Análisis y Decisiones Tomadas

Grupo: Desirée Vera, Felipe Gómez, Harmynn Garrido, Diego Granados.

Parte A: Descripción del Problema

El problema es el no pago de las deudas (default) por parte de los clientes, por lo que se busca determinar qué cliente va a incurrir en default. Por su parte, el Default es una variable binaria, por lo que se requiere un modelo de clasificación para predecir si un cliente no pagará sus obligaciones.

Los datos con los que se cuenta están en un archivo Excel llamado "Tabla Trabajo Grupal N°2", que se compone de dos hojas, Modelamiento y Predicción con los siguientes campos:

i. Id_Cliente: Número único del cliente
ii. Edad: Campo cuantitativo que detalla la edad del cliente.
iii. Nivel Educacional: Campo categórico que detalla el nivel educacional del cliente.
iv. Años Trabajando: Campo cuantitativo con el detalle de los años trabajando del cliente.
v. Ingresos: Campo cuantitativo que detalla el monto encriptado del ingreso del cliente. Deuda Comercial: Campo cuantitativo que detalla la deuda comercial del cliente.
vi. Deuda Crédito: Campo cuantitativo que detalla la deuda consumo en crédito del cliente.
vii. Otras Deudas: Campo cuantitativo que detalla el monto deudas, no comerciales ni consumo del cliente.
viii. Ratio Ingresos Deudas: Campo cuantitativo que detalla la proporción de ingresos sobre deudas totales del cliente.
ix. Default: Campo cuantitativa binaria, 1 si el cliente incurre en default y 0 cliente cumple con el pago. (variable objetivo).

Parte B: Inspección y Limpieza de Datos

Los datos de la hoja Modelación se componen de 12356 filas y 10 columnas. En las columnas podemos encontrar a las variables descritas anteriormente.

La base no presenta datos faltantes ni repetidos, por lo que no requiere imputación de datos.

Parte C: Análisis Exploratorio de Datos

Al inspeccionar los datos restantes, se encuentra lo siguiente:

- existencia de datos outliers en algunas variables, dado que el promedio y la mediana son muy distintos, la desviación estándar es mayor que el promedio y que el valor máximo es más del doble que el tercer cuartil (Q3) en casi todas. Son variables con sesgo a la derecha y puede que no sean normales.
    - 'Años_Trabajando',
    - 'Ingresos',
    - 'Deuda_Comercial',
    - 'Deuda_Credito',
    - 'Otras_Deudas'
    - 'Ratio_Ingresos_Deudas'

