Predicción de Ingresos: Análisis del Mercado Laboral en Bogotá

Harold Stiven Acuña

José David Cuervo
José David Dávila
César Augusto Alfaro

3 de marzo de 2025

Resumen

Este estudio emplea técnicas de modelado estadístico y aprendizaje automático para analizar los determinantes del ingreso laboral en Bogotá, utilizando datos de la Gran Encuesta Integrada de Hogares (GEIH) de 2018. Mediante el desarrollo de modelos predictivos, exploramos el perfil edad-salario, la brecha salarial de género y la capacidad predictiva de diferentes especificaciones estadísticas. Nuestros resultados revelan una relación no lineal entre edad e ingresos, con un pico salarial alrededor de los 48 años, y evidencian una persistente brecha salarial de género que permanece incluso después de controlar por características educativas y laborales. Los modelos basados en ensambles (Random Forest) muestran el mejor desempeño predictivo.

Palabras clave: predicción de ingresos, brecha salarial, economía laboral, aprendizaje automático
Clasificación JEL: J31, C53, J16

Repositorio GitHub:
https://github.com/alfarocesar/BDML_Predicting-income_Equipo8

1. Introducción

En el sector público, la precisión en la declaración de ingresos individuales es fundamental para el cálculo de impuestos. Sin embargo, el fraude fiscal de todos los tipos ha sido siempre un problema importante. Según el Servicio de Impuestos Internos (IRS), aproximadamente el 83,6 % de los impuestos en Estados Unidos se pagan voluntaria y puntualmente (Internal Revenue Service, 2019). Una de las causas de esta brecha es la subdeclaración de ingresos por parte de los individuos. Un modelo de predicción de ingresos podría potencialmente ayudar a identificar casos de fraude que podrían llevar a la reducción de esta brecha. Además, permite identificar a individuos y familias vulnerables que necesitan asistencia adicional.

El objetivo principal de este estudio es construir un modelo de predicción del salario horario individual: $w = f(X) + u$, donde w es el salario horario y X es una matriz de variables explicativas o predictores. En este trabajo, nos enfocamos en la forma funcional $f(X) = X\beta$, siguiendo el enfoque de mínimos cuadrados ordinarios, así como técnicas más avanzadas de machine learning.

El análisis se centra en tres áreas principales: (1) el perfil edad-salario, donde exploramos cómo los ingresos evolucionan a lo largo del ciclo de vida laboral; (2) la brecha salarial de género, examinando las diferencias de ingresos entre hombres y mujeres; y (3) la predicción de ingresos mediante diversos modelos estadísticos y de aprendizaje automático.

Para abordar estos objetivos, empleamos datos de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 para Bogotá, recopilados por el Departamento Administrativo Nacional de Estadística (DANE). Esta encuesta proporciona información detallada sobre la situación laboral, ingresos, educación y características sociodemográficas de los individuos, lo que la hace adecuada para nuestros propósitos analíticos.

El resto del documento está organizado de la siguiente manera: la Sección 2 describe los datos utilizados, incluyendo el proceso de adquisición, limpieza y variables seleccionadas;

la Sección 3 analiza el perfil edad-salario; la Sección 4 examina la brecha salarial de género; la Sección 5 desarrolla modelos predictivos de ingresos y evalúa su desempeño; y finalmente, la Sección 6 presenta las conclusiones del estudio.

2. Datos

2.1. Descripción y Adquisición de los Datos

Los datos utilizados en este análisis provienen de la Gran Encuesta Integrada de Hogares (GEIH) de 2018, específicamente para la ciudad de Bogotá. La GEIH es una encuesta de carácter continuo realizada por el DANE con el propósito de proporcionar información básica sobre el tamaño y estructura de la fuerza de trabajo del país, así como las características sociodemográficas de la población colombiana.

