IMPAGO DE PRÉSTAMOS: ANÁLISIS DE DATOS Y ELABORACIÓN DE UN MODELO PREDICTIVO

ANEXOS

Gonzalo Fernández de Córdoba García

Tutor del proyecto: Jaime Martín García-Cuerva

Proyecto de Investigación de Bachillerato de Excelencia

IES Arquitecto Ventura Rodríguez
Curso: 2022/2023

1. MARCO TEÓRICO

ILUSTRACIÓN 1.1

Función de coste y algoritmo de gradiente descendente.

(Arriba) Función de coste que expresa la diferencia entre las predicciones y los datos de entrenamiento, es decir, el error.

(Abajo) Algoritmo de gradiente descendente que actualiza los coeficientes $(\theta_0 \dots \theta_m)$ *y que se repite hasta que se minimiza la función de coste.*

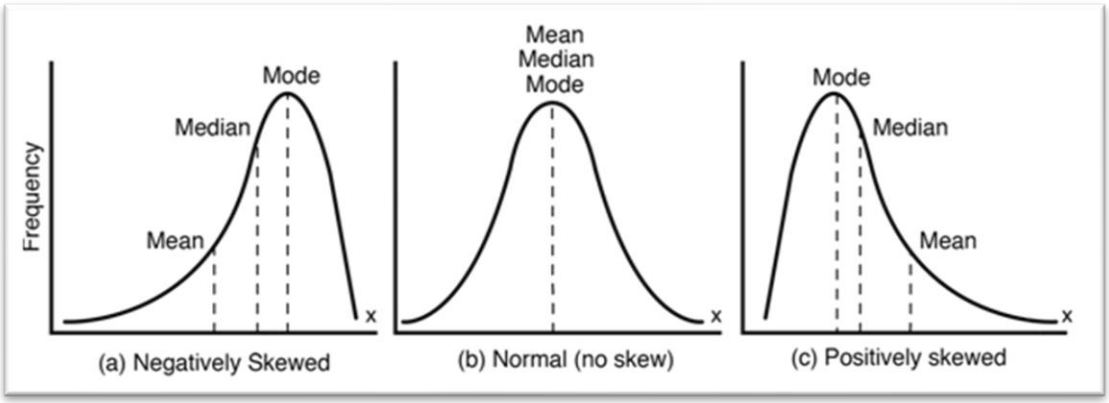
$$J(\theta_0 \dots \theta_m) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log\, h_\theta(x) + \left(1 - y^{(i)}\right)\left(1 - \, h_\theta(x)\right)$$

$$\theta_j := \, \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0 \dots \theta_m)$$

(*para j = [0, m]*)

ILUSTRACIÓN 1.2

Explicación visual de la asimetría de una distribución (en inglés).



Obtenido de: https://www.researchgate.net/figure/a-Negative-skewness-b-Normal-curve-c-Positive-skewness-Durkhure-and-Lodwal-2014_fig5_294890337

ILUSTRACIÓN 1.3

Ejemplos de histograma (izquierda), diagrama de cajas y bigotes (derecha)

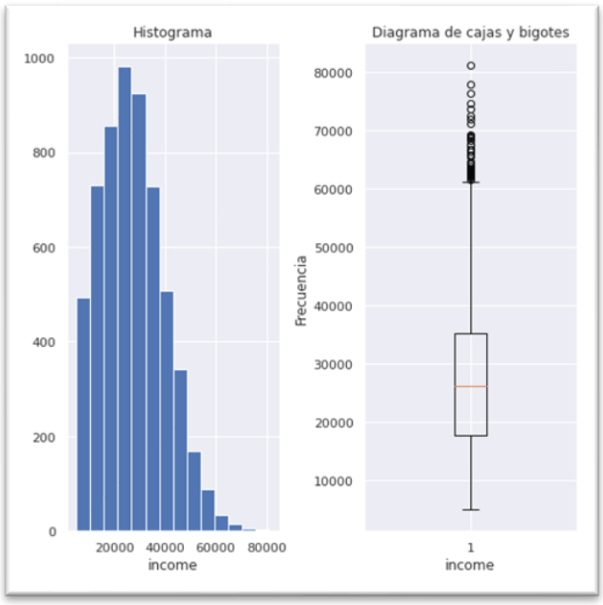


ILUSTRACIÓN 1.4

Expresión de la estandarización.

(x) = datos sin estandarizar, (μ) = la media de los datos sin estandarizar, (σ) = desviación estándar.

$$x_{estand} = \frac{x - \mu}{\sigma}$$

ILUSTRACIONES 1.5, 1.6

Ejemplo de la estandarización de unos datos.

Los datos estandarizados presentan un rango menor, como se puede observar en la parte inferior de la imagen, pero conservan la misma forma.

(Izquierda) Distribución de datos no estandarizados.

(Derecha) Distribución de datos estandarizados.

