

Informe

Grupo 4

Integrantes:

- Benancio, Paola
- De la Cruz, Zarit
- Rázuri, Nickole

I. ANÁLISIS: MODELO DE REGRESIÓN

1. Análisis estadístico inicial

Análisis de las variables del dataset. En primer lugar, es importante clasificarlas según su tipo: numéricas o categóricas, pues esto les permitirá realizar análisis más pertinentes según tipo.

a. **Numéricas:** Se identificaron 4 variables numéricas

	count	mean	std	min	25%	50%	75%	max
Age	10000.0	44.0217	15.203998	18.0	31.0	44.0	57.0	70.0
Number_of_Dependents	10000.0	2.5270	1.713991	0.0	1.0	3.0	4.0	5.0
Work_Experience	10000.0	24.8588	14.652622	0.0	12.0	25.0	37.0	50.0
Household_Size	10000.0	3.9896	2.010496	1.0	2.0	4.0	6.0	7.0

b. **Categóricas:** Se identificaron 9 variables categóricas y para cada una de ellas, se contabilizaron las categorías que contienen.

Variables categóricas	Cardinalidad
Occupation	5
Education_Level	4
Primary_Mode_of_Transportation	4
Marital_Status	3
Location	3
Employment_Status	3
Type_of_Housing	3
Homeownership_Status	2
Gender	2

Asimismo, es importante identificar los valores que dominan las categorías pues tendrán influencia en los resultados del modelo.

Respecto a los niveles de educación, el nivel de bachillerato es el mayoritario, albergando al 41% de la población. Por otro lado, las ocupaciones de salud y tecnología se posicionan como las dos más comunes, ocupando el 30% y 24% de la población estudiada, respectivamente.

Un hallazgo notorio

Top 10 valores más frecuentes de Education_Level		
...	Education_Level	Total
	% sobre el total	
	Bachelor's	4058
	High School	2959
	Master's	2482
	Doctorate	501

Top 10 valores más frecuentes de Occupation		
Occupation	Total	% sobre el total
Healthcare	3035	30.35
Technology	2407	24.07
Finance	1525	15.25
Others	1521	15.21
Education	1512	15.12

Top 10 valores más frecuentes de Location		
Location	Total	% sobre el total
Urban	7037	70.37
Suburban	1951	19.51
Rural	1012	10.12

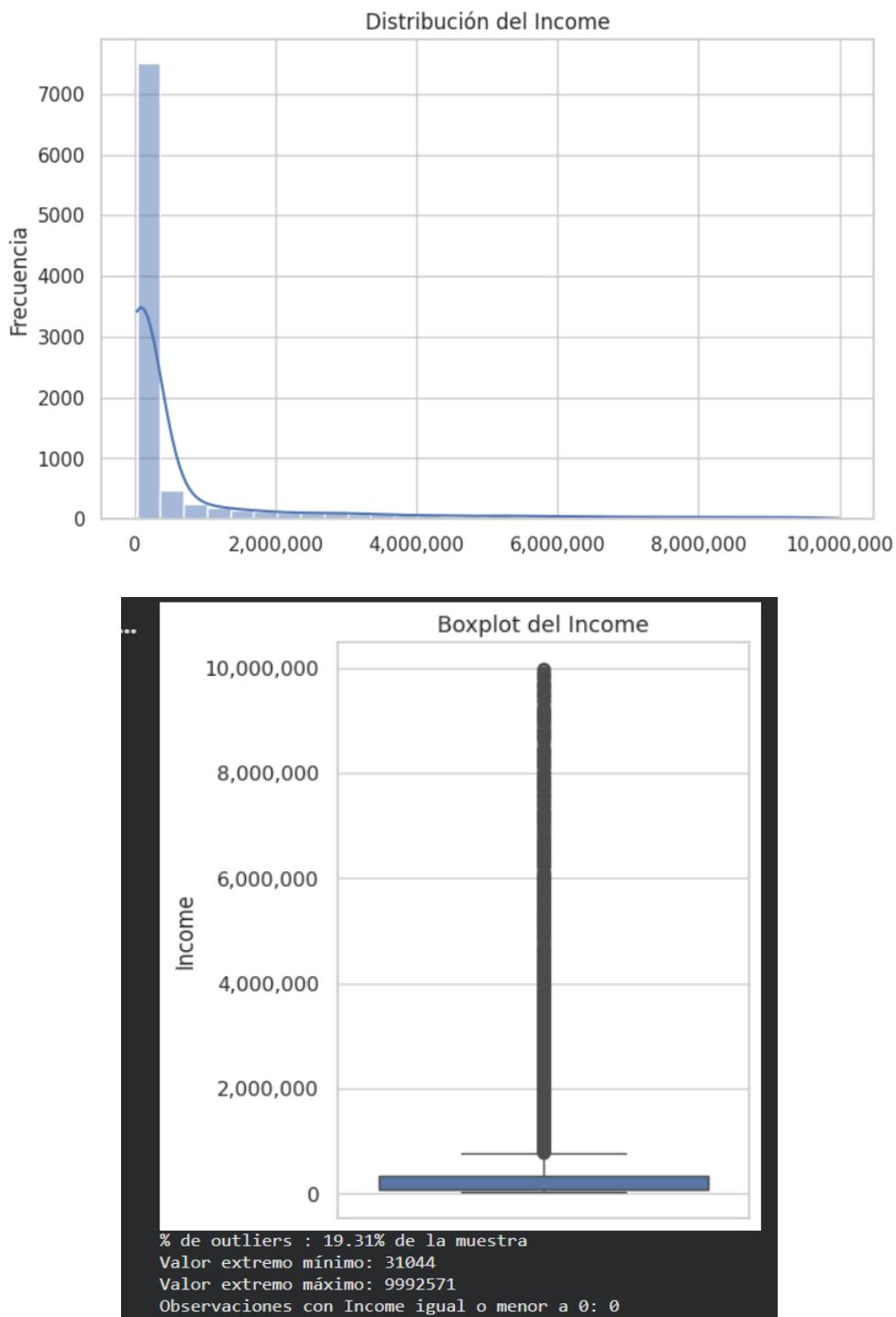
Top 10 valores más frecuentes de Marital_Status		
Marital_Status	Total	% sobre el total
Married	5136	51.36
Single	3900	39.00
Divorced	964	9.64