Para la adquisición de los datos, desarrollamos un algoritmo automatizado que identifica y extrae las tablas contenidas en cada una de las diez páginas HTML disponibles en el sitio https://ignaciomsarmiento.github.io/GEIH2018_sample/. El proceso consistió en realizar solicitudes HTTP a cada URL, procesar el contenido HTML recibido y extraer las estructuras tabulares utilizando las bibliotecas `httr` y `rvest` en R. El código implementado respeta las políticas del sitio web, incluyendo pausas entre solicitudes para evitar sobrecargar el servidor.

El conjunto de datos original contiene información de 32,177 individuos y 178 variables que abarcan aspectos demográficos, educativos, laborales y económicos. Esta riqueza de información permite un análisis detallado de los determinantes del ingreso laboral en la capital colombiana.

2.2. Proceso de Limpieza y Variables Incluidas

El proceso de limpieza de datos comprendió varias etapas orientadas a obtener una muestra analítica adecuada para la modelación de ingresos laborales. Siguiendo los cri-

terios establecidos en las instrucciones del problema, filtramos la muestra para incluir únicamente a individuos mayores de 18 años y que estuvieran empleados al momento de la encuesta. Esta restricción redujo la muestra a aproximadamente 16,682 observaciones.

Posteriormente, creamos las variables clave para nuestro análisis:

1. **Salario por hora (hourly_wage)**: Calculado como el ingreso laboral mensual dividido por el número de horas trabajadas mensualmente (considerando 4.345 semanas por mes).
2. **Logaritmo del salario por hora (log_hourly_wage)**: Transformación que permite normalizar la distribución de los salarios y facilita la interpretación de los coeficientes como elasticidades o cambios porcentuales.
3. **Variable indicadora de género (female)**: Recodificada para que tome el valor de 1 para mujeres y 0 para hombres.
4. **Variables para el análisis del perfil edad-salario**: Incluimos tanto la edad como su término cuadrático.

Para garantizar la calidad de los datos, implementamos un proceso de detección y manejo de valores atípicos (outliers) en los salarios. Calculamos los z-scores para la variable de salario por hora y excluimos las observaciones con valores absolutos superiores a 3. Finalmente, eliminamos las observaciones con valores faltantes en las variables clave para el análisis, resultando en un conjunto de datos final con 16,015 observaciones.

Las variables seleccionadas para nuestro análisis incluyen:

- **Variable dependiente**: log_hourly_wage (logaritmo natural del salario por hora)
- **Variables independientes clave**: age (edad), age_squared (edad al cuadrado), female (género)
- **Variables de control**: educ_level (nivel educativo), formal_work (formalidad laboral), totalHoursWorked (horas trabajadas), sizeFirm (tamaño de la empresa)

El análisis descriptivo de las variables muestra que la edad promedio de los trabajadores en la muestra es de aproximadamente 34.4 años, con una desviación estándar de 20.9 años. La proporción de mujeres es del 47.9 %, y el promedio de horas trabajadas semanales es de 47.2 horas. La tasa de formalidad laboral es del 58.2 %, y el salario por hora presenta una distribución altamente heterogénea, lo que justifica nuestra transformación logarítmica.

3. Perfil Edad-Salario

Para analizar cómo evoluciona el salario a lo largo del ciclo de vida laboral, estimamos la siguiente regresión:

$$\log(w) = \beta_1 + \beta_2 \cdot Age + \beta_3 \cdot Age^2 + u$$

Los resultados indican que la constante (β_1) de 13.060 representa el valor esperado del logaritmo del ingreso cuando la edad es cero (un valor teórico sin interpretación práctica). El coeficiente de la edad (β_2) es de 0.038, estadísticamente significativo, lo que indica que, inicialmente, cada año adicional de edad está asociado con un aumento del 3.8 % en el salario horario, manteniendo constantes otros factores.

El coeficiente del término cuadrático de la edad (β_3) es de -0.0004, también estadísticamente significativo, lo que confirma una relación cóncava entre edad y salario. Esto implica que los ingresos aumentan con la edad hasta llegar a un punto máximo, para luego comenzar a decrecer.