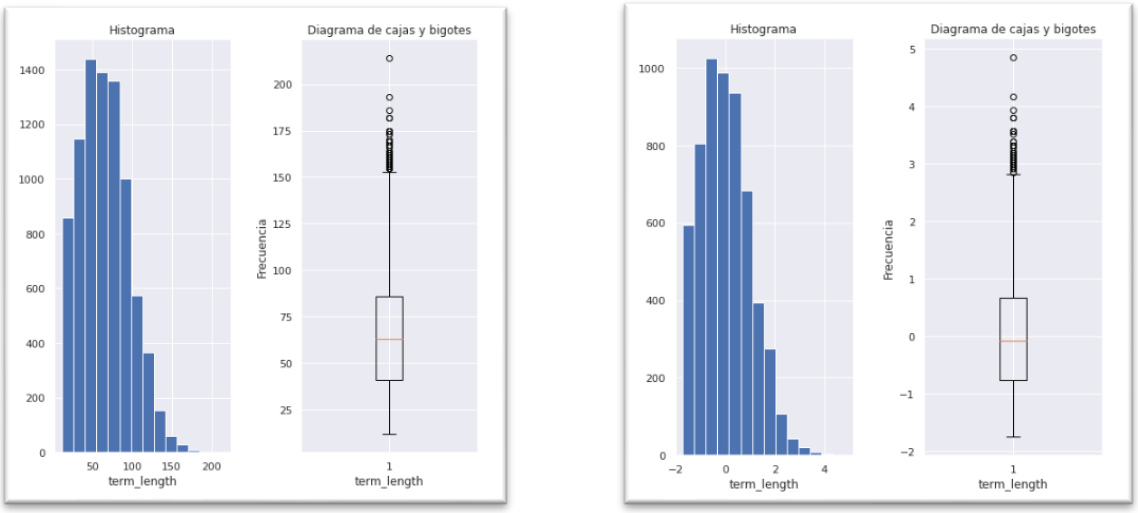
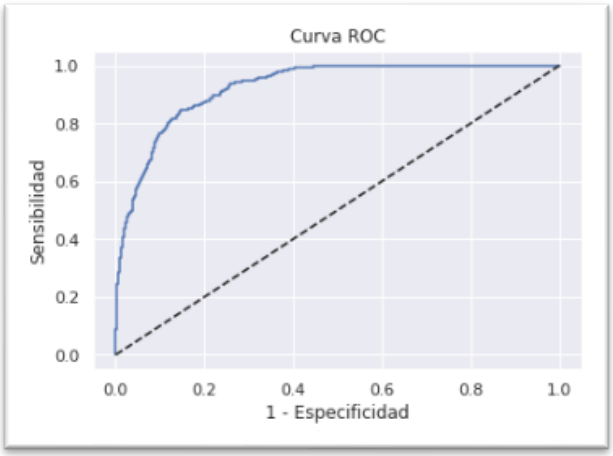


ILUSTRACIÓN 1.7

Ejemplo de curva ROC

**La recta diagonal discontinua representa un modelo sin capacidad predictiva.*



2. MARCO PRÁCTICO

ILUSTRACIÓN 2.1

Muestra de parte del conjunto de datos

	income	loan_amount	term_length	install_to_inc	occup	marital	schufa	num_applic	OBS_DATE	target_var
0	18785.517943694504	12300	67.0	0.009772532761805456	nan	Single	7106.014471346445	1	18JUL2018 - 00:00:00	nan
1	12861.495159877606	Not avail.	113.0	0.003509136597401741	Employee	Divorced	7694.806893720367	1	16JUL2018 - 00:00:00	0
2	14886.776341632107	10700	Not avail.	0.013310346283231944	Unemployed	Single	7142.496337537019	1	21DEC2010 - 00:00:00	1
3	Not avail.	33000	112.0	0.004760874885112928	Employee	Single	7446.1706118576485	1	05NOV2015 - 00:00:00	0
4	15897.753806874589	19900	59.0	0.02121608748572823	Unemployed	Separated	7241.656646188094	1	13JUL2015 - 00:00:00	1

ILUSTRACIÓN 2.2

Traducción de las variables del conjunto

	variables	explicación / traducción
0	income	Ingresos del cliente
1	loan_amount	Cantidad prestada
2	term_length	Longitud del préstamo
3	install_to_inc	Ratio cuota a ingresos
4	occup	Situación laboral
5	marital	Estado civil
6	schufa	Puntuación externa
7	num_applic	Número de solicitantes
8	OBS_DATE	Fecha de solicitud del crédito
9	target_var	Variable de salida. Si se impagó o no

ILUSTRACIÓN 2.3

En rojo, valores vacíos en el conjunto de datos.

	income	loan_amount	term_length	install_to_inc	occup	marital	schufa	num_applic	OBS_DATE	target_var
0	18785.517943694504	12300	67.0	0.009772532761805456	nan	Single	7106.014471346445	1	18JUL2018 - 00:00:00	nan
1	12861.495159877606	nan	113.0	0.003509136597401741	Employee	Divorced	7694.806893720367	1	16JUL2018 - 00:00:00	0
2	14886.776341632107	10700	nan	0.013310346283231944	Unemployed	Single	7142.496337537019	1	21DEC2010 - 00:00:00	1
3	nan	33000	112.0	0.004760874885112928	Employee	Single	7446.1706118576485	1	05NOV2015 - 00:00:00	0
4	15897.753806874589	19900	59.0	0.02121608748572823	Unemployed	Separated	7241.656646188094	1	13JUL2015 - 00:00:00	1

ILUSTRACIÓN 2.4

Muestra del desbalance del conjunto de datos

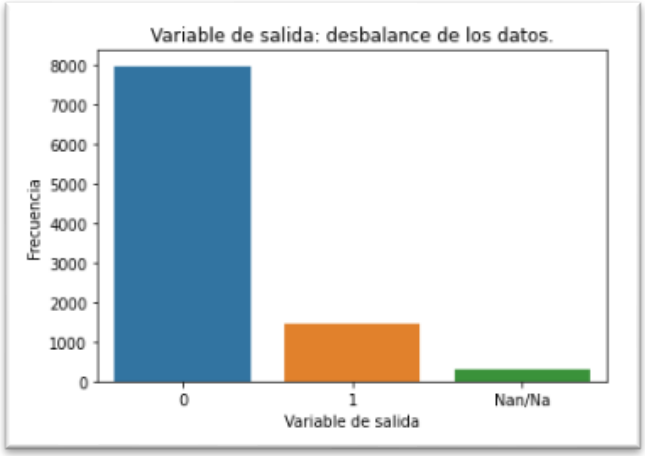


ILUSTRACIÓN 2.5

Clases iniciales de la variable situación laboral.

