TP4 – Regresión y Clasificación

## Integrantes: David Agudelo, Juan Tomasello y Franco Bustelo

**Tabla 2. Estimación por regresión lineal de salarios usando la base de entrenamiento**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Var. Dep: *salario\_semanal*  Variables | Modelo 1  (1) | Modelo 2  (2) | Modelo 3  (3) | Modelo 4  (4) | Modelo 5  (5) |
| *edad* | 207.924\*\*\*  (29.92) | 1141.612\*\*\*  (142.49) | 1142.524\*\*\*  (142,50) | 1178.789\*\*\* (140,585) | 1047.242\*\*\* (155.17) |
| *edad2* |  | -11.736\*\*\*  (1.66) | -11,750\*\*\* (1,66) | -12.173\*\*\* (1,63) | -11.008\*\*\* (1.75) |
| *educ* |  |  | 65,945 (72,39) | 56,626 (71,38) | 58.049 (70.87) |
| *Mujer* |  |  |  | -5028,269\*\*\* (698,55) | -4972.265\*\*\* (709.91) |
| *Estado civil\_2*  *Estado civil\_3*  *Estado civil\_4*  *Estado civil\_5* |  |  |  |  | 4066.256\*\*\* (973.65)  3042.810 \*\* (1521.96)  2612.815 (2325.59)  1852.749\* (1072.42) |
| escribe |  |  |  |  | -13900\*\*\*  (3876.988) |
|  |  |  |  |  |  |
| N (observaciones) | 2557 | 2557 | 2557 | 2557 | 2557 |
| *R2* | 0,026 | 0,046 | 0,046 | 0,074 | 0,09 |

Nota: destaque con \*, \*\*, y \*\*\* cuando el p-valor de los coeficientes reportados sean menor que 0.1, 0.05 y 0.001 respectivamente.

**Tabla 3. Performance por regresión lineal de la predicción de salarios usando la base de testeo**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Var. Dep: *salario\_semanal* | Modelo 1  (1) | Modelo 2  (2) | Modelo 3  (3) | Modelo 4  (4) | Modelo 5  (5) |
| *MSE test* | 591364718 | 581252431 | 586252962 | 574227462 | 561771502 |
| *RMSE test* | 24317 | 24109 | 24212 | 23963 | 23701 |
| *MAE test* | 14857 | 14631 | 14800 | 14654 | 14656 |

En esta tabla vemos los resultados de varios modelos que intentan explicar el salario semanal según variables/características de las personas. En general, la variable más fuerte es la edad: a más edad, mayor salario, pero como también se incluye la edad al cuadrado (edad2), eso muestra que llega un punto donde la edad ya no impacta tanto o incluso puede bajar el salario (porque el coeficiente pasa a ser negativo).