Para identificar la edad a la que se alcanza el salario máximo, derivamos la ecuación con respecto a la edad e igualamos a cero:

$$\frac{d(\log(w))}{dAge} = \beta_2 + 2\beta_3 \cdot Age = 0$$

Despejando, obtenemos:

$$Age_{max} = -\frac{\beta_2}{2\beta_3} = -\frac{0{,}038}{2(-0{,}0004)} = 47{,}5$$

Por lo tanto, según nuestro modelo, los salarios aumentan hasta aproximadamente los 48 años y posteriormente disminuyen. Este resultado es consistente con la literatura económica, en particular con **?**, quien sostiene que los ingresos tienden a ser relativamente bajos en los primeros años laborales, incrementan gradualmente con la experiencia adquirida, y alcanzan su punto máximo entre los 45 y 54 años.

El coeficiente de determinación (R^2) del modelo es de 0.022, lo que indica que la edad por sí sola explica una pequeña proporción de la variabilidad en los salarios. Esto sugiere la necesidad de incluir variables adicionales para capturar otros determinantes importantes del ingreso laboral, como la educación, la experiencia específica y las habilidades.

Finalmente, las pruebas de diagnóstico revelaron la presencia de heterocedasticidad en el modelo (p-valor <0.05), lo que señala la necesidad de utilizar errores estándar robustos o metodologías alternativas para abordar este problema en la estimación de los salarios.

4. Brecha Salarial de Género

Para analizar la brecha salarial de género, estimamos dos modelos: uno incondicionado, que captura la diferencia bruta en salarios entre hombres y mujeres, y otro condicionado, que controla por diversas características individuales y laborales.

En el modelo incondicionado, estimamos:

$$\log(w) = \beta_0 + \beta_1 \cdot Female + u$$

Los resultados indican que, en promedio, las mujeres ganan un 13.1 % menos que los hombres (coeficiente de -0.131, estadísticamente significativo al 1 %). Esta brecha bruta refleja las diferencias totales en salarios sin considerar posibles factores explicativos

relacionados con características observables de los individuos.

Para examinar si esta brecha se mantiene después de controlar por diferencias en características individuales y laborales, estimamos un modelo condicionado:

$$\log(w) = \beta_0 + \beta_1 \cdot Female + \beta_2 \cdot Age + \beta_3 \cdot Age^2 + \beta_4 \cdot X + u$$

donde X representa un vector de variables de control que incluye nivel educativo, experiencia laboral, formalidad laboral, sector económico, ocupación y horas trabajadas.

Los resultados del modelo condicionado muestran que, incluso después de controlar por estas características, persiste una brecha salarial de género del 7.8 % (coeficiente de -0.078, estadísticamente significativo al 1 %). Esta brecha residual podría atribuirse a factores no observables o a discriminación en el mercado laboral.

Utilizando la técnica de Frisch-Waugh-Lovell (FWL) para estimación por etapas, obtuvimos resultados similares, confirmando la robustez de nuestras estimaciones. Adicionalmente, implementamos un procedimiento de bootstrap con 1,000 repeticiones para obtener intervalos de confianza robustos para la brecha salarial condicionada. El intervalo de confianza al 95 % para el coeficiente de la variable Female es [-0.102, -0.054], lo que confirma la significancia estadística de la brecha salarial.

Al examinar cómo la brecha varía a lo largo del ciclo de vida, encontramos que ésta es más pronunciada durante los años de mayor potencial productivo (entre los 30 y 45 años) y se reduce en edades más avanzadas. Esto podría estar relacionado con factores como la maternidad y las responsabilidades familiares, que afectan de manera desproporcionada a las mujeres durante las etapas medias de su carrera laboral.

También analizamos la interacción entre género y nivel educativo, encontrando que la brecha salarial es menor entre individuos con niveles educativos más altos, aunque no desaparece completamente. Esto sugiere que la educación superior es un factor mitigante, pero no eliminador, de las disparidades salariales de género.