occup	loan_amount	target_var	customer	PD
Employee	32923230.000	32	1614	0.020
2	1268100.000	7	63	0.111
1	1311600.000	8	62	0.129
Student	21858110.000	140	1067	0.131
Not avail.	6970740.000	66	358	0.184
Unemployed	54667990.000	647	2663	0.243
3	1055250.000	13	52	0.250

ILUSTRACIÓN 2.6, 2.7

Contrastes de hipótesis de la variable situación laboral.

	eje	customer	target_var	PD	eje-1	customer-1	target_var-1	PD-1	Z	pvalor
1	2	63	7	0.111	Employee	1614.000	32.000	0.020	4.716	0.000
0	1	62	8	0.129	2	63.000	7.000	0.111	0.308	0.758
5	Student	1067	140	0.131	1	62.000	8.000	0.129	0.049	0.961
4	Not avail.	358	66	0.184	Student	1067.000	140.000	0.131	2.474	0.013
6	Unemployed	2663	647	0.243	Not avail.	358.000	66.000	0.184	2.452	0.014
2	3	52	13	0.250	Unemployed	2663.000	647.000	0.243	0.117	0.907

	eje	customer	target_var	PD	eje-1	customer-1	target_var-1	PD-1	Z	pvalor
1	Student	1550	221	0.143	Employee	1614.000	32.000	0.020	12.726	0.000
2	Unemployed	2715	660	0.243	Student	1550.000	221.000	0.143	7.799	0.000

ILUSTRACIÓN 2.8

Clases resultantes de la variable situación laboral.

occup_agg	loan_amount	target_var	customer	PD
Employee	32923230.000	32	1614	0.020
Student	31408550.000	221	1550	0.143
Unemployed	55723240.000	660	2715	0.243

ILUSTRACIÓN 2.9

Clases iniciales de la variable estado civil.

marital	loan_amount	target_var	customer	PD
Married	15760280.000	22	784	0.028
Living together	16447330.000	88	801	0.110
Not avail.	7271560.000	47	354	0.133
Single	46297080.000	407	2273	0.179
Divorced	17071820.000	167	825	0.202
Separated	17206950.000	182	842	0.216

ILUSTRACIONES 2.10, 2.11

Contrastes de hipótesis de la variable estado civil.

	eje	customer	target_var	PD	eje-1	customer-1	target_var-1	PD-1	Z	pvalor
1	Living together	801	88	0.110	Married	784.000	22.000	0.028	6.407	0.000
3	Not avail.	354	47	0.133	Living together	801.000	88.000	0.110	1.117	0.264
5	Single	2273	407	0.179	Not avail.	354.000	47.000	0.133	2.143	0.032
0	Divorced	825	167	0.202	Single	2273.000	407.000	0.179	1.480	0.139
4	Separated	842	182	0.216	Divorced	825.000	167.000	0.202	0.689	0.491

	eje	customer	target_var	PD	eje-1	customer-1	target_var-1	PD-1	Z	pvalor
1	Living together	801	88	0.110	Married	784.000	22.000	0.028	6.407	0.000
3	Single	2627	454	0.173	Living together	801.000	88.000	0.110	4.275	0.000
0	Divorced_sep	1667	349	0.209	Single	2627.000	454.000	0.173	2.992	0.003

ILUSTRACIÓN 2.12

Clases resultantes de la variable estado civil.

marital_agg	loan_amount	target_var	customer	PD
Married	15760280.000	22	784	0.028
Living together	16447330.000	88	801	0.110
Single	53568640.000	454	2627	0.173
Divorced_sep	34278770.000	349	1667	0.209

ILUSTRACIÓN 2.13

Clases iniciales de la variable número de solicitantes.

num_applic	loan_amount	target_var	customer	PD
2	32273760.000	190	1605	0.118
Not avail.	7042500.000	53	354	0.150
1	80738760.000	670	3920	0.171

ILUSTRACIONES 2.14, 2.15

Contrastes de hipótesis de la variable número de solicitantes.

	ejes	customer	target_var	PD	ejes-1	customer-1	target_var-1	PD-1	Z	pvalor
2	Not avail.	354	53	0.150	2	1605.000	190.000	0.118	1.619	0.105
0	1	3920	670	0.171	Not avail.	354.000	53.000	0.150	1.019	0.308

	eje	customer	target_var	PD	eje-1	customer-1	target_var-1	PD-1	Z	pvalor
0	1	3920	670	0.171	2	1959.000	243.000	0.124	4.678	0.000

ILUSTRACIÓN 2.16

Clases resultantes de la variable número de solicitantes.

num_applic_agg	loan_amount	target_var	customer	PD
2	39316260.000	243	1959	0.124
1	80738760.000	670	3920	0.171

ILUSTRACIÓN 2.17

Mapa de calor del análisis de correlación entre las variables.

