Tema 6: Comparación, selección y evaluación de modelos

Minería de Datos

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades

¿Para que sirve la evaluación?

- Hemos construido un modelo o varios basados en los datos, pero necesitamos saber como de bueno es.
- Necesitamos comparar los datos que hemos obtenido.
- ¿Nuestro modelo generaliza?
- ¿Este modelo es útil en fase de explotación?

Selección de métricas

- El experto en aprendizaje automático de Fayrix habla de las **métricas de rendimiento** que se utilizan comúnmente en la ciencia de datos para evaluar y realizar los modelos de aprendizaje automático.
- 1º Entender la tarea:
  - Según los requisitos previos, debemos comprender qué tipo de problemas estamos tratando de resolver:
    - Clasificación
    - Regresión
    - Categorización

Clasificación

- Matriz de confusión

Esta matriz se utiliza para evaluar la precisión de un clasificador y se presenta en la tabla a continuación.

		Resultado de la predicción		
		Positivo	Negativo	
Valor actual	Positivo	TP	FN	TP + FN
	Negativo	FP	TN	FP + TN

Clasificación

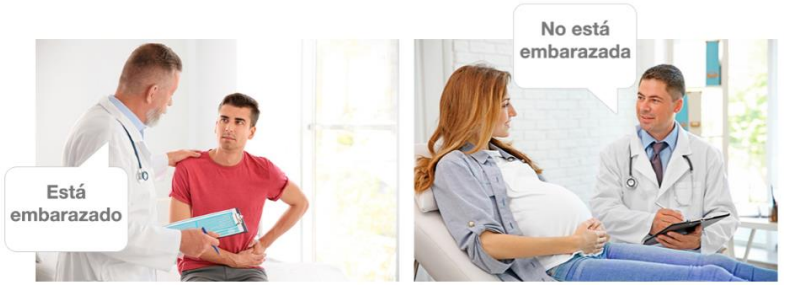
Matriz de confusión

Esta matriz se utiliza para evaluar la precisión de un clasificador y se presenta en la tabla a continuación.

		Resultado de la predicción: Positivo	Resultado de la predicción: Negativo	
Valor actual	Positivo	TP	FN	TP + FN
	Negativo	FP	TN	FP + TN

Error tipo I
(falso positivo)

Error tipo II
(falso negativo)



Clasificación

Exactitud

Indica el número de elementos clasificados correctamente en comparación con el número total.

Tenga en cuenta que la métrica de exactitud tiene limitaciones: no funciona bien con las clases desequilibradas que pueden tener muchos elementos de la misma clase e incluir algunas otras clases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP = total positivos

TN = total negativos

FP = falsos positivos

FN = falsos negativos

Clasificación

Exhaustivididad / Sensibilidad

La métrica de exhaustividad muestra la cantidad de verdaderos positivos que el modelo ha clasificado en función del número total de valores positivos.

$$recall = \frac{TP}{TP + FN}$$

TP = total positivos

TN = total negativos

FP = falsos positivos

FN = falsos negativos

Clasificación

Exhaustivididad / Sensibilidad

La métrica de exhaustividad muestra la cantidad de verdaderos positivos que el modelo ha clasificado en función del número total de valores positivos.

$$recall = \frac{TP}{TP + FN}$$

TP = total positivos

TN = total negativos

FP = falsos positivos

FN = falsos negativos

Clasificación

Precisión

Esta métrica representa el número de verdaderos positivos que son realmente positivos en comparación con el número total de valores positivos predichos.

$$precision = \frac{TP}{TP + FP}$$

TP = total positivos

TN = total negativos

FP = falsos positivos

FN = falsos negativos

Clasificación

Puntuación F1

Esta métrica es la combinación de las métricas de precisión y exhaustividad y sirve de compromiso entre ellas. La mejor puntuación F1 es igual a 1 y la peor a 0.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

Regresión

Error Medio Absoluto (EMA)

Esta métrica de regresión es el valor medio de la diferencia absoluta entre el valor real y el valor predicho.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |original_t - predict_t|$$

Regresión

Error Cuadrático Medio (ECM)

El error cuadrático medio (ECM) calcula el valor medio de la diferencia al cuadrado entre el valor real y el predicho para todos los puntos de datos.

En esta métrica, el impacto de los errores es mayor. Cuanto menor sea el ECM, más precisas serán nuestras predicciones. ECM = 1 es el punto óptimo.

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}(original_t - predict_t)^2$$

El MSE tiene algunas ventajas frente al MAE:

1. El ***MSE*** destaca grandes errores entre los pequeños.
2. El ***MSE*** es diferenciable, lo que ayuda a encontrar los valores mínimos y máximos utilizando los métodos matemáticos de manera más efectiva.

Regresión

Raíz del Error Cuadrático Medio (RECM)

El RECM es la raíz cuadrada del ECM. Es fácil de interpretar en comparación con el ECM y utiliza valores absolutos más pequeños, lo que es útil para los cálculos informáticos.

$$\text{RMS}E = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(original_t - predict_t)^2}$$

Clasificación por categorias

Coefficiente Tau de Kendall

El coeficiente tau de Kendall muestra la correlación entre las dos listas de elementos clasificados según el número de pares concordantes y discordantes: en cada caso tenemos dos rangos (máquina y predicción humana). En primer lugar, los elementos clasificados se convierten en una matriz de comparación por pares con la correlación entre el rango actual y otros. Un par concordante significa que el rango de algoritmo se correlaciona con el rango humano. En el caso opuesto será un par discordante. Por lo tanto, este coeficiente se define de la siguiente manera

$$\tau = \frac{(número\ de\ pares\ coincidentes) - (número\ de\ pares\ no\ coincidentes)}{n * (n-1)/2}$$

Los valores de τ varían de 0 a 1. Cuanto más |τ| se aproxime a 1, tanto mejor será el ranking. Por ejemplo, cuando el valor de τ se aproxima a -1, la clasificación es igual de precisa, sin embargo, el orden de sus ítems debería ser inverso. Esto es bastante consistente con los indicadores de estimación que asignan el rango más alto a los mejores valores, mientras que durante el ranking humano los mejores reciben los rangos más bajos. τ = 0 indica la falta de correlación entre los rangos.