Top 10 valores más frecuentes de Employment_Status		
Employment_Status	Total	% sobre el total
Full-time	5004	50.04
Part-time	3016	30.16
Self-employed	1980	19.80
Top 10 valores más frecuentes de Homeownership_Status		
Homeownership_Status	Total	% sobre el total
Own	6018	60.18
Rent	3982	39.82
Top 10 valores más frecuentes de Type_of_Housing		
Type_of_Housing	Total	% sobre el total
Single-family home	4055	40.55
Apartment	4001	40.01
Townhouse	1944	19.44
Top 10 valores más frecuentes de Gender		
Gender	Total	% sobre el total
Male	5123	51.23
Female	4877	48.77
Top 10 valores más frecuentes de Primary_Mode_of_Transportation		
Primary_Mode_of_Transportation	Total	% sobre el total
Public transit	4047	40.47
Car	2986	29.86
Biking	1940	19.40
Walking	1027	10.27

- c. Variable target: Es la variable ingreso (income), la cual es de tipo entero (int64) y cuenta con 10,000 observaciones. Al realizar un análisis más detallado de sus categorías y composición, encontramos que el valor promedio es de 816,838. Sin embargo, este valor podría estar sobreestimado debido a la existencia de outliers muy altos. Esto se evidencia al analizar estadísticos descriptivos de la variable. Los valores van de 31 mil a los 9 millones; sin embargo, se observa una gran distancia entre el valor máximo y los valores del percentil 75% (350,667). Además, la existencia de estos outliers se confirma con el análisis de cuantiles desde el 1% al 99%, que presenta una diferencia de 3 millones a más desde el percentil 95% al 90%.

	count	mean	std	min	25%	50%	75%	max	
Income	10000.0	816838.1667	1.821089e+06	31044.0	68446.0	72943.0	350667.5	9992571.0	
	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99
Income	36121.95	64149.7	65766.0	68446.0	72943.0	350667.5	2890510.4	5343905.45	8836918.82

Asimismo, los gráficos de tipo histograma y boxplot comprueban la existencia de outliers. En particular el boxplot considera que el 19.31% de la muestra es outliers.



2. Análisis de calidad:

En el análisis descriptivo se comprobó que los datos originales pueden no ser ideales para usar en un modelo de regresión.

a. Numéricas

Por un lado, verificamos que las variables explicativas de este tipo no registren valores nulos a los que deberíamos someter a tratamiento.

Por otro lado, la variable target (ingreso), registra outliers.