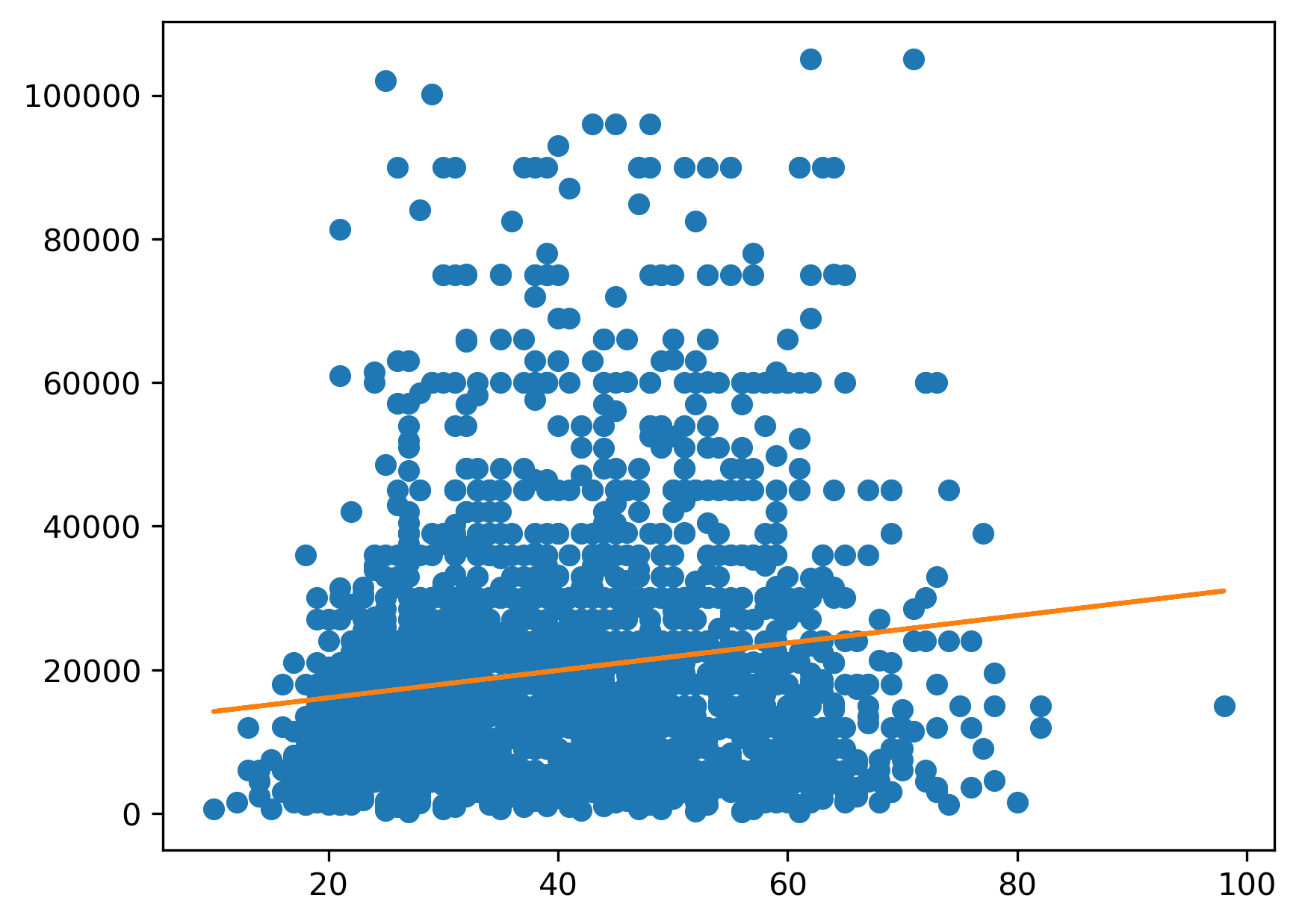
Después, estar casado (estado civil 2) o separado (estado civil 3) parece estar asociado con salarios más altos que quienes no están en esos estados, especialmente casado que tiene un coeficiente más grande y significativo. En cambio, ser viudo no tiene efecto claro (nada significativo) y ser soltero (estado civil 5) tiene un impacto positivo pero más chico.

El nivel educativo (educ) tiene un efecto positivo pero no llega a ser tan fuerte en estos modelos. Ser mujer parece tener un impacto negativo en el salario (coeficiente negativo y significativo), lo cual muestra una posible brecha salarial.

Lo más raro es la variable “escribe”: los que saben escribir (en esta codificación al revés, parece ser 1 si no sabe escribir) ganan mucho menos según el modelo, pero el valor es poco confiable por lo grande del error estándar.

Por último, los R² son bastante bajos (entre 0.026 y 0.065), o sea que los modelos explican poco del salario solo con estas variables.