Clasificación

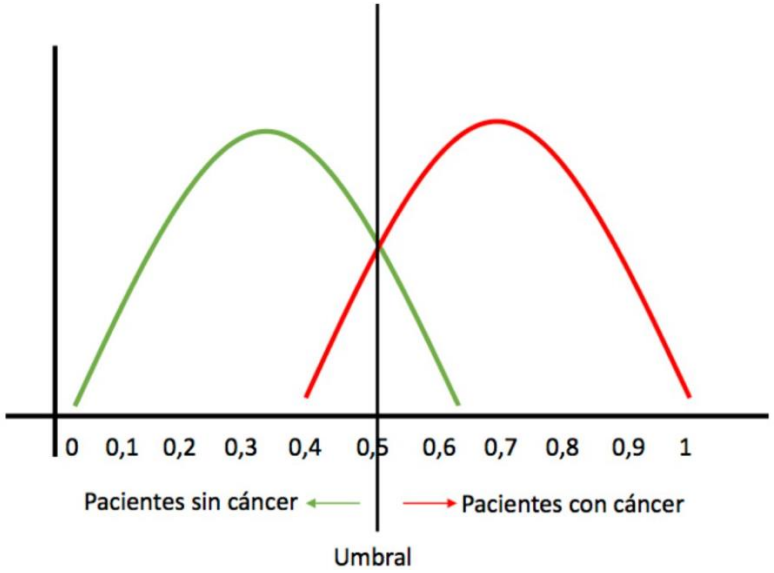
Curva ROC y Área bajo la curva (AUC)

- Esta es una de las métricas de evaluación más importante para verificar el rendimiento de cualquier modelo de clasificación.
- ROC viene de las características de funcionamiento del receptor y AUC del área bajo la curva.
- La curva ROC nos dice qué tan bueno puede distinguir el modelo entre dos cosas. Mejores modelos pueden distinguir con precisión entre los dos, mientras que un modelo pobre tendrá dificultades para distinguir entre los dos.

Clasificación

Curva ROC y Área bajo la curva (AUC)

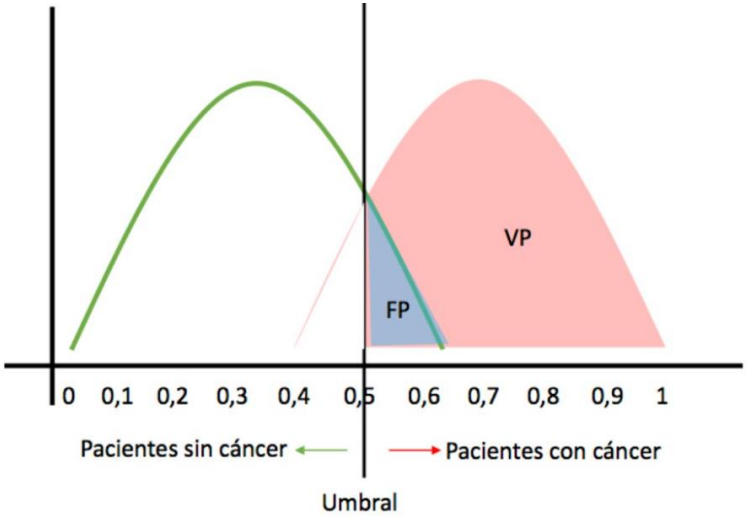
Supongamos que tenemos un modelo que predice si un paciente tiene cáncer o no, el resultado es el siguiente:



Clasificación

Curva ROC y Área bajo la curva (AUC)

Ahora debemos elegir un valor en donde establecemos el corte o un valor umbral, por encima del cual predeciremos a todos como positivos, tienen cáncer, y por debajo del cual predeciremos como negativos, NO cáncer. Este umbral lo establecemos en 0.5.

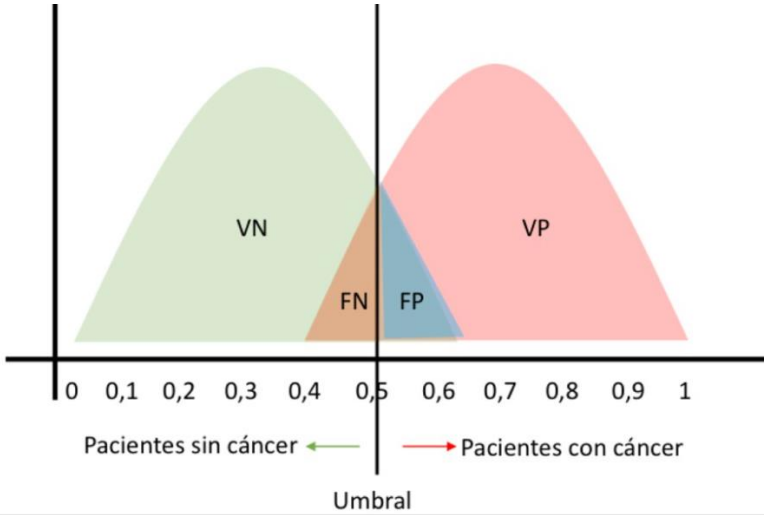


Clasificación

Curva ROC y Área bajo la curva (AUC)

Tomando los conceptos aprendidos en la matriz de confusión, todos los valores positivos por encima del umbral serán “verdaderos positivos” y los valores negativos por encima del umbral será “falsos positivos”, ya que se predicen incorrectamente como positivos.

Todos los valores negativos por debajo del umbral serán “verdaderos negativos” y los valores positivos por debajo del umbral serán “falsos negativos”, ya que se pronostican incorrectamente como negativos



Clasificación

Curva ROC y Área bajo la curva (AUC)

