EPH DSOCUPACION, OCUPACION Y SARIO SEMANAL DE LA PATAGONIA (2004 Y 2024)

MEDIA DE EDAD, EDACION MUJER:

Valores NO nulos por variable:

edad 5523

edad2 5523

educ 2407

mujer 5523

ocupación 5523

dtype: int64

📌 Valores únicos en 'ocupación':

ocupación

NaN 3264

1.0 2407

3.0 2357

4.0 644

2.0 115

Name: count, dtype: int64

📊 Tabla de diferencia de medias entre Train y Test:

Media Train Media Test Diferencia

edad 41.242 40.638 0.604

edad2 1849.440 1793.362 56.078

educ 0.996 1.011 -0.015

mujer 0.446 0.461 -0.015

La tabla muestra que las medias de las variables en los conjuntos de entrenamiento y testeo son bastante similares. Las diferencias son relativamente pequeñas en todas las variables. Por ejemplo, la edad difiere en menos de un año (0.604), lo que indica que la distribución etaria es consistente en ambos sets. De manera similar, la educación (educ) y la proporción de mujeres (mujer) presentan diferencias mínimas (-0.015 en ambos casos), sugiriendo que la división de datos no ha introducido sesgos significativos en estas características. La variable edad2 (edad al cuadrado) presenta una diferencia numérica mayor (56.078), pero en términos proporcionales a su magnitud, sigue siendo un desvío aceptable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Modelo i | Modelo ii | Modelo iii | Modelo iv | Modelo v |
| Edad | 3199.737 (889.83) | 17875.539 (5273.26) | 17094.704 (5291.43) | 19284.784 (5259.79) | 18884.674 (5218.49) |
| Edad2 |  | -170.386 (60.35) | -165.015 (60.4) | -190.753 (60.05) | -186.188 (59.57) |
| Educación |  |  | -30736.72 (18504.9) | -20693.488 (18431.39) | -10550.719 (18385.3) |
| mujer |  |  |  | -119527.465 (21512.14) | -124786.039 (21378.12) |
| Obra social |  |  |  |  | -30391.369 (6398.83) |
| Jefe/a de hogar |  |  |  |  | 66565.176 (25826.38) |

Acá lo que hicimos fue intentar predecir el salario semanal usando varias cosas de las personas. La tabla nos muestra qué tanto "tira para arriba" o "para abajo" cada cosa en el salario, y si ese efecto es confiable.

Estos modelos nos ayudan a ver qué "pesa" más en el salario semanal: la edad y la edad2 nos dicen que el salario tiene una curva con la edad, ser mujer se asocia a un salario más bajo, y ser jefe/a lo dispara para arriba. La educación y la obra social tienen un efecto más complejo o menos claro en este caso particular.

Performance por regresión lineal de la predicción de salarios usando la base de testeo:

Performance por regresión lineal de la predicción de salarios usando la base de testeo

Var. Dep: salario semanal

Modelo MSE test RMSE test MAE test

i 1.121826e+11 334936.77 293084.68

ii 8.030168e+10 283375.52 228818.53

iii 9.134845e+10 302239.05 221376.65

iv 8.827014e+10 297102.91 219200.90

v 8.733093e+10 295518.07 214112.2

**MSE (Error Cuadrático Medio):** Cuanto más chico, mejor. Mide el promedio de los errores al cuadrado. Números grandes como 1.12e+11 significan que los errores pueden ser muy grandes.