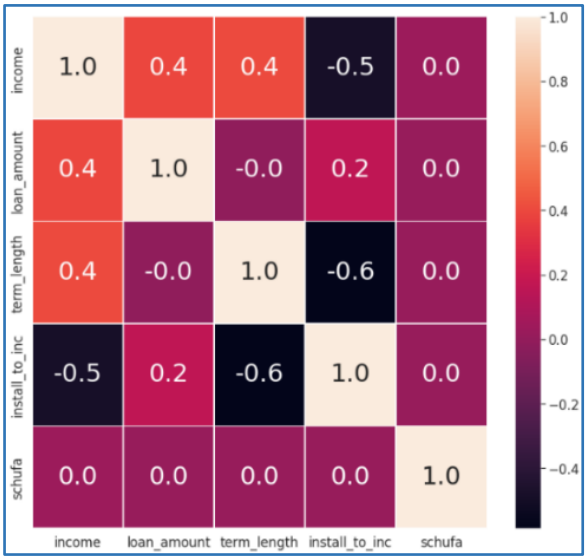
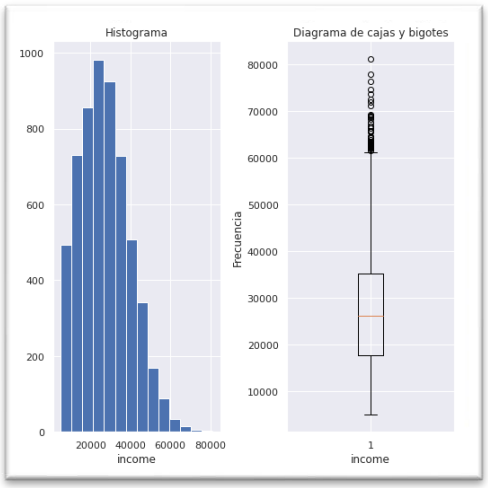


ILUSTRACIÓN 2.18

Distribución de la variable ingresos sin transformar.



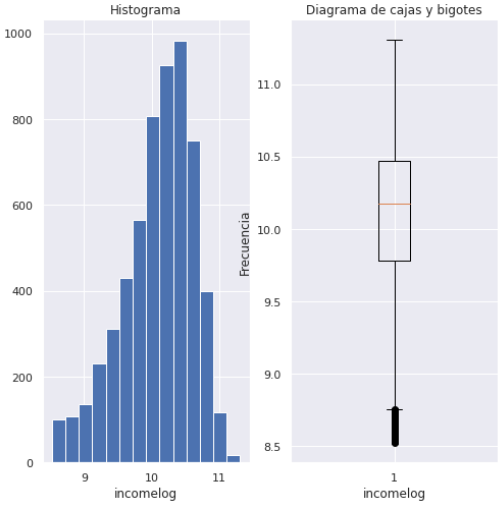
ILUSTRACIONES 2.19, 2.20, 2.21

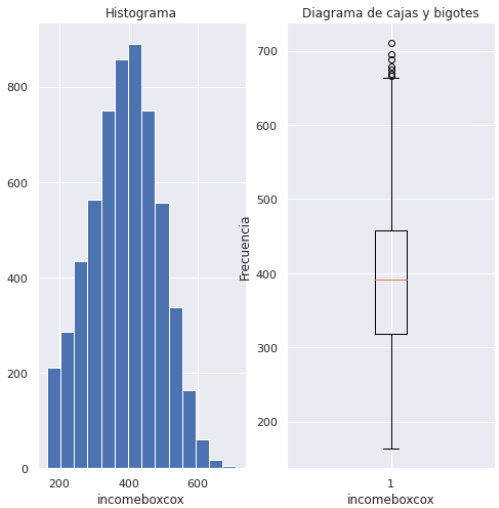
Transformaciones de la variable ingresos.

(Arriba derecha) Transformación logarítmica

(Abajo derecha) Transformación raíz cuadrada.

(Abajo izquierda) Transformación box-cox.





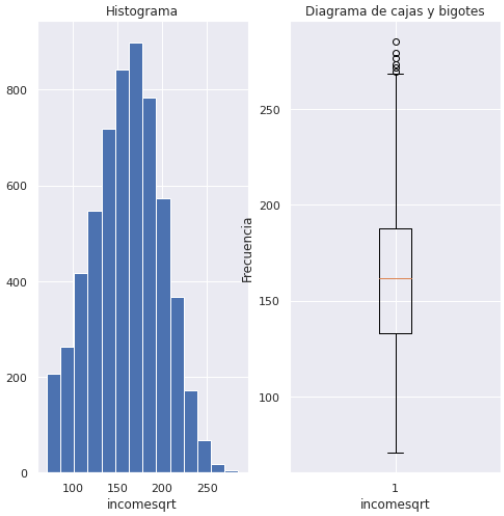
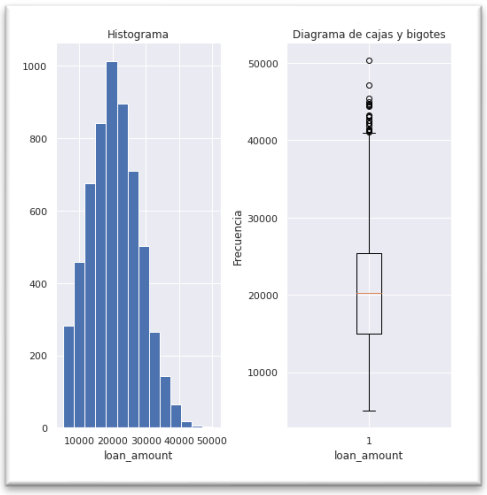


ILUSTRACIÓN 2.22

Distribución de la variable cantidad prestada sin transformar.