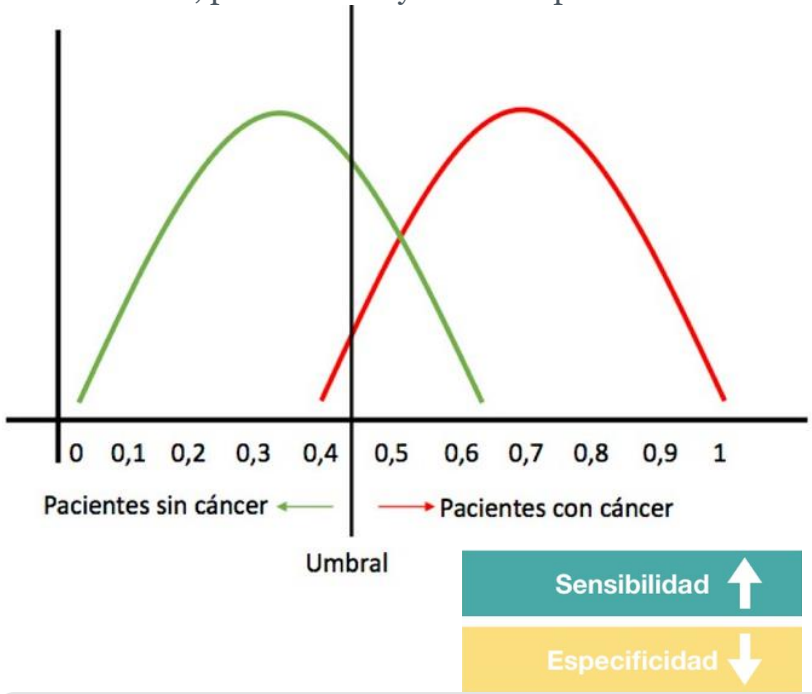
Aquí, tenemos una idea básica de que el modelo predice valores correctos e incorrectos con respecto al conjunto de umbrales.

Recordamos dos conceptos previos:

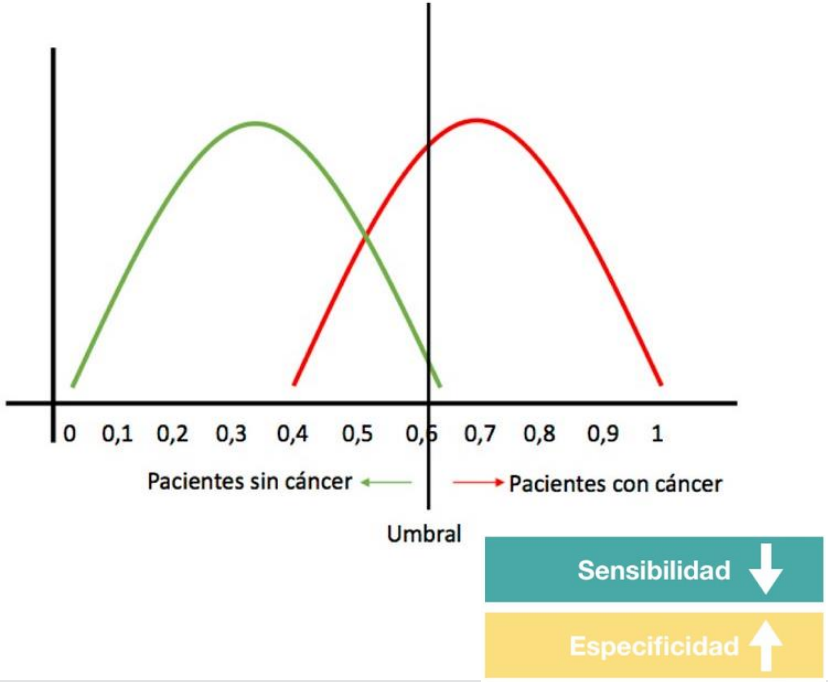
La sensibilidad o recall, es la proporción de pacientes que se identificaron correctamente por tener cáncer, es decir verdadero positivo, sobre el número total de pacientes que realmente tienen la enfermedad.

Por su parte, especificidad es la proporción de pacientes que se identificaron correctamente por no tener cáncer, verdadero positivo, sobre el número total de pacientes que no tienen la enfermedad.

Si volvemos a nuestra gráfica anterior, si disminuimos el valor del umbral, obtenemos más valores negativos, aumentando la sensibilidad, pero disminuyendo la especificidad.



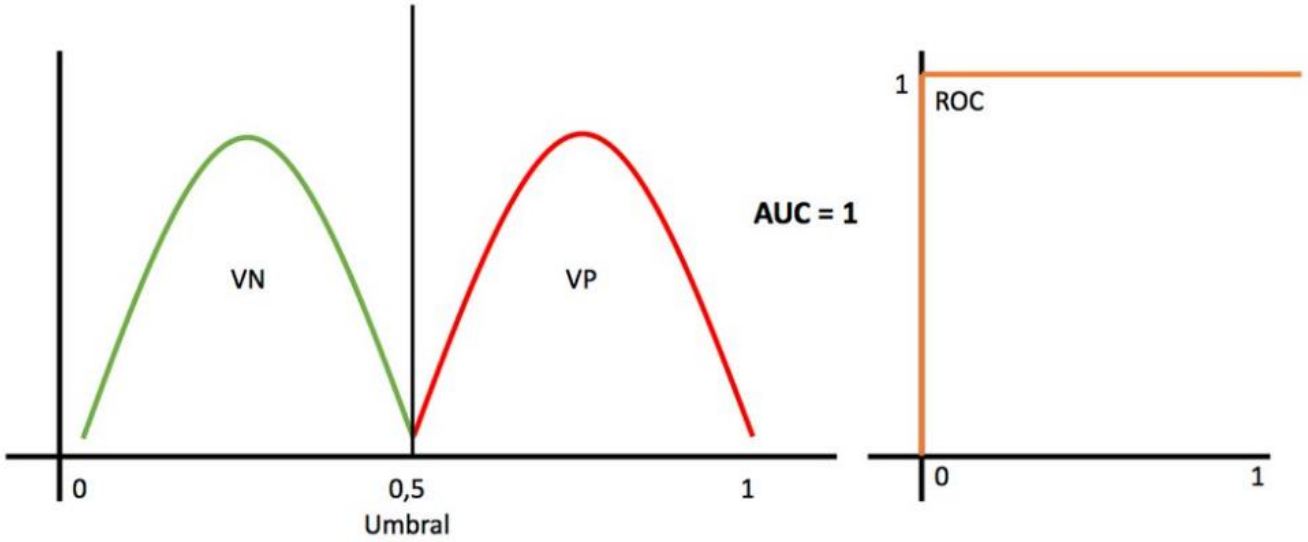
En cambio, si aumentamos el umbral, obtenemos más valores negativos, lo que aumenta la especificidad y disminuye la sensibilidad.



