual por hora, salario de ocupación principal, salario mensual por hora, entre otros) y un conjunto de características claves que funcionarán como predictoras tales como la edad, el género y el nivel educativo. Por otra parte, el número de observaciones es de más de 32.000 casos, lo que permite tener una muestra grande para posteriormente dividirla en conjunto de entrenamiento y conjunto de prueba.

El trabajo fue realizado en el lenguaje de programación Python dada su potencia y versatilidad para trabajar con grandes volumenes de datos. A continuación, se detalla el proceso de obtención, limpieza, transformación y descripción de este insumo.

2.1. Web Scraping de datos

La obtención de la base completa requirió la realización de un proceso de Web Scraping del repositorio donde estaban alojados los datos, especificamente se realiza un scrapeo horizontal, ya que se extrae información de diversas páginas. Cabe destacar que no se presentaron restricciones que dificultaran o imposibilitaran el Web Scraping.

2.2. Limpieza y transformación de los datos

Para el análisis predictivo, los requisitos de permanencia en la muestra son los siguientes: (i) que sea mayor de edad, y; (ii) que cuente con ingreso positivo, es decir, mayor estricto a 0. Se procedió a filtrar por la variable edad, lo que implicó eliminar de la base a todo aquel que no tenga 18 años. Por otra parte, se utilizó la variable de salario mensual por hora para quedarnos con todos aquellos que percibieran una remuneración.

Adicionalmente, el conjunto de datos contaba con varios datos faltantes. Primeramente, realizamos una selección de columnas para verificar que esta modificación no genere inconsistencias, dado que el cero es un valor de respuesta factible. Estas fueron: el directorio del individuo, el identificador de hogar, el identificador de persona, el género, la edad, el estrato socioeconómico, si cuenta con empleo formal, el máximo nivel educativo, el ponderador de frecuencias, el salario horario mensual y el ingreso total mensual. Varias de estas variables cumplirán un rol central en nuestro modelo. Una vez seleccionadas, eliminamos los valores faltantes, quedandonos con un total de 9.844 registros.

Por otra parte, se llevó a cabo a una serie de transformaciones en algunas variables. En primer lugar, y que será de utilidad para el análisis descriptivo, se aplica logaritmo natural al salario mensual por hora para aproximar su distribución a una normal. Asimismo, observamos que la GEIH pregunta directamente si se percibe algún tipo de subsidio, los cuales pueden ser

alimentario, en transporte, familiar o educativo. Por lo tanto, se procedió a generar una única variable dicotómica donde 1 representaría que un individuo sí percibió alguno de los auxilios mencionados antes, y 0 para el caso contrario.

2.3. Análisis descriptivo

Para culminar esta sección, realizamos un análisis descriptivo con los datos resultantes, el cual va desde cuadros con estadísticas relevantes hasta gráficos para observar la distribución del salairo por hora. La tabla 1 realiza un resumen de la media, el desvío, el total de observaciones, y el mínimo y máximo valor. Cabe destacar el uso del ponderador "fweight"para que los datos sean representativos de la población.

En primer lugar, el salario por hora promedio es de $COL 7.968,09, con un desvío estandar de $COL 11.690,91, para una población representada de 2.459.723. El valor mínimo en el rango es de $COL 151,91 y el máximo es de $COL 291.666.66. Estas estadísticas nos muestran que existe una dispersión importante en los datos, producto de valores extremos en la cola superior. Esto refuerza la convención que las distribuciones de ingresos (y salarios) presentan una asimetría negativa. El gráfico 1 confirma nuestras apreciaciones.

Para facilitar la interpretación del gráfico de densidad, utilizamos la variable logarítmica del salario por hora (véase gráfico 2).

Tal como venimos mencionando, observamos que la curva tiene un sesgo hacia la derecha, indicando que la media es mayor que la mediana salarial producto de los valores en la cola superior. Evidenciamos un pico alrededor de un salario horario de $COL 2.980, y una mayor dispersión entre este último y $ln(10)$ que es igual a $COL 22.026, lo que indica que la mayoría de las observaciones se encuentran dentro de este rango.