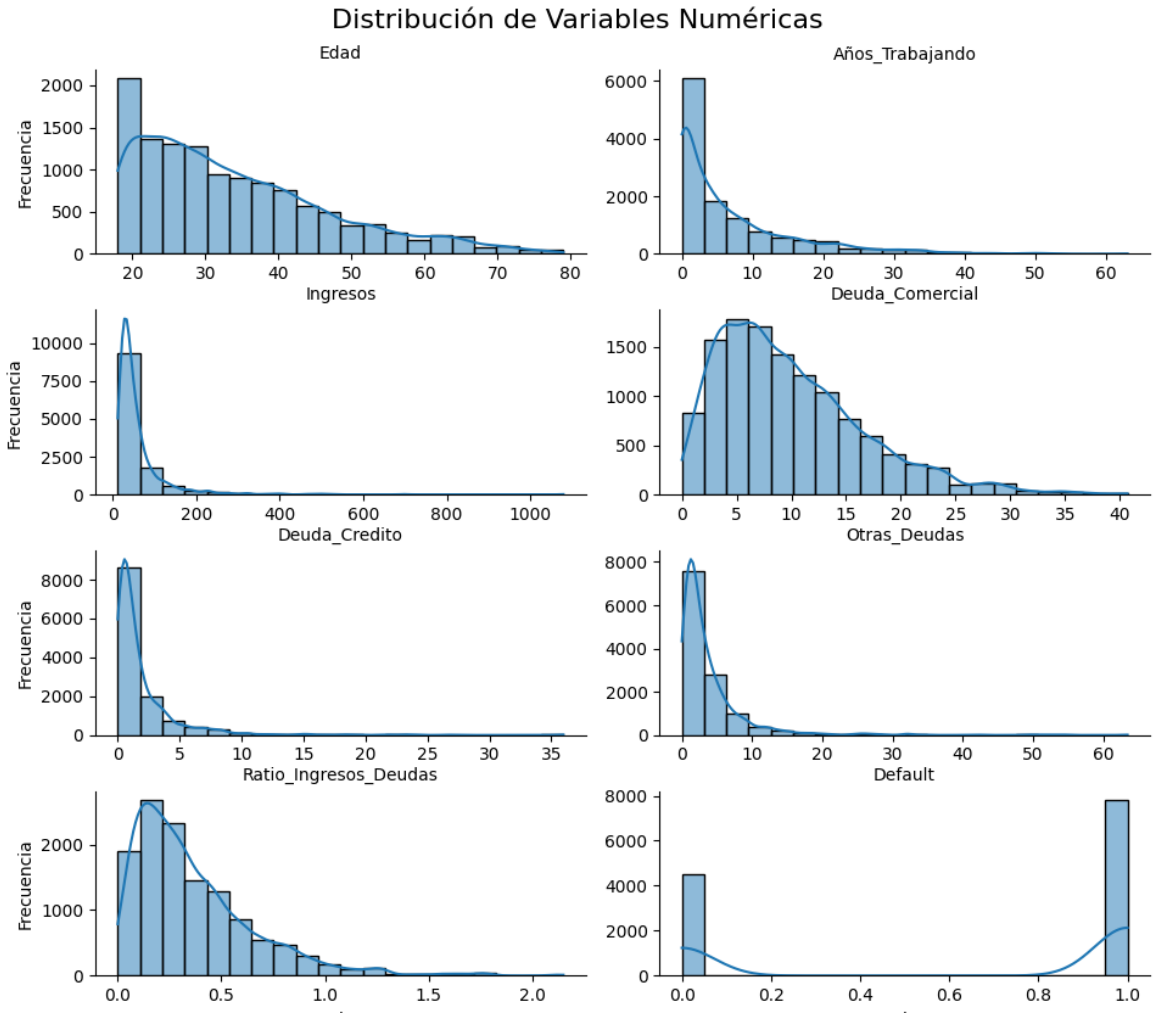
	Edad	Años Traba-jando	Ingresos	Deuda Comer-cial	Deuda Credito	Otras Deudas	Ratio Ingresos Deudas	Default
Count	12356	12356	12356	12356	12356	12356	12356	12356
Mean	34.165	6.945	59.747	9.945	1.959	3.871	0.365	0.632
Std	13.136	8.994	67.204	6.734	3.025	5.439	0.296	0.482
Min	18.000	0.000	12.000	0.000	0.000	0.000	0.000	0.000
25%	24.000	0.000	27.00	4.800	0.420	1.090	0.150	0.000
50%	31.000	4.000	40.000	8.500	1.000	2.210	0.290	1.000
75%	42.000	10.000	64.000	13.600	2.210	4.590	0.500	1.000
max	79.000	63.000	1079.000	40.700	35.970	63.470	2.150	1.000

- la variable 'Default' es binaria y contiene 63,28% de defaults (1), muestra un desbalanceo leve.
- respecto a la variable 'Nivel_Educacional' tiene cinco categorías, entre las cuales Med es la más común con casi un tercio de las observaciones.

Nivel_Educacional	
Med	4320
SupInc	2766
SupCom	2580
Bas	2005
Posg	685

- los clientes tienen un promedio de 34 años.
- los clientes tienen mayoritariamente menos de 1 año de experiencia, pero el promedio es 7 años.

- el ingreso se concentra en las 40 unidades monetarias.
- la deuda comercial se concentra en 8.5 unidades.
- la deuda de consumo se concentra bajo las 1.0 unidades.
- las otras deudas se concentran bajo las 2.2 unidades.
- la razón de ingresos a deuda se concentra bajo los 0,3.
- en la variable objetivo, puede considerarse hacer un balance.



Al revisar las relaciones entre las variables se encontró que:

- Edad está correlacionada directa con Años_trabajando e Ingresos
- Años_Trabajando además se correlaciona directamente con Otras_Deudas
- Ingresos además se correlaciona directamente con Deuda_Credito y Otras_Deudas
- Deuda_Comercial se correlaciona directamente con Ratio_Ingresos_Deudas

- Deuda_Credito además se correlacioan directamente con Otras_Deudas
- Default solo tiene correlaciones débiles con el resto de las variables numéricas.

Parte D: Decisiones Metodológicas

No se hará nada con los outliers, presentes en las variables, dado que pueden aportar información importante al modelo, al generar entrenamiento en situaciones extremas pese a que puede generar más errores dada la varianza de los datos.

No se hará nada con el desbalance en la variable ‘Default’ al ser leve, sin embargo, se pondrá énfasis en el F1 Score para un mejor manejo del mismo.

La variable Nivel_Educacional se codificará con Target Encoder una vez que se hayan separado las muestras de entrenamiento y test.