Clasificación

Área bajo la curva (AUC)

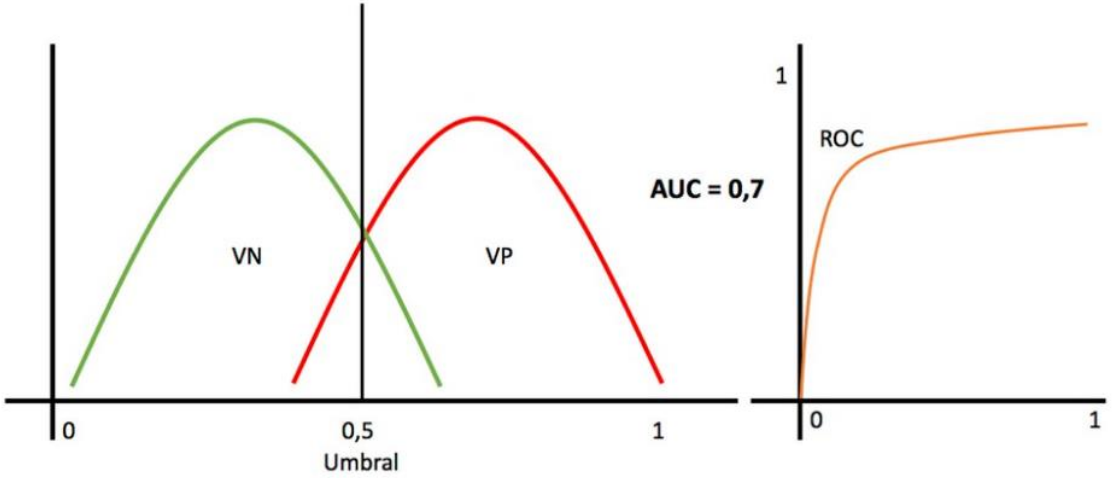
El AUC es el área bajo la curva ROC. Este puntaje nos da una buena idea de qué tan bien funciona el modelo.



Clasificación

Área bajo la curva (AUC)

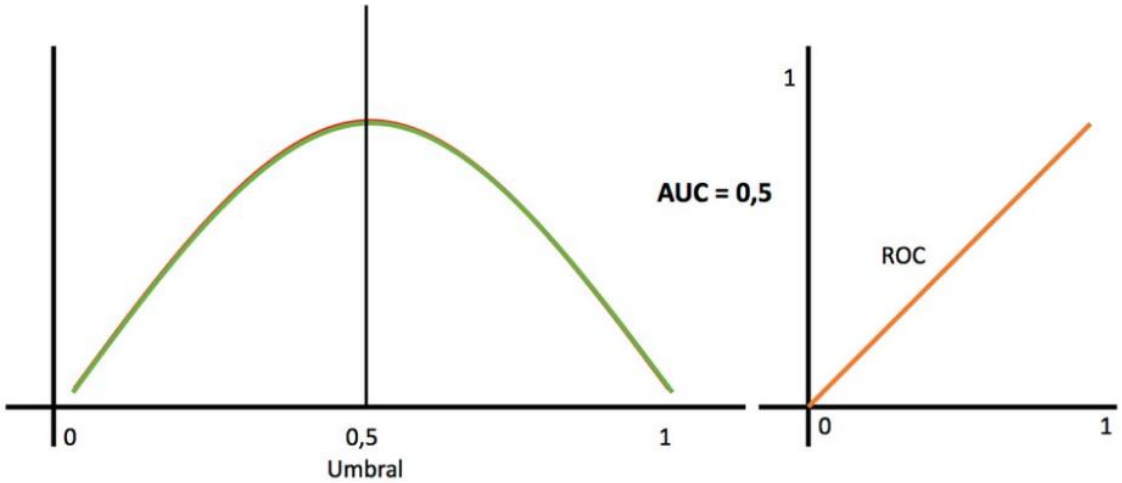
Cuando dos distribuciones se superponen, introducimos errores. Dependiendo del umbral, podemos minimizarlos o maximizarlos. Cuando AUC es 0.7, significa que hay 70% de probabilidad de que el modelo pueda distinguir entre clase positiva y clase negativa.



Clasificación

Área bajo la curva (AUC)

Esta es la peor situación. Cuando el AUC es aproximadamente 0.5, el modelo no tiene capacidad de discriminación para distinguir entre clase positiva y clase negativa.



Jupyter

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades

Interpretación de los resultados de la regresión lineal mediante el resumen OLS

OLS Regression Results

Dep. Variable:	money	R-squared (uncentered):	0.929
Model:	OLS	Adj. R-squared (uncentered):	0.929
Method:	Least Squares	F-statistic:	2.782e+04
Date:	Tue, 22 Jun 2021	Prob (F-statistic):	0.00
Time:	19:52:41	Log-Likelihood:	-2.9610e+05
No. Observations:	27494	AIC:	5.922e+05
Df Residuals:	27481	BIC:	5.923e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	110.1530	5.604	19.657	0.000	99.169	121.137
workclass	81.8499	71.835	1.139	0.255	-58.951	222.650
fnlwgt	0.1460	0.001	227.934	0.000	0.145	0.147
education	-85.9176	22.169	-3.876	0.000	-129.369	-42.466
education-num	768.3089	25.392	30.258	0.000	718.540	818.078
marital-status	-104.2694	70.370	-1.482	0.138	-242.199	33.660
occupation	-462.1639	19.770	-23.377	0.000	-500.914	-423.414
relationship	-679.6150	50.752	-13.391	0.000	-779.091	-580.139
race	-814.9227	121.905	-6.685	0.000	-1053.863	-575.983
sex	-5889.0633	161.446	-36.477	0.000	-6205.506	-5572.621
capital-gain	1.2994	0.009	138.942	0.000	1.281	1.318
capital-loss	3.6170	0.171	21.136	0.000	3.282	3.952
hours-per-week	33.3916	5.623	5.939	0.000	22.371	44.412

Omnibus:	3147.001	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3842.564
Skew:	0.885	Prob(JB):	0.00
Kurtosis:	2.528	Cond. No.	5.04e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.04e+05. This might indicate that there are strong multicollinearity or other numerical problems.

- **Variable dependiente (Dep. Variable):** La variable dependiente es aquella que va a depender de otras variables. En este análisis de regresión Y es nuestra variable dependiente porque queremos analizar el efecto de X (**todas las variables de la tabla**) sobre (**money**) Y.
- **Modelo**: El método de mínimos cuadrados ordinarios (MCO) es el modelo más utilizado debido a su eficiencia. Este modelo da la mejor aproximación de la verdadera línea de regresión de la población. El principio de OLS es minimizar el cuadrado de los errores ($\sum ei2$).
- **Número de observaciones (No. Observations):** El número de observaciones es el tamaño de nuestra muestra, es decir, N = 27494.