Ahora bien, veamos cómo se modifican los estadísticos si desagregamos por un conjunto de variables. La tabla 2 resume esta información por género.

Para el caso de las mujeres, el salario por hora medio es de $COL 7.808,22, mientras que para los varones es de $COL 8.123,18, es decir que en promedio el género masculino gana un 4 % que el femenino según la muestra trabajada. Sin embargo, la dispersión para los primeros es mayor que para estas últimas, como así también los valores mínimo y máximo. Por lo tanto, existe una brecha de género que la media no termina de capturar, pero que se demuestra en la variabilidad de estos datos. El gráfico 3 ilustra la distribución por sexo utilizando el logaritmo del salario.

Como podemos observar, la curva de densidad para los varones presenta un grado mayor de dispersión que la equivalente para mujeres, además de tener un mayor sesgo hacia la derecha.

Por otro lado, la tabla 3 muestra las estadísticas descriptivas dividido si la persona tiene un empleo formal, o dicho de otra manera, si cuenta con un registro en la seguridad social.

Un empleado formal tiene un salario horario medio que equivale a más de dos ingresos de los informales. Especificamente, la remuneración esta por encima en un 119 %. Sin embargo, hay una diferencia significativa entre el total de formales, 1.886.900, y el total de no formales, 572.823. Esto nos indica 2 cosas: (i) con los filtros realizados, capturamos una porción importante de los empleados registrados, pero no ocurre lo mismo con los no registrados en la seguridad social, y (ii) los informales pueden tener ingresos provenientes de otras ocupaciones, incluso más de una, por ende la variable de referencia seleccionada, el salario por hora mensual, no sería la adecuada para capturar este segmento de la población. El gráfico 4 vislumbra la distribución de lo mencionado previamente.

Tomando el logaritmo natural del salario por hora mensual, los formales presentan un sesgo hacia la derecha, y los salarios están concentrada entre $ln(8)$ y $ln(10)$, mientras que para los

no formales la dispersión es más alta, indicando una mayor heterogeneidad, y también cuentan con una asimetría positiva pero menos marcada.

Por último, incorporamos al análisis dos gráficos más donde vemos la distribución del logaritmo del salario por hora mensual según el estrato socioeconómico y el máximo nivel educativo alcanzado. Tomando el primero, la GEIH define 6 grupos para el *estrato socioeconómico* en función de la zona donde se encuentre la persona encuestada. Si esta reside en una de las 13 áreas metropolitanas, entonces se define su grupo social en función del consumo energético, mientras que, si vive en las cabeceras o en las zonas rurales, para la clasificación se utiliza el Índice de Calidad de Vida, integrando factores de bienestar vinculados a accesibilidad de servicios, condiciones de hábitat e ingresos (véase gráfico 5).

En primer lugar, observamos mayor presencia de los 3 primeros estratos socioeconómicos, es decir, los más vulnerables. Estos grupos se encuentran concentrados alrededor de un salario horario mensual de $COL 2.980, aunque el tercer sextil presenta una mayor dispersión. Por otra parte, los 3 sextiles superiores tienen un volumen de observaciones mucho menor, presentan una mayor duspersión en comparación a los estratos inferiores, y su moda esta alrededor de una remuneración por hora mensual de $COL 22.026.

El gráfico 6 ilustra la distribución salarial por máximo nivel educativo alcanzado. Las categorías van desde *sin instrucción* hasta *nivel superior*. Cabe destacar que, dado el filtrado de datos, el nivel faltante es *preescolar completo*.

Estas curvas de densidad tienen en común que todas tienen un pico en $ln(8)$ que es igual a $COL 2.980. Sin embargo, para los niveles educativos más bajos la densidad es mucho más baja, mientras que es mayor para los dos niveles más altos, *secundario completo y superior*. Este último, cuenta con una dispersión más alta que el resto, producto de la heterogeneidad de salarios entre personas con nivel superior, ya sea que este fuera completado o no.

En síntesis:

1. La muestra presenta importante variabilidad en los salarios por hora mensuales, además de contar con una asimetría positiva en la distribución.