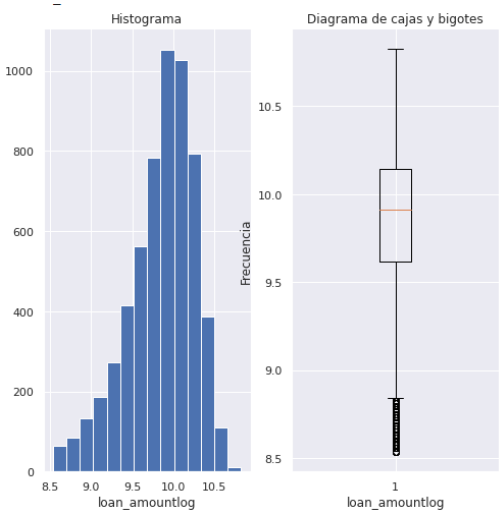
ILUSTRACIONES 2.23, 2.24, 2.25

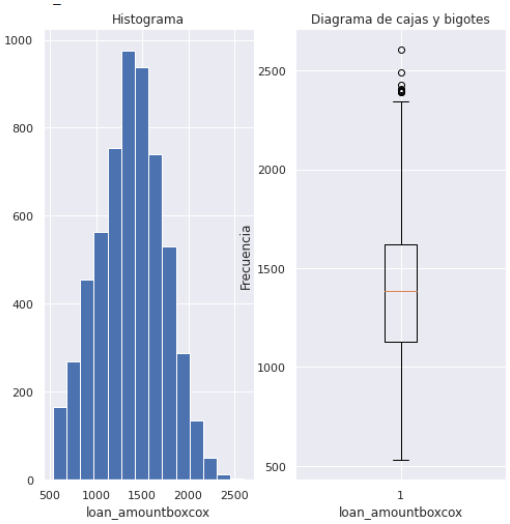
Transformaciones de la variable cantidad prestada.

(Arriba derecha) Transformación logarítmica

(Abajo derecha) Transformación raíz cuadrada.

(Abajo izquierda) Transformación box-cox.





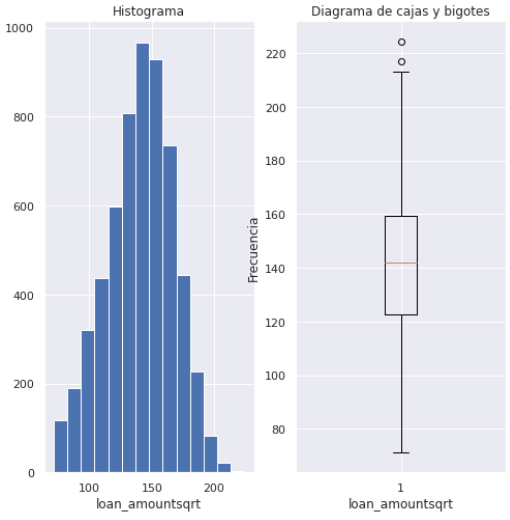
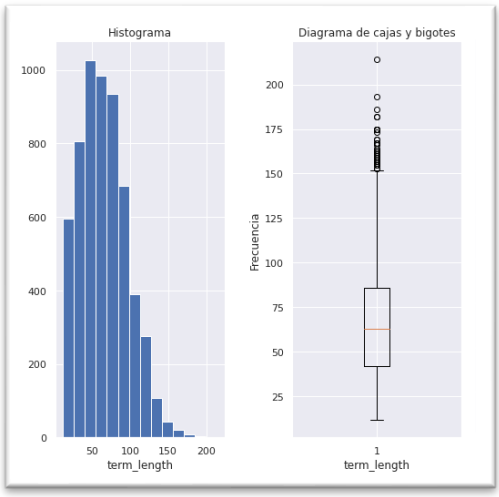


ILUSTRACIÓN 2.26

Distribución de la variable longitud del préstamo sin transformar.



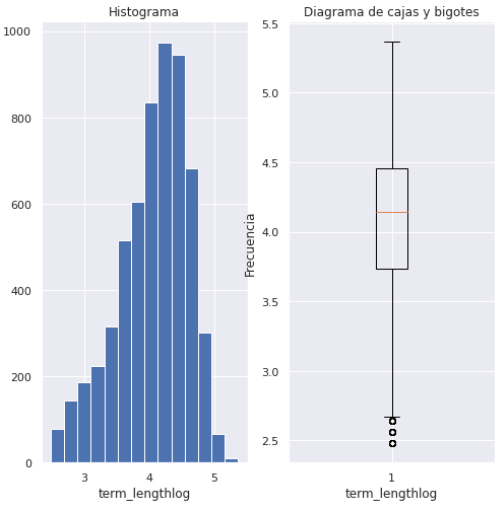
ILUSTRACIONES 2.27, 2.28, 2.29

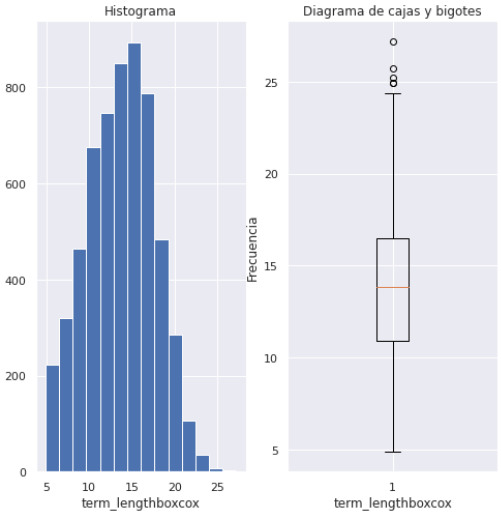
Transformaciones de la variable longitud del préstamo.

(Arriba derecha) Transformación logarítmica

(Abajo derecha) Transformación raíz cuadrada.

(Abajo izquierda) Transformación box-cox.