**Punto 2 i)**

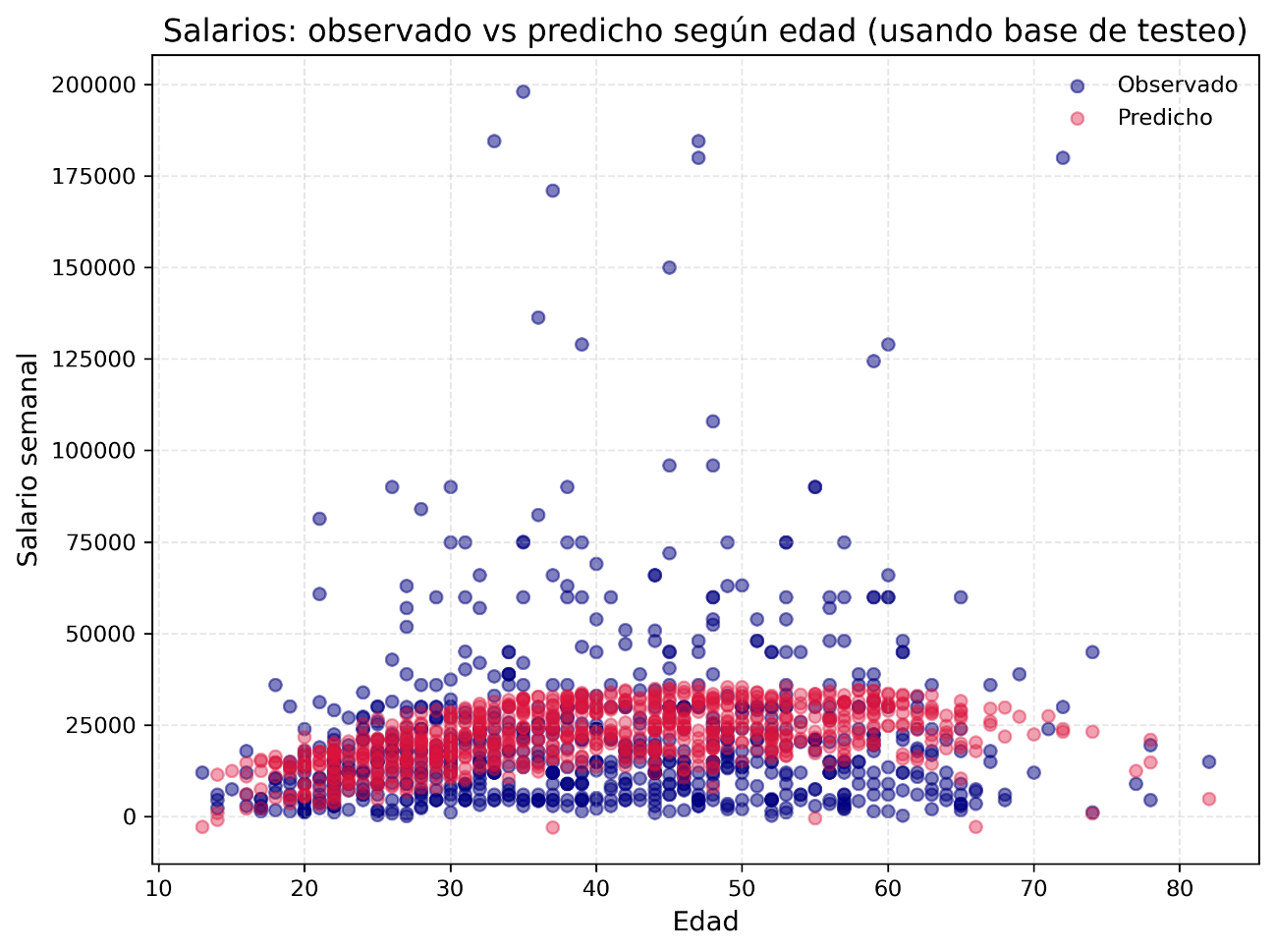


En el gráfico de arriba se muestra la relación entre la edad (eje X) y el salario semanal (eje Y). Cada punto azul representa a una persona, y la línea naranja es lo que aparece al aplicar una regresión lineal.

Para este histograma se usó el 70% de los datos para entrenar el modelo y el 30% para probarlo. Aunque la línea naranja muestra que a mayor edad tiende a subir el salario, la dispersión de los puntos indica que la edad no explica muy bien el salario semanal por sí sola. Esto sirve para ver si el modelo puede mejorar agregando más predictores.