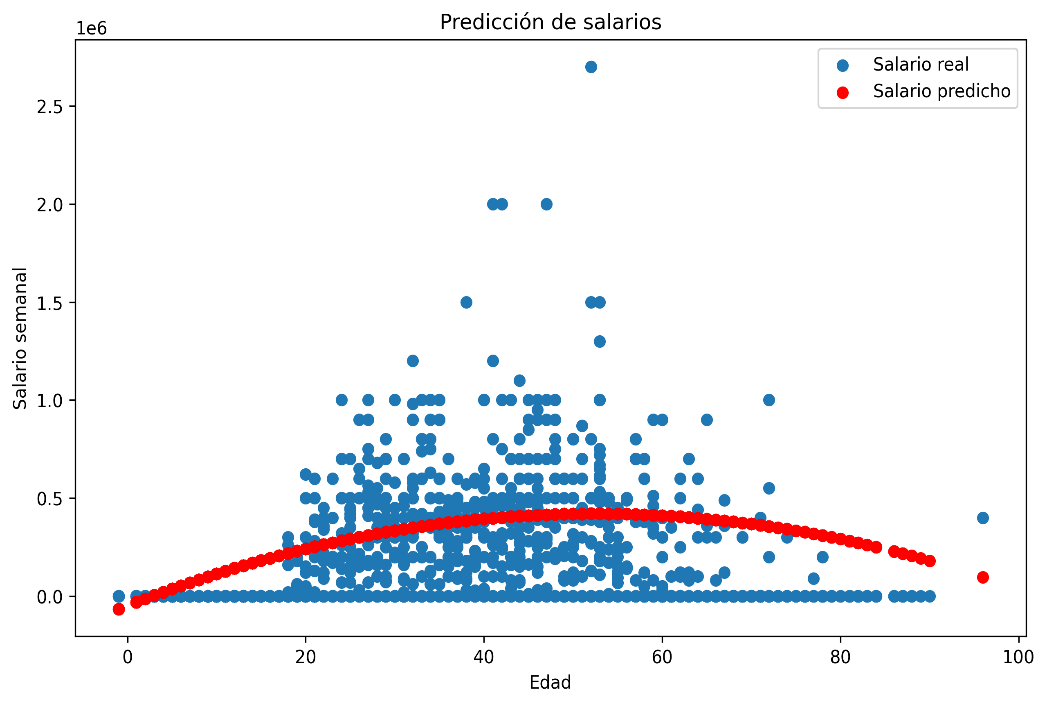
**RMSE (Raíz del Error Cuadrático Medio):** Es la raíz cuadrada del MSE, lo que lo pone en las mismas unidades que nuestro salario (pesos). También, cuanto más chico, mejor. Es más fácil de interpretar que el MSE.

**MAE (Error Absoluto Medio):** Cuanto más chico, mejor. Nos da el error promedio en términos absolutos, es decir, cuánto se equivocó el modelo en promedio, sin importar si fue por arriba o por abajo.

Conclusión de edad2 (Modelo ii) fue clave para mejorar sustancialmente la capacidad predictiva del modelo. Luego, la adición de mujer, Obra social y jefe/a en los modelos posteriores (especialmente el Modelo v) afinó aún más las predicciones, logrando los menores errores promedio. Aunque los errores siguen siendo numéricamente grandes (hablamos de salarios en unidades de pesos que pueden ser altos), la tendencia es a la mejora a medida que se incorporan más variables relevantes.

Gráfico de Dispersión de Salarios Predichos vs. Edad:

1)