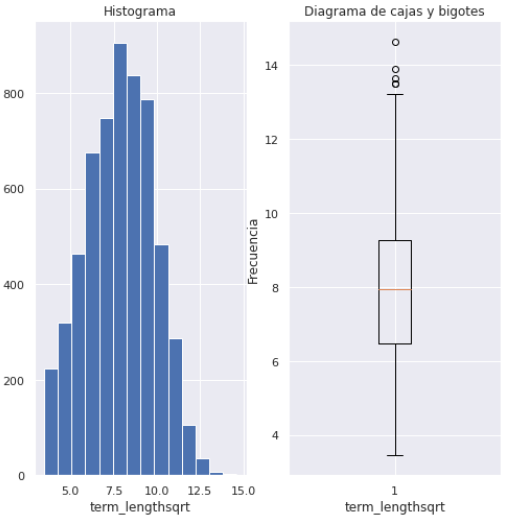
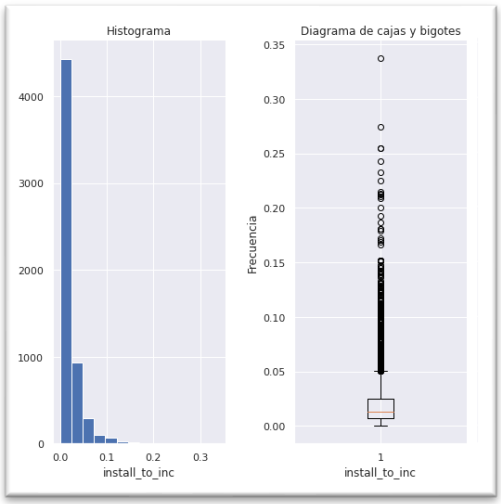


ILUSTRACIÓN 2.30

Distribución de la variable longitud del préstamo sin transformar.



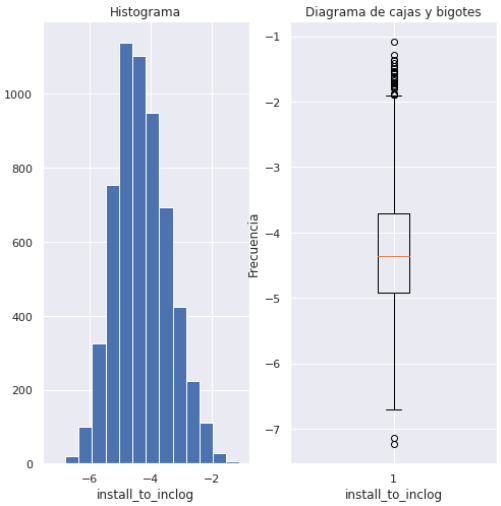
ILUSTRACIONES 2.31, 2.32, 2.33

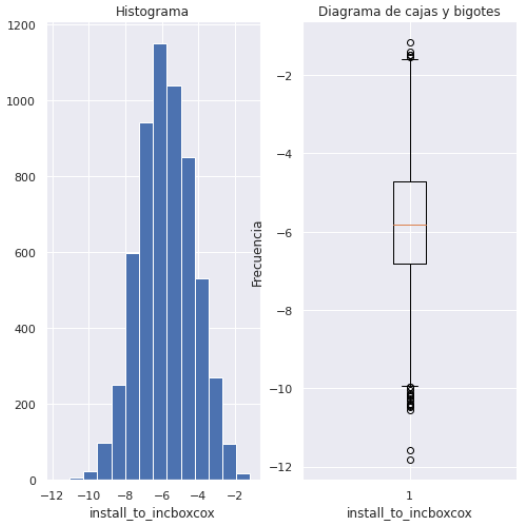
Transformaciones de la variable longitud del préstamo.

(Arriba derecha) Transformación logarítmica

(Abajo derecha) Transformación raíz cuadrada.

(Abajo izquierda) Transformación box-cox.





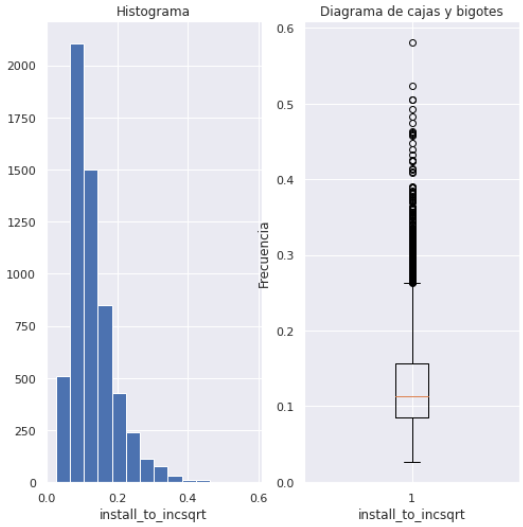
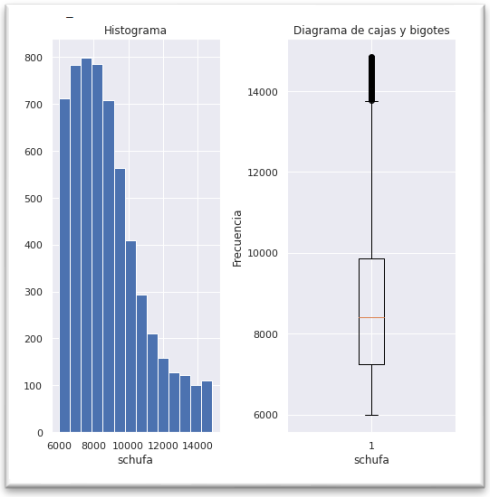


ILUSTRACIÓN 2.34

Distribución de la variable puntuación externa sin transformar.