5. Predicción de Ingresos

Para evaluar la capacidad predictiva de diferentes especificaciones estadísticas, dividimos la muestra en conjuntos de entrenamiento (70 %) y prueba (30 %), estableciendo una semilla (10101) para garantizar la reproducibilidad. Implementamos diversos modelos predictivos, desde especificaciones lineales simples hasta métodos avanzados de aprendizaje automático.

Comenzamos con un modelo lineal básico (Modelo 1) que incluye solo la edad y su término cuadrático como predictores:

$$\log(w) = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Age^2 + u$$

Luego, incrementamos progresivamente la complejidad de las especificaciones. El Modelo 2 incorpora variables adicionales como género, horas trabajadas, formalidad laboral y nivel educativo:

$$\log(w) = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Age^2 + \beta_3 \cdot Female + \beta_4 \cdot Hours + \beta_5 \cdot Formal + \beta_6 \cdot Educ + u$$

El Modelo 3 explora interacciones entre variables, permitiendo que el efecto de la edad varíe según el nivel educativo, y que el impacto del género difiera según la formalidad laboral:

$$\log(w) = \beta_0 + \beta_1 \cdot Age \cdot Educ + \beta_2 \cdot Female \cdot Formal + u$$

Además de estos modelos lineales, implementamos técnicas de aprendizaje automático. El Modelo 4 utiliza Random Forest, un método de ensamble que construye múltiples árboles de decisión y promedia sus predicciones, mientras que el Modelo 5 emplea una

red neuronal (Nnet) con cinco nodos en la capa oculta.

Para cada modelo, calculamos el Error Cuadrático Medio (RMSE) en el conjunto de prueba. Los resultados muestran un desempeño diferenciado entre las distintas especificaciones:

- Modelo 1 (Lineal básico): RMSE = 0.853
- Modelo 2 (Lineal expandido): RMSE = 0.684
- Modelo 3 (Con interacciones): RMSE = 0.697
- Modelo 4 (Random Forest): RMSE = 0.617
- Modelo 5 (Red Neuronal): RMSE = 0.649

El Modelo 4 (Random Forest) emerge como la especificación con mejor capacidad predictiva, con un RMSE significativamente menor que los modelos lineales. Esto sugiere que las relaciones entre las variables predictoras y el salario son complejas y no lineales, beneficiándose del enfoque flexible que ofrece Random Forest.

Examinamos también los errores de predicción del mejor modelo, identificando patrones interesantes. Los mayores errores de predicción tienden a concentrarse en los extremos de la distribución salarial, especialmente para individuos con ingresos muy altos o muy bajos. Los trabajadores informales y aquellos en ocupaciones atípicas también presentan mayores errores de predicción, lo que podría reflejar la mayor variabilidad y menor regularidad en sus patrones de ingresos.

Para validar adicionalmente nuestros resultados, aplicamos Leave-One-Out Cross-Validation (LOOCV) a los dos mejores modelos. Esta técnica, que ajusta el modelo n veces (dejando una observación fuera cada vez), proporciona una evaluación más robusta del error de predicción. Los resultados de LOOCV confirmaron la superioridad del Random Forest, aunque también revelaron que la diferencia en desempeño entre este modelo y la red neuronal se reduce cuando se considera un criterio de validación más riguroso.

Un análisis de la influencia de las variables en el modelo Random Forest (a través de la importancia de las variables) reveló que el nivel educativo, la edad y la formalidad laboral son los predictores más importantes del salario, seguidos por el tamaño de la empresa y el género. Esta jerarquía de importancia es consistente con la literatura económica sobre determinantes del ingreso laboral.

6. Conclusiones

Este estudio ha analizado los determinantes del ingreso laboral en Bogotá, explorando el perfil edad-salario, la brecha salarial de género y desarrollando modelos predictivos mediante diversas técnicas estadísticas y de aprendizaje automático. Las principales conclusiones son:

1. El perfil edad-salario muestra una relación cóncava, con salarios que aumentan hasta aproximadamente los 48 años y luego disminuyen. Este patrón es consistente con la teoría del capital humano y estudios previos sobre ciclos de vida laborales.