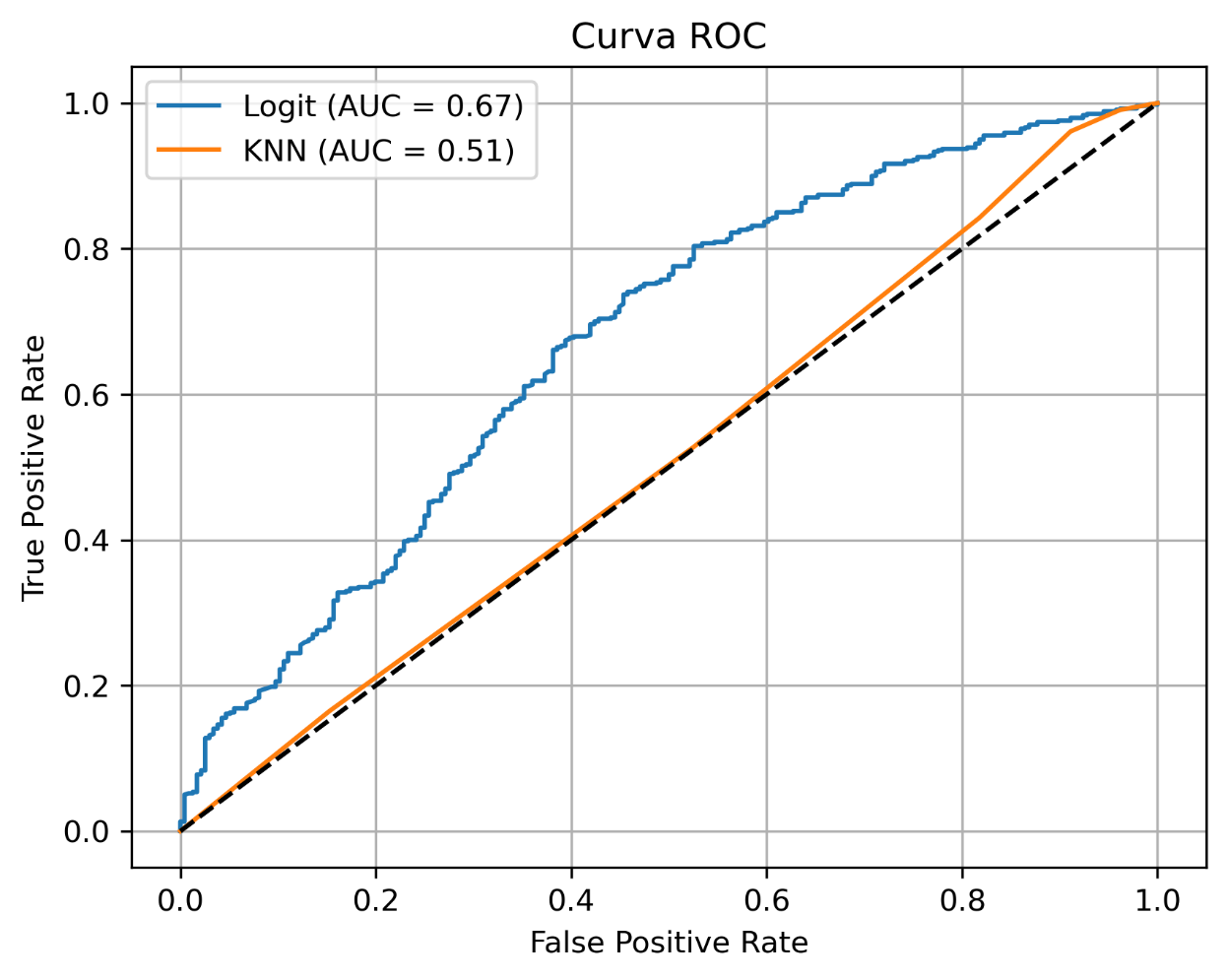
Antes de este grafico había otro que hicimos en Python pero a este le agregamos lo del percentil 99 para evitar outliers o los valores extremos.

**Punto 4**

****

En este gráfico se comparan los salarios reales observados con los que predijo el modelo, en función de la edad. Se ve que el modelo (los que están en rojo) sigue más o menos la tendencia general de los datos reales (los que están en azul), pero no llega a capturar bien los valores extremos (una especie de outliers). En general, el modelo predice mejor los sueldos más comunes, pero le cuesta con los casos atípicos. También se puede ver que los salarios no aumentan con la edad como uno esperaría, lo cual puede indicar que está relacionado con otros factores además de la edad.

**Punto 5**



En este grafico se ve que el modelo KNN tuvo un desempeño más bajo, con una precisión del 64% y un Auc de 0.51 (parecido al de tirar una moneda 0.5), lo cual vendría a indicar que no fue muy útil que digamos. Por el otro lado se ve que la regresión logística clasificó correctamente el 77% de los casos y el área bajo la curva ROC fue de 0.67, lo cual indica un desempeño, dentro de lo que cabe, **aceptable**.

También se ve claramente que la curva de **logit** está más alejada de la línea diagonal (que representa un modelo aleatorio), mientras que la curva de **KNN** está muy cerca de esa línea. Esto confirma que la regresión logística funciona mejor para este caso.

Con estas variables elegidas, la **regresión logística funciona mucho mejor que KNN** para predecir quién gana más de $10.000 semanales. La información como tener más educación o saber escribir está relacionada con salarios más altos, y la regresión logística fue mejor para ver eso.