Interpretación de los resultados de la regresión lineal mediante el resumen OLS

OLS Regression Results

Dep. Variable:	money	R-squared (uncentered):	0.929
Model:	OLS	Adj. R-squared (uncentered):	0.929
Method:	Least Squares	F-statistic:	2.782e+04
Date:	Tue, 22 Jun 2021	Prob (F-statistic):	0.00
Time:	19:52:41	Log-Likelihood:	-2.9610e+05
No. Observations:	27494	AIC:	5.922e+05
Df Residuals:	27481	BIC:	5.923e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	110.1530	5.604	19.657	0.000	99.169	121.137
workclass	81.8499	71.835	1.139	0.255	-58.951	222.650
fnlwgt	0.1460	0.001	227.934	0.000	0.145	0.147
education	-85.9176	22.169	-3.876	0.000	-129.369	-42.466
education-num	768.3089	25.392	30.258	0.000	718.540	818.078
marital-status	-104.2694	70.370	-1.482	0.138	-242.199	33.660
occupation	-462.1639	19.770	-23.377	0.000	-500.914	-423.414
relationship	-679.6150	50.752	-13.391	0.000	-779.091	-580.139
race	-814.9227	121.905	-6.685	0.000	-1053.863	-575.983
sex	-5889.0633	161.446	-36.477	0.000	-6205.506	-5572.621
capital-gain	1.2994	0.009	138.942	0.000	1.281	1.318
capital-loss	3.6170	0.171	21.136	0.000	3.282	3.952
hours-per-week	33.3916	5.623	5.939	0.000	22.371	44.412

Omnibus:	3147.001	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3842.564
Skew:	0.885	Prob(JB):	0.00
Kurtosis:	2.528	Cond. No.	5.04e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.04e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

- **Grado de libertad (df) de los residuos:**

El grado de libertad es el número de observaciones independientes a partir de las cuales se calcula la suma de los cuadrados.

D.f Residuales = 27494 - (13) = 27481

El grado de libertad (D.f) se calcula como,

Grados de libertad, D . f = N - K

Donde, N = tamaño de la muestra (número de observaciones) y K = número de variables + 1

Interpretación de los resultados de la regresión lineal mediante el resumen OLS

OLS Regression Results

Dep. Variable:	money	R-squared (uncentered):	0.929
Model:	OLS	Adj. R-squared (uncentered):	0.929
Method:	Least Squares	F-statistic:	2.782e+04
Date:	Tue, 22 Jun 2021	Prob (F-statistic):	0.00
Time:	19:52:41	Log-Likelihood:	-2.9610e+05
No. Observations:	27494	AIC:	5.922e+05
Df Residuals:	27481	BIC:	5.923e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	110.1530	5.604	19.657	0.000	99.169	121.137
workclass	81.8499	71.835	1.139	0.255	-58.951	222.650
fnlwgt	0.1460	0.001	227.934	0.000	0.145	0.147
education	-85.9176	22.169	-3.876	0.000	-129.369	-42.466
education-num	768.3089	25.392	30.258	0.000	718.540	818.078
marital-status	-104.2694	70.370	-1.482	0.138	-242.199	33.660
occupation	-462.1639	19.770	-23.377	0.000	-500.914	-423.414
relationship	-679.6150	50.752	-13.391	0.000	-779.091	-580.139
race	-814.9227	121.905	-6.685	0.000	-1053.863	-575.983
sex	-5889.0633	161.446	-36.477	0.000	-6205.506	-5572.621
capital-gain	1.2994	0.009	138.942	0.000	1.281	1.318
capital-loss	3.6170	0.171	21.136	0.000	3.282	3.952
hours-per-week	33.3916	5.623	5.939	0.000	22.371	44.412

Omnibus:	3147.001	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3842.564
Skew:	0.885	Prob(JB):	0.00
Kurtosis:	2.528	Cond. No.	5.04e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.04e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

- **Df of model:**

Df of model = *K – 1 = 14 – 1 = 13 ,*

Donde, ***K*** = *número de variables + 1*

Término constante: Los términos constantes son el intercepto de la línea de regresión. De la línea de regresión (ec...1) el intercepto en este caso no existe. En la regresión omitimos algunas variables independientes que no tienen mucho impacto en la variable dependiente, el intercepto indica el valor medio de estas variables omitidas y el ruido presente en el modelo.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.669
Model:	OLS	Adj. R-squared:	0.667
Method:	Least Squares	F-statistic:	299.2
Date:	Mon, 01 Mar 2021	Prob (F-statistic):	2.33e-37
Time:	16:19:34	Log-Likelihood:	-88.686
No. Observations:	150	AIC:	181.4
Df Residuals:	148	BIC:	187.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.2002	0.257	-12.458	0.000	-3.708	-2.693
x1	0.7529	0.044	17.296	0.000	0.667	0.839

Omnibus:	3.538	Durbin-Watson:	1.279
Prob(Omnibus):	0.171	Jarque-Bera (JB):	3.589
Skew:	0.357	Prob(JB):	0.166
Kurtosis:	2.744	Cond. No.	43.4

Interpretación de los resultados de la regresión lineal mediante el resumen OLS

OLS Regression Results

Dep. Variable:	money	R-squared (uncentered):	0.929
Model:	OLS	Adj. R-squared (uncentered):	0.929
Method:	Least Squares	F-statistic:	2.782e+04
Date:	Tue, 22 Jun 2021	Prob (F-statistic):	0.00
Time:	19:52:41	Log-Likelihood:	-2.9610e+05
No. Observations:	27494	AIC:	5.922e+05
Df Residuals:	27481	BIC:	5.923e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	110.1530	5.604	19.657	0.000	99.169	121.137
workclass	81.8499	71.835	1.139	0.255	-58.951	222.650
fnlwgt	0.1460	0.001	227.934	0.000	0.145	0.147
education	-85.9176	22.169	-3.876	0.000	-129.369	-42.466
education-num	768.3089	25.392	30.258	0.000	718.540	818.078
marital-status	-104.2694	70.370	-1.482	0.138	-242.199	33.660
occupation	-462.1639	19.770	-23.377	0.000	-500.914	-423.414
relationship	-679.6150	50.752	-13.391	0.000	-779.091	-580.139
race	-814.9227	121.905	-6.685	0.000	-1053.863	-575.983
sex	-5889.0633	161.446	-36.477	0.000	-6205.506	-5572.621
capital-gain	1.2994	0.009	138.942	0.000	1.281	1.318
capital-loss	3.6170	0.171	21.136	0.000	3.282	3.952
hours-per-week	33.3916	5.623	5.939	0.000	22.371	44.412

Omnibus:	3147.001	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3842.564
Skew:	0.885	Prob(JB):	0.00
Kurtosis:	2.528	Cond. No.	5.04e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.04e+05. This might indicate that there are strong multicollinearity or other numerical problems.