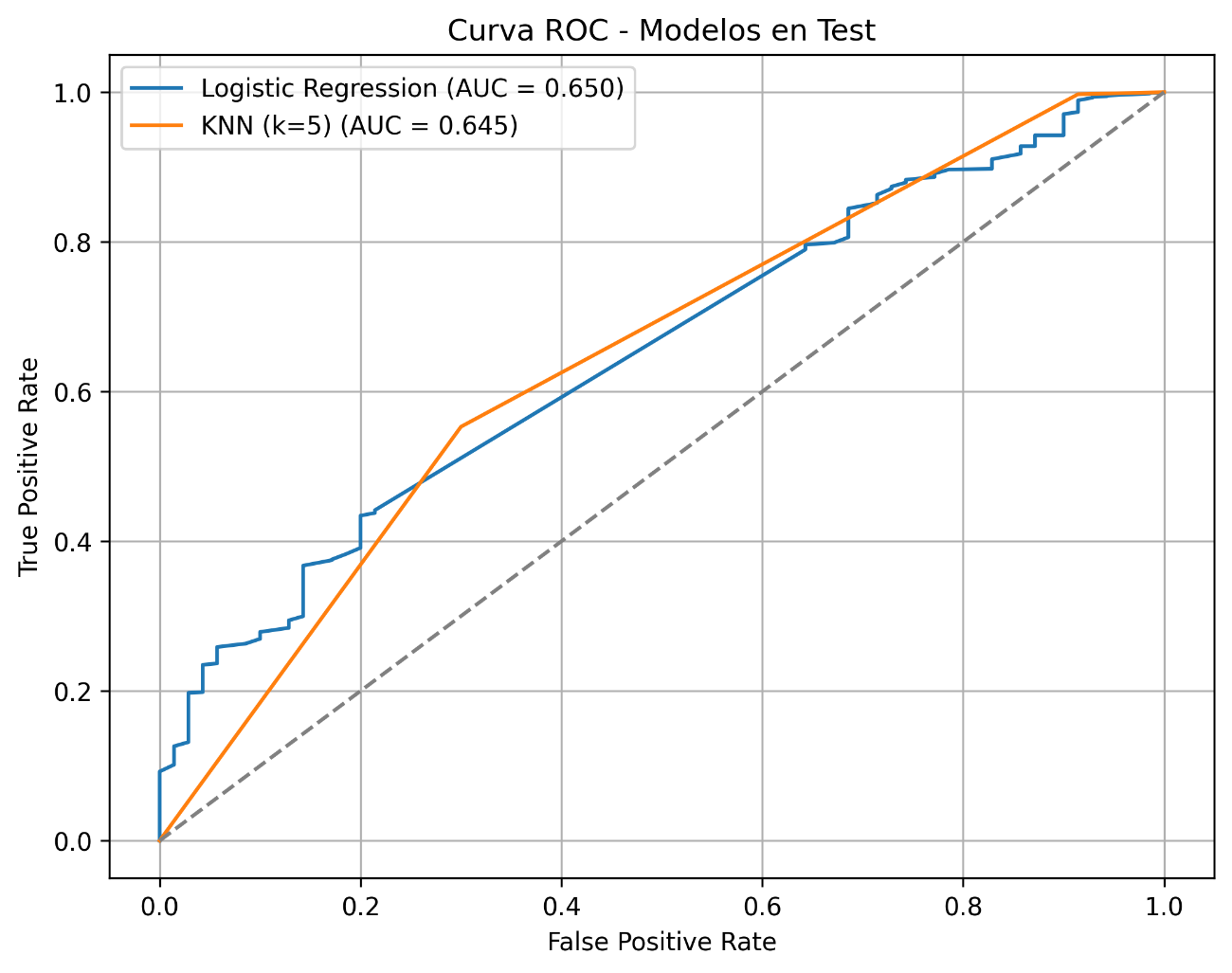
Puntos azules (Salario real): Cada punto azul representa el salario semanal real de una persona en nuestra base de datos de testeo, según su edad. Vemos que hay una gran dispersión en estos puntos. Hay personas de diferentes edades con salarios muy variados, y algunos con salarios extremadamente altos que se salen del patrón general. Esto nos dice que el salario real tiene mucha variabilidad y que no solo depende de la edad, sino de muchos otros factores no mostrados aquí.

Puntos rojos (Salario predicho): La línea formada por los puntos rojos representa el salario semanal que nuestro modelo predice para cada edad. Lo más importante aquí es la forma de arco o parábola que dibuja.

Esta curva roja ascendente y luego descendente refleja exactamente lo que habíamos comentado con los coeficientes de edad y edad2 en el Modelo 5. El modelo predice que los salarios aumentan con la edad hasta un punto (alrededor de los 40-50 años, donde la curva alcanza su punto más alto), y luego, a partir de esa edad, los salarios predichos empiezan a disminuir suavemente.

DESUCUPADOS Y OCUPADOS :

2)



Este gráfico nos muestra qué tan bien los modelos (Regresión Logística y KNN) pueden distinguir a los "desocupados" de los "no desocupados".

Línea Gris Punteada (diagonal): Esta línea representa un modelo que adivina al azar.

Líneas Azul (Regresión Logística) y Naranja (KNN). Cuanto más se alejen estas líneas de la línea gris y se acerquen a la esquina superior izquierda del gráfico, mejor es el modelo para diferenciar. Lalíneas azul y naranja están un poco por encima de la gris, lo cual es bueno, pero no están cerca de la esquina superior izquierda. Esto significa que tus modelos son un poco mejores que adivinar al azar, pero no son muy buenos para distinguir eficazmente a los desocupados.

Logistic Regression - Accuracy: 0.940

Logistic Regression - AUC: 0.650

KNN (k=5) - Accuracy: 0.940

KNN (k=5) - AUC: 0.645

Matriz de confusión:

[[ 0 70] [ 0 1094]]