Variable	MIN	P1	P50	P99	MAX
Age	18.00	18.00	44.00	70.00	70.00
Number_of_Dependents	0.00	0.00	3.00	5.00	5.00
Work_Experience	0.00	0.00	25.00	50.00	50.00
Household_Size	1.00	1.00	4.00	7.00	7.00

b. Categóricas

Se verifica que las variables explicativas de este tipo no registran valores nulos. Luego, se identifican categorías con baja frecuencia para evaluar la inestabilidad.

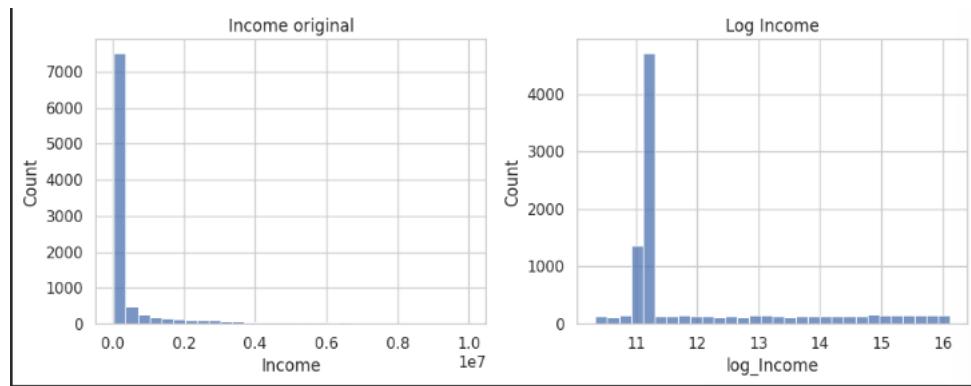
3. Preprocesamiento

Después de realizar el diagnóstico de la calidad de datos, concluimos que aplicando

- a. De acuerdo al diagnóstico, el dataset actual no cuenta con valores nulos (NaN). Sin embargo, la imputación se realiza para garantizar que el pipeline se encuentre matemáticamente bien definido para cualquier entrada válida.
 - b. A futuro, si en el dataset aparecen observaciones con valores nulos, el pipeline ya se encuentra preparado para tratarlas. Esto permite que el modelo sea más robusto y reutilizable.
 - c. Esto permite que el modelo esté definido para cualquier observación que respete el esquema de datos, incluso si esa observación tiene valores nulos.
- 3.1. Numéricas: Se imputará la mediana de valores de una variable como reemplazo de los missing. Sin embargo, no hay valores nulos, así que no se aplica.
- 3.2. Categóricas: De acuerdo al diagnóstico, el dataset actual no cuenta con valores nulos (NaN) ni nulos codificados. Sin embargo, la imputación se realiza para garantizar que el pipeline se encuentre matemáticamente bien definido para cualquier entrada válida.
A futuro, si en el dataset aparecen observaciones con valores nulos, el pipeline ya se encuentra preparado para tratarlas. Esto permite que el modelo sea más robusto y reutilizable.
Esto permite que el modelo esté definido para cualquier observación que respete el esquema de datos, incluso si esa observación tiene valores nulos y nulos codificados.

3.3. Variable target

La variable Income presenta una distribución asimétrica con una cola derecha extensa. Por ello, se aplicó la transformación logarítmica a modo de reducir la asimetría y la influencia de valores extremos. Esto permite que la variable sea más adecuada para el modelado de la regresión.



(Por continuar)

II. ANÁLISIS: MODELO DE CLASIFICACIÓN

1. Análisis estadístico inicial

Análisis de las variables del dataset “lending_dataset”. En primer lugar, es importante clasificarlas según su tipo: numéricas (9 variables) y categóricas (10 variables), pues esto les permitirá realizar análisis más pertinentes según tipo.

a. **Numéricas:** Los estadísticos descriptivos muestran media y la distribución de valores según percentil. Se identifica que el valor medio de la variable podría estar sesgado por outliers grandes. Al examinar los percentiles de cada variable, resulta evidente que podrían haber casos de outliers o valores faltantes.

En primer lugar, el análisis de valores missing muestra que la variable pub_rec_bankruptcies ocupa el 1.6% de sus valores en missing. Es sumamente relevante realizar una corrección de los datos.

	% Missing
pub_rec_bankruptcies	1.612417
revol_util	0.140647
dti	0.055254
delinq_2yrs	0.045208
annual_inc	0.015069
index	0.000000
loan_amnt	0.000000
installment	0.000000
term	0.000000
int_rate	0.000000

- b. **Categóricas:** Se emplea Cardinalidad para identificar las categorías por cada variable de este tipo.

2. Análisis de calidad

- a. Por continuar (...)