- **Término de coeficiente**: El término del coeficiente indica el cambio en Y para una unidad de cambio en X, es decir, si X aumenta en 1 unidad, Y aumenta en 110,1560 (en el caso de age vs money). Si está familiarizado con las derivadas, puede relacionarlo con la tasa de cambio de Y con respecto a X.
- **Error estándar de los parámetros**: El error estándar también se llama desviación estándar. El error estándar muestra la variabilidad muestral de estos parámetros. El error estándar se calcula como –
- **Error estándar del término de intercepción (b1):**

$$se\,(b_1) = \sqrt{\left(\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\right)\sigma^2}$$

- **Error estándar del término de coeficiente (b2):**

$$se\,(b_2) = \sqrt{\frac{\sigma^2}{\sum(x_i-\bar{x})}}$$

- **σ2 es el error estándar de regresión (SER). Y σ2 es igual a RSS (suma residual de cuadrados, es decir, ∑ei2).**

Interpretación de los resultados de la regresión lineal mediante el resumen OLS

OLS Regression Results

Dep. Variable:	money	R-squared (uncentered):	0.929
Model:	OLS	Adj. R-squared (uncentered):	0.929
Method:	Least Squares	F-statistic:	2.782e+04
Date:	Tue, 22 Jun 2021	Prob (F-statistic):	0.00
Time:	19:52:41	Log-Likelihood:	-2.9610e+05
No. Observations:	27494	AIC:	5.922e+05
Df Residuals:	27481	BIC:	5.923e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	110.1530	5.604	19.657	0.000	99.169	121.137
workclass	81.8499	71.835	1.139	0.255	-58.951	222.650
fnlwgt	0.1460	0.001	227.934	0.000	0.145	0.147
education	-85.9176	22.169	-3.876	0.000	-129.369	-42.466
education-num	768.3089	25.392	30.258	0.000	718.540	818.078
marital-status	-104.2694	70.370	-1.482	0.138	-242.199	33.660
occupation	-462.1639	19.770	-23.377	0.000	-500.914	-423.414
relationship	-679.6150	50.752	-13.391	0.000	-779.091	-580.139
race	-814.9227	121.905	-6.685	0.000	-1053.863	-575.983
sex	-5889.0633	161.446	-36.477	0.000	-6205.506	-5572.621
capital-gain	1.2994	0.009	138.942	0.000	1.281	1.318
capital-loss	3.6170	0.171	21.136	0.000	3.282	3.952
hours-per-week	33.3916	5.623	5.939	0.000	22.371	44.412

Omnibus:	3147.001	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3842.564
Skew:	0.885	Prob(JB):	0.00
Kurtosis:	2.528	Cond. No.	5.04e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.04e+05. This might indicate that there are strong multicollinearity or other numerical problems.

- **t – statistics:**

En teoría, suponemos que el término de error sigue la distribución normal y por ello los parámetros b1 y b2 también tienen distribuciones normales con la varianza calculada en la sección anterior.

Es decir,

b1 ~ N(B1, σb12)

b2 ~ N(B2 , σb22)

Aquí B1 y B2 son las verdaderas medias de b1 y b2.

Los estadísticos t - se calculan asumiendo la siguiente hipótesis

- H0 : B2 = 0 (la variable X no tiene influencia en Y)

- Ha : B2 ≠ 0 (X tiene un impacto significativo en Y)

Cálculos para los estadísticos t - :

t = (b1 - B1) / s.e (b1)

De la tabla de resumen, b1 = 110.1530 y se(b1) = 5.604 por lo que

t = 110.1530 / 5.604 = 19.657

Del mismo modo, b2 = 81.8499 , se(b2) = 71.835

t = (81.8499 – 0.255) / 71.835 = 1.139 81.8499

Interpretación de los resultados de la regresión lineal mediante el resumen OLS

OLS Regression Results

Dep. Variable:	money	R-squared (uncentered):	0.929
Model:	OLS	Adj. R-squared (uncentered):	0.929
Method:	Least Squares	F-statistic:	2.782e+04
Date:	Tue, 22 Jun 2021	Prob (F-statistic):	0.00
Time:	19:52:41	Log-Likelihood:	-2.9610e+05
No. Observations:	27494	AIC:	5.922e+05
Df Residuals:	27481	BIC:	5.923e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	110.1530	5.604	19.657	0.000	99.169	121.137
workclass	81.8499	71.835	1.139	0.255	-58.951	222.650
fnlwgt	0.1460	0.001	227.934	0.000	0.145	0.147
education	-85.9176	22.169	-3.876	0.000	-129.369	-42.466
education-num	768.3089	25.392	30.258	0.000	718.540	818.078
marital-status	-104.2694	70.370	-1.482	0.138	-242.199	33.660
occupation	-462.1639	19.770	-23.377	0.000	-500.914	-423.414
relationship	-679.6150	50.752	-13.391	0.000	-779.091	-580.139
race	-814.9227	121.905	-6.685	0.000	-1053.863	-575.983
sex	-5889.0633	161.446	-36.477	0.000	-6205.506	-5572.621
capital-gain	1.2994	0.009	138.942	0.000	1.281	1.318
capital-loss	3.6170	0.171	21.136	0.000	3.282	3.952
hours-per-week	33.3916	5.623	5.939	0.000	22.371	44.412

Omnibus:	3147.001	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3842.564
Skew:	0.885	Prob(JB):	0.00
Kurtosis:	2.528	Cond. No.	5.04e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.04e+05. This might indicate that there are strong multicollinearity or other numerical problems.

- **p – values:**

En teoría, leemos que el valor **p** es la probabilidad de obtener los estadísticos **t** al menos tan contradictorios con H0 como los calculados a partir de la suposición de que la hipótesis nula es verdadera.