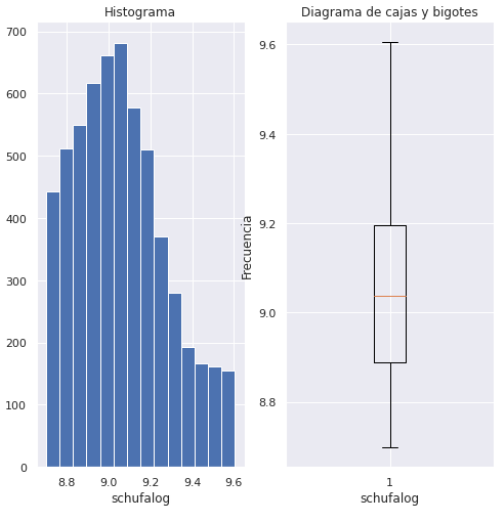
ILUSTRACIONES 2.35, 2.36, 2.37

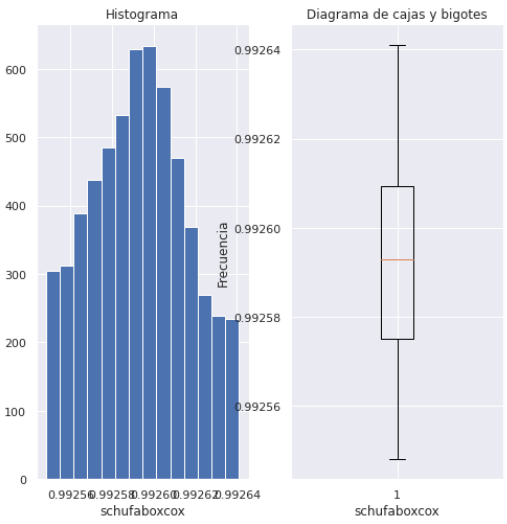
Transformaciones de la variable cantidad prestada.

(Arriba derecha) Transformación logarítmica

(Abajo derecha) Transformación raíz cuadrada.

(Abajo izquierda) Transformación box-cox.





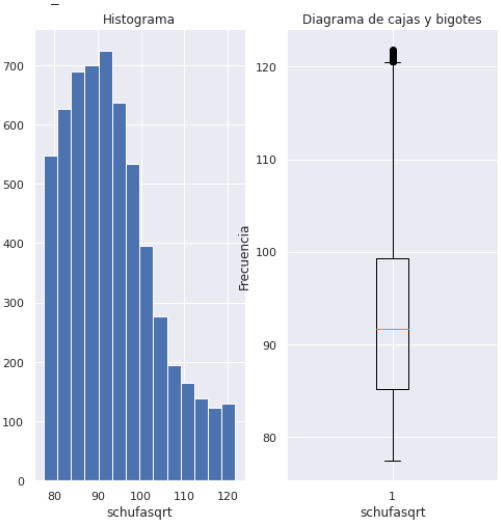


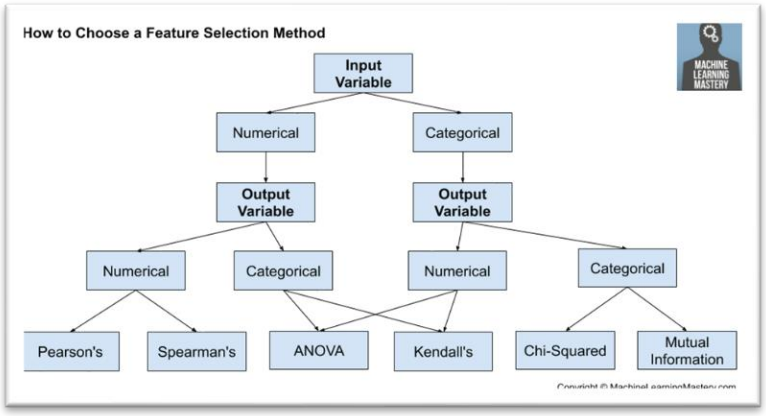
ILUSTRACIÓN 2.38

Características de los datos transformados.

	incomesqrt	incomeboxcox	loan_amountsqrt	loan_amountboxcox	term_lengthsqrt	term_lengthboxcox	install_to_inclog	install_to_incboxcox	schufaboxcox	schufalog
count	5879.000	5879.000	5879.000	5879.000	5879.000	5879.000	5879.000	5879.000	5879.000	5879.000
mean	160.014	387.651	140.358	1373.594	7.843	13.655	-4.288	-5.777	0.993	9.059
std	38.887	98.866	26.845	356.618	1.946	3.880	0.882	1.504	0.000	0.217
min	71.022	164.187	71.414	532.513	3.464	4.922	-7.224	-11.808	0.993	8.700
25%	133.055	318.699	122.474	1125.813	6.481	10.940	-4.916	-6.817	0.993	8.888
50%	161.689	391.327	142.127	1383.888	7.937	13.844	-4.362	-5.820	0.993	9.037
75%	187.437	457.165	159.374	1622.082	9.274	16.507	-3.705	-4.727	0.993	9.196
max	284.984	710.140	224.277	2604.647	14.629	27.174	-1.087	-1.165	0.993	9.605

ILUSTRACIÓN 2.39

Esquema de métodos de selección de variables (en inglés)



Obtenido de: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

ILUSTRACIÓN 2.40

Puntuaciones de las variables numéricas.