2. Existen diferencias por género no captadas en las medias salariales, pero sí en su dispersión.

3. Con los filtros realizados y la variable de ingresos elegida captamos bien a los trabajadores formales, pero no así a los informales. Además, un trabajador formal tiene un salario por hora medio hasta dos veces más grande que un informal.

4. Los 3 sextiles inferiores tienen mayor presencia en el conjunto de datos y están más concentrados, mientras los 3 sextiles superiores están más dispersos y cuentan con menor volumen de observaciones. Estos últimos evidencian un salario mayor que los grupos más vulnerables.

5. Secundario completo y nivel superior son los dos máximos niveles educativos con mayor densidad en la distribución. Respecto a todos los niveles, estos están concentrados alrededor del mismo punto, pero quienes cuentan con formación terciaria/universitaria presentan mayor heterogeneidad.

3. Predicting wages

El modelo con el que buscamos predecir el salario por hora mensual es el siguiente:

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

Como ya mencionamos, w es el salario por hora mensual que percibe un trabajador. $f(X)$ es el conjunto de predictores, donde las variables elegidas son: la edad, el género, si es trabajador formal, el estrato socioeconómico, el máximo nivel educativo y si percibe subsidio. Por último u es un término error.

En primer lugar, vamos a dividir aleatoriamente el conjunto de datos en dos: un conjunto de entrenamiento, que cumplirá la función de insumo para que el modelo aprenda de los datos y ajuste sus parámetros, y un conjunto de prueba, utilizado para evaluar la capacidad del modelo. Para el primero será asignado un 70 % de la muestra, y para el testing el 30 % restante.

Segundo, estimamos el modelo a través de una iteración donde vamos agregando los predictores mencionados arriba. En total se corren 6 modelos lineales. La tabla 4 resume los resultados especificando el RMSE (raiz cuadrada del error cuadrático medio) tanto del conjunto de entrenamiento como del de prueba. Adicionalmente, se incorpora una columna llamada *Error*, la cual se obtiene de dividir el RMSE de los datos de testing respecto de la media de la variable dependiente del mismo conjunto, permitiendo verificar cuánto la estimación se aleja de la media (expresado en porcentajes).

Partiendo de un modelo de regresión lineal simple, con la edad como único predictor, el RMSE en el entrenamiento es de 11.508,82 unidades frente a las 11.521,53 unidades del RMSE del testing. El coeficiente de error es de 1,5 %. Si vamos incorporando variables, observamos que los errores de ambos conjuntos se reducen hasta el último modelo lineal que cuenta con 6 predictores y tiene un RMSE de 9.485,92 para el conjunto de entrenamiento y de 9.766,92 para el conjunto de prueba, siendo el coeficiente de error de 1,27 %.

Adicionalmente, generamos una transformación polinómica en el modelo con todos las variables explicativas para encontrar una versión que alcance el RMSE mínimo en el conjunto de testing. Una simplificación de la ecuación sería la siguiente:

$$w = \beta_1 X + \beta_2 X^2 + ... + \beta_p X^p + u$$

Los grados del polinomio se modificarán iterativamente iniciando por un grado 2 hasta un modelo con grado 6. La tabla 5 ilustra los resultados.

Primero, evidenciamos que el número de parámetros se incrementa exponencialmente donde, por ejemplo, en el último modelo la cantidad de predictores se incrementa a 924. Sin embargo, aquí empiezan a diverger los RMSE de ambos conjuntos. Para el entrenamiento, la raiz del error cuadrático medio se reduce hasta tender a un valor levemente superior a las 7.000 unidades, mientras que en el testeo el RMSE disminuye en el modelo de grado 2, pero a partir de la transformación de grado 3 este último se incrementa, teniendo un salto exponencial en el último de grado 6. El gráfico 7 ilustra la complejidad de todos los modelos frente a la raiz del error cuadrático medio.

Los modelos con mejor performance predictiva, es decir, menor RMSE son el modelo 7, que tiene un grado polinómico igual a 2, y el modelo 8, con un grado polinómico igual a 3. Ambos tienen los menores RMSE con 9.133,29 y 9.228,33 unidades, respectivamente. Además, el coeficiente de error es de 1,19 % para el primero, y 1,20 % para el segundo.

Referencias