En la tabla de resumen, podemos ver que el valor **p** para el primer parámetro es igual a 0 . Esto no es exactamente 0, pero como tenemos estadísticas muy grandes **(19.657)** el valor p será aproximadamente 0.

Si conoce los niveles de significación, podrá ver que podemos rechazar la hipótesis nula en casi todos los niveles de significación.

Intervalos de confianza [0.025 0.975]:

Hay muchos enfoques para probar la hipótesis, incluido el enfoque del valor p mencionado anteriormente. El enfoque del intervalo de confianza es uno de ellos. El 5% se realizan los I.C. $(b_1 - t_{\alpha/2}\, s.e(b_1) , b_1 + t_{\alpha/2}\, s.e(b_1))$

C.I para B_1 es

Con $\propto = 5\ \%$, $b_1 = 110.1530$, $s.e(b_1) = 5.604$, de la tabla t , $t_{0.025,27481} = 99.169$.

Lo mismo puede hacerse para b2 también.

Al calcular los valores de p rechazamos la hipótesis nula y podemos ver lo mismo en C.I. también. Como 0 no se encuentra en ninguno de los intervalos, rechazamos la hipótesis nula.

Interpretación de los resultados de la regresión lineal mediante el resumen OLS

OLS Regression Results

Dep. Variable:	money	R-squared (uncentered):	0.929
Model:	OLS	Adj. R-squared (uncentered):	0.929
Method:	Least Squares	F-statistic:	2.782e+04
Date:	Tue, 22 Jun 2021	Prob (F-statistic):	0.00
Time:	19:52:41	Log-Likelihood:	-2.9610e+05
No. Observations:	27494	AIC:	5.922e+05
Df Residuals:	27481	BIC:	5.923e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	110.1530	5.604	19.657	0.000	99.169	121.137
workclass	81.8499	71.835	1.139	0.255	-58.951	222.650
fnlwgt	0.1460	0.001	227.934	0.000	0.145	0.147
education	-85.9176	22.169	-3.876	0.000	-129.369	-42.466
education-num	768.3089	25.392	30.258	0.000	718.540	818.078
marital-status	-104.2694	70.370	-1.482	0.138	-242.199	33.660
occupation	-462.1639	19.770	-23.377	0.000	-500.914	-423.414
relationship	-679.6150	50.752	-13.391	0.000	-779.091	-580.139
race	-814.9227	121.905	-6.685	0.000	-1053.863	-575.983
sex	-5889.0633	161.446	-36.477	0.000	-6205.506	-5572.621
capital-gain	1.2994	0.009	138.942	0.000	1.281	1.318
capital-loss	3.6170	0.171	21.136	0.000	3.282	3.952
hours-per-week	33.3916	5.623	5.939	0.000	22.371	44.412

Omnibus:	3147.001	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3842.564
Skew:	0.885	Prob(JB):	0.00
Kurtosis:	2.528	Cond. No.	5.04e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.04e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

- **R - valor al cuadrado:**

R2 es el coeficiente de determinación que nos dice qué porcentaje de variación de la variable independiente puede ser explicado por la variable independiente. En este caso, el 92,9% de la variación de Y puede ser explicada por X. El valor máximo posible de R2 puede ser 1, lo que significa que cuanto mayor sea

- **Estadística F:**

La prueba F indica la bondad del ajuste de una regresión. La prueba es similar a la prueba **t** u otras pruebas que hacemos para la hipótesis. El estadístico F se calcula como sigue el valor de R2, mejor será la regresión.

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Valores de R^2, n and k, F = (0.929/1) / (0.071/27481) = 2.78 e+04

Interpretación de los resultados de la regresión lineal mediante el resumen OLS

OLS Regression Results

Dep. Variable:	money	R-squared (uncentered):	0.929
Model:	OLS	Adj. R-squared (uncentered):	0.929
Method:	Least Squares	F-statistic:	2.782e+04
Date:	Tue, 22 Jun 2021	Prob (F-statistic):	0.00
Time:	19:52:41	Log-Likelihood:	-2.9610e+05
No. Observations:	27494	AIC:	5.922e+05
Df Residuals:	27481	BIC:	5.923e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	110.1530	5.604	19.657	0.000	99.169	121.137
workclass	81.8499	71.835	1.139	0.255	-58.951	222.650
fnlwgt	0.1460	0.001	227.934	0.000	0.145	0.147
education	-85.9176	22.169	-3.876	0.000	-129.369	-42.466
education-num	768.3089	25.392	30.258	0.000	718.540	818.078
marital-status	-104.2694	70.370	-1.482	0.138	-242.199	33.660
occupation	-462.1639	19.770	-23.377	0.000	-500.914	-423.414
relationship	-679.6150	50.752	-13.391	0.000	-779.091	-580.139
race	-814.9227	121.905	-6.685	0.000	-1053.863	-575.983
sex	-5889.0633	161.446	-36.477	0.000	-6205.506	-5572.621
capital-gain	1.2994	0.009	138.942	0.000	1.281	1.318
capital-loss	3.6170	0.171	21.136	0.000	3.282	3.952
hours-per-week	33.3916	5.623	5.939	0.000	22.371	44.412

Omnibus:	3147.001	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3842.564
Skew:	0.885	Prob(JB):	0.00
Kurtosis:	2.528	Cond. No.	5.04e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.04e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

- **R-cuadrado adjunto**: Es la versión modificada de R-cuadrado que se ajusta al número de variables de la regresión. Sólo aumenta cuando una variable adicional añade poder explicativo a la regresión.

- **Prob(Estadística F):** Indica la significación global de la regresión. Se trata de evaluar el nivel de significación de todas las variables juntas, a diferencia del estadístico **t**, que lo mide para las variables individuales. La hipótesis nula es "todos los coeficientes de la regresión son iguales a cero". El estadístico **F** indica la probabilidad de que la hipótesis nula sea cierta. Según los resultados anteriores, la probabilidad es cercana a cero. Esto implica que, en general, las regresiones son significativas.

- **AIC/BIC:** Son las siglas de Akaike's Information Criteria y se utiliza para la selección de modelos. Penaliza el modo de los errores en caso de que se añada una nueva variable a la ecuación de regresión. Se calcula como el número de parámetros menos la probabilidad del modelo global. Un AIC más bajo implica un modelo mejor. Por su parte, BIC significa criterio de información bayesiano y es una variante de AIC en la que las penalizaciones son más severas.