2. Existe una persistente brecha salarial de género en el mercado laboral bogotano. Las mujeres ganan, en promedio, un 13.1 % menos que los hombres. Esta brecha se reduce al 7.8 % al controlar por características individuales y laborales, pero sigue siendo estadísticamente significativa, lo que sugiere la presencia de factores no observables o discriminación.

3. En términos de capacidad predictiva, los modelos basados en técnicas de aprendizaje automático, particularmente Random Forest, superan a las especificaciones lineales tradicionales. Esto indica que las relaciones entre características individuales e ingresos son complejas y no lineales.

4. Los principales determinantes del ingreso laboral son el nivel educativo, la edad, la formalidad laboral y el tamaño de la empresa. El género, aunque significativo, tiene un impacto relativamente menor en comparación con estas variables.

5. Los mayores errores de predicción se concentran en los extremos de la distribución salarial y en sectores con mayor informalidad, lo que sugiere áreas donde los modelos podrían mejorarse en futuros estudios.

Estos hallazgos tienen implicaciones importantes tanto para la detección de posibles casos de subdeclaración de ingresos como para la identificación de grupos vulnerables. La capacidad de predecir ingresos con razonable precisión podría ayudar a las autoridades fiscales a identificar discrepancias significativas entre ingresos declarados y esperados, contribuyendo a reducir la evasión fiscal. Asimismo, la persistencia de la brecha salarial de género señala la necesidad de políticas específicas para promover la equidad laboral.

Futuras investigaciones podrían expandir este análisis incorporando datos longitudinales para capturar la dinámica temporal de los ingresos, explorando técnicas de aprendizaje profundo para mejorar aún más la capacidad predictiva, y desarrollando análisis más detallados de subgrupos específicos dentro del mercado laboral.

Referencias

Anthropic (2023). Conversaciones con claude ai assistant. Utilizado para estructura de documentos LaTeX, revisión de código y refinamiento de texto.

Departamento Administrativo Nacional de Estadística (DANE) (2018). Gran encuesta integrada de hogares (geih) 2018. Datos de encuesta para Colombia.

Internal Revenue Service (2019). The tax gap: Tax gap estimates for tax years 2008-2010. Consultado el 25 de febrero de 2025.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, NY, 2 edition.

OpenAI (2023). Conversaciones con chatgpt. Utilizado para verificación de sintaxis, correcciones ortográficas y asistencia en la redacción.

Sarmiento-Barbieri, I. (2023a). The bootstrap. Notas de clase, Big Data y Machine Learning para Economía Aplicada, Universidad de los Andes.

Sarmiento-Barbieri, I. (2023b). Computing the ols coefficients. Notas de clase, Big Data y Machine Learning para Economía Aplicada, Universidad de los Andes.

Sarmiento-Barbieri, I. (2023c). Introduction to resampling. Notas de clase, Big Data y Machine Learning para Economía Aplicada, Universidad de los Andes.

Sarmiento-Barbieri, I. (2023d). Lógica básica regresión líneal. Notas de clase, Big Data y Machine Learning para Economía Aplicada, Universidad de los Andes.

Sarmiento-Barbieri, I. and Martínez-Gutiérrez, G. (2023a). Regularization. Notas de clase, Big Data y Machine Learning para Economía Aplicada, Universidad de los Andes.

Sarmiento-Barbieri, I. and Martínez-Gutiérrez, G. (2023b). Resampling methods and training with caret. Notas de clase, Big Data y Machine Learning para Economía Aplicada, Universidad de los Andes.

Sarmiento-Barbieri, I. and Martínez-Gutiérrez, G. (2023c). Subset selection. Notas de clase, Big Data y Machine Learning para Economía Aplicada, Universidad de los Andes.

Sarmiento-Barbieri, I. and Rojas, J. (2023). Coordinate descent. Notas de clase, Big Data y Machine Learning para Economía Aplicada, Universidad de los Andes.