	Variables	Puntuación
0	schufaboxcox	866.262
1	schufalog	789.661
2	incomesqrt	739.725
3	incomeboxcox	737.757
4	schufa	689.432
5	income	677.334
6	install_to_incboxcox	476.947
7	install_to_inclog	475.352
8	term_length	338.179
9	term_lengthsqrt	295.072
10	term_lengthboxcox	295.033
11	install_to_inc	242.319
12	loan_amount	59.233
13	loan_amountboxcox	59.073
14	loan_amountsqrt	58.723

ILUSTRACIÓN 2.41

Puntuaciones de las variables categóricas.

	features	score
0	occup	607.812
1	occup_agg	228.106
2	marital	130.840
3	marital_agg	73.286
4	num_applic	13.505
5	num_applic_agg	7.291

ILUSTRACIÓN 2.42

Datos estandarizados.

	schufaboxcox	incomesqrt	install_to_incboxcox	term_length	loan_amountboxcox	occup	marital	num_applic
count	5879.000	5879.000	5879.000	5879.000	5879.000	5879.000	5879.000	5879.000
mean	-0.000	0.000	0.000	-0.000	-0.000	2.967	2.694	1.394
std	1.000	1.000	1.000	1.000	1.000	2.156	1.557	0.886
min	-1.933	-2.289	-4.011	-1.737	-2.359	0.000	0.000	0.000
25%	-0.768	-0.693	-0.691	-0.759	-0.695	0.000	1.000	0.000
50%	0.006	0.043	-0.028	-0.075	0.029	4.000	3.000	2.000
75%	0.715	0.705	0.699	0.675	0.697	5.000	4.000	2.000
max	2.092	3.214	3.068	4.846	3.452	6.000	5.000	2.000

ILUSTRACIÓN 2.43

Gráfica de la eliminación recursiva de las variables con validación cruzada

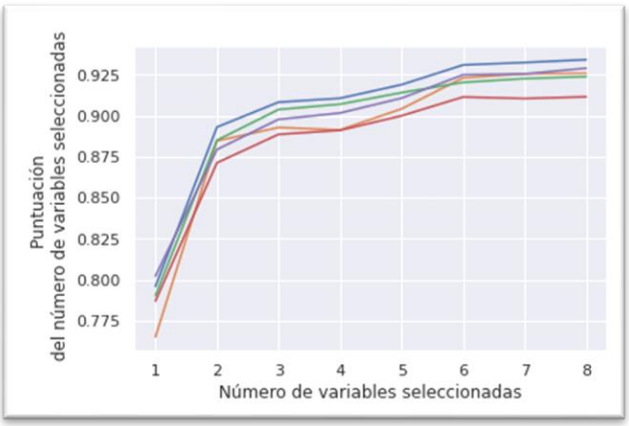


ILUSTRACIÓN 2.44

Curva ROC.

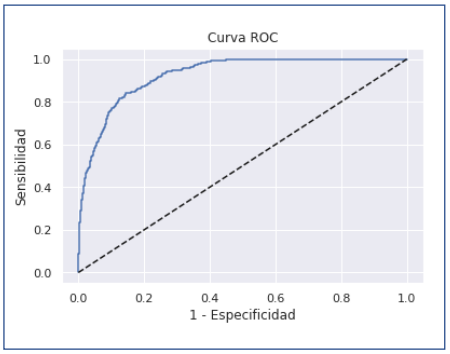


ILUSTRACIÓN 2.45

Matriz de confusión.

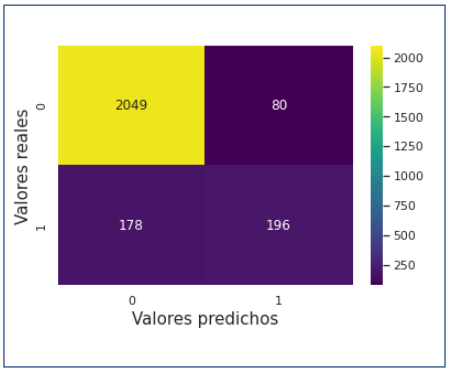


ILUSTRACIÓN 2.46

Puntuaciones del modelo obtenidas.

	Puntuaciones
Gini	0.851
AUROC	0.925
Exactitud	0.896
Precisión	0.705
Sensibilidad	0.524
Puntuación F1	0.601

ILUSTRACIÓN 2.47

Valor p obtenido del test ANOVA.

	income	loan_amount	term_length	install_to_inc	schufa
valor p	0.000	0.000	0.000	0.000	0.000

ILUSTRACIÓN 2.48

Probabilidades de impago de las clases de la variable estado laboral.

occup_agg	PD
Employee	0.021561
Student	0.142339
Unemployed	0.239305

ILUSTRACIÓN 2.49

Probabilidades de impago de las clases de la variable estado civil.

marital_agg	PD
Married	0.025087
Living together	0.105727
Single	0.170653
Divorced_sep	0.212236

ILUSTRACIÓN 2.50

Probabilidades de impago de las clases de la variable número de solicitantes.

num_applic_agg	PD
2	0.122318
1	0.169173

ILUSTRACIÓN 2.51

Características de la variable cantidad prestada.

	cantidad prestada
count	8382.000000
mean	20523.962658
std	7451.741917
min	5100.000000
25%	15100.000000
50%	20300.000000
75%	25500.000000
max	50300.000000

ILUSTRACIÓN 2.52

Características de la variable ingresos.

	ingresos
count	8382.000000
mean	27236.863932
std	12627.145673
min	5021.691061
25%	17627.947804
50%	26157.884298
75%	35385.686494
max	81215.721992

ILUSTRACIÓN 2.53

Ingresos por año del conjunto de datos de 2008 a 2018.

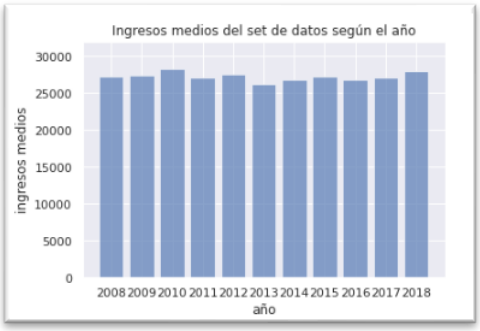


ILUSTRACIÓN 2.54

Ingresos por año de España de 2008 a 2018.