Interpretación de los resultados de la regresión lineal mediante el resumen OLS

OLS Regression Results

Dep. Variable:	money	R-squared (uncentered):	0.929
Model:	OLS	Adj. R-squared (uncentered):	0.929
Method:	Least Squares	F-statistic:	2.782e+04
Date:	Tue, 22 Jun 2021	Prob (F-statistic):	0.00
Time:	19:52:41	Log-Likelihood:	-2.9610e+05
No. Observations:	27494	AIC:	5.922e+05
Df Residuals:	27481	BIC:	5.923e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	110.1530	5.604	19.657	0.000	99.169	121.137
workclass	81.8499	71.835	1.139	0.255	-58.951	222.650
fnlwgt	0.1460	0.001	227.934	0.000	0.145	0.147
education	-85.9176	22.169	-3.876	0.000	-129.369	-42.466
education-num	768.3089	25.392	30.258	0.000	718.540	818.078
marital-status	-104.2694	70.370	-1.482	0.138	-242.199	33.660
occupation	-462.1639	19.770	-23.377	0.000	-500.914	-423.414
relationship	-679.6150	50.752	-13.391	0.000	-779.091	-580.139
race	-814.9227	121.905	-6.685	0.000	-1053.863	-575.983
sex	-5889.0633	161.446	-36.477	0.000	-6205.506	-5572.621
capital-gain	1.2994	0.009	138.942	0.000	1.281	1.318
capital-loss	3.6170	0.171	21.136	0.000	3.282	3.952
hours-per-week	33.3916	5.623	5.939	0.000	22.371	44.412

Omnibus:	3147.001	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3842.564
Skew:	0.885	Prob(JB):	0.00
Kurtosis:	2.528	Cond. No.	5.04e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.04e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

- **Prob(Omnibus):** Uno de los supuestos de MCO es que los errores se distribuyen normalmente. La prueba ómnibus se realiza para comprobarlo. Aquí, la hipótesis nula es que los errores se distribuyen normalmente. Se supone que la Prob(Omnibus) debe ser cercana a 1 para que se cumpla el supuesto de MCO. En este caso, la Prob(Omnibus) es 3147.001.

- **Durbin-watson**: Otro supuesto de los MCO es el de la homocedasticidad. Esto implica que la varianza de los errores es constante. Se prefiere un valor entre 1 y 2. En este caso, es ~1.9, lo que implica que los resultados de la regresión son fiables desde el punto de vista de la interpretación de esta métrica.

- **Prob(Jarque-Bera):** Está en línea con la prueba Omnibus. También se realiza para el análisis de la distribución de los errores de regresión. Se supone que coincide con los resultados del test Ómnibus. Un valor grande de la prueba JB indica que los errores no se distribuyen normalmente.

Los términos como Skewness y Kurtosis nos hablan de la distribución de los datos. La asimetría y la curtosis para la distribución normal son 0 y 3 respectivamente. La prueba de Jarque-Bera se utiliza para comprobar si un error tiene una distribución normal o no.

Curva ROC y el AUC

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades

Curva ROC y el AUC en Python

```
#Importamos
from sklearn.datasets import make_classification
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from matplotlib import pyplot
# Generamos un dataset de dos clases
X, y = make_classification(n_samples=1000, n_classes=2,
random_state=1)
# Dividimos en training y test
trainX, testX, trainy, testy = train_test_split(X, y, test_size=0.5,
random_state=2)
#Generamos un clasificador sin entrenar , que asignará 0 a
todo
ns_probs = [0 for _ in range(len(testy))]
# Entrenamos nuestro modelo de reg log
model = LogisticRegression(solver='lbfgs')
```

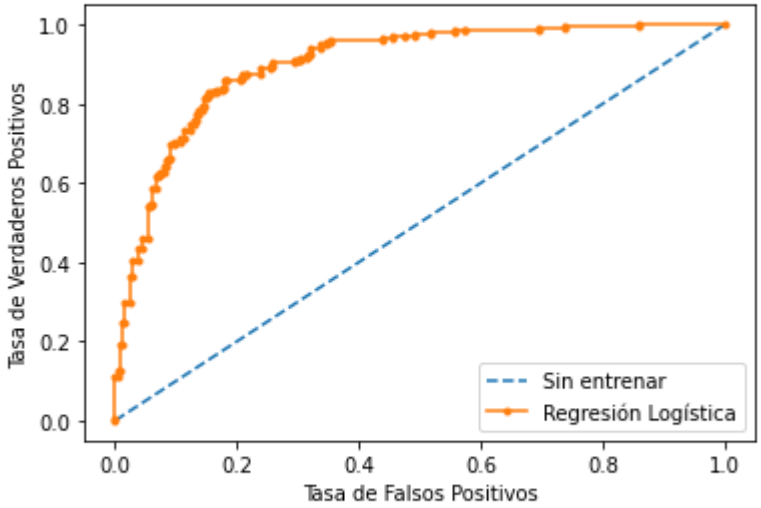
Curva ROC y el AUC en Python

```
model.fit(trainX, trainy)
# Predecimos las probabilidades
lr_probs = model.predict_proba(testX)
#Nos quedamos con las probabilidades de la clase positiva (la probabilidad de 1)
lr_probs = lr_probs[:, 1]
# Calculamos el AUC
ns_auc = roc_auc_score(testy, ns_probs)
lr_auc = roc_auc_score(testy, lr_probs)
# Imprimimos en pantalla
print('Sin entrenar: ROC AUC=%.3f' % (ns_auc))
print('Regresión Logística: ROC AUC=%.3f' % (lr_auc))
# Calculamos las curvas ROC
ns_fpr, ns_tpr, _ = roc_curve(testy, ns_probs)
lr_fpr, lr_tpr, _ = roc_curve(testy, lr_probs)
```

Curva ROC y el AUC en Python

```
# Etiquetas de los ejes
pyplot.xlabel('Tasa de Falsos Positivos')
pyplot.ylabel('Tasa de Verdaderos Positivos')
pyplot.legend()
pyplot.show()
```

```
Sin entrenar: ROC AUC=0.500
Regresión Logística: ROC AUC=0.903
```



Curva ROC (Salida del modelo de ejemplo)

Bibliografía

- https://scikit-learn.org/stable/modules/model_evaluation.html
- https://medium.com/analytics-vidhya/evaluation-metrics-for-regression-algorithms-along-with-their-implementation-in-python-9ec502729dad

